Afslutningsrapport

Simpel Bioinformatik

Mejeribrugets ForskningsFond
Rapport nr. 2009-95

Marts 2009

MEJERIFORENINGEN

Afslutningsrapport

SIMPEL BIOINFORMATIK

Projektleder

Professor Rasmus Bro
Kvalitet og Teknologi
Institut for Fødevarevidenskab
Københavns Universitet
Rolighedsvej 30
1958 Frederiksberg C
Tlf: 3533 3296
Fax: 3533 4532
E-mail: rb@life.ku.dk

Øvrige deltagere

Ph.d.-studerende Karin Kjeldahl, KU-LIFE, e-mail: kkd@life.ku.dk
Postdoc: Bonnie Schmidt, KU-LIFE, e-mail: bosc@life.ku.dk
Lektor: Charlotte Møller Andersen, KU-LIFE, e-mail: cma@life.ku.dk

Øvrige finansieringskilder

KU-LIFE

Resume af det samlede projekt

Data indsamlet fra biologiske systemer er typisk store og komplekse. Dette stiller store krav til dataanalysen og kræver betydelig viden inden for både det specifikke fagområde samt om dataanalyse for at sikre, at de rigtige konklusioner drages. Dette projekt søger at udvikle fremgangsmåder og visualiseringsteknikker, der muliggør at en bredere gruppe af relevante eksperter inden for ostemodning bringes i stand til på simpel vis at identificere den vigtigste information i data og dermed foretage kvalificerede diskussioner på detaljeret, videnskabeligt grundlag.

Der er fokuseret på DNA-microarrays, hvor forskellige datasæt er blevet inddraget omhandlende mælkesyrebakterier og humane celler. Derudover er der set på et enkelt datasæt bestående af syrningskurver opnået ved en mælkefermentering med *Lactococcus lactis*.

I projektet er udviklet tre fremgangsmåder til analyse af DNA-microarray data. Den ene består i en adskillelse af gener i to grupper, hvoraf den ene udgøres af såkaldte unikke gener, og den anden indeholder en større gruppe af gener, der har en fælles underliggende struktur. Metoden baserer sig på kemometriske metoder som principal komponent analyse (PCA) og multivariat kurve resolution (MCR). Den er blevet udviklet ved at bruge data fra en mælkefermentering med bakterien *Lactococcus lactis* indeholdende målinger fra ca. 2500 gener. Resultaterne stemmer overens med resultater opnået ved en traditionel men mere omstændelig databehandling, hvor der ses på hvert gen for sig. En anden fordel ved den udviklede metode er, at *alle* gener beskrives, visualiseres og tilforordnes en af tre typer af variation: fejlagtig måling, unikt gen, eller del af underliggende struktur.

En anden metode blev udviklet på et større datasæt indeholdende information fra mere end 50.000 gener/prober. Denne metode indeholder først en reduktion af data og udnytter derefter trevejsstrukturen i data i en PARAFAC-model. Ud fra denne kan man rent visuelt opnå en forståelse for sammenhængen i data.

Den sidste metode blev udviklet i samarbejde med Ana Conesa, Bioinformatics Department, Centro de Investigacion Principe Felipe, Valencia, Spanien og baserer sig på kendskab til genernes funktionelle annoteringer. Generne samles i grupper i forhold til disse, hvorefter data reduceres yderligere vha. PCA. Hver funktionel gruppe vil ideelt set kun blive repræsenteret én gang og vil derfor have en større chance for at påvirke den efterfølgende dataanalyse som fx kan foretages med partial least squares (PLS) regression.

Samlet set har projektet vist eksempler på, hvorledes meget komplicerede datastrukturer kan analyses og fortolkes med basale kemometriske metoder. På sigt vil dette have potentiale, bl.a. inden for mejeriforskningen og dermed mejeriindustrien, idet de udviklede metoder kan give lettilgængelige oplysninger om detaljerede molekylærbiologiske og fysiologiske forhold. Projektet har imidlertid også vist, at mange af de data, der i dag opsamles, lider af et meget specifikt problem. Mængden af data (observationer) er oftest ikke tilstrækkelig til at opnå de resultater, som oprindeligt var ønsket. En meget konkret løsning på dette problem er at inddrage den kemometriske og statistiske ekspertise på et langt tidligere tidspunkt i det eksperimentelle arbejde.

English summary

Data sets from biological systems are typically huge and complex. This entails large demands to the data analysis and requires knowledge within the specific scientific subject

as well as knowledge about data analysis to secure that the correct conclusions are drawn. This project seeks to develop simple data analytical methods and visualization techniques that make experts within cheese ripening able to extract the relevant information from data and thus take part in the discussions on a detailed, scientific basis.

The focus of the project is on DNA microarrays. A couple of data sets are included dealing with lactic acid bacteria and human cells. In addition, a data set consisting of acidification curves obtained from fermentation of milk by the bacteria, *Lactococcus lactis* is contained in the project.

Three data analytical techniques for analysis of DNA microarrays are developed. One method consists of a separation of genes into two groups. One group is made up by so-called unique genes. The other contains a larger group of genes having similar underlying structure. The method is based on chemometric techniques such as principal component analysis (PCA) and multivariate curve resolution (MCR). It is developed using DNA microarray data obtained from a milk fermentation process by the bacteria, *Lactococcus lactis* containing measurements from approximately 2500 genes. An advantage of the method is that *all* genes are described, visualized and appointed to one of three types of variation: erroneous measurement, unique gene or part of the underlying structure. Another method was developed using a larger data set with more than 50,000 genes/probes. This method contains a data reduction step followed by PARAFAC modeling which takes advantage of the three-way structure in data. From this model, it is possible to visually identify the information in data.

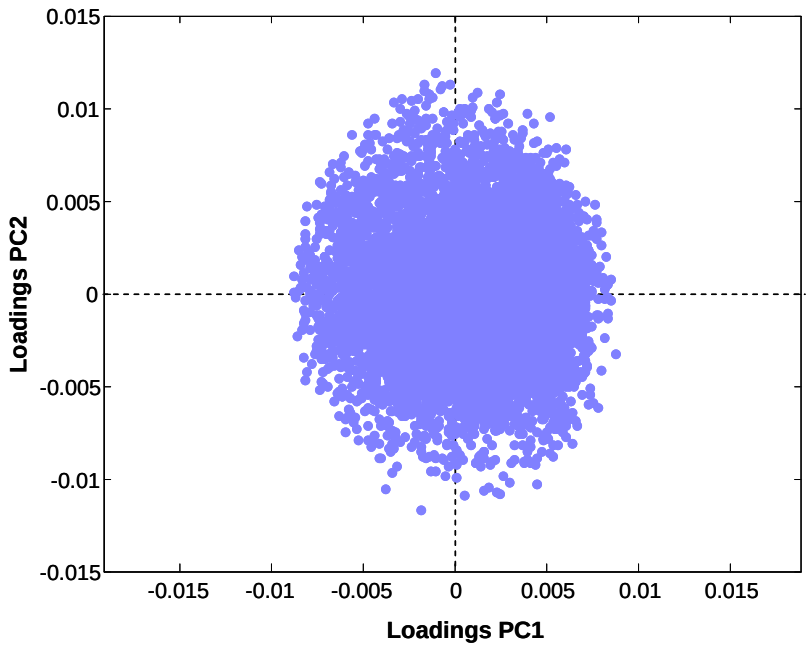
The third method is developed together with Ana Conesa, Bioinformatics Department, Centro de Investigación Principe Felipe, Valencia, Spain and is based on the knowledge of functional annotations. The genes are collected in groups in relation to these annotations, whereupon the size of the data is reduced using PCA. Each functional group is then represented only once and will thus have a more equal chance of influencing the following data analysis performed by e.g. PLS regression.

The project has shown examples on how very complicated data structures can be analyzed using fundamental chemometric methods. This shows great potentials within dairy research and thus the dairy industry, since the methods developed makes the information about detailed molecular biological and physiological relationships easier to obtain. At the same time, the project has shown that many of the collected data suffer from a very specific problem. The amount of data (observations) is often inadequate to give the required results. A solution to this problem is to include the chemometric and statistical knowledge earlier in the experimental work.

Introduktion

Forståelse for processer og tilstande i biologiske systemer er et vigtigt forskningsområde og kan være med til at give et helhedsbillede af de samspil, der foregår i naturen. Målinger på sådanne systemer indeholder typisk en mængde informationer, som gør det svært at uddrage den relevante information ved at se på hvert enkelt datapunkt for sig selv. En betydelig udfordring er derfor at kunne håndtere den kompleksitet, der ofte kendetegner sådanne data, og som stiller store krav til dataanalysen. Dette er vigtigt for at sikre, at den relevante information identificeres, for at undgå overfortolkning og for ikke at risikere at der drages fejlagtige konklusioner.

Data opnået fra biologiske systemer er i dag ofte kendetegnet ved at have mange variable og få prøver. De kan være meget støjfyldte, og der kan også være flere tilfælde, hvor målinger for nogle datapunkter mangler. Derudover varierer timingen af biologiske processer, hvilket gør det svært at opnå replikater. Sådanne faktorer er med til at gøre data endnu mere komplekse og bidrager til kravet om en omhyggelig dataanalyse. Et eksempel på dette er vist i figur 1, der illustrerer loadings opnået ved principal komponent analyse (PCA). Datasættet er DNA-microarray-data fra humane celler og indeholder målinger fra 34 prøver og over 50.000 prober/gener. Loadings for principal komponent ét (PC1) og to (PC2) ligger i en stor gruppe omkring centrum, og det er således ikke muligt at identificere de gener, der er vigtige for at kunne beskrive forskellen mellem prøverne. Figuren illustrerer også, at de enkelte multivariate metoder i sig selv ikke er i stand til at give en fyldestgørende beskrivelse og tolkning af data. For at kunne gøre dette er det nødvendigt med mere dedikerede analysemetoder. Gennem de senere år er der sket en udvikling af sådanne, men der mangler stadig simple, brugervenlige metoder, der samtidig er tilstrækkelig kraftfulde til at kunne give en kvalificeret fortolkning af de opnåede data.



Figur 1. Loadings fra en PCA foretaget på DNA-microarray-data fra humane celler. Datasættet indeholder 54.675 prober og 34 prøver.

Dette projekt fokuserer på dataanalysen og søger at udvikle metoder, der kan drage nytte af nyere analysemetoder og samtidig give en statistisk forståelse. Det primære formål er, at de udviklede metoder er i stand til at visualisere resultaterne på en måde, så de bliver

overskuelige uden at brugeren behøver være ekspert i databehandling. Dette vil sikre, at en bredere gruppe af videnskabelige eksperter inden for fagområdet vil være i stand til at forstå og analysere data.

Projektet skal ses i sammenhæng med andre MFF-finansierede projekter: "Starterkulturers fysiologiske og genetiske status som indikator for modningsforløb i ost" og "Fermentering af mælk kortlagt ud fra fysiologiske ændringer i *Lactococcus lactis*, undersøgt ved DNA-microarray og proteomanalyse". Formålet med samarbejdet har været at bruge data fra ovennævnte projekter som eksempler i metodeudviklingen. Samtidig er data relevante i mejerimæssig sammenhæng. Derudover er der inddraget humandata, der indeholder lignende problemstillinger.

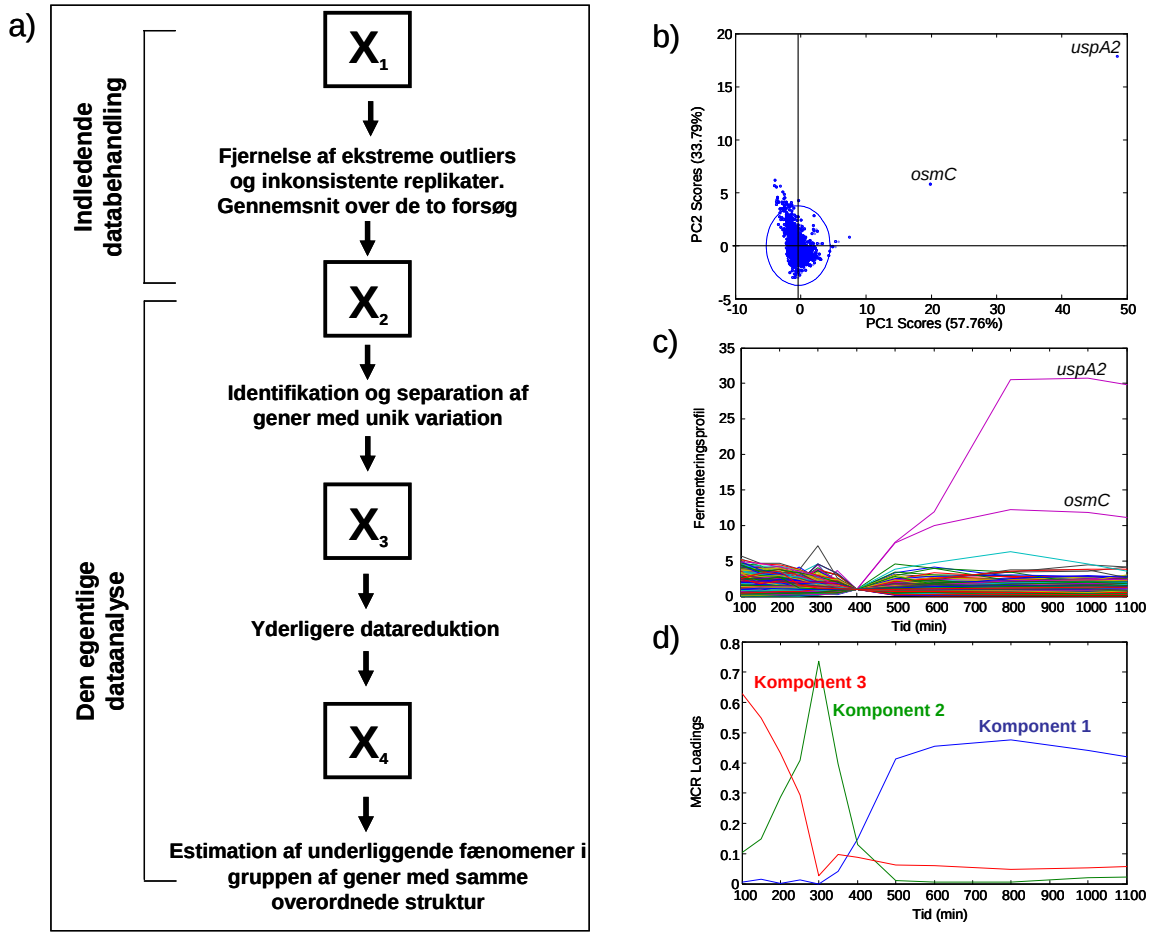
Resultater og diskussion

Projektet har inddraget flere forskellige datasæt til analyse og metodeudvikling. Det har primært drejet sig om data fra DNA-microarrays. Dette er et attraktivt værktøj, da udtryk af flere tusinde gener kan måles på én gang til det ønskede tidspunkt. Hermed et det muligt at undersøge det genetiske respons til en given behandling. Overordnet set benyttes microarray- chips, der indeholder et antal prober. Hver probe beskriver niveauet af et bestemt mRNA. I det følgende benævnes dette som gener. Ud over data fra DNA-microarrays indgår også analyse af syrningskurver.

I projektet er udviklet tre metoder til analyse af data fra DNA-microarrays. Alle metoder kombinerer basale kemometriske teknikker, således at de vigtigste konklusioner kan identificeres og visualiseres på en simpel måde. Metoderne er beskrevet nedenfor.

Metode 1: Identifikation af unik og generel variation i microarraydata
Princippet i denne metode er en separation af individuelle gener i to grupper. Den ene indeholder gener, der tydeligt har en unik individuel variation, og som derfor skal tolkes i sig selv. Den anden indeholder gener med fælles karakteristiske, underliggende fermenteringsprofiler.

Data brugt i metodeudviklingen er fra en mælkefermentering med *Lactococcus lactis*, hvor der blev udtaget prøver fra to parallelle mælkekulturer ved 12 tidspunkter under fermenteringen. Formålet var at undersøge, hvordan forskellige grupper af gener opfører sig under en syrningsproces. Dette er interessant, idet en given kulturs syrningskapacitet afhænger af dens respons på ændringer i det omgivende miljø. En grundig forståelse heraf er derfor helt central i forbindelse med optimering af mælkefermenteringer. Forsøgsomstændigheder og resultater fra disse data er også beskrevet i afslutningsrapporten for projektet: "Fermentering af mælk, kortlagt ud fra fysiologiske ændringer i *Lactococcus lactis*, undersøgt ved DNA-microarray og proteomanalyse".



Figur 2. a) Skematisk forløb af den udviklede metode, b) Scoreplot af en PCA foretaget på X_3, c) fermenteringsprofiler for alle gener i X_3 og d) underliggende fermenteringsprofiler estimeret med MCR.

Der arbejdes kun med gener, hvor genprodukternes funktion i cellens biologiske processer kendes (annoteringer). Datasættet har således dimensionen (2 × 1.219) × 12 svarende til de to fermenteringsforsøg, 1.219 annoterede gener og 12 tidspunkter. Data fra de to forsøg er placeret efter hinanden, således at der opnås en data-matrix med dimensionen 2.438 × 12. Figur 2.a viser et skematisk forløb af metoden. Indledningsvis fjernes outliers og inkonsistente replikater fra datamatrix X_1 ved brug af PCA og den poolede standardafvigelse mellem de to forsøg (beregnet for hvert enkelt gen). Dette giver en reduceret datamatrix, X_2, hvor der endvidere er taget gennemsnit over værdierne fra de to forsøg. Denne matrix indeholder således information fra alle gener, der er vurderet pålidelige og som efterfølgende indgår i den egentlige dataanalyse.

Ved PCA identificeres outliers. Disse er ikke forkerte observationer, men gener der er ekstreme i forhold til de øvrige. Disse gener indeholder unik variation, som har stor betydning for bakteriens genudtryk. I dette tilfælde er der identificeret to gener *uspA* og *osmC* (Figur 2.b). Disse gener er udtrykt kraftigere end alle andre gener især i slutningen af fermenteringen (Figur 2.c). Værdierne fra de resterende gener samles i datamatrix, X_3, hvorefter den underliggende struktur identificeres efter yderligere datareduktion. I dette tilfælde er det gjort med multivariat kurve resolution (MCR) (de Juan & Taulor 2003), men der findes andre metoder afhængigt af formålet med dataanalysen. Ved MCR identificeres tre komponenter (Figur 2.d). Den ene beskriver gener, der udtrykkes kraftigere fra starten af den eksponentielle fase og indtil slutningen af fermenteringen. Den anden beskriver

gener, der udtrykkes kraftigt umiddelbart før den eksponentielle fase indledes, men ellers ikke er udtrykt i nævneværdig grad. Den sidste gruppe har fra start høje mRNA-niveauer, der falder umiddelbart efter forsøgets opstart.

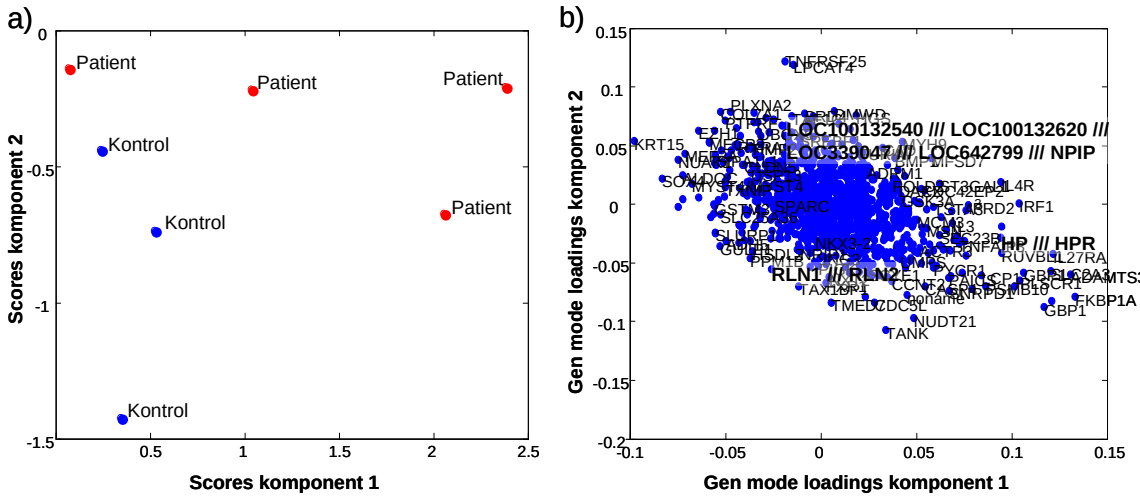
Metode 2: PARAFAC på humangenom data
En metode, der minder om ovenstående, er udviklet og afprøvet på DNA-microarray-data fremkommet ved et samarbejde med lektor Jørgen Olsen, Institut for Cellulær og Molekylær Medicin, Københavns Universitet. Datasættet omhandler genudtryk i forbindelse med kontaktallergi på mennesker. Det ønskes at opnå viden om variation i genudtrykket mellem nikkelallergikere og raske personer, samt hvilke gener der op- eller nedreguleres ved nikkel-eksponering. Denne problemstilling findes helt tilsvarende inden for mejeribrugsforskningen. Dette datasæt egner sig derfor udmærket som modeldatasæt til metodeudvikling.

Hudbiopsier er udtaget fra 12 personer eksponeret for nikkel. Microarray-analyse er foretaget til fire tidspunkter (0, 7, 48 og 96 timer). Syv af patienterne har nikkelallergi (patienter). De øvrige er kontroller. Imidlertid har det pga. mange manglende værdier været nødvendigt at reducere datasættet til syv personer, hvoraf kontrolgruppen udgør de tre og nikkelallergikere de fire. Datasættet er i øvrigt beskrevet yderligere i Pedersen et al. (2007).

Fra start er der information om over 50.000 gener/prober, hvilket er en stor udfordring i forbindelse med dataanalyse. Det skyldes dels, at støjniveauet er højt, og dels at langt de fleste variable ikke er specielt relevante for problemstillingen. Data reduceres i tre omgange. Først udvælges de gener, hvortil der er knyttet annoteringer. Dette reducerer antallet til 21.212. Derefter reduceres antallet af gener ved at beregne den poolede standardafvigelse over tid for hvert gen og for hver patientgruppe og sætte en tærskelværdi. Herved medtages kun de ca. 12.000 gener med den bedste repeterbarhed. Den sidste datareduktion går ud på at lave en partial least squares discriminant analyse (PLS-DA) hvorved de gener, der ikke synes at være relateret til forskellen mellem patienter og kontroller fjernes. Efter denne selektion er der 1.191 gener tilbage. Denne reduktion af data var ikke mulig i forbindelse med data fra mælkesyrefermenteringen, idet der kræves to grupper for at kunne lave en PLS-DA. Man kunne således for humangenomdata også have udeladt at benytte den poolede standardafvigelse, ligesom man kunne have udeladt PLS-DA og brugt en lavere tærskelværdi ved udvælgelse af gener ud fra den poolede standardafvigelse.

Data arrangeres i en trevejsstruktur med personer i en retning, gener i en anden og tidspunkter i en tredje retning, hvilket giver dimensionerne $7 \times 1.191 \times 4$. Med syv individer og fire tidspunkter skulle der være udtaget i alt 28 hudbiopsier for at have et fuldfaktorforsøg. Af forskellige årsager er der imidlertid kun data for 22 biopsier, og en del datapunkter har derfor værdien "missing", hvilket kan have betydning for den efterfølgende dataanalyse. Imidlertid er PARAFAC (Bro 1997) i stand til at håndtere og modellere data, selv når flere datapunkter har manglende værdier. Resultatet er, at en tokomponent model synes optimal. Figur 3 viser scores og loadings for prøve og gen/probe-mode. Der ses en forskel mellem kontrolgruppen og patientgruppen i de første to komponenter. Man burde derfor kunne sammenligne dette med loadings, men disse indeholder stadig mange variable og må reduceres, inden konklusioner kan drages. Dette kunne eksempelvis gøres ved at ændre tærskelværdien for repeterbarheden eller ved yderligere variabelselektion. Et andet forhold, der komplicerer tolkningen af dette datasæt, er at de tilknyttede annoteringer har en hierarkisk struktur. Ideelt set burde man kunne genfinde denne hierarkiske struktur i loadingplottet, men så simpelt er det desværre ikke. Dels pga. målefejl, men også fordi den hierarkiske struktur ikke er ren top-down, men også indeholder en række forgreninger på

tværs, som vanskeliggør entydige definitioner af hierarkiets niveauer. Det kunne være en mulighed at reducere data enten efterfølgende eller som gjort ovenfor. Endvidere kunne det være relevant at analysere data på en måde, der specifikt tager højde for den hierarkiske struktur. Arbejdet med at raffinere denne metode pågår således stadig.

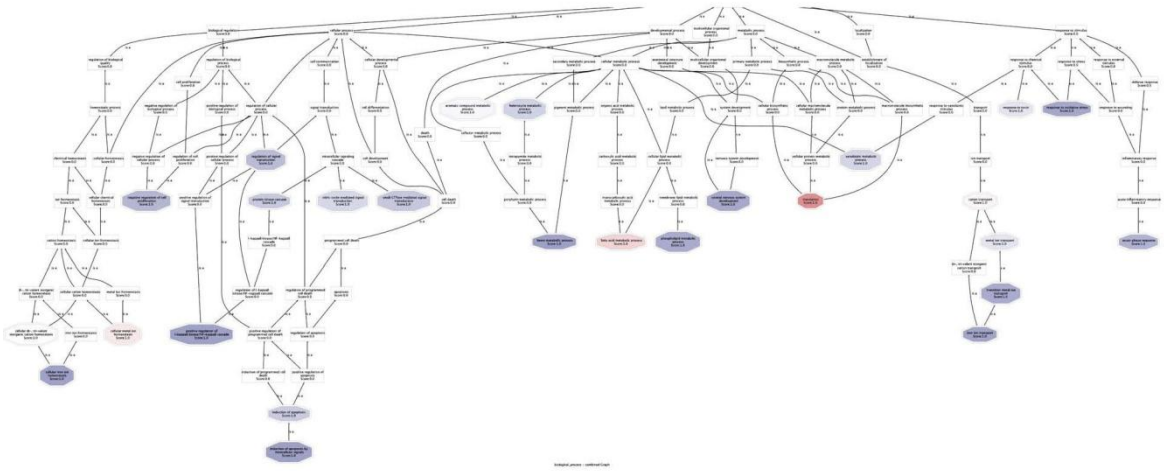


Figur 3. Resultater fra en PARAFAC-model med to komponenter a) Scores for komponent 1 og 2 og b) Loadings (gen/probe-mode) for komponent 1 og 2.

Både denne metode og metoden indeholdende PCA og MCR giver en identifikation af de underliggende profiler. Den største forskel er, at den ene håndterer multivariate data og den anden multi-vejs data. I datasættet indeholdende målinger fra en mælkesyrefermentering blev der taget gennemsnit over målinger fra to forsøg. Dette er også muligt at gøre for de analyserede humangenomdata. Det forventes. at resultaterne vil stemme nogenlunde overens med ovenstående. En anden forskel på de to metoder er identifikationen af unikke gener. Der var ingen unikke gener for humangenomdata, men hvis det havde været tilfældet kunne de have været identificeret på en lignende måde ligesom yderligere datareduktion kunne være foretaget, så PARAFAC-modellen kun blev udviklet ud fra fx 100 prøver.

Metode 3: Funktionel multivariat analyse

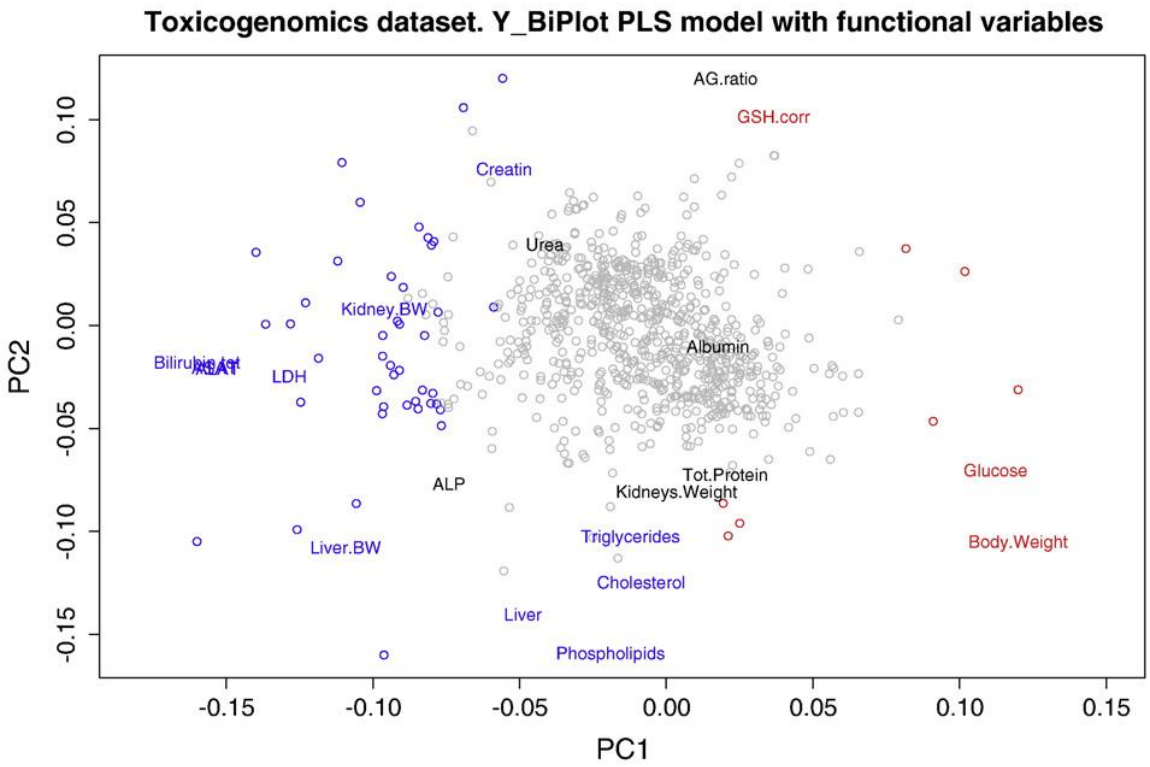
Den sidste af de udviklede fremgangsmetoder opstod via et samarbejde med Ana Conesa, Bioinformatics Department, Centro de Investigación Principe Felipe, Valencia, Spanien (Conesa *et al.* 2008). Metoden baserer sig på, at der til en væsentlig del af proberne er knyttet annoteringer, og disse inddrages i højere grad end de foregående metoder til at komprimere data. Annoteringerne er ekstern information, der fortæller noget om, hvordan variablene funktionelt er knyttet sammen. Første skridt er at strukturere data efter annoteringerne. Her er Gene Ontology's bibliotek (http://geneontology.org) benyttet, da det er det største mht. genfunktion. Annoteringerne betegnes derfor også GO-termer. Pga. strukturen i Gene Ontology's annoteringer vil et gen, der har en annotering, der findes i den nedre del af strukturen i figur 4 automatisk også have samtlige annoteringer fra de højere lag i denne gren. Listen over repræsenterede GO-termer reduceres ved at fjerne termer, der repræsenterer samme gruppe af gener som dens underliggende forgrening(er), dvs. at det primært er lavereliggende termer, der udvælges til videre analyse.



Figur 4. Gene Ontology Direct Acyclic Graph (DAG) for en gruppe udvalgte GO-termer fra grenen "Biological Process" til illustration af den hierarkiske struktur i Gene Ontology annoteringerne. I hver boks er angivet en funktionel annotering. Jo længere ned i hierarkiet, jo mere specifik en proces. De farvede heksagonale termer er udvalgt.

For hver af GO-termerne laves et datasæt bestående af genekspressiondata for de gener, der har tilknyttet denne annotering. Ved at lave PCA på hvert af disse datasæt repræsenteres data nu ved 1-2 nye funktionelle variable. For eksempel viste Conesa *et al.* (2008), at et datasæt, der indledningsvis indeholdt resultater fra 2.665 gener, blev reduceret til først 1.140 GO-termer og siden til 823 funktionelle komponenter. Det er herefter muligt at identificere signifikante funktionelle variable med partial least squares regression (PLS) og relatere disse til relevante responsvariable. Et eksempel på dette er vist i figur 5, der illustrerer, at de funktionelle variable markeret som røde eller blå (annotering ikke vist) er relateret til responsvariablene ligeledes markeret som røde eller blå. Endvidere kan man ved at gå baglæns gennem den foretagne analyse identificere interessante GO-termer, og derefter kan de mest relevante gener findes.

Metoden integrerer således genudtryk, funktionel annotering og fænotype-karakteristika i én analyse og giver en direkte sammenhæng mellem genfunktionen og responsvariablen. Fordelen er, at gener, der ikke er vigtige, er reduceret betydeligt i antal. Samtidig er informationen udtrykt i en reduceret form som scores. Ideelt set vil hver funktionel gruppe kun blive repræsenteret én gang, og alle grupper vil derfor have en chance for at påvirke analysen. Et vigtigt element i denne metode er netop udnyttelsen af ekstern information til at reducere informationsmængden. Variabelselektion kan være et vanskeligt element, når mængden af irrelevante variable er stor og antallet af observationer få. Det skyldes, at variabelselektionen ofte på den ene eller anden måde er baseret på korrelationer, og med få observationer vil der indimellem tilfældigt findes gode korrelationer. Når der er rigtig mange variable, kan dette forekomme mange gange.



Figur 5. Eksempel på resultat af funktionel ekspressionsanalyse (Conesa et al. 2008). Figuren er et bi-plot fra PLS med funktionelle variable (o) som prediktorvariable og en række fysiologiske variable (angivet med navn) som responsvariable. Grå/sort er ikke-siknifikant. Heraf ses, hvilke funktionelle variable, der er relateret til hvilke responsvariable.

Syrningskurver

Analyse af syrningskurver er inddraget som en del i et større og fra start planlagt samarbejde med det MFF-støttede projekt: "Fysiologisk status og genekspression i starterkulturer som indikator for modningsforløbet i ost". Nærværende projekt har været inddraget i forbindelse med forsøgsplanlægning samt indledende dataanalyse omkring form af syrningskurve som funktion af behandling og podemængde. Det er hypotesen, at en multivariat tilgang ved at arbejde med hele syrningskurven kan give en mere nuanceret indsigt i, hvordan forskellige parametre påvirker syrningen.

Syrningskurver opnået fra fermenteringer med forskellige køle- og/eller frysebehandling blev analyseret med PCA- og PLS-regression. Resultaterne fra de indledende forsøg antydede, at podemængden kunne have stor indflydelse på formen af kurven, men da denne podemængde og behandlingsfaktoren ikke var helt adskilt i forsøgsdesign var det ikke muligt at drage nærmere konklusioner på det foreliggende grundlag, og nye forsøg er igangsat. Det er aftalt, at i det omfang, der er ressourcer til det, vil nærværende projekt efter sin afslutning fortsat assistere, når der foreligger relevante data omkring syrningskurverne.

Samlet konklusion

Projektet har udviklet tre fremgangsmetoder til analyse af DNA-microarray-data ved at benytte basale kemometriske teknikker. Den ene metode separerer og beskriver unikke gener og gener med en fælles underliggende struktur vha. PCA og MCR. Dette er illustreret på et datasæt fra en mælkesyrefermentering. Den anden metode benytter PARAFAC og

inddrager derfor trevejs-strukturen i de analyserede humangenomdata. Den sidste metode udnytter, at nogle gener beskriver samme funktion i cellen, hvilket bruges til at reducere datasættet via PCA, således at hvert fænomen kun beskrives én gang. Dette gør at data efterfølgende bliver lettere at analysere.

Referencer

Bro R. (1997) PARAFAC. Tutorial and applications. *Chemometrics and Intelligent Laboratory Systems*, 38, 149-171.

Conesa, A.; Bro, R.; García-García, F.; Prats,J.M.; Götz, S.; Kjeldahl, K.; Montaner, D.; Dopazo, J. (2008) Direct functional assessment of the composite phenotype through multivariate projection strategies. *Genomics*, 92, 373-383.

de Juan, A.; Tauler, R. (2003) Chemometrics applied to unravel multicomponent processes and mixtures. Revisiting latest trends in multivariate resolution. *Analytica Chimica Acta*, 500, 195-210.

Pedersen, M.B.; Skov, L.; Menn, T.; Johansen J.D.; Olsen, J. (2007) Expression time course in the human skin during elicitation of allergic contact dermatitis. *Journal of Investigative Dermatology*, 127, 2585-2595.

Publikationer og præsentationer

Conesa, A.; Bro, R.; García-García, F.; Prats,J.M.; Götz, S.; Kjeldahl, K.; Montaner, D.; Dopazo, J. (2008) Direct functional assessment of the composite phenotype through multivariate projection strategies. *Genomics*, 92, 373-383.

Andersen, C.M.; Schmidt, B.; Kjeldahl, K.; Kilstrup, M., Bro, R. A simplified approach for identifying unique and bulk variations in microarray data *Under udarbejdelse.*

Kjeldahl, K; Bro. R. (2008) Simpel bioinformatik. *Mælkeritidende*, 13/14, 324-326.

Poster: Karin Kjeldahl, Rasmus Bro,"Simple bioinformatics". Vist på "The 1st Arla Foods Research Seminar", 6. december 2007.

Forskeruddannelse

Ph.d.-projekt, Karin Kjeldahl: Mathematical chromatography of complex biological samples.

Samarbejdsrelationer national og internationalt

Projektet har samarbejdet med lektor Jørgen Olsen, Institut for Cellulær og Molekylær Medicin, Københavns Universitet i forbindelse med analyse af humangenomdata.

Projektet har endvidere samarbejdet med Ana Conesa, Bioinformatics Department, Centro de Investigacion Principe Felipe, Valencia, Spanien i forbindelse med udvikling af en metode til analyse af genudtryksdata, der udnytter sammenhængen mellem gener med samme funktion for at opnå den bedste prædiktion af fænotype og samtidig give en forståelse for sammenhængen mellem disse.

Resultaternes praktiske og videnskabelige betydning for mejeribruget

Projektet har givet en uddybning af resultaterne allerede opnået i de omtalte MFF-støttede projekter ved at uddrage yderligere information fra data. Dette kan bruges i forbindelse med optimeringen af osteproduktionen for at sikre optimal syrningskapacitet etc.

Endvidere kan resultaterne opnået i dette projekt benyttes i andre sammenhænge både inden for mejeribranchen men også mange andre steder. Projektet har vist, hvordan man kan håndtere komplicerede datastrukturer som genudtryksdata opnået fra DNA microarrays på en relativ simpel måde. Metoderne vil således kunne anvendes til nye lignende forsøg og derved sikre at den relevante information findes. Der er potentiale for at analysere yderligere syrningskulturer eller samme syrningskultur under andre forsøgsbetingelser, hvilket kan være en vigtig parameter i forbindelse med produktudvikling og procesoptimering.

Relationer til andre/nye mejerirelaterede samarbejdsprojekter

Som nævnt har projektet samarbejdet med to andre MFF-finansierede projekter: "Starterkulturers fysiologiske og genetiske status som indikator for modningsforløb i ost" og "Fermentering af mælk kortlagt ud fra fysiologiske ændringer i *Lactococcus lactis*, undersøgt ved DNA-microarray og proteomanalyse".

MEJERIFORENINGEN

Mejeribrugets ForskningsFond
Frederiks Allé 22 · DK-8000 Århus C
Tel 8731 2000 · ddb@mejeri.dk
www.mejeri.dk/forskning