

# Multivariate tiltag i proteomanalyse

I den postgenome æra er fokus nu rettet mod proteomanalyse, der giver de bedste forudsætninger for kortlægning af cellulære mekanismer i biologiske processer. Håndtering af store mængder data kan dog besværliggøre processen. Multivariat analyse kan være løsningen på det problem

Af David Mark Gottlieb, Statens Serum Institut, Jakob Schultz, Foss Analytical AB, Susanne W. Bruun, Susanne Jacobsen og Ib Søndergaard, BioCentrum-DTU

Proteomanalyse er en krævende proces, bl.a. pga. de store mængder data der genereres. Resultaterne vurderes ofte subjektivt. Derved risikerer man forudindtagne konklusioner, idet man ved visuel inspektion meget sjældent kan skabe sig et overblik over sine data. Vi har implementeret multivariat analyse i proteomanalysen for at opnå en objektiv, nemmere og hurtigere behandling af data.

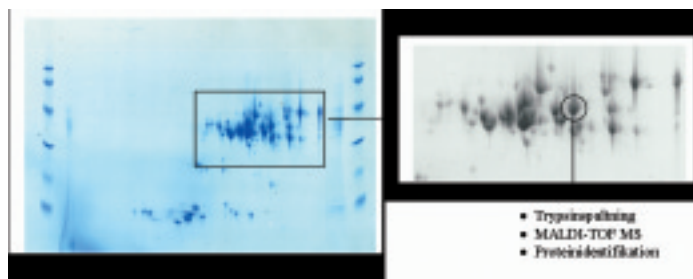
Vores interesse for anvendelse af multivariat analyse som led i proteomanalysen, er baseret på erfaringer inden for hvedes proteinkemi. Hvedesorter kan kvalitetsinddeles efter om de er velegnede til brødbagning eller dyrefoder. Gliadinene, som er de alkoholopløselige lagerproteiner fra hvedes glutenkompleks, spiller en vigtig rolle. Set ud fra en proteinkemisk synsvinkel er det interessant at kunne redegøre for kvalitetsforskelle i hvedesorter.

De ekstraherede gliadiner er først blevet separeret vha. to-dimensional polyacrylamid gelelektroforese (2-DE) og herefter sammenlignet gel for gel. Der er således fundet enkelte proteiner, der kan have betydning for hvedesorters kvalitet.

Fremgangsmåden i proteomanalysen kan være særdeles omstændig, og behovet for at kunne spotte interessante proteiner på en hurtigere og nemmere måde lægger op til anvendelsen af multivariat analyse.

## Proteomanalysens styrker og svagheder

Formålet med proteomanalyse er at fokusere på proteiner, der er involveret i en bestemt biologisk funktion. Første skridt er



Figur 1. 2-DE-separation af alkoholopløselige hvedeproteiner. Interessante proteiner identificeres efter trypsinspaltning med MS og søgning i databaser. Det er et tidskrævende arbejde, når denne proces skal gentages for alle proteinpletter på gelen, og ikke mindst når mange geler skal sammenlignes.

som regel en ekstraktion af proteiner, efterfulgt af en separation med høj opløselighed. Ved 2-DE adskilles proteinerne ift. deres isoelektriske punkt (pI) i første dimension på gelen, efterfulgt af en adskillelse ift. proteinernes molekylvægt i anden dimension på gelen. Man får hermed en unik kortlægning af de se-

parerede proteiner, hvilket giver mulighed for at karakterisere hvert enkelt protein.

I den »klassiske« proteomanalyse påbegynder man herefter en proces for at finde proteiner med særlig interesse, jf. figur 1. Disse proteiner skæres ud fra gelen og undergår enzymatisk spaltning, som oftest med enzymet trypsin. Man får derved en række peptidrester, der efterfølgende kan undersøges for molekylvægt ved massespektrometri. Herved har man et unikt »fingeraftryk« af proteinet, som kan identificeres ved søgning i databaser på internettet.

Fremgangsmåden er effektiv, hvis man ved, hvor i sine data man skal lede. Problemet opstår ofte mellem separations- og karakteriseringsfasen, hvor mængden af data bliver uoverskuelig, og sammenligningsgrundlaget forsvinder. Det er meget arbejdskrævende at skulle sammenligne mange geler eller massespektre, og det medfører subjektive vurderinger (du ser hvad du vil se), reproduktionsvanskeligheder etc. [1].

## Styrkerne ved anvendelse af multivariat analyse

Multivariat analyse bygger på statistiske og matematiske metoder og omhandler analyse af data med mange vigtige typer variation. Man kan tale om to grundlæggende principper ved udførelse af multivariat analyse: 1) Beskrivelse af et eksperiment før analyse (planlægning) og 2) reduktion af et givent problem under og efter analyse (modellering) [2].

I traditionelle statistiske termer formulerer man først en hypotese og udfører derefter eksperimenter for at bekræfte eller afkræfte denne hypotese (deduktiv analyse). Med multivariat analyse forholder det sig stik modsat. Her er tale om en induktiv analyse, hvor en hypotese opsættes, efter at man har udført indledende computereksemperimenter på sine data. Explorativ dataanalyse er et vigtigt element, der giver forskeren mulighed for at tage en række vigtige forbehold, før analysen igangsættes.

Der findes mange teknikker inden for multivariat analyse. Her nævnes kun to, hvoraf den ene vil blive fremhævet, nemlig *principal component analysis* (PCA), der er en af grundstenene i multivariat analyse [3].

Separationsteknikker som gelelektroforese og massespektrometri skaber mange variable og gør det svært at sammenligne flere objekter. Meningen med PCA er at nedbryde det multidimensionelle koordinatsystem og genopføre et nyt og mere struktureret koordinatsystem.

Man ønsker altså at danne et rearrangeret multidimensionalt rum med principielle komponenter, baseret på en bi-lineær model af den originale datamatrix X.

X nedbrydes til en strukturel del og en fejldele. De principielle

komponenters forhold til objekterne (data-rækkerne,  $t_i$ ) kaldes *scores* og variableerne (data-kolonnerne,  $p_i$ ) kaldes *loadings*.

Den strukturelle del (af den ustrukturerede X-matrix) består af en *scorematrix*,  $T$ , en transponeret *loadingmatrix*,  $P^T$  og fejldelen,  $E$ . Det matematiske skelet for PCA kan således sammenfattes til:

$$X = T \cdot P^T + E \Leftrightarrow$$

$$X = (t_1 \cdot p_1^T) + (t_2 \cdot p_2^T) + \dots + (t_i \cdot p_i^T) + E = PC_1 + PC_2 + \dots + PC_i + E$$

PCA bruges altså til at transformere et sæt observerede variable til et nyt sæt variable, som så er ukorrigerede ift. hinanden. De nye ukorrigerede variable repræsenteres i faldende orden af vigtighed mht. variation, hvilket betyder, at den første principielle komponent (PC) dækker så meget af variationen i datasættet som muligt, og hver efterfølgende komponent dækker så meget som muligt af den resterende variation.

Nedbrydningen af datasættet fortsætter, indtil så meget systematisk variation som muligt (eller ønsket) er forklaret. Fordelen ved PCA er, at man nu har mulighed for at koncentrere sig om to eller tre dimensioner ad gangen og dermed skabe sig et overskueligt billede af sine data.

### Eksempel på anvendelse af multivariat analyse i proteomanalyse

Baseret på proteomanalyse med 2-DE har man på BioCentrum-DTU fundet et specifikt gliadin, der kun er til stede i hvedesorter, der ikke egner sig til brødbagning (også nævnt som foderhveder). Proteinets identitet er blevet indsnævret ved N-terminal sekventering, og nærmere undersøgelser skal vise om proteinet har en inhiberende effekt på brødkvalitet [4].

Interessen har også været fokuseret på, om resultatet kunne nås på en hurtigere måde ved at implementere multivariat

analyse, før selve karakteriseringsfasen i proteomanalysen indledes. Der er udført multivariat analyse på data fra nærinfra-rød (NIR) spektroskopi, 2-DE og matrix assisted laser desorption/ionisation time-of-flight massespektrometri (MALDI-TOFMS) [1,4,5]. Her nævnes kun resultater baseret på massespektrometri.

Der indgår otte forskellige hvedesorter i datamaterialet, hvorfra gliadinerne er ekstraheret og separeret ved MALDI-TOFMS i intervallet 14-45 kDa. I alt 120 spektre (objekter), med over 8000 variable, danner grundlaget for brug af de multivariate værktøjer PCA og *interval partial least squares* (iPLS).

PCA er indledningsvis blevet brugt til explorativ analyse. Udeliggende objekter er identificeret og fjernet. Dvs. objekter der som følge af f.eks. apparaturfejl ved analysen vil have negativ indflydelse på den samlede modellering. Præprocesseringen har til formål at *aligne* data, dvs. finde en fællesnævner for data, så forudsætningerne for modellering er optimale. Når man arbejder med massespektre er det f.eks. vigtigt at udføre basislinje-korrektion.

En vigtig detalje i den explorative del af analysen er at bestemme det optimale antal principielle komponenter (PC'er), der skal medtages i modelleringen. Med 120 objekter vil man kunne lave et rearrangeret koordinatsystem bestående af 119 PC-akser. Hele ideen i PCA er dog netop at reducere antallet af PC-akser til det lavest nødvendige, så man kun fokuserer på den variation, der indeholder en skjult, men struktureret og betydningsfuld information.

Hele PCA-proceduren er en iterativ proces, så i virkeligheden er der ikke noget kvantespring fra den explorative del af analysen til selve modelleringsfasen. Man skal huske, at PCA ikke i sig selv giver noget svar på ens problemer, men *hjælper* ►

Science  
imaging  
SCANDINAVIA AB

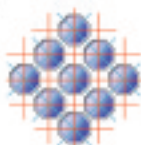
Scandinavian reseller of FUJIFILM SCIENCE IMAGING SYSTEMS

FUJIFILM  
I&I - Imaging & Information



#### THE LAS-3000 IMAGING SYSTEM CAN SEE IT.

Discovering the most subtle images in the world of bio-analysis can be a bit like trying to see a chameleon in full camouflage. But, with Fujifilm's new LAS-3000 imaging system featuring Super CCD technology, you'll have the advantage of more pixels of information and faint-light image definition. You'll have the advantage of discovery ... at its finest.



Intersecting pixel layout of new Super CCD.



#### New Super CCD

More pixels. Better resolution.

#### New Binning Mode

Unprecedented faint-image sensitivity. Improved resolution.

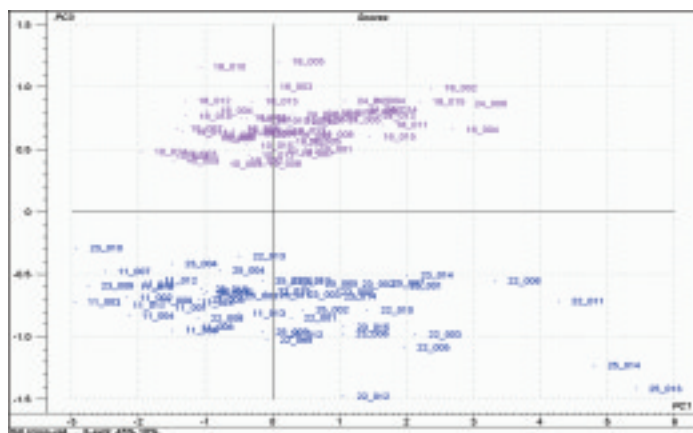
#### New Filter Options and Light Sources

Allows even more applications.

#### New User-friendly Operation

All configuration and imaging functions are controlled remotely through an easy-to-use Mac™ or Windows® interface.

email: info@scienceimaging.se web: www.scienceimaging.se phone: +46 8 55 60 40 70 fax: +46 8 55 60 40 77

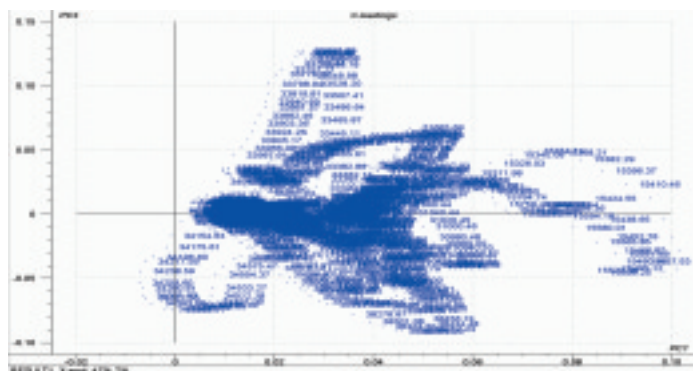


Figur 2. Scoreplot af PC1 vs. PC3. Foderhvederne (øverste gruppe) adskilles fra bagehvederne. Den tredje principielle komponent beskriver således kvaliteten af hvedesorterne i dette tilfælde.

med at belyse dem i form af skjulte strukturer, dvs. er hypotese-genererende.

Ved modellering af de MS-baserede gliadindata kunne vi på baggrund af otte forskellige hvedesorter differentiere mellem kvalitet og i en vis udstrækning også mellem sorterne. Figur 2 viser scoreplottet af PC1 vs. PC3. Det fremgår af figuren, at ved sammenligning af de to PC-akser, inddeles objekterne i to grupper ift. PC3. De to grupper består af hhv. bage- og foderhveder. Men hvorfor er der en forskel? Forklaringen findes blandt de variable, der i dette tilfælde beskriver de separerede gliadineres molekylvægt.

PCA giver i en vis udstrækning også mulighed for at studere de variable i et loadingplot. Figur 3 viser loadingplottet, der korresponderer til scoreplottet i figur 2. Her ses det, at over



Figur 3. Loadingplot af PC1 vs. PC3. En mindre gruppe variable i intervallet 33,4-34,0 kDa trækker i retningen af foderhvederne, der var grupperet øverst i det korresponderende scoreplot i figur 2.

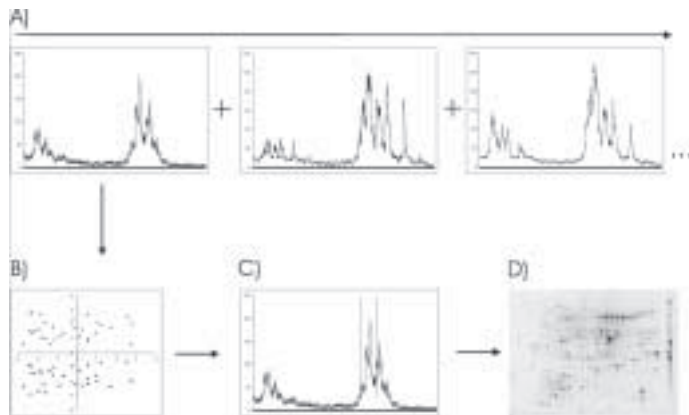
8000 variable giver et samlet uoverskueligt billede, bortset fra gruppen af variable (33,4-34,0 kDa) der trækker op ad PC3-aksen. Denne gruppe variable forklarer muligvis adskillelsen af foder- fra bagehveder i scoreplottet (figur 2).

Med iPLS har man mulighed for yderligere undersøgelse af de variable [6]. Ud fra scoreplottet i PCA, hvor hvedesorterne kunne adskilles i to kvaliteter, er det nu muligt at benytte en af kvaliteterne som reference. På baggrund af dette har vi fundet et molekylvægtinterval blandt hvedesorter, der er uegnet til brødbagning. Det inkluderer det »kvalitetsinhiberende« protein.

## Opsummering

Fremgangsmåden ved det multivariate tiltag i proteomanalyse

er illustreret i figur 4. Man fokuserer i første omgang på en bestemt separationsmetode (her MS), der underkastes multivariat analyse. Hypoteser genereres og efterprøves med PCA. Hermed skabes grobund for undersøgelse af de variable med f.eks. iPLS. I sidste ende har man fra et tilsyneladende



Figur 4. Det multivariate arbejdsflow i proteomanalysen.

(A) Massespektre danner her grundlag for multivariat analyse.

(B) PCA udføres for at sammenligne objekternes indbyrdes forhold i et scoreplot (hvert spektrum repræsenteres ved et spot).

(C) Interessante variable detekteres ved iPLS.

(D) Med den nye viden er fokus nu kendt, og den høje opløsningsevne, man opnår ved 2-DE, kan udnyttes til at udføre en optimeret proteomanalyse som illustreret i figur 1.

uoverskueligt datamateriale samlet så meget information, at man kan returnere til 2-D-gelen og udnytte dennes høje opløsningsevne til at udpege og undersøge interessante proteiner. Til forskel fra den klassiske fremgangsmåde i proteomanalyse, hvor man på en uoverskuelig måde skulle tage stilling til mange geler på én gang, kan man nu stille og roligt fokusere på et bestemt område af gelen. Multivariat analyse kan således lette en stor del af karakteriseringsarbejdet, som i en tidlig fase ofte har tendens til at blokere for det videre arbejde, der ligger og venter.

E-mail-adresse:

David Mark Gottlieb: dmgt@ssi.dk

## Referencer

- Gottlieb, D.M., Schultz, J., Bruun, S.W., Jacobsen, S., Søndergaard I., 2004. Review: Multivariate approaches in plant science. Phytochemistry, (in press).
- Martens, H., Martens, M., 2001. Multivariate Analysis of Quality – An Introduction. John Wiley & Sons Ltd, Chichester.
- Esbensen, K.H., Guyot, D., Westad, F., 2000. Multivariate Data Analysis – In Practice. Camo ASA, Oslo.
- Gottlieb, D.M., Schultz, J., Petersen, M., Nesic, L., Jacobsen, S., Søndergaard, I., 2002. Determination of wheat quality by mass spectrometry and multivariate data analysis. Rapid Communications In Mass Spectrometry 16, 2034-2039.
- Schultz, J., Gottlieb, D.M., Petersen, M., Nesic, L., Jacobsen, S., Søndergaard, I., 2004. Explorative data analysis of 2-D electrophoresis gels. Electrophoresis 25, 502-511.
- Nørgaard, L., Saudland, A., Wagner, J., Nielsen, J.P., Munck, L., Engelsen, S.B., 2000. Interval partial least squares regression (iPLS): a comparative chemometric study with an example from near-infrared spectroscopy. Applied Spectroscopy 54, 413-419.