(12) **PATENT**

(19) NO (11) **337745** (13) **B1**

NORGE (51) Int Cl.

G06F 19/24 (2011.01)
G06K 9/62 (2006.01)
G01N 30/86 (2006.01)
G06F 19/18 (2011.01)

Patentstyret

(21)	Søknadsnr	20084615	(86)	Int.inng.dag og søknadsnr	2007.03.26 PCT/US2007/007467
(22)	Inng.dag	2008.10.31	(85)	Videreføringsdag	2008.10.31
(24)	Løpedag	2007.03.26	(30)	Prioritet	2006.03.31, US, 11/396,328
(41)	Alm.tilgj	2008.12.19			
(45)	Meddelt	2016.06.13			

(73) Innehaver Biodesix Inc, P O Box 774872, US-CO80477 STEAMBOAT SPRINGS, USA
(72) Oppfinner Heinrich Roder, 22105 West Whitewood Drive, US-CO80487 STEAMBOAT SPRINGS, USA
Maxim Tsypin, 3360 Covey Circle, US-CO80487 STEAMBOAT SPRINGS, USA
Julia Grigorieva, 3360 Covey Circle, US-CO80487 STEAMBOAT SPRINGS, USA
(74) Fullmektig Zacco Norway AS, Postboks 2003 Vika, 0125 OSLO, Norge

(54) Benevnelse **Fremgangsmåte og apparat for å bestemme om et medikament vil være effektivt eller ikke for en pasient med en sykdom**

(56) Anførte publikasjoner
WO 2005098445 A2
US 2005164218 A1
US 2006029574 A1
BHANOT, G. et al., A robust meta-classification strategy for cancer detection from MS data, Proteomics 2006, Vol.6, side 592-602.
COOMANS, D. et al., Alternativ k-nearest neighbour rules in supervised patteren recognition, Analytical Chimica Acta, 1982, vol. 136, side 15-27, ISSN: 0003-2670.
CHAURAND, P. et al., Imaging mass spectrometry: principles and potentials, Toxicologic Pathology, 2005, Vol.33, side 92¿101, ISSN: 0192-6233 print / 1533-1601 online.

(57) Sammendrag

En prosess for å bestemme hvorvidt en pasient med en sykdom eller lidelse vil være mottakelig for et medikament, benyttet for å behandle sykdommen eller lidelsen, inkluderende å oppnå et testspektrum produsert ved et massespektrometer fra et serum produsert fra pasienten. Testspektrumet kan bli prosessert for å bestemme et forhold til en gruppe av klassemerkede spektre produsert fra respektivt serum fra andre pasienter som har det samme eller liknende kliniske trinn av sykdom eller lidelse og kjent for å ha respondert eller ikke-respondert til medikamentet. Basert på forholdet av testspektrumet vil gruppen av klassemerkede spektre, kan en bestemmelse bli gjort i forhold til hvorvidt pasienten vil være mottakelig for medikamentet.

Oppfinnerne av den foreliggende oppfinnelsen har funnet en ny metode for å bestemme om en pasient vil respondere til en behandling ved å teste pasientens biomarkører ved massespektroskopi. Som et eksempel på en utførelsesform av denne oppfinnelsen har oppfinnerne benyttet deres teknikk til en cancer, ikke-småcelle lungecancer (NSCLC).

Ikke-småcelle lungecancer er en ledende årsak av død fra cancer hos både menn og kvinner i US. Det er minst fire (4) forskjellige typer av NSCL, inkluderende adenokarcinom, platecelle, stor celle og bronkiealdeolar karcinom. Platecelle (epidermoid) karcinom av lungen er en mikroskopisk type cancer typisk relatert til røyking. Adenokarcinom av lungen utgjør over 50% av lungecancertilfeller i USA. Denne canceren er mer alminnelig hos kvinner og er fortsatt den hyppigste typen sett hos ikke-røykere. Stor cellekarcinom, spesielt dem med neuroendokrine trekk, er alminnelig assosiert med en spredning av tumorer til hjernen. Når NSCLC går inn i blodstrømmen kan det spre seg til forskjellige steder slik som leveren, ben, hjerne og andre plasser i lungen.

Behandling av NSCLC har vært relativt dårlig over årene. Kjemoterapi, hovedbehandlingen for avanserte cancere, er bare marginalt effektivt, med unntaket av lokaliserte cancere. Mens kirurgi er den mest potensielle kurerende terapeutiske muligheten for NSCLC, er det ikke alltid mulig avhengig av trinnet av canceren.

Nyere metoder for å utvikle anti-cancermedikamenter for å behandle NSCLC pasienten fokuserer på å redusere eller eliminere muligheten for cancerceller til å vokse og dele seg. Disse anti-cancermedikamentene er benyttet for å ødelegge signalene til cellene for å fortelle dem at de enten skal vokse eller dø. Normalt er cellevekst tett kontrollert ved signalene som cellene mottar. I cancer går imidlertid disse signalene feil og cellen fortsetter å vokse og dele seg på en ukontrollert måte, som derved danner en tumor. En av disse signalveiene begynner når et kjemikalie i kroppen, kalt epidermal vekstfaktor, binder til en reseptor som er funnet på overflaten av mange celler i kroppen. Reseptoren, kjent som den epidermale vekstfaktoriserte reseptoren (EGFR) sender signaler til cellene, gjennom aktiveringen av et enzym kalt tyrosin kinase (TK), som er funnet i cellene, signalene er benyttet for å telle cellene at de skal vokse og dele seg.

To anti-cancermedikamenter som ble utviklet og preskribert til NSCLC pasientene er kalt gefitinib (varenavn "Iressa") og erlotinib (varenavn "Tarceva"). Disse anti-cancermedikamentene målsøker EGFR veien og har vist seg lovende i å være effektive mot å behandle NSCLC cancer. Iressa inhiberer enzym tyrosin kinasen som er til stede i

lungecancerceller, så vel som andre cancere i normalt vev, og som viser seg å være viktig for veksten av cancerceller. Iressa har blitt benyttet som et enkelt middel for behandlingen av NSCLC som har utviklet seg etter, eller sviktet å respondere til, to andre typer av kjemoterapier.

Responsrater har imidlertid vært mellom 10% og 20% i kaukasiske populasjoner, og har ført til at Federal Drug Administration (FDA) i 1995 trakk tilbake støtte for benyttelsen av Iressa som en andrelinjebehandling. Overraskende har responsraten i Asia vært betraktelig høyere og Iressa er fortsatt benyttet. Traceva er fortsatt godkjent og rutinemessig gitt til pasienter, men har fortsatt responsratebekymringer. Imens det viser seg at Iressa og Traceva har muligheten til å være effektive i noen pasienter, er kanskje ikke generiske medikamenter effektive i behandling av alle pasienter. Det kan være mange faktorer involvert i en pasients mulighet til å respondere til disse medikamentene som nå er ukjent. Om en bestemmelse av faktorer som kan bli benyttet for å forutsi effektiviteten av en NSCLC pasient til å respondere til disse anti-cancermedikamentene, kan imidlertid FDA tillate disse anti-cancermedikamentene å bli preskribert til de pasientene som har tilstander som indikerer at de vil være mottakelige for disse medikamentene. Leger kan så preskribere disse medikamentene til de pasientene beregnet til å respondere til anti-cancermedikamentene med kunnskapen om deres pasienter vil være mottakelige for behandlingene.

WO 2005/098445 A2 beskriver en fremgangsmåte for anvendelse innen diagnostikk av lungekreft. Fremgangsmåten omfatter å ta en biologisk prøve fra pasienten som antas å lide av lungekreft og detektere minst en proteinbiomarkør i nevnte prøve. Videre kan fremgangsmåten også anvendes for å bestemme medikament respons status på en spesiell medisinsk behandling.

US 2005/164218 A1 beskriver en fremgangsmåte for å bestemme om en pasient som lider av for eksempel NSCLC vil respondere positivt på en behandling som omfatter inhibering av EGFRreaksjonsveien, så som medikamentene Gefitinib og Erlotinib. Fremgangsmåten omfatter bestemmelse av nivåene av RNA-transkripter i en biologisk prøve fra en pasient. Nivået av RNA-transkriptene kan bestemmes ved RT-PCR, immunohistokjemi eller proteomikk, kravsettet.

Sammendrag

For å overvinne problemet av de lave ratene av behandlingssuksess ved å benytte medikamenter tilveiebringer prinsippene av den foreliggende oppfinnelsen en

diagnostisk test for å bestemme hvorvidt en pasient vil respondere til disse medikamentbehandlingene. Bestemmelsen er gjort ved å detektere differensierende topper av et spektrum produsert ved et massespektrometer fra serum ekstrahert fra en pasients blod. Biomarkører er målbare og kvantifiserbare biologiske parametere som kan bli evaluert som en indikator for normale eller unormale biologiske prosesser eller patogene prosesser. Massespektrometre produserer et spektrum som har bestemte topper som kan bli benyttet for å sammenlikne med spektret produsert fra serum fra pasienter som var mottakelige og ikke-mottakelige for medikamentbehandlingene. Det er ofte ikke nødvendig å virkelig bestemme hvilken kjemisk forbindelse som er lokalisert i toppen. Spektrum i seg selv er en verdifullt fingeravtrykk som kan karakterisere behandlingspotensialet for medikamentet i en spesifikk pasient. Noen utførelsesformer av den foreliggende oppfinnelsen omslutter å isolere materialet som er i toppene og bestemme hvilket materiale som er forhøyet eller forminsket i prøven.

I et aspekt vedrører foreliggende oppfinnelse en fremgangsmåte for å bestemme hvorvidt det vil være sannsynlig at en pasient som lider av ikke-småcelle lungecancer vil ha fordel av behandling med gefitinib eller erlotinib som målsøker en epidermal vekstfaktor reseptor vei, eller hvorvidt det ikke vil være sannsynlig at nevnte pasient har fordel av behandling med gefitinib eller erlotinib, kjennetegnet ved at fremgangsmåten omfatter følgende trinn:

a) oppnå et massespektrum fra en blodbasert prøve fra pasienten;

b) utføre ett eller flere forhåndsdefinerte pre-prosesseringstrinn på massespektrumet oppnådd i trinn a);

c) oppnå integrerte intensitetsverdier for valgte trekk i nevnte spektrum på ett eller flere forhåndsdefinerte m/z områder etter at pre-prosesseringstrinnene av massespektrumet i trinn b) har blitt utført;

d) benytte verdiene oppnådd i trinn c) i en klassifiseringsalgoritme ved å bruke et treningssett som omfatter klassemerkede spektre oppnådd fra blodbaserte prøver fra andre pasienter med ikke-småcelle lungecancer for å bestemme om det er sannsynlig eller ikke sannsynlig at pasienten drar fordel av gefitinib eller erlotinib;

hvori nevnte ett eller flere forhåndsdefinerte m/z områder omfatter ett eller flere m/z områder valgt fra gruppen av m/z områder bestående av:

5732 til 5795

5811 til 5875

6398 til 6469

11376 til 11515

11459 til 11599

11614 til 11756
11687 til 11831
11830 til 11976
23183 til 23525
23279 til 23622
65902 til 67502.

I et ytterligere aspekt vedrører foreliggende oppfinnelse et apparat konfigurert for å bestemme hvorvidt det vil være sannsynlig at en pasient som lider av ikke-småcelle lungecancer vil ha fordel av behandling med gefitinib eller erlotinib som målsøker en epidermal vekstfaktor reseptor vei, eller hvorvidt det ikke vil være sannsynlig at nevnte pasient har fordel av behandling med gefitinib eller erlotinib, kjennetegnet ved at det omfatter:
en lagringsanordning som lagrer et massespektrum av en blodbasert prøve fra pasienten, og
en prosessor som utfører programvareinstruksjoner konfigurert for å:
a) oppnå integrerte intensitetsverdier av trekk i nevnte massespektrum ved et eller flere m/z områder, hvor m/z områdene er valgt fra gruppen av m/z områder bestående av:
5732 til 5795
5811 til 5875
6398 til 6469
11376 til 11515
11459 til 11599
11614 til 11756
11687 til 11831
12375 til 12529
23183 til 23525
23279 til 23622
65902 til 67502; og

b) benytte en klassifiseringsalgoritme som opererer på verdiene av trekkene i spektrumet ved det valgte ene eller flere m/z områder og ved å benytte et treningssett omfattende klassemerkede spektre oppnådd fra blodbaserte prøver fra andre pasienter med ikke-småcelle lungecancer for å bestemme om det er sannsynlig eller ikke sannsynlig at pasienten drar fordel av gefitinib eller erlotinib som målsøker en epitel vekstfaktor reseptorvei.

Det beskrives en prosess for å bestemme hvorvidt en pasient med en sykdom eller lidelse vil være mottakelig for et medikament eller behandling benyttet for å behandle sykdommen eller lidelsen. Prosessen inkluderer å oppnå et testspektrum produsert ved et massespektrometer fra et serum fra en pasient. Testspektrumet kan bli prosessert for å bestemme et forhold til en gruppe av klassemerkede spektre produsert fra respektivt serum fra andre pasienter fra samme eller liknende klinisk trinn av sykdom eller lidelse og kjent for å ha respondert eller ikke respondert for medikamentet. Basert på forhold av testspektrumet mot gruppen av klassemerkede spektre, kan en bestemmelse bli gjort angående hvorvidt pasienten vil være mottakelig for medikamentet eller behandlingen. I prosessering av testspektrumet kan bakgrunnsreduksjon, normalisering og oppstilling av testspektrumet bli utført for å bedre matche testspektrumet med gruppen av klassemerkede spektre, som har blitt prosessert på den samme måten eller på liknende måte. Ved prosessering av råspektret for å generere de klassemerkede spektrene, kan bestemmelsen av hvorvidt medikamentet vil være effektivt bli gjort uavhengig av de bestemte klinikkene og massespektrometrene benyttet for å prosessere serumet fra pasienten.

Det beskrives også systemer for å bestemme hvorvidt en pasient vil være mottakelig for et medikament eller behandling. Systemene kan inkludere en lagringsanordning konfigurert for å lagre et testspektrum produsert ved et massespektrometer fra et serum produsert fra en pasient med en sykdom eller lidelse og en gruppe av klassemerkede spektre produsert fra respektivt serum fra andre pasienter med det samme eller liknende klinisk trinn av sykdom eller lidelse og kjent for å ha respondert eller ikke respondert ved et medikament eller behandling. Slike systemer kan videre inkludere en prosessor i kommunikasjon med en lagringsanordning, hvor prosessoren iverksetter programvare til å (i) oppnå et testspektrum produsert ved et massespektrometer fra et serum produsert fra en pasient som har en sykdom eller lidelse, (ii) prosessere testspektrumet for å bestemme et forhold til en gruppe av klassemerkede spektre produsert fra respektivt serum fra andre pasienter som har det samme eller liknende klinisk trinn av en sykdom eller en lidelse og kjent for å ha respondert eller ikke respondert for et medikament eller behandling, (iii) bestemme, basert på forholdet av testspektrumet til gruppen av klassemerkede spektre, hvorvidt pasienten vil være mottakelig for medikamentet. Systemet kan være i kommunikasjon med et nettverk, slik som internett, for å kommunisere med laboratorier og klinikker som meddeler testspektret for testing. Bestemmelsen av forholdet av testspektret til gruppen av klassemerkede spektre kan inkludere å utmate en indikator eller klassemerket representativ av potensiell mottakelighet av pasienten til medikamentet eller behandlingen. Indikatoren kan være

positiv, negativ eller ufullstendig, slik at en medisinsk fagperson kan bestemme hvorvidt det skal preskriberes medikament eller behandling. Sykdommen eller lidelsen kan være cancer. Cancertypen kan være ikke-småcelle lungecancer. Systemet kan bli benyttet for å bestemme hvorvidt medikamentet gefitinib og/eller erlotinib vil være effektivt i behandling av ikke-småcelle lungecancer pasienter.

Fig. 1 er et blokkdiagram av et eksemplarisk forhold mellom et laboratorie testprosesseriengssenter, cancer forskningsklinikker og cancer pasientklinikker;

Fig. 2 er et blokkdiagram av et eksemplarisk system for å kommunisere og produsere informasjon mellom laboratoriet testproduseringssentre, cancer forskningsklinikkene og cancer pasientklinikkene i fig. 1;

Fig. 3 er et flytdiagram av en eksemplarisk arbeidsflytprosess for å utvikle en tekst for å bestemme hvorvidt en cancerpasient vil være mottakelig for et anti-cancermedikament i overensstemmelse med prinsippene av den foreliggende oppfinnelsen;

Fig. 4 er et bilde av et eksemplarisk gelplot av alle spektre benyttet i en testutvikling;

Fig. 5 er et histogram som viser et eksemplarisk sett av datapunkt resultater fra et spektrometer som har bakgrunns og signalkomponenter;

Fig. 6A og 6B er grafer som viser et spektrum med henholdsvis bakgrunn og uten bakgrunn etter at bakgrunnen har blitt trukket ut av spektrum;

Fig. 7A er en graf som viser multiple spektre som blir fullstendig pre-prosessert for å forenkle sammenlikning av spektrene som vist i fig. 7B;

Fig. 8A og 8B er grafer som viser multiple prøvespektre som blir oppstilt;

Fig. 9 er en graf av en eksemplarisk prosess for å velge et trekk ved å lokalisere en topp alminnelig i mer enn x spektre som har en bestemt bredde;

Fig. 10 er en graf representativ for de gjennomsnittlige spektrene i kliniske grupper PD, PD-tidlig, PR, SD-kort og SD-lang beregnet i gjennomsnitt av over alle de tilgjengelige testutviklingsprøvene i deres respektive gruppe;

Fig. 11 er en graf som viser en eksemplarisk gruppe av klassemerkede spektre indisier representative for to forskjellige klasser av sykdomsprogresjon og testspektrum indisier som skal bli klassifisert;

Fig. 12 er et Kaplan-Meier plot av testdata som overlevelsesrater og grupper av pasienter som klassifisert i overensstemmelse med prinsippene av den foreliggende oppfinnelsen som oppnådd fra å benytte italienske prøver som et treningssett og japanske prøver som et testsett;

Fig. 13 et Kaplan-Meier plot av testdata som viser overlevelsesrater av grupper av pasienter som klassifisert i overensstemmelse med prinsippene av den foreliggende oppfinnelsen som oppnådd fra å benytte de japanske prøvene som et treningssett og de italienske prøvene som et testsett;

Fig. 14 er et Kaplan-Meier plot av testdata som viser overlevelsesrater av grupper av pasienter som klassifisert i overensstemmelse med prinsippene av den foreliggende oppfinnelsen som generert ved en klassifiserende algoritme for et fullstendig blindet sett av prøver; og

Fig. 15 er et blokkdiagram av en eksemplarisk prosess for å bestemme hvorvidt en cancerpasient vil være mottakelig for et anti-cancermedikament i overensstemmelse med prinsippene av den foreliggende oppfinnelsen.

Detaljert beskrivelse av figurene

Fig. 1 er et blokkdiagram av et eksemplarisk forhold mellom et laboratorie testprosesseringssenter 102, cancer forskningsklinikker 104a-104n (kollektivt 104) og cancer pasientklinikker 106a-106m (kollektivt 106). Laboratorie testprosesseringssentret 102 opererer for å prosessere tester fra cancer forskningsklinikk 104 og cancer pasientklinikk 106. I en utførelsesform er cancer forskningsklinikk 104 og cancer pasientklinikk 106 en del av den samme organisasjonen, slik som et sykehus. Cancer forskningsklinikken 104 utfører medikamentforsøk og tester for å bestemme effektivitet av enkelte medikamenter for å behandle pasienter. Pasienter med ikke-småcelle lungecancer som har gått gjennom kliniske studier og tester av ulike anti-cancermedikamenter for å kontrollere vekst og spredning av cancercellene har for eksempel forskjellige responser for anti-cancermedikamentene. Disse anti-cancermedikamentene kan inkluderer gefitinib og erlotinib og målsøker epidermal vekstfaktor reseptorveien. Under kliniske studier og ikke-kliniske studier monitorerer

cancer forskningsklinikken 104 forsiktig ulike aspekter av behandlingene, inkluderende cancertrinn, blodkomponenter, cancerprogresjon, total helse av pasient og andre faktorer indikative av pasienten for å bestemme effektiviteten av anit-cancermedikamentet.

Cancer forskningsklinikken 106 kan være enhver fasilitet som utfører kliniske studier eller på annen måte administrerer cancermedsineringer til cancerpasienter og monitorerer effektivitet av medisineringene. Cancer forskningsklinikken 104 kan ta blodprøver og prosessere dem for å produsere serum, som er blodplasma (den flytende komponenten av blod hvor blodcelle er suspendert) som har koagulasjonsfaktorer fjernet, slik som fibrin. Serumet kan bli prosessert og benyttet for å produsere spektrum ved et massespektrometer slik at biomarkører innenfor spektrumet kan bli detektert. I en utførelsesform er massespektrometeret et "time-of-flight" (TOF) massespektrometer som benytter matrise-assosiert laser desorpsjon/ionisering (MALDI). Spektrumet kan inkludere surrogatmarkører eller tapper i spektrumet (se fig. 11) indikative av bestemte kjemikalier eller materialer i serumet.

Som et resultat av massespektrometer produksjonen av spektret av pasienter, kan effektivitet av anti-cancermedikamentene som blir administrert til cancerpasienten for å produsere kliniske resultater bli registrert og observert. Laboratorie testprosesseringssentre 102 kan benytte de registrerte (kvantitative) og observerte (generelle helse) resultatene av pasientene for å bestemme klassifiseringer for hver av cancerpasientene med hensyn til hvorvidt hver er mottakelig for anti-cancermedikamentet/medikamentene.

Ved å fortsette med fig. 1, som et resultat av massespektrometer produksjonen av spektret av pasienter, kan effektivitet av anti-cancermedikamentene som blir administrert til cancerpasienten for å produsere kliniske resultater blir registrert og observert. Laboratorie testprosesseringssentret 102 mottar råspektrene med assosierte kjente kliniske resultater 108 fra cancer forskningsklinikkene og utfører en klassifisering av hvert spektrum. Klassifiseringen av hvert spektrum, beskrevet i detalj her senere, klassifiserer hvert spektrum assosiert med en cancerpasient som mottar anti-cancermedikamenter til å være mottakelige, ikke-mottakelige eller delvis mottakelige. Klassifiseringen av spektrene muliggjør laboratorie testprosesseringssentret 102 å motta testspektret 110a-110m (kollektivt 110) fra cancer pasientklinikk 106 og utføre analyse på disse testspektre 110 for å bestemme hvilken klassifisering hvert testspektrum (dvs. hver pasient) mest sannsynlig likner. I stedet for å motta råspektret kan laboratorie

testprosesseringssentret 102 alternativt motta blodprøver eller serumprøver for å prosessere og produsere råspektrene for prosessering og klassifisering.

I klassifisering av råspektrene er en bestemmelse gjort med hensyn til hvorvidt hvert spektrum er "godt" eller "dårlig" basert på hvorvidt cancerpasienten hadde en positiv respons, ingen respons eller begrenset respons til anti-cancermedikamentet. Ved å sammenlikne testspektrumet fra hver cancerpasient med de klassemerkede spektrene, kan en bestemmelse bli gjort med hensyn til sannsynligheten for at en cancerpasient hvorfra et testspektrum er generert vil ha en positiv respons til anti-cancermedikamentet. En mer detaljert beskrivelse av sammenlikningsprosessen er tilveiebrakt her senere. Med en gang laboratorie testprosesseringssentret 102 har klassifisert testspektrumet 110, og gjør eventuelt bestemmelsen med hensyn til hvorvidt cancerpasienten vil ha en positiv respons til anti-cancermedikamentet, kan klassifiseringsresultater 112a-112m (kollektivt 112) bli levert til for eksempel den respektive cancer pasientklinikken 108a. I en utførelsesform er klassifiseringsresultatene klassemerket produsert ved en klassifiseringsfunksjon som videre beskrevet her nedenfor.

Selv om vist separat kan laboratorie testprosesseringssentret 102 være del av cancer forskningsklinikker 104 eller cancer pasientklinikker 106. I en utførelsesform er laboratorie testprosesseringssentret 102 funksjonelt inkorporert i testutstyret, slik som et massespektrometer eller prosesseringssystem som opererer sammen med testutstyret. Alternativt kan funksjonaliteten være inkorporert til et computersystem eller annet prosesseringssystem som er konfigurert for å utføre de ulike prosesseringene benyttet i prosessering og klassifisering av spektrene og ikke del av eller assosiert med testutstyret. For eksempel kan computersystemet være en server operert ved laboratorie testprosesseringssentret 102, klinisk forskningsklinikk 104 og/eller cancer pasientklinikk 106.

Selv om fig. 1 beskriver cancerklinikker, burde det bli forstått at disse klinikkene kan være alminnelige klinikker eller klinikker spesifikke for en bestemt sykdom eller lidelse. Følgelig er laboratorie testprosesseringssentret 102 konfigurert for å motta og teste den bestemte sykdommen eller lidelsen som blir sendt i overensstemmelse med prinsippene av den foreliggende oppfinnelsen.

Fig. 2 er et blokkdiagram av et eksemplarisk system 200 for å kommunisere og prosessere informasjon mellom laboratorie testprosesseringssentret 101, cancer

forskningsklinikker 104 og cancer pasientklinikker 106 av fig. 1. Et laboratorie testprosesseringsenter beregningssystem 202 kan bli operert ved laboratorie testprosesseringssentret 104. Cancer forskningsklinikk servere 204a-204n (kollektivt 204) kan bli operert ved cancer forskningsklinkken 104 og cancer pasientklinikk servere 206a-206m (kollektivt 206) kan bli operert ved cancer pasientklinikkene 106. Hver av beregningssystem 202 og servere 204 og 206 kan kommunisere over nettverk 208 via digitale datapakker 209a-209b eller annen kommunikasjonsteknikk som forstått i fagfeltet. Nettverket 208 kan være internett eller annet offentlig eller ikke-offentlig kommunikasjonsnettverk.

Laboratorie testprosesseringssenter beregningssystemet 202 kan inkludere en prosessor 210 som utfører programvare 212 for prosessering av råspektrene og testspektrene for å bestemme klassifiseringer av alle eller en del derav i overensstemmelse med prinsippene av den foreliggende oppfinnelsen som beskrevet videre her nedenfor. Beregningssystemet 202 kan videre inkludere minnet 214, hvor programvaren 212 kan oppholde seg når den blir utført, tilførsel/utførelse (I/O) enhet 216, som kan utføre kommuniseringen over nettverket 208, og lagringsanordning 218 hvortil prosessoren 210 kommuniserer. Lagringsanordningen 218 kan inkludere en eller flere databaser 220a-220n (kollektivt 220) hvor råspektrene, testspektrene og andre relaterte data er lagret for å muliggjøre laboratorie testprosesseringssentret 102 og bestemme hvorvidt en cancerpasient vil være mottakelig for et anti-cancermedikament. Det burde bli forstått at lagringsanordningen 218 kan inkludere en eller flere lagringsanordninger og lokalisert i eller eksternt fra beregningssystemet 202. det burde videre bli forstått at prosessoren 210 kan inkludere en eller flere prosessorer. Videre burde det bli forstått at beregningssystemet 202 kan være direkte indirekte i kommunikasjon med nettverket 208.

I overensstemmelse med fig. 1 kan cancer forskningsklinikk serverne 204 kommunisere råspektret med assosierte kjente kliniske resultater basert på kliniske forsøk av anti-cancermedikament til laboratoriet testprosesseringssenter beregningssystemet 202. Prosessoren 210, automatisk eller halvautomatisk med assistansen ved en forsker eller på annen måte, kan utføre klassifiseringsprosessering for å klassifisere hvert råspektrum for å klassifisere råspektrene for å danne en gruppe av klassifiserte spektre. På liknende måte kan cancerpasient klinikkserverne 206 kommunisere testspektret 110 til laboratoriet for prosessoren 210 for automatisk eller halvautomatisk klassifisere testspektrene 110 for cancer pasientklinikkene 108. Laboratoriet testprosesseringssenter beregningssystemet 202 kan prosessere testspektrene 110 og kommunisere

klassifiseringsresultatet 112 (fig. 1) tilbake til cancer pasientklinikk serverne 206. Som et resultat av klassifisering av råspektrene og testspektrene 112, kan beregningssystemet 2020 lagre klassifiseringsresultater og benytte resultatene for å generere statistisk informasjon som kan bli benyttet for ulike andre formål, slik som rater for suksess og mislykkethet av anti-cancermedikamentet.

Dataanalyse spiller en sentral rolle i oppdagelsen av å detektere topper som differensierer spektre fra pasienter med forskjellig klinisk resultat og deres anvendelse enten som ledetråder for immunhistokjemiske tester eller indirekte i massespektrometri basert diagnostikk. I utviklingen av teste og analyseprosedyrer i overensstemmelse med prinsippene av den foreliggende oppfinnelsen har et integrert analysesystem inneholdende algoritmer designet for komparativ analyse av massespektret blitt utviklet. Det integrerte analysesystemet inkluderer et antall av verktøy som fremmer deteksjonen av differensierende topper i spektrene fra massespektret, imens det samtidig tilveiebringer strenge verktøy for bestemmelsen av deres signifikans og validering av resultatene.

Fig. 3 er et flytdiagram av en eksemplarisk arbeidsstrømprosess 300 for å utvikle og utføre en test for å bestemme hvorvidt en cancerpasient vil være mottakelig for anti-cancermedikament i overensstemmelse med prinsippene av den foreliggende oppfinnelse. Prosessen starter ved trinn 302 når prøver er tatt fra cancerpasienter. Avhengig av typen av cancer eller annen sykdom, kan spottet vev, cellelysater eller kuttede celler bli benyttet som prøver for å generere spektret via et massespektrometer 304. Massespektrometeret kan være et ABI Voyager, et ABI 4700, et Bruker Autoflex eller et Bruker Ultraflex massespektrometer. Andre massespektrometere kan på liknende måte bli benyttet. I tilfelle av ikke-småcelle lungecancer kan serum bli benyttet for å generere spektret, ved å benytte serum kan lungecancer pasienter i fremskredne trinn av lungecancer, hvor det er vanskelig eller umulig å ta en vevsprøve av pasienten, bli diagnostisert uten en invasiv prosedyre. I tillegg kan kroppsvæske slik som urin bli benyttet for prøver i å detektere topper i et massespektrum for å bestemme hvorvidt bestemte anti-cancermedikamenter vil være effektive i behandling av en cancerpasient med ikke-småcelle lungecancer. Ved å benytte ikke-invasive prosedyrer for å samle serum eller andre væsker, er kostnaden for diagnose signifikant lavere enn om en vevsprøve fra en lunge var nødvendig.

Generering og prosessering av serum benyttet for en teststudie kan inkludere å benytte råstoff serumprøver fra individuelle sykehus. I en utførelsesform kan råstoff

serumprøvene bli tint på is og sentrifugert ved 1500 rpm i 5 min. ved 4°C. Videre kan serumprøvene bli fortynnet 1:10, som utført ved University of Colorado Health Sciences Center (UCHSC) eller 1:5, som utført ved Vanderbilt University medical Center (VUMC), i MilliQ vann. Fortynnede prøver kan bli spottet i tilfeldig lokaliserte posisjoner på en MALDI plate i triplikat (dvs. på tre forskjellige MALDI mål). Etter at 0,75 µl av fortynnet serum er spottet på en MALDI plate kan 0,75 µl av 35 mg/ml sinapinsyre (i 505 acetonitril og 0,1% TFA) bli tilsatt og blandet ved pipettering opp og ned fem ganger. Plater kan bli tillatt å tørke ved romtemperatur. Det burde bli forstått at andre teknikker og prosedyrer kan bli benyttet for å preparere og prosessere serum i overensstemmelse med prinsippene av den foreliggende oppfinnelsen.

Massespektre kan bli ervervet for positive ioner i lineær modus ved å benytte en Voyager DE-PRO (UCHSC) eller DE-STR (VUMC) med automatisert eller manuell samling av spektrene. I en studie ble 75 (UCHSC) eller 100 (VUMC) spektre samlet fra syv (UCHSC) eller fem (VUMC) posisjoner med hver MALDI spot for å generere et gjennomsnitt på 525 (UCHSC) eller 500 (VUMC) spektre for hver serumprøve. Spektre ble eksternt kalibrert ved å benytte en blanding av proteinstandarder (insulin (bovin), tioredoxin (E.coli) og apomyglobin (hest)). For formål av validering ble tre replikater av den samme prøven kjørt for alle prøveeksemplarer resulterende i en total på 717 spektre (329 prøveeksemplarer ganger 3) underlagt analyse for den foreliggende studien.

I utføring av dataanalysen er det generelt akseptert at cancerøse celler har forskjellig ekspresjonsnivå av spesifikke proteiner som er forskjellig fra normale celler. Forskjellige trinn av sykdom er ledsaget ved endringer i spesifikke proteiner, f.eks. endringer i ekspresjonsnivået av cellebindende proteiner i tilfelle av metastatisk cancer. I tilfelle av serumprøver, og for å skjelne serumtesting fra vevsprøvetesting, er det ikke sannsynlig at direkte tumorekskresjoner er målt på grunn av fortynning av disse ekskresjonene i blodet. Differensieringstoppene i serum (eller andre kroppsvæsker) prøver oppstår sannsynligvis på grunn av en vertsrespons reaksjon avhengig av sykdomstilstanden, slik som autoimmune reaksjoner. På denne måten kan det bli forventet at tester basert på vevsprøver er sterkt spesifikke, men ikke nødvendigvis veldig signifikante, og serumbaserte massespektrometer tester burde være sterkt signifikante, men ikke så spesifikke. Dette er vist ved resultatene presentert her nedenfor. Ved å detektere differensieringstopper i spektrene, kan korrelasjon i endringer med klinisk relevante spørsmål bli utført. For å generere differensieringstopper i spektrene av verdi, uavhengig av deres videre anvendelse, enten direkte eller som et diagnostisk verktøy eller som ledetråder for immunhistokjemisk basert testing, kan de

følgende spørsmålene bli analysert under oppdagelsesprosessen av differensieringstopper, inkluderende datanalysetrinnet:

Reproduserbarhet: resultatene av en analyse skal være reproduserbare. Biomarkører kan bli identifisert gjennom differensieringstopper som kan bli repetert funnet i de ulike sykdoms og kontrollgruppene, og verdiene tildelt disse differensieringstoppene kan ikke variere så mye innenfor en gruppe. Som en forenklet måling av reproduserbarhet, kan koeffisienter for variasjoner (CV), som har blitt en standard for å bedømme diagnostiske tester, bli tilveiebrakt ved en programvare utført på en prosessor. Variasjonene av markører i en gruppe, og til og med i den samme prøven, kan bli målt, karakterisert og benyttet i nedstrømsanalyse og klassifisering.

Robusthet: differensieringstopper skal være robuste mot uunngåelige variasjoner i prøvepreparering og håndtering, så vel som mot variasjoner som oppstår fra tendenser i massespektrometer egenskaper. En annen årsak for variabilitet fra pasient til pasient oppstår fra irrelevante forskjeller i den biologiske tilstanden av en pasient, for eksempel fordøyelsestilstanden ved tidspunktet for prøvetaking. Kriterier kan bli utviklet for å skjelne de irrelevante endringer fra biologiske signifikante endringer. I designen av klassifiseringer (dvs. klassifiseringsfunksjoner eller algoritmer), som er funksjoner som avbildes fra rom av multidemensjonalt trekk (f.eks. 12 differensieringstopper) til rom for klassemerke (f.eks. "god", "dårlig" eller "ikke definert") og under trekkekstraksjon, burde virkelig differensieringstopper ikke endres veldig mye imens de gjør små endringer til datanalyse parametere. Liknende lokaliserte differensieringstopper burde bli funnet i forskjellige datasett.

Fortolkbarhet: de resulterende differensieringstoppene kan bli puttet i sammenheng av biologisk fortolkbarhet. Først er identifiserte differensieringstopper generelt visuelt bemerkelsesverdige i massespektrene. M/Z posisjonene av differensieringstopper gir verdifull informasjon om den biologiske relevansen av underliggende biomarkører som forårsaker disse differensieringstoppene. Dette tillater fortolkningen og filtreringen av differensieringstoppene som oppstår fra biologisk irrelevante prosedyrer. For eksempel målingen av forskjellig hemoglobininnhold av cancerøse versus normale prøver, som bare er en artifakt av prøvepreparering. I noen tilfeller kan det vise seg at klinisk relevante differensieringstopper av spektrumet er av ikke-lineære kombinasjoner av multiple trekk i spektrumet, og er ikke enkle opp/ned reguleringen. Til og med i dette tilfellet burde differensieringstoppene som utgjør trekket i spektrene være synlig (fig. 4), og funksjonene for å evaluere markører burde bli konstruert eksplisitt.

Sensitivitet: stor anstrengelse er vanligvis foretatt for å samle prøver og generere massespektre. Stor forsiktighet er også gjort for å unngå å overse relevante differensieringstopper i massespektrometer spektrene ved å benytte dataanalyse algoritmer som ikke er selektive eller sensitive nok for å virkelig finne disse differensieringstoppene i et spektrum. Om et m/z område for eksempel er definert til å være relevant for et trekk, må dette område være stort nok til å inneholde trekket, og burde ikke klumpe seg sammen i andre trekk til stede i spektrumet. Algoritmer for å velge område avleder sine parametre fra dataene i seg selv, eventuelt på en lokal måte, og kan ikke avhenge av eksterne glattende og bindingsparametere.

Oppgaven for å sammenlikne massespektre for ekstraksjonen av differensieringstopper er gjort vanskelig ved den spesifikke naturen av disse spektrene på grunn av indre intensitetsvariasjoner. Ioniserings sannsynligheten av individuelle ioner avhenger av den lokale prøvekjemien (f.eks. ioneunderstrykkings effekter), og selv om masseoppløsningen av moderne massespektrometre nesten er tilstrekkelig, kan den absolutte masseskalaen variere fra spektrum til spektrum.

I overensstemmelse med prinsippene av den foreliggende oppfinnelsen kan massespektrometer spesifikke variasjoner bli målt for å redusere eller eliminere disse variasjonene (i tilfelle av bakgrunnsvariasjoner) eller tilveiebringe målinger for å bestemme den relevante signifikansen av signaler ved å estimere det lokale støynivået. Unngåelse av å introdusere ytterligere variasjoner som oppstår fra data pre-prosessering og analyse kan bli utført. Programvare for lukking av topp som ofte er satt sammen med mange massespektrometre har for eksempel blitt funnet til å være upålitelige for å direkte benytte disse tappene i komparativ spektral analyse. Tidlige forsøk ved spektral sammenlikning har i stedet ført til å benytte hele massespektrene i seg selv i deres sammenliknings og klassifiseringsalgoritmer. Hele spektret inkluderer imidlertid mange tusenvis av individuelle datapunkter, hvor de fleste er målinger av instrumentstøy med bare relevant informasjon er begrenset til toppene i massespektrene. Videre er fortolkningen av trekk i spektrene komplisert og noen ganger ikke-lineære i tilfelle av nøytralt nettverkbaserte klassifiseringsalogritmer, og blir veldig tungvinte. Som et resultat har anvendelsen av disse forsøkene for å klassifisere serumprøver ført til overdrevne krav som ikke kan bli reprodusert i andre laboratorier.

Fig. 4 er en avbildning av et eksemplarisk gel-plot 404 av et markørresultat ved et spektrometer. Spektrene er klinisk merket ved å benytte standard progresjonsmerker fra

World Health Organization (WHO), inkluderende stabil sykdom (SS), progressiv sykdom (PS), og partielle respondere (PR), Raffinerte kliniske merker, som separerer de kliniske hovedmerkene til ekstreme kliniske merker, er imidlertid laget for å inkludere tre ytterligere merker av SS-kort, SS-lang og PS-tidlig. Et gel-plot er et plot hvor hver linje korresponderer til et massespektrum av en klinisk prøve, den horisontale aksen er masse/ladningsaksen, og gråskalaen avbilder intensiteten. De kliniske merkene 402 er tilveiebrakt på gel-plotet 404 med horisontale linjer 404 som avgrenser de forskjellige kliniske merker. Gelplottet 404 er det av alle spektre (dvs. spektre mottatt fra en cancer forskningsklinikk av en kontrollgruppe av ikke-småcelle lungecancer pasienter i Italia og Japan som mottok Iressa som en cancerbehandling) benyttet for å trene en klassifiseringsalgoritme. Differensieringstopper kan visuelt bli sett på hvert av spektrene ved 406 og 408, men er kvantitativt målt for presishet og kvantitative formål.

For å unngå noen av disse målingsproblemene kan råmassespektret blir pre-prosessert for å fjerne og måle irrelevante artifakter av massespektrometri prosessen, og for å registrere dem på en liknende m/z og amplitudeskala.

Ved å fortsette med fig. 3 utfører prosessen av trinn 306 data pre-prosessering. Pre-prosesseringen kan inkludere hver eller alle av bakgrunnssubstraksjon, estimering av støy, normalisering, plukking av topp og spektral oppstilling. Disse prosessene er illustrert fig. 5-10 og beskrevet her nedenfor.

Fig. 5 er et histogram 500 som viser et eksemplarisk sett av datapunktresultater fra et spektrometer som har støy og signalkomponenter. Bakgrunn eller grunnlinje er en sakte varierende komponent av et massespektrum – det gradvis totale skifte av dataene over m/z område. Som funksjonelle definisjoner: bakgrunn er jevne variasjoner av signalstyrken som kan oppstå fra ladnings akkumuleringseffekter eller ikke-lineære detektoregenskaper eller partiell ionesvekkelse osv., som det motsatte til støy som oppstår fra elektronikk, tilfeldige ioner og fluktuerer raskt (i m/z).

Bakgrunn kan bli modellert og derfor subtrahert. Støy er en statistisk situasjon og bare dens styrke kan bli målt. Videre kan bakgrunn være forårsaket ved uoppløste "søppel" ioner og kan bli estimert og subtrahert før ytterligere dataprosesseringstrinn, slik at toppdeteksjon meningsfullt kan bli utført. Bakgrunnen kan bli estimert ved å benytte robuste lokale statistiske estimatorer. Å oppnå et pålitelig estimat for styrken av støyen i dataene er benyttet for påfølgende deteksjon av topp basert på signal-til-støy (S/S) forholdskriterie. Slike estimatorer er også benyttet i hver oppgave av spektral

sammenlikning for å tilveiebringe et mål for feil. Som i bakgrunnsestimeringen kan asymmetriske robuste estimatorer bli benyttet for å utføre denne oppgaven.

Bakgrunnen er vist til å inkludere det høyeste antallet av datapunkter og signalet inkluderer færre datapunkter. Bakgrunnen kan bli bestemt ved gjentagelse ved å benytte korrelasjonsanalyse og optimal separasjon. Ettersom bakgrunn ikke inneholder biologisk relevant informasjon og varierer fra spektrum til spektrum, kan amplitudeinformasjon bli gjort mer sammenliknbar ved å subtrahere verdien av bakgrunnen fra hvert spektrum. Denne prosessen er beskrevet i samtidig pågående patentsøknad serie nr. 10/887,138 og arkivert 7. juli 2004, som er inkorporert her i sin helhet.

Figur 6A og 6B er grafer 600a og 600b som viser et spektrum med en bakgrunn 602 og uten bakgrunn etter at bakgrunnen har blitt trukket ut av spektrumet 604. Som vanligvis i serum er det topper som er sterkt variable på grunn av naturlige fluktueringer i overfloden av serum proteom. Videre kan mengden av prøve ionisert fluktuere fra spektrum til spektrum på grunn av endringer i laserkraft, variasjoner i mengden av ioniserbare prøve, og variasjoner i posisjoneringen av laseren på MALDI platen. Denne fluktueringen gjør standardnormalisering rutinemessig, slik som total ionestrøm normalisering (dvs. normalisering over hele spektrumet), som er mindre nyttig ettersom fluktueringene i disse toppene er preparert til toppene av interesse. En partiell normalisering (dvs. normalisering over spektrene som identifiserer og ekskluderer disse variable toppene og regionene) kan bli benyttet for å unngå resultater som fluktuerer, som derved tilveiebringer reproduserbare resultater.

Mer spesielt kan partiell ionstrøm normalisering bli derivert som følger. Massespektrum inkluderer datapunkter, par (m/z, amplitude), arrangert i stigende rekkefølge i m/z. Ettersom spektrum er oppnådd på et "time-of-flight" instrument, kan m/z aksen bli betraktet segmentert til beholdere. Hvert datapunkt representerer den korresponderende beholderen og dets amplitude representerer (eller proporsjonalt til) ionetellingen i beholderen (dvs. ionestrøm i beholderen).

Summen av alle amplituder i spektrumet er derfor "den totale ionestrømmen" (TIS). Den korresponderer til det totale antallet av ioner som kommer ved en detektor av massespektrometre. Normalisering til den totale ionestrømmen betyr at for hver spektrum er en normaliseringsfaktor valgt slik at de korresponderende normaliserte

spektrene (m/z = opprinnelig m/z, aplitude = (norm faktor)*(opprinnelig amplitude)) har den samme (preskriberte) totale ionestrømmen, slik som 100.

Generelt har normalisering av total ionestrøm bare betydning etter bakgrunnssubstraksjon. På annen måte er den totale ionestrømmen dominert ved den integrerte bakgrunnen, i stedet for ved ionestrøm i de betydningsfulle signalene, slik som topper. Med andre ord integrerer total ionestrøm alle tilgjengelige ioner og er dominert ved store topper. I tilfelle hvor toppene er sterkt variable, er den totale ionestrømmen også sterkt variabel, som derved forårsaker normaliseringsvariasjon, som kan føre til falsk positiv deteksjon av differensieringstrekk.

I overensstemmelse med prinsippene av den foreliggende oppfinnelsen, deteksjon av "trekk" – intervaller av m/z akse som viser seg å være "ikke tomme", dvs. ikke "ren bakgrunn" på grunn av at de inneholder noen signaler, slik som topper. Et trekk er en topp som er synlig i mer enn et brukerdefinert antall av spektre av en kontrollgruppe av pasienter. Å ha et sett av trekk (en samling av ikke-overlappende m/z intervaller) tilveiebringer for definering av en mer fleksibel normaliseringsmetode, "normalisering til partiell ionestrøm" (PIS)". Partiell ionestrøm er summen av amplituder i spektrumet for alle datapunkter som tilhører det spesifiserte settet av trekk (vanligvis en delmengde av det fulle settet av trekk). Normalisering til den partielle ionestrømmen betyr at for hvert spektrum kan en normaliseringsfaktor bli valgt slik at de korresponderende normaliserte spektrene (m/z = opprinnelig m/z, amplitude = (norm faktor)*(opprinnelig amplitude)) har den samme (preskriberte) partielle ionestrømmen. Generelt benytter partiell ionestrøm stabile topper for normalisering, ettersom de sterkt variable ikke er inkludert i kalkuleringene. Ved å benytte stabile topper resulterer stabilitet i normasliseringsprossen.

Topper fra spektre i en kontrollgruppe av pasienter er inkludert i en liste, og en splittende sammenknyttende algoritme, som forstått i fagfeltet, kan bli benyttet for å finne klynger av topper.

ID	m/z senter	ID	m/z senter	ID	m/z senter	ID	m/z senter
0	3085.867	33	9157.859	70	17168.815	97	28102.608
1	3102.439	34	9371.936	71	17272.049	98	28535.715
2	3107.451	35	9424.089	72	17391.032	99	28889.368
3	3129.212	36	9432.518	73	17412.315	100	28896.086
4	4154.918	37	9446.061	74	17590.691	101	28902.778

5	4187.865	38	9635.796	75	17620.442	102	33277.541
6	4711.48	39	9638.7	76	18629.158	103	33340.741
7	5104.862	40	9659.863	77	18824.353	104	33839.223
10	6433.973	41	9717.098	78	19104.212	105	38830.258
11	6588.426	42	9738.03	79	19460.971	106	43474.948
12	6591.603	43	9941.016	80	20868.776	107	50722.939
13	6632.237	44	10220.11	81	21040.264	108	56307.899
14	6839.537	45	10504.31	82	21063.912	109	57257.535
15	6883.021	46	10841.693	83	21275.194	110	59321.131
16	6941.514	53	12579.848	84	22690.405	111	65392.98
17	7390.573	54	12771.505	85	22844.388	112	66702.45
20	7673.52	55	12861.925	86	22927.864	113	67633.769
22	8206.572	56	12868.575	88	23215.972	114	68328.628
23	8230.679	57	13082.443	89	23354.353	115	73363.308
24	8697.822	58	13765.804	90	23451.251	116	77948.338
27	8822.777	59	13885.668	91	24917.423	117	91016.846
28	8880.021	60	14050.987	92	25147.019	118	96444.862
29	8920.239	61	14157.312	93	25185.861	119	98722.464
30	8940.18	62	14651.73	94	25466.131		
31	9135.182	68	16206.683	95	25582.933		
32	9138.189	69	17143.885	96	25813.574		

TABELL I. Trekk for PIS normalisering

Tabell I inkluderer en liste av de 80% (PIS = 0,8) av alle trekk (gjenværende sett av trekk) som ble holdt i en PIS normalisering. m/z verdiene er i Dalton med en usikkerhet på 1000 ppm (etter oppstilling).

Et ekstremt tilfelle av partiell ionestrøm normalisering er når det fulle settet av trekk er benyttet for å beregne den partielle ionestrømmen. Dette tilfelle er analogt med total ionestrøm normalisering, hvor forskjellen er at de "tomme" regionene av spektrumet bidrar til den totale ionestrømmen, men ikke til den partielle ionestrømmen. Bidrag av støy i den "tomme" regionen er derfor ikke inkludert i den partielle ionestrømmen. Et annet ekstremt tilfelle er når bare et trekk er benyttet for å beregne den partielle ionestrømmen. Om dette er trekket inneholdende den sterkeste toppen, er basal toppnormalisering bestemt.

I sammenlikning av spektrum er begrunnelsen ved anvendelsen av partiell ionestrøm normalisering som følger. Betrakt to grupper av spektre, slik som sykdom og kontroll. Spektrene inneholder i størrelsesordenen av 100 signaler (topper), og det meste av signalene er forventet til å være uendret mellom grupper, mens noen signaler kan bli opp- eller nedregulert. I massespektre er de unormaliserte intensitetene ikke direkte sammenliknbare mellom spektre. Når det benyttes total ionestrøm normalisering, er en antagelse gjort om at opp- eller nedregulerte signaler er færre og svake, slik at de ikke signifikant forvrenger den totale ionestrømmen, som antageligvis dominerer signalene som er uendret mellom grupper. I virkeligheten er dette imidlertid ikke nødvendigvis tilfellet. Om for eksempel det oppregulerte signalet er sterkt nok til å signifikant bidra til den totale ionestrømmen, viser andre signaler i de normaliserte dataene seg å være nedregulert, til og med om de virkelig er uendret. Om spektrene inneholder sterke og sterkt varierende signaler, viser analogt andre signaler i det normaliserte spektrumet økte koeffisienter av variasjon, til og med om de er i og for seg stabile. Ved å benytte den partielle ionestrøm normalisering i stedet for total ionetelling, og ved å benytte delmengden av trekk som inneholder de mest stabile trekkene, imens det unngås oppregulerte, nedregulerte eller sterkt variable trekk, kan det kompenseres for problemet av økte koeffisienter av variasjon. Hovedspørsmålet er hvordan å velge denne delmengden.

For å velge delmengden av partiell ionestrøm, kan den følgende prosedyre bli benyttet. Om flere grupper av spektre er oppnådd, for formålene av denne prosedyren, kan gruppene av spektret bli kombinert i et kombinert sett.

Først er delmengden av trekk lik en full liste av trekk. Videre kan den følgende prosedyren bli gjentatt et antall ganger for å produsere den nye delmengden av "minst variable" trekk inneholdende et trekk mindre enn det opprinnelige.

Prosessen kan bli fortsatt som følger:

- Ved å benytte den opprinnelige delmengden av trekk, normalisere alle trekkverdier (fullstendig sett) til den partielle ionestrømmen
- For hvert trekk beregne koeffisienten av variasjon = (standard avvik)/(middelverdi)
- Sortere trekkene i henhold til den absolutte verdien for KV
- Velge den nye delmengden av trekk fra denne sorterte listen – dem med den minste absolutte (KV); inkludere et trekk mindre enn i den opprinnelige delmengden

- Erstatte den opprinnelige delmengde med den nye delmengden

Termineringskriteriene er de følgende. Brukeren spesifiserer to verdier:

- den laveste tillatte fraksjonen av ionestrømmen
- den laveste tillatte fraksjonen av antallet av trekk

Prosessen er terminert når enhver av kriteriene er brutt. Om brukeren spesifiserer begge verdier (dvs. den laveste tillatte fraksjonen av ionestrømmen og antall av trekk) som 0,8. er derfor den resulterende delmengden av trekk garantert til å inneholde minst 80% av ionestrømmen (som beregnet fra det fullstendige sett av trekk), så vel som minst 80% av trekkene. Å spesifisere 1,0 for enhver av verdien resulterer i at det fullstendige settet av trekk blir benyttet. Vanligvis er 0,8 omkring den rette verdien å benytte for optimale resultater. Avhengig av anvendelsen kan imidlertid høyere og lavere verdier bli benyttet. Trekkverdier normalisert til den partielle ionestrømmen kan så bli benyttet for klassifisering og andre formål.

Som oppsummering kan partiell ionestrøm bli bestemt som følger:

- kalkulere KV
- droppe toppen med høyeste KV
- stoppe når maksimal KV er mindre enn et spesifisert nivå

Implementering av den partielle ionestrømmen kan bli kalkulert ved å benytte to operasjoner. Den første operasjonen beregner en liste av trekk for anvendelse i PIC denominatoren. Denne operasjonsmarkøren fusjonerer først de to valgte gruppeverdiene til en todimensjonal oppstilling, hvor rader er spektre (dvs. prøver) og kolonner er trekkverdier korresponderende i størrelsesorden til trekklisten sortert ved senter m/z. Denne operasjonen foregår i to parametere i tillegg til de fusjonerte trekkverdiene. Disse to parameterne er MinAllowedFracOfIC og MinAllowedFracOfFeatures. MinAllowedFracOfIC – minimum tillatt fraksjon av ionestrømmen i den målte delmengden av trekk. Å beholde disse trekkene korresponderer til verdien på 1. MinAllowedFracOfFeatures – minimum tillatt fraksjon av trekk i den beholdte delmengden av trekk. Å beholde disse trekkene korresponderer til verdien på 1. Denne operasjonen gir en ArrayList av heltall, som representerer indeksene av trekkene som skal bli benyttet i denominatoren.

En utførelsesform av en algoritme benyttet for å motta listen av trekk ved å benytte PIC normalisering er oppsummert i den følgende pseudokoden:

```
int n_prøver  = antall av spektre i de 2 valgte gruppene;
int n_trekk = antall av trekk i listen av trekk;

// Bygg listen av alle trekk
ArrayList NFList = new ArrayList();
for (int j = 0; j < n_trekk; j++)
{
        NFList.Add(j);
}

//resulterer i 1, ettersom dette er den fullstendige listen av trekk i NFList.
Dobbel frac_ic = FracIonCurrent(f, NFList);
//resulterer også i 1.
Double frac_f = ((double)NFList.Count)/n_features;
ArrayList NFList_old = (ArrayList)NFList.Clone();

//imens fraktsjonen av ionestrom er større enn eller lik
//til den spesifisert ved brukeren og prosentandelen av benyttede
//trekk er større enn det spesifisert ved brukeren.

Imens (frac_ic >= MinAllowedFracOfIC &&
         frac_f >= MinAllowedFracOfFeatures)
{
        NFList_old = (ArrayList)NFList.Clone();
        //renormalisert basert på foreliggende NFList, beregn så
    // koeffisienten av variasjon for hvert sett av normaliserte
    // trekk, sorter så ved koeffisienten av variasjon,
    // den høyeste koeffisienten av variasjon er fjernet fra
    //NFList.
        OneStep(f, ref NFList);
    //nå er det minst et trekk i listen av indekser,
    //og fraksjon av ionestrøm kan bli beregnet som resultatet av
    // PIC / TIC
    //hvor PIC er summen av alle trekkverdier for
    //trekkindeksene spesifisert ved  NFList. Og hvor
    //TIC er summen av alle trekkverdier.
        Frac_ic = FracIonCurrent(f, NFList);
    //frac f er i all enkelthet prosentandelen av spektre som nå blir benyttet
    //i NFList
        frac_f = ((dobbel)NFList.Count)/n_features;

}
returner NFList_gammel;
```

Tallrike ytterligere mindre og større variasjoner i denne algoritmen vil være tydelig for en fagperson.

Med en gang denne kalkulasjonen er fullført er listen av trekk som skal bli benyttet i den partielle ionestrøm denominatoren bestemt.

Den andre operasjonen er å renormalisere alle trekkverdien for de spesifiserte gruppene ved å benytte den partielle ionestrøm denominatoren. Først er normaliseringsverdier mottatt for hvert spektrum/prøve ved å benytte trekkverdiene spesifisert ved listen av indeksresultater fra den forrige operasjonen. Så er disse normaliseringsverdiene benyttet for å modifisere listen av trekkverdier spesifisert i den todimensjonale oppstillingen av trekkverdier.

Denne funksjonen er oppnådd ved å utføre en algoritme representert ved den følgende pseudokoden:

```
Int n_prøver  = antall av spektre i de valgte gruppene;
int n_trekk = antallet av trek i listen av trekk;

//initialiser resultatoppstilling.
Dobbel[,] f2 = ny dobbel[n_prøver, n_trekk];


//oppstilling for normaliseringsverdiene.
Dobbel[] norm = ny dobbel[n_prøver];

// finn normaliseringsfaktoren for hver prøve
for (int I = 0; I < n_prøver; I++)
{
        norm[I] = 0;
        for hver (int k in NFList)
        {
                //sett norm[I] til sum av alle trekkverdier for
            //trekkindekser spesifisert i  NFList.
                Norm[I] += f[I,k];
        }
        //delt ved antallet av spesifiserte trekk.
        Norm[I] /= NFList.Count;

        for (int j = 0; j < n_features; j++)
        {
                //normaliser ved å dele trekkverdiene ved normaliserings
            //verdien for det spesifiserte trekket.
                F2[I,j] = f[I,j]/norm[I];
        }
```

}
//returner resultater.
Returner f2;

Tallrike ytterligere mindre og større variasjoner i denne algoritmen vil være tydelig for en fagperson.

Etter at disse to trinnene er fullført er partiell ionestrøm normalisering fullført. Partiell ionestrøm normalisering kan føre til en temmelig drastisk reduksjon i KV av individuelle topper. For urin reproduserbarhetsdata hvor man måler variabiliteten av prøve preprodusering via fraksjonering (resin for å fjerne salter) er reduksjonen i KV omkring en faktor på to.

Fig 7A er en graf 700a som viser multiple spektre 702 og 704 som blir normalisert for å i all enkelhet sammenlikne spektrene som vist i fig. 7B. Som vist er trekk (f.eks. topper) av de to spektrene 702 og 704 relativt oppstilt, men har forskjellige amplituder. Denne amplitudeforskjellen resulterer i de forskjellige intensitetene av de forskjellige spektrene 702 og 704. Med normalisering av de to spektrene 702 og 704 ved å benytte partiell ionenormalisering eller annen normaliseringsalgoritme, overlapper de to spektrene 702 og 704 vesentlig, og kan bli korrekt sammenliknet som vist i grafen 700b av fig. 7B.

Fig. 8A og 8B er grafer 800a og 800b som viser multiple prøvespektre 802a-802n (fig. 8A) som blir oppstilt 802a'-802n' (fig. 8B). Den absolutte masseskalaen av spektre kan variere betraktelig. Spektre kan bli skiftet med hensyn på hverandre, og til og med den interne masseskalaen er ikke konstant. I standard proteomikoppgaver er spesielle forbindelser tilsatt som gir opphav til topper ved kjente m/z verdier. Spektret kan så bli rekalibrert (dvs. at m/z verdiene kan bli reskalert i henhold til disse eksterne kalibreringsmidlene) og absolutte massepresisjoner på noen få tiltalls ppm kan bli oppnådd i det lave masseområdet, hvor peptider er forventet. I tilfelle av ufordøyde prøver er det noen ganger vanskelig å tilsette kalibreringsmidler til vev, og ofte ikke ønskelig, ettersom kalibreringsmidler kan undertrykke relevante topper på grunn av ioneundertrykkingseffekter. For spektral sammenlikning er det imidlertid tilstrekkelig å oppstille spektrene med en alminnelig masseskala og det er ikke så viktig at denne masseskalaen virkelig korresponderer til et absolutt mål av masse (dvs. databasesøk er ikke utført). Identifisering av felles topper kan bli utført, som beskrevet med hensyn på fig. 9.

For å oppstille spektre kan felles topper bli identifisert over grupper av spektre. Topper fra spektre er satt på en linje og splittende sammenklyngende algorimter kan bli benyttet for å separere denne store listen til en liste av klynger på den følgende måten:

Initialisering: Posisjoner av topper i spektrene er arrangert i en ordnet liste (ved m/z verdier).

Første separasjonstrinn: Hvor en minimal separasjon (vanligvis 30 Da) kan bli benyttet for å splitte denne lange listen til klynger av topper, hvor hver individuelle topp er nærere sammen enn den ønskede minimale separasjonen. Som et resultat kan en liste av klynger av nære topper bli oppnådd.

Fin separasjon: For hver av disse klyngene kan et histogram av toppforskjeller bli generert. Klyngen ved fremstikkende distanse, som er definert som to ganger den mediane separasjonen av topper i klyngen kan bli splittet, og den splittede forskjellen er mindre enn to ganger toppbredden eller mindre enn instrumentoppløsningen ved dette m/z område, så er klyngene ikke splittet. Om en splitting forekommer, så kan den samme analysen av de to resulterende klyngene bli rekursivt utført inntil ingen ytterligere splittinger forekommer. Om ingen splittinger forekommer, så gå til den neste klyngen.

Som et resultat er en liste av klynger som er nære i m/z og vel separert oppnådd. Hver klynge kan bli karakterisert ved dens sentrum (medianen av m/z posisjonen av alle topper i klyngen) og dens bredde (den 25 og 75 percentilen av disse posisjonene). Alternativt, men mindre robust, kan middelverdien og standardavvik bli benyttet som et mål for lokaliseringen og spredningen.

Seleksjon vanligvis i størrelsesordenen av ti klynger av passende gjennomsnittlig intensitet og så uniformt spredd som mulig over m/z område kan bli utført. En lineær (kvadratisk) regresjon på hvert spektrum for å oppstille masseskalaene av alle spektre av disse til toppene kan også bli utført. I en utførelsesform kan følgende sentrum av klynger bli benyttet: 6434.50, 6632.18, 11686.94, 12864.88, 15131.14, 15871.47, 28102.55.

En oppstilling kan bli utført med en toleranse på 5000 ppm, dvs. om det i hvert spektrum ikke blir funnet et oppstillingspunkt ved de spesifiserte posisjonene innenfor denne toleransen kan dette punktet bli ignorert. Om en oppstilling imidlertid ikke er

utført, er det følgende ikke detektert som trekk: :5764, 8702, 9426, 11443, 11686, 21066, 28102, 28309. Som et resultat er det mediane standardavviket av trekk redusert fra 4,63 Da til 3,68 Da for toppene som er synlig i de ikke-oppstilte spektrene.

Denne seleksjonen av disse felles toppene kan bli benyttet for å registrere spektret til en felles m/z skala, som vist i fig. 8B.

Ekstraksjon av trekk

Fortsette med fig. 3 er en ekstraksjonsprosess for trekk ved trinn 308 benyttet for å ekstrahere trekk (f.eks. topper) fra spektrene. Ved å gjøre dette er en bestemmelse gjort i forhold til hvilke trekk som skal bli ekstrahert.

Imens en visuell inspeksjon av spektre, deres gjennomsnitt og gruppeforskjeller, tilveiebringer noen veiledning om muligheten til å skille ulike tilstander eller kliniske trinn av sykdom ved å benytte massespektroskopi, kan en mer kvantitativ analyse bli utført. En differensieringstopp er basert på m/z posisjonene av topper i spektrene. En slik posisjon er en tentativ markør om den er felles for noen brukerdefinerte antall av spektre i en gitt gruppe av trekk. Med en gang en liste av disse trekkene er skapt for hver gruppe, kan hvert trekk bli gitt en definisjonsmessig verdi. Ved å benytte topp-bredde innstillingene av en algoritme som finner topper, kan de normaliserte og bakgrunnssubtraherte amplitudene bli integrert over dette området og tildelt denne integrerte verdien (dvs. området under kurven mellom bredden av trekket) til et trekk. For spektret hvor ingen topp har blitt detektert i dette m/z område, kan integrasjonsområdet bli definert som intervallet omkring den gjennomsnittlige m/z posisjonen av dette trekket med en bredde korresponderende til toppbredden av den foreliggende m/z posisjonen.

Verdiene av trekk kan variere betraktelig fra spektrum til spektrum, til og med innenfor den samme prøven (f.eks. serum eller vev), eller innenfor forskjellige prøver fra den samme celletypen. Imens m/z posisjonen av toppene er veldig reproduserbar, viser amplitudene store fluktuasjoner.

Som tidligere beskrevet er et mål for variasjonen av verdiene for trekk deres koeffisienter for variasjon (KV). Koeffisientene for variasjon er definert som forholdet av trekkene standardavvik over deres gjennomsnittlige verdi. Andre definisjoner er mulig, slik som forholdet av persentilområdet mellom den 25 og 75 persentilen over deres mediane verdi. En typisk distribusjon av KV verdier for spektrene benyttet er

tilveiebrakt i et histogram. Imens det er trekkverdier som er sterkt reproduserbare med KV verdier mindre enn 0,5, viser størsteparten av trekk en større variasjon. Dette poengterer hvorfor ekstraksjon ikke er viktig og fluktuasjoner og distribusjoner av trekk skal bli analysert før identifisering av trekket som en potensiell differensieringstopp med en skjelnelig egenskap.

Ved å fortsette med fig. 3 er en seleksjonsprosess for trekk utført ved trinn 310 for å velge trekkene som er benyttet i å utføre klassifiseringsanalysen. Seleksjonsprosessen for trekk kan bli illustrert som vist i fig. 9.

Fig. 9 er en graf av en eksemplarisk prosess for å velge et trekk (kandidattrekk) ved å lokalisere en topp felles i mer enn "x" spektre som har en bestemt bredd, hvor bredden er definert som oppstillingsfeil pluss toppbredde. Ulike seleksjonsteknikker kan bli benyttet i å utføre seleksjonen av trekk. Som vist er det tre spektre 902a-902c (kollektivt 902). Disse spektrene 902 er benyttet for å lokalisere et trekk (f.eks. topp) 904. Som vist strekker en midtre vertikal linke 906 seg gjennom sentrum av trekket 904, som er felles på en mer enn et spekter 902, og sidevertikale linjer 908a og 908b definerer bredden av trekket (oppstillingsfeil + toppbredde).

Valget av differensieringstrekk kan bli utført i tre trinns prosess: først er alle trekk ordnet ved en univariert p-verdi oppnådd fra en enkel hypotesetest som antar at alle trekk er uavhengig. I noen implementeringer kan en Mann-Whitney test for å oppnå en p-verdi for hvert trekk bli benyttet. Andre metoder er mulig, men mindre robuste, slik som to-prøve t-tester, Kolomogorov Smirnov tester, eller andre. Ved å benytte Bonferroni korreksjoer er videre de topprangerte (minste p-verdi) trekkene inspisert ved å sammenlikne de gjennomsnittlige gruppespektrene (gjennomsnittet av spektrene i en klinisk gruppe). Om et trekk ikke skjelner mellom grupper, er det droppet som en kandidat. I et tredje og endelig trinn kan seleksjon av trekk bli utført ved å benytte kryss-valideringsfeil som et kriterie for suksess. Ulike implementeringer ved å utføre dette er beskrevet nedenfor:

Seleksjonen av relevante trekk er mer et spørsmål i gen mikromatrise eksperimenter ettersom det er tusenvis av trekk og få prøver. Seleksjon av trekk er også et spørsmål for identifiseringen av biomarkører når det undersøkes en massespektrale data og det er ingen bevis for at seleksjon av trekk ikke påvirker yteevnen av noen klassifiseringer veldig mye. Ikke desto mindre er det vanskelig å fortolke klassifiseringsresultater om

det er mange titalls av trekk, og i virkeligheten er det ingen forventning om at alle av disse trekkene er relevante.

Rangering av trekk ved deres viktighet kan bli utført for å differensiere ulike trinn av sykdom. Det er enkelt å velge et trekk ved et tidspunkt, men når det er mange titalls av trekk, er det mer vanskelig å bestemme hvilke av trekkene som er viktige for det bestemte trinnet av sykdom. For å sammenlikne biomarkører og spektre over laboratorier, skal de samme trekkene være identifiserbare, og de trekkene som viser seg på grunn av usikkerheter i prøvepreparering, instrumentbruk og populasjonsvariasjoner skal være skjelnbare.

Seleksjon av trekk står ovenfor algoritmiske bestemmelser. Den første bestemmelsen er bare kombinasjonsmessig. Et fullstendig søk av alle mulige kombinasjoner av l trekk av en total på m tilgjengelige (målte) trekk fører til $\binom{m}{l} = \frac{m!}{l!(m-l)!}$ kombinasjoner, f.eks. for m=20, 1=5 er dette tallet 15504. Som typisk i massespektre er det et kobbel av hundrevis av tilgjengelige trekk, dette tallet av kombinasjoner kan være fro stort for et fullstendig søk. Det er også heller ikke innlysende hvilken verdi for l som er optimal. Spesielle heuristiske søkstrategier kan derfor bli benyttet. Den andre bestemmelser oppstår fra mangelen på et unikt kvalitetsmål som bestemmer hvilke trekk av sett som er bedre enn et annet. Ettersom en kriterie for seleksjon av trekk kan være klassifiseringspresetasjonen, inneslutter "wrapper metoder" seleksjon av trekk som del av klassifiseringsalgoritmen. Disse metodene benytter en estimering av klassifiseringsfeilen, ideelt et mål for generaliseringsfeilen, som er vanlig å bestemme, og er vanligvis tilnærmet ved "leave-one out crosss-validation" (LOOCV) eller marginbaserte feilgrenser i tilfelle av Support Vector Machines (SVM) lærdom. Alternativer inkluderer filtermetoder som utfører seleksjon av trekk før klassifiseringen er generert. Hver av disse metodene har deres egne utfall, og benytter spesiell håndtering med hensyn på validering.

Strategier for søk er diskutert nedenfor først, og så er et sett av kvalitetsmål alminnelig benyttet listet opp.

Strategier for søk av trekk

De fleste strategier for søk er basert på en "del og vinn" metode, som optimaliserer kriteriet for seleksjon av trekk. For spesifikke valg av kriterien for seleksjon av trekk, kan det være mulig å benytte probabilistisk prøvetaking i ånden av "importance

sampling" Monte Carlo, eller spesielle optimaliseringsteknikker, slik som dynamisk programmering.

Som benyttet kan tre-basert klynging starte med alle trekk og trekk kan bli deletert ett etter ett, alternativt kan prosessen starte med et trekk og tilføye andre trekk ett etter ett. Som illustrasjon kan fire trekk eksistere $\{x_1, x_2, x_3, x_4\}$.

Topp-ned søk:

- Kalkuler verdien av trekk seleksjonskriterien for $\{x_1, x_2, x_3, x_4\}$ som gir C_4.
- Kalkuler verdien av trekk seleksjonskriterien for hver av $\{x_1, x_2, x_3\}$, $\{x_1, x_2, x_4\}$, $\{x_1, x_3, x_4\}$, $\{x_2, x_3, x_4\}$, og velg den bestem si $\{x_1, x_2, x_3\}$ med verdi C_3.
- Kalkuler verdien for trekk seleksjonskriterien for hver av $\{x_1, x_2\}$, $\{x_1, x_3\}$, $\{x_2, x_3\}$, velg den beste, si $\{x_1, x_2\}$, med verdi C_2.
- Og til slutt plukk det beste enkle trekket fra $\{x_1, x_2\}$ med verdi C_1.
- Den beste verdien av $\{C_1, C_2, C_3, C_4\}$ definerer den (sub)optimale settet av trekk.

På liknende måte ved å starte fra et trekk, og tilføye flere ett etter ett definerer et bunn-opp søk. Dette gir ikke nødvendigvis en optimal løsning, og det er ikke en garanti for at det optimale lavere (høyere) tallet av trekk utvikles i henhold til disse tre. En måte å forbedre disse enkle prosedyrene er å overveie trekk tidligere forkastet, eller å forkaste tidligere valgte trekk. Denne algoritmen er kalt metoden for flytende søk, og forstått i fagfeltet som følger:

Flytende metode for søk:

Det følgende beskriver et søk for et satt tall l av m trekk. En løkke over l kan bli utført for å optimalisere antallet av trekk. Metoden for flytende søk er basert på enten topp-ned eller bunn-opp søk. Flytende beskrevet er basert på bunn-opp metoden.

Betrakt et sett av m trekk. Ideen er å søke for den beste delmengden av k av dem for $k = 1, 2, \ldots, l \leq m$ ved å optimalisere C. La $X_k = \{x_1, \ldots, x_k\}$ være det optimale settet for k trekk og Y_{m-k} settet av de gjenværende av m-k trekkene. De nedre dimensjonale beste delmengdene $X_2, X_3, \ldots, X_{k-1}$ av 2,3,…,k-1 trekk er holdt for lagring. Ved det neste trinnet er den (k+1)th optimale delmengden X_{k+1} dannet ved å ta et element av Y_{m-k}. Så er en sjekk utført gjennom alle lavere dimensjonale delmengder for å se hvorvidt dette

forbedrer C, og erstatter det tidligere valgte trekket. Algoritmen kjøres som følger (C er slik at større er bedre):

- Velge det beste enkle trekket, som gir X_1 med C_1.
- Tilføy en annen basert på C, som gir X_2 og C_2.

Nå gjenta over k:

- Trinn I, inklusjon: Velg det elementet fra Y_{m-k} som kombinert med X_k gir beste C, dvs. $x_{k+1} = \arg\max_{y \in Y_{m-k}} C(X_k, y)$ som definerer $X_{k+1} = \{X_k, x_{k+1}\}$ som i bunn-opp algoritmen.

- Trinn II, test:
  1. Finn trekket x_r som har den minste effekten på kostnad C når fjernet fra X_{k+1}, dvs. $x_r = \arg\max_{x_r \in X_{k+1}} C(X_{k+1} / \{x_r\})$.
  2. Om $r = k+1$, $k = k+1$, $C_{k+1} = C$ og gå til trinn I.
  3. Om $r \neq k+1$ og $C(X_{k+1} / \{x_r\}) < C_k$ gå til trinn I, dvs. om fjerning av x_r ikke forbedrer den tidligere valgte gruppen ikke gjør et søk bakover.
  4. Spesielt tilfelle for k=2: If k=2 sett $X_2 = X_3 / \{x_r\}$ og $C_2 = C(X_3 / \{x_r\})$.

- Trinn III, eksklusjon (søk bakover):
  1. $X_k^{'} = X_{k+1} / \{x_r\}$ dvs. fjerne x_r.
  2. Finn det minst signifikante trekket x_s i det nye settet via $x_s = \arg\max_{y \in X_k^{'}} C(X_k^{'} / \{y\})$.
  3. Om $C(X_k^{'} / \{x_s\}) < C_{k-1}$ når $X_k = X_k^{'}$, tilbakefør C_k og gå til trinn I som terminerer søket bakover.
  4. Sett $X_{k-1}^{'} = X_k^{'} / \{x_s\}$ og k=k-1.
  5. Spesialtilfelle k=2: Sett $X_2 = X_2^{'}$ og $C_2 = C\left(X_2^{'}\right)$ og gå til trinn I.
  6. Gå til trinn III.

Denne algoritmen opererer generelt hovedsakelig bedre enn den enkle bunn-opp algoritmen, og den kan bli kjørt opp til m for igjen plukke opp det maksimale (minimale) kriteriesettet.

Tilfeldige trekkeseleksjons algoritmer

Den tilfeldige trekkseleksjons algoritmen er en optimaliseringsstrategi basert på telling av hyppigheten av konfigurasjoner fra tilfeldig prøvetaking. I bygging av hierarkimessige agglomererende klynger fra en intiell konfigurasjon (k-medianer, k-

middelverdier, uklar klynging) kan for eksempel algoritmen bli startet mange ganger om igjen, lagre de individuelle konfigurasjoner fra hver kjøring, og bygge et frekvenshistogram. Dette kan ofte bli kombinert med kryssvalidering.

Generering av klassifiserer

Ved å fortsette med fig. 3, ved trinn 312, er generering av klassifiserer utført. Genereringen av klassifiserer kan inkludere noen få funksjoner, inkluderende (i) inspisert læring, (ii) kryssvalidering og (iii) blind klassifisering eller testing. De første to funksjonene, inspisert læring og kryssvalidering, kan bli utført på råspektrene med assosierte kjente kliniske resultater 108 tilveiebrakt ved cancer forskningsklinikker 104, som beskrevet i fig. 1.

Imens rangering av trekk gir noen ide omkring viktigheten av trekk for å diskriminere grupper, benytter en mer gjennomgående analyse en spesiell læringsprosedyre. Inspisert læring er prosessen hvorved kategorimerker er tilveiebrakt for hvert tilfelle, i en treningssett (dvs. hvert spektrum) og søker etter å reduser antallet av misklassifisering. En annen mer spesifikk definisjon av inspisert læring er kartleggingen fra et høydimensjonalt rom av trekk til rom av merke fra trekk/differensieringstopp ekspresjon til sykdomsmerket eller responsmerket (ellers betegnet som klassemerke). Merket er en funksjon av massespektrometer toppene og assosierte parametere. En forsker eller annen person som har spektret fra, og klinisk informasjon omkring cancerpasienten hvorfra spektrumet ble produsert kan utføre den inspiserte læringsprosessen. Prosessen kan bli utført ved standard algoritmer fra teorien av inspisert læring. Resultatet fra det inspiserte klassifiserer algoritmene er en algoritme av klassifiserer (avhengig av treningssettet) som genererer til klassemerket fro et nytt tilfelle eller spektrum. I en utførelsesform kan en k næreste naboer (KNN) algoritme bli benyttet for klassifiseringen.

Algoritme for K nærmeste naboer

Metoden for k-nærmeste naboer er en enkel metode for tetthetsestimering. Sannsynligheten for at et punkt x' faller innenfor et volum V sentrert på x er:

$$p \int_V dx\, p(x)$$

For et lite volum $p \quad p(x)V$. Sannsynligheten kan bli tilnærmet ved proporsjonen av prøver som faller innenfor volumet V. Om k derfor er antall av prøver ut av en total på n som faller innenfor V så

$$p \quad \frac{k}{n} \text{ and } p(x)=\frac{k}{nV}$$

Tilnærmingen for k-nærmeste-nabo er å sette sannsynligheten k/n (eller for et satt antall av prøver for å sette k) og å bestemme volumet som inneholder k prøver. Det er i motsetning til histogramestimater hvor beholderbreddene er satt, og antallet av punkter er telt. Det er noen spørsmål med regulariteten av denne definisjonen, men den kan bli vist til å være uten skjevhet og samsvarende om $\lim_{n\to\infty} k(n) = \infty$ and $\lim_{n\to\infty} k(n)/n = 0$.

En bestemmelsesregel kan bli konstruer t på den følgende måten. Anta at det var k_m prøver i klasse ω_m, og det totale antallet av ω_m er n_m. Så er den klassekondisjonelle sannsynligheten:

$$p(x \mid \omega_m) = \frac{k_m}{n_m V}$$

Den forrige er n_m/n (om det er en prøve totalt over alle klasser).

Bayesian bestemmelsesregelen er å tilskrive x til ω_m om

$$p(\omega_m \mid x) \geq p(\omega_i \mid x) \ \forall i$$

og benytte Bayes teorem resulterer dette i denne seleksjonen

$$\frac{k_m}{n_m V}\frac{n_m}{n} \geq \frac{k_i}{n_i V}\frac{n_i}{n} \ \forall i \ \Rightarrow \ k_m \geq k_i$$

I tilfelle av "tie breaker" kan en bryting av knute bli gjort via den nærmeste middelverdien, nærmeste medlemmet eller på annen måte. Alternativt kan "tie breaker" bli registrert til oddetall k. minste k fører til irregulære overflater imens K til glatte overflater. Den asymptotiske misklassifiseringsraten er avgrenset fra ovenfor ved to ganger Bayes feil, som er en veldig god asymptotisk utførelse for en slik enkel algoritme. KNN klassifisering gir støtte til anvendelsen av prototyper, dvs. en data kondensasjonsteknikk. Men her er anvendelsen av KNN klassifisering mer benyttet for

reduksjonen i nødvendig lagring. Valget av en distansefunksjon kan bli benyttet. Alternativt kan euklidiske forskjeller, som ikke er optimale, også bli benyttet. Voteringsprosessen får et enkelt eksempel av et todimensjonalt rom av trekk er i justert i fig. 11.

Fig. 11 er en graf 1100 som viser en eksemplarisk gruppe av klassemerkede spektreindisier representative for to forskjellige klasser av sykdomsprogresjon og et testspektrum indisier som skal bli klassifisert. De grafisk representerte differensieringstoppene i rom av trekk, illustrerer her et todimensjonalt trekk av rom, grafen 1100 er en todimensjonal graf som har en x-akse og en y-akse. Om rommet av trekk var et 12-dimensjonalt rom av trekk (dvs. 12 trekk eller topper) ble valgt som differensieringstopper indikative for å skjelne mellom egenskaper som klassifiserte et spektrum til å være klassemerket som "godt" eller "dårlig"), så vil det ikke være mulig og lette grafisk representere spektrene, så et todimensjonalt rom av trekk er benyttet som et eksempel.

I dette tilfellet er spektrene klassifisert med klassemerke som "god" 102 og "dårlig" 1104, hvor de "gode" klassemerkede spektraindisiene 1102 er representert på grafen 1100 som et mønster og de "dårlige" klassemerkede spektraindisiene 1104 er representert som et annet mønster. Som tidligere beskrevet kan de klassemerkede spektrene bli utviklet fra en cancer forskningsklinikk og benyttet som en kontrollprøve for klassifiseringsformål basert på de kliniske resultatene av en cancerpasient i å respondere til et anti-cancermedikament slik som Iressa. Et testspektrum indisie 1106 kan bli plassert på graf 1100 i en lokalisering representativ av et prøvespektrum fra en ny cancerpasient hvorfra en behandlingsplan blir bestemt. Lokaliseringen av testspektrum indisiene 1106 er basert på amplitudene av de to trekkene (dvs. x og y amplitudene). Som vist, og i overensstemmelse med sannsynlighetsalgoritmen KNN, er de næreste tre klassemerkede spekterindisiene 1108a, 1108b og 1108c potensielle kandidater for testspektrumet som skal bli assosiert.

En eksemplarisk sannsynlighetstest for klassifiseringsprosessen for et testpunkt av det todimensjonale rommet av trekk er:

$$P(\bar{x} \in A) = \frac{k_A + 1}{k_A + k_B + 2} \; {}_2F_1(1, k_B + 1, k_A + k_B + 3, 1 - \frac{N_A}{N_B})$$

Om sannsynlighetsforskjellen mellom to klasser overskrider en bestemt bruker-levert terskel delta-p, så kan sannsynligheten bli betraktet signifikant og en klassifisering av "god" eller "dårlig" kan bli gjort. Om sannsynlighetsforskjellen er under en bestemt terskel, så kan en klassifisering av "ubestemt" bli gjort.

Imens en KNN algoritme kan bli benyttet som en algoritme for klassifiserer, kan andre klassifiseringsalgoritmer bli benyttet. En annen algoritme utviklet i overensstemmelse med prinsippene av den foreliggende oppfinnelsen er en probabilistisk k nærmeste nabo algoritme, som en modifisert KNN algoritme som tilveiebringer ytterligere fleksibilitet og tilveiebringe ytterligere informasjon for kliniske anvendelser.

Modifisert (probabilistisk) k nærmeste nabo algoritme

I overensstemmelse med prinsippene av den foreliggende oppfinnelsen kan en modifisert k nærmeste nabo algoritme bli benyttet for klassifisering. I dens enkleste implementering søker den modifiserte KNN algoritmen for de k nærmeste naboene i rom av trekk og tildeler et klassemerke i henhold til en enkel majoritetsavstemming over merkene av disse nærmeste naboene. Rom av trekk er definert til å være antallet av trekk (f.eks. 12 trekk) som blir benyttet for å definere et spektrum. I en utførelsesform er det ingen eksplisitt treningsfase og alle tilfeller er benyttet i klassifiseringen av spektret. Vanligvis er bare enkle auklidiske distanser benyttet for å bestemme naboene, men andre definisjoner er mulig (f.eks. Mahanolobis distanser fra egnede definerte kovarians matriser).

I det tradisjonelle K-nærmeste naboer (KNN) rammeverket er klassifisering utført som følger:

Ethvert objekt, eller tilfelle, som skal bli klassifisert (her – massespektrumet) er karakterisert ved *d* antall x_i, I = 1....D (her – verdiene av d trekk), og er derfor representrert ved et punkt i *d*-dimensjonalt rom. Distansen mellom de to tilfellene er definert ved vanlig euklidisk mål $\sqrt{\sum_i (x_i - x_i')^2}$. Selvfølgelig kan ethvert likhetsmål også bli benyttet her.

I tillegg kan en implementering benytte en "winsorized" Mahalanobis distanse for å bestemme distansen mellom to spektre.

Et treningssett kan inkludere tilfeller med kjente klassebetegnelser. Gi treningssettet og et positivt oddetall k, er klassifisering av testobjektet utført som følger:

1. I treningssettet finn k nærmeste naboer av testobjektet (dvs. spektrum) i det d-dimensjonale rommet.
2. Hver av disse k naboene tilhører en av klassene (f.eks. god eller dårlig). Finn hvilken klasse som har den største antall av representativer.
3. Klassifiser testobjektene til å tilhøre denne klassen.

Denne KNN klassifiseringen har to ulemper. Først tilveiebringer den ingen informasjon om konfidensen av klasseanvisning. Det er intuitivt tydelig at i tilfelle k=15 og to klasser, er konfidensen av klassetildeling i 15:0 mye høyere enn i 8:7 situasjonen. I kliniske anvendelser er karakterisering i konfidensnivået av hver individuelle klassetildeling relevant og benyttet for å diagnostisere pasienter. Dette nivå kan virkelig bli definert ved starten.

For det andre tar den ikke korrekt i betraktning antallet av tilfeller av hver klasse i treningssettet. Bare tilføye noen tilfeller av den gitte klassen til treningssettet tenderer til å påvirke klassifiseringsresultater i favør av denne klassen.

For å korrigere for disse problemene har en e"probabilistisk KNN" klassifiserer blitt utviklet som starter fra informasjonen av klassene av k nærmeste naboer fra treningssettet, men i stedet for klassetildelingsprosedyrer sannsynligheter av testtilfellet som tilhører hver av klassene. Nedenfor er en konsis beskrivelse av begrunnelsen og deriveringen av hovedformlene for probabilistisk KNN.

KNN metoden for klassifisering av spektrumprøver kan bli sett på som følger: betrakt en ball med en bestemt radius i det *d*-dimensjonale rommet og sentrert ved testtilfellet. Radiusen av ballen er bestemt ved kravet at den inneholder eksakt k tilfeller fra treningssettet. Så observer hvor mange medlemmer av hver klasse som er blant disse k tilfellene, og benytt denne informasjonen for å tildele klassemerket (i standardmetoden) eller beregn sannsynligheter for at testtilfellet tilhører denne eller en annen klasse (i den probabilistiske metoden).

Treningssettet kan være en prøve trukket fra en eller annen (ukjent) sannsynlighets distribusjon. Mer presist, for hver klasse, er delmengden av treningssettet som tilhører klassen betraktet til å være en prøve trukket fra den korresponderende sannsynlighets distribusjonen, som er forskjellig for hver klasse.

Betrakt samspillet av treningssett trukket fra den samme sannsynlighets distribusjonen. I KNN metoden for klassifisering er radiusen av ballen omkring testtilfellet forskjellig for hver realisering av treningssettet for å sikre at det alltid inneholder eksakt k nærmeste naboer. Se også beskrivelsen av KNN metoden i forrige avsnitt:

De følgende tilnærmingene kan bli gjort:

1. Ballen omkring testtilfellet kan bli betraktet satt, som betyr at det er avhengig av posisjonen av testtilfellet og av sannsynlighets distribusjonene hvorfra treningssettet er trukket, men det samme for hver realisering av treningssettet. Denne tilnærmingen er gyldig når k ikke er for liten.
2. For hver klasse er antallet av tilfeller for hver klasse i ballen trukket fra Poisson distribusjonen. Denne tilnærmingen er gyldig når ballen bare består av en liten fraksjon av den totale sannsynligheten for denne klassen.
3. Sannsynlighetstetthetene for klassene er tilnærmet konstante i ballen.

Betrakt tilfellet av to klasser. Hvert tilfelle er representert ved et punkt $\bar{x}$ i d-dimensjonalt rom. Det fullstendige d-dimensjonale rommet er betegnet ved Ω.

Klasse 1 er karakterisert ved sannsynlighets distribusjonen $p_1(\bar{x})$, $\int_{\Omega} p_1(\bar{x})d\bar{x} = 1$.

Klasse 2 er karakterisert ved sannsynlighets distribusjonen $p_2(\bar{x})$, $\int_{\Omega} p_2(\bar{x})d\bar{x} = 1$.

Et treningssett kan bli dannet ved en N_1 punkter trukket fra klasse 1, og N_2 punkter grukket fra klasse 2. Nærheten av testpunktene kan bli betegnet som ω. Dette er virkelig en ball sentrert ved testpunktet, men dette er irrelevant for det følgende. For en gitt realisering av treningssettet er det k_1 punkter i ω fra klasse 1 og k_2 punkter i ω fra klasse 2. Det er antatt at $k_1 << N_1$.

$\int_{\omega} p_1(\bar{x})d\bar{x} << 1$. Det samme som for klasse 2.

Dette sikrer gyldigheten av Poisson tilnærmingen; k_1 kommer fra Poisson distribusjonen med forventningsverdien λ_1,

$$\lambda_1 = N_1 \int_{\omega} p_1(\bar{x}) d\bar{x}$$

k_2 kommer fra Poisson distribusjonen med forventningsverdien λ_2,

$$\lambda_2 = N_2 \int_{\omega} p_2(\bar{x}) d\bar{x}$$

Nå er testpunktet (sentrum av ω) behandlet som "ennå et annet punkt". Med andre ord er det k_1+k_2+1 punkter i ω, i stedet for k_1+k_2, og det er ikke kjent i hvilken klasse testpunktet tilhører. Sannsynligheten fro at testpunktet tilhører klasse 1 og klasse 2 kan bli betegnet som følger:

$$\frac{p(class1)}{p(class2)} = \frac{\int_{\omega} p_1(\bar{x}) d\bar{x}}{\int_{\omega} p_2(\bar{x}) d\bar{x}}$$

derfor

$$p(class1) = \frac{\int_{\omega} p_1(\bar{x}) d\bar{x}}{\int_{\omega} p_1(\bar{x}) d\bar{x} + \int_{\omega} p_2(\bar{x}) d\bar{x}} = \frac{\frac{\lambda_1}{N_1}}{\frac{\lambda_1}{N_1} + \frac{\lambda_2}{N_2}}$$

Ved behandling av testpunktet (sentrum av ω) som "ennå et annet punkt", er det implisitt forutsatt at både $p_1(\bar{x})$ og $p_2(\bar{x})$ ikke endres signifikant i ω.

Problemet er at λ_1 og λ_2 er virkelig ukjente. Deres sannsynligheter kan imidlertid bli estimert i Bayesian måten. Både k_1 og k_2 er antatt å adlyde Poisson distribusjonen,

$$p(k \mid \lambda) = \frac{\lambda^k}{k!} e^{-\lambda}$$

Betegn den tidligere distribusjonen for λ ved $p_0(\lambda)$,

$$p(k) = \int d\lambda p(k \mid \lambda) p_0(\lambda).$$

Ved standard Bayesian begrunnelsen,

$$p(\lambda \mid k) = \frac{p(k \mid \lambda) p_0(\lambda)}{\int d\lambda p(k \mid \lambda) p_0(\lambda)}$$

Ved å anta fra nå på den flate tidligere distribusjonen av λ, $p_0(\lambda) = 1$ kan det følgende bli oppnådd

$$p(\lambda \mid k) = p(k \mid \lambda) = \frac{\lambda^k}{k!} e^{-\lambda}$$

Eventuelt er det følgende oppnådd

$$p(class1) = \int_0^\infty d\lambda_1 \int_0^\infty d\lambda_2 \frac{\lambda_1}{\lambda_1 + \frac{N_1}{N_2}\lambda_2} p(\lambda_1) p(\lambda_2),$$

hvor

$$p(\lambda_1) = \frac{\lambda_1^{k_1}}{k_1!} e^{-\lambda_1},$$

$$p(\lambda_2) = \frac{\lambda_2^{k_2}}{k_2!} e^{-\lambda_2}$$

Beregning av disse integralene gir det følgende:

$$p(class1) = \frac{k_1 + 1}{k_1 + k_2 + 2} \, {}_2F_1(1, k_2 + 1, k_1 + k_2 + 3, 1 - \frac{N_1}{N_2}).$$

For de like størrelsene av prøvene benyttet i treningssettet (N_1=N_2) forenkles dette til det følgende:

$$p(class1) = \frac{k_1 + 1}{k_1 + k_2 + 2}$$

$$\frac{p(class1)}{p(class2)} = \frac{k_1 + 1}{k_2 + 1}$$

For mer enn to klasser og forskjellige prøvestørrelser i treningssettet er det vanskelig å oppnå *p*(klasse *I*) i lukket form. I dette tilfellet kan det følgende mye forenklede esitmatet bli benyttet:

$$\frac{p(class \quad i)}{p(class \quad j)} = \frac{k_i + 1}{k_j + 1} \cdot \frac{N_j}{N_i},$$

Eller, ekvivalent, er hver *p(klasse I)* proporsjonalt med $\frac{1}{N_i}(k_i + 1)$ imens

$$\sum_{i=1}^{N_{classes}} p(class \quad i) = 1$$

Parameteren som karakteriserer robustheten av resultater til feilaktige spektere er en brukerlevert parameter, p-diff, som definerer hvor forskjellig klassesannsynlighetene må

være for å assosiere et merke i et spektrum. Om for eksempel p-diff er satt til 0,1 og sannsynligheten for klasse A er 0,6 og for klasse B er 0,4, så er forskjellen 2 større enn 0,1, og klasse A vil bli valgt. På den andre siden, om klassesannsynligheten for klasse A er 0,52 og for klasse B er 0,48, så er forskjellen 0,04 mindre enn 0,1 og, og klassifisereren returnerer et klassemerke til å være "udefinert".

Alternativt kan hypotesetesting ha klassifiseringen til å være signifikant med en eksternt spesifisert signifikant α. I en standard hypotese testformulering kan klassifiseringen bli beskrevet som følger:

Data: Et testtilfelle kan inkludere to klasser A og B, k_A og k_B nærmeste naboer av klasse A og klasse B, og populasjonen av N_A tilfeller av klasse A og N_A tilfeller av klasse B.

Teststatistikk: I all enkelhet antallet av naboer i klasse A:

$$T = k_A$$

Null distribusjon: Nullen er antatt til å være antallet av A naboer som er forventet fra populasjonsforholdene alene, dvs. k_A under nullen er en binominal tilfeldig variabel med parameterne $k = k_A + k_B$ and $p^* = N_A / N_B$.

Hypotese: (to-halet) Dette er en implementering av en binominal test, som forstås i fagfeltet.

$$H_0 : p_A = p^*$$
$$H_1 : p_A \neq p^*$$

I tilfelle av testutvikling er antallet av nærmeste naboer sjelden større enn 20, slik at anvendelsen av den normale tilnærmingen er ikke benyttet. For en gitt total signifikans er α løst fra en tabell (eller kjørt på en computer) $P(Y \leq t_1) = \alpha_1$ og $P(Y \leq t_2) = 1 - \alpha_2$ for t_1 for t_1 og t_2 hvor Y er en binominal tilfeldig variabel som definert under nullen, og hvor α_1 og α_2 er tilnærmelsesvis $\alpha/2$ og beløper seg til α. Forkastningsregionene er verdiene av T mindre enn t_1 eller større enn t_2. Konfidensregioner kan også bli estimert for p^* ved å følge prosedyren beskrevet i avsnittet binominal test.

Imens den modifiserte KNN algoritmen kan bli benyttet som klassifiseringsalgoritmen som beskrevet ovenfor, kan alternative klassifiseringsargoritmer bli benyttet i overensstemmelse med prinsippene av den foreliggende oppfinnelsen. Slike klassifiseringsalgoritmer kan inkludere uklar KNN, Kernel metoder (f.eks. SWM), ikke

overrasket klassifisering, sektralklynging, kernel PCA, ikke-parametrisk klynging, k-middelverdier, k-histogrammer, hierarkisk klynging og "random forests". Disse klassifiseringsalgoritmene tilveiebringer muligheten til å klassifisere spektrum i overensstemmelse med klassemerkede spektre (f.eks. spektre som har blitt klassifisert og merket fra en kontrollgruppe av cancerpasienter), men mangler transparent og letthet av anvendelse av de ovenfor beskrevne KNN algoritmene.

Ved å fortsette med fig. 3, trinn 312, kan lærdom bli benyttet for å generere klassifiserer for et treningssett av spektre. I tilfelle av prøvetaking av serum for å detektere hvorvidt et anti-cancermedikament vil være effektivt på ikke-småcelle lungecancer, ble kontrollgrupper av pasienter benyttet, inkluderende ved å benytte tre sett av pasienter hvis cancer utviklet seg kjemoterapi. Hver av pasientene ble behandlet med Iressa og informasjon, inkluderende overlevelsestider av disse pasientene ble registrert. Kontrollprøvene var fra pasienter av mindre kraftige tilfeller (cancertrinn III og IV) som ikke mottok behandling med EGFR-K1 inhibitorer og serum ble produsert under behandlingen. En oppsummering av datasett benyttet i flere studier er tilveiebrakt i tabell III. Hvert datasett representerer cancer forskningssentret hvorfra spektret og assosiert pasientinformasjon ble mottatt.

Datasett	**Størrelse**	**Pasientdata**	**Anvendelse**
Italiensk 1	70	Fullstending	Treningssett for FIG. 12, testsett for FIG. 13
Japansk 1	43	Prognose, overlevelse	Treningssett for FIG. 13, testsett for FIG. 12
Japansk 2	26	Prognose, overlevelse	Treningssett for FIG. 13, testsett for FIG. 12
VUMC	100	Overlevelse	Kontrollsett
Italiensk 2	69	Overlevelse	Fullstendig blindet testsett for FIG. 14

Tabell III. Datasett benyttet i studie

Tabell III er en oppsummering av datasett attributtene benyttet i en studie for å bestemme hvorvidt en klassifiseringsalgoritme kan være effektiv i å bestemme hvorvidt en cancerpasient vil være mottakelig for Iressa. Datasett italiensk 1, italiensk 2, japansk 1 og japansk 2 ble behandlet med Iressa etter prøvesamling. Trening og testing i utviklingsfasen ble gjort kryssvis fra italiensk 1 settene og de to japanske settene. Pasientdataene inkluderte overlevelsesdata, hvor det italienske settet hadde veldig fullstendig pasienthistorie sammen med behandling og cancertype, de japanske settene inkluderte bar prognostisk informasjon relatert til WHO definisjonene av kliniske merker, inkluderende stabil sykdom (SS), progressiv sykdom (PS) og partielle

respondere (PR) målt ved CT avbildning. Med en gang klassifisereren var etablert, ble en fullstendig blindet test utført på italiensk 2 settet.

Fig. 10A er en graf 1000a representativ for en eksemplarisk prosess for klassifisering av et testspektrum i forhold til en gruppe av klassemerkede spektre i overensstemmelse med prinsippene av den foreliggende oppfinnelsen. Et testspektrum er betraktet til å ha et forhold til et klassemerket spektrum om det er bestemt ved en klassifiserer at testspektrumet er klassemerket det samme som minst et klassemerket spektrum fra de klassemerkede spektrene. Kurvene er spektre for gjennomsnitt av gruppe. Som vist er det en klynge av differensieringstopper omkring 11700 Dalton (Da) benyttet i klassifiseringen. Forskjellene mellom gruppene er mellom de fine klinisk merkede grupper PD-tidlig 1002 og SS-lang 1004 spekter gjennomsnittene. Selv om ikke vist er det 11 differensieringstopper benyttet for å konstruere en klassifiserer (dvs. algoritmen for klassifiserer ved å benytte modifisert k-nærmeste nabo klassifisereren) fra det italienske datasettet (tabell III) og dens parametere er optimalisert ved å benytte kryssvalidering. Det er tydelig i sammenlikning av de to grupper gjennomsnittlige spektrene at et nærvær av biomarkører resulterende i differensieringstoppene i spektra av pasienter med raskt utviklende cancer (PS-tidlig 1002) er nesten fraværende i de pasientene som overlever en lang tid og klassifisert med SS-lang cancer (SS-lang 1004).

Fig. 10B og 10C er grafer 1000b og 1000c som viser eksemplariske plot fra italiensk og to japanske treningssett. I fig. 10B varier grafen 1000b fra 5500-6000 Da og fig. 10C varierer grafen 1000c fra 11000-13000 Da. Som vist i disse to grafene 1000a og 1000b, er mange differensieringstopper mellom de forskjellige gruppene vist. Plottet av gruppene er beregnet gjennomsnitt over hver gruppe av spektre. Det betyr at plottet ikke er av individuelle spektre.

Den uvanlige fine klassifiseringen av standard differensieringstoppene er virkelig reflektert i styrken av de indikerte differensieringstoppene. En liste av differensieringstoppene benyttet er vist i tabell IV. Tabell V er den samme listen av differensieringstopper som tabell IV, men inkluderer også verdier for trekk inneholdende gruppe gjennomsnitt av trekkverdiene for oppdagelsesfase prøvene (italiensk 1, japansk 1 og 2). Et sett av dominante klynger er vist som gruppe gjennomsnitt i fig. 10. Det burde blitt forstått at differensieringstoppene vist er eksemplariske og at de samme eller andre differensieringstopper kan bli benyttet i overensstemmelse med prinsippene av den foreliggende oppfinnelsen for å forutsi cancerpasient respondere av medikamentet Iressa. Om beregninger for andre anti-cancer

eller andre medikamenter skal bli gjort, kan differensieringstopper andre enn dem listet opp bli benyttet for slike beregninger.

Den optimale k-NN klassifisereren resulterer i en "leave-on-out" kryssvalidering (LOOCV) feil, mens 6 av 26 spektre ikke kan bli klassifisert. Ved å øke kravene for de probabilistiske k-NN klassifisererene, er det mulig å knytte denne feilmerking til tilfellet av et uklassifiserbart spektrum. Om det er fornuftig antatt at den fine klassifiseringen er korrelert til prognose, hvor PS-tidlig tilfeller er den verste progresjonen og SS-lang tilfeller er de lengste stabile sykdommen, kan det tentativt bli konkludert at det er mulig å oppnå forestående medikament responsinformasjon fra pre-behandling serumspektret.

mz_senter	mz_lav	mz_høy	bredde = mz høy – mz lav
5763.791	5732.131	5795.45	63.3
5843.241	5811.097	5875.384	64.3
6433.973	6398.186	6469.759	71.6
11445.75	11376.15	11515.34	139.2
11529.52	11459.32	11599.73	140.4
11685.37	11614.03	11756.71	142.7
11759.16	11687.28	11831.04	143.8
11903.24	11830.3	11976.18	145.9
12452.38	12375.37	12529.4	154
23354.35	23183.57	23525.13	341.6
23451.25	23279.53	23622.97	343.4
66702.45	65902.02	67502.88	1600

TABELL IV. Liste av differensieringstopper

m/z	Verdi (God)	Std (God)	Verdi (dårlig)	Std (dårlig)	Bredde
5763.791	25.387	11.038	113.79	129.02	63.3
5843.241	22.617	11.595	120.27	199.13	64.3
6433.973	402.09	142.69	397.01	165.53	71.6
11445.75	22.334	16.645	353.57	756.68	139.2
11529.52	36.524	39.911	951.3	1401.1	140.4
11685.37	40.505	43.465	1019.9	2135.8	142.7
11759.16	29.745	22.773	341	472.6	143.8
11903.24	20.727	9.2393	158.1	290.08	145.9
12452.38	16.825	10.226	73.804	83.106	154
23354.35	31.089	12.447	63.381	39.39	341.6
23451.25	28.718	13.185	55.475	31.4	343.4
66702.45	342.98	250.02	369.86	203.21	1600

TABELL V. Liste av differensieringstopper inneholdende parametere av trekkverdier

5

I testing av klassifiseringsalgoritmen kan responsmarkører for Iressa bli laget med de følgende assosiasjonene: SS og PR tilfeller er gruppert sammen i en gruppe som har et klassemerke av ”god” og PS tilfeller er klassemerket som ”dårlig”. Klassifisereren

utviklet fra den fine klassifiseringen ovenfor ble så igjen assosiert "dårlig" med SS-lang og "dårlig" med PS-tidlig. Denne klassifiseringen ble så benyttet for de japanske tilfellene (tabell I) hvor 18 av disse spektrene ikke kan bli klassifisert, som etterlater 51 spektre for klassifisering. Av disse 51 spektrene hadde 37 klassemerke "gode", og 14 hadde klassemerket "dårlig". Testresultatene er oppsummert i tabell VI:

Testresultat	Opprinnelig klassemerke "god"	Opprinnelig klassemerke "dårlig"
"god"	32	6
"dårlig"	5	8

Tabell VI. Klassemerke

Denne testen har en sensitivitet på 90% og en spesifisitet på 57%. For formålene av å benytte Iressa ble 6 tilfeller, hvor det var ingen respons, dvs. "dårlig" merket til å ha en respons, som ga en positiv forutsiende verdi på 0,84. På liknende måte ble 5 tilfeller feilmerket som "dårlig", som gir en negativ forutsiende verdi på 0,61.

For å oppsummere, ved å benytte en serumbasert massespektrometer test for å filtrere ikke-respondere fra respondere i den japanske populasjonen responsraten av Iressa fra 65% til 90%, imens 5 av 51 pasienter, som kanskje har hatt fordel av Iressa vil bli utelatt. Av disse 5 pasienten ble 1 merket SS og 4 ble merket PR. Generelt er klassifiseringen til PS verst på grunn av en høy variabilitet i denne gruppen. Dette påvirker ikke seleksjonen av de "gode" tilfellene, men resulterer i den lave spesifisiteten. Denne økningen indikerer at en lege kan oppnå uventet bedre beregninger av prognosen ved å benytte Iressa tidlig i behandlingstrinnet for en bestemt gruppe av pasienter. For disse pasienten kan Iressa bli fortsatt mens pasienter beregnet til å ha en dårlig prognose kan bli byttet til en alternativ anti-cancer terapi. Dette tillater en bedre langtids overlevelsesrate siden tidligere en alternativ anti-cancer terapi er benyttet, jo mer sannsynlig vil det føre til en fordelaktig effekt.

Ved å fortsette med fig. 3, trinn 312, kan blindtesting av klassifiseren bli utført. Dette betyr at klassifiseringsalgoritmen benytter de klassemerkede spektrene for klassifisering av testspektre (f.eks. fra nye cancerpasienter) for å bestemme hvorvidt cancerpasienten som har den samme canceren som cancerpasientene fra de klassemerkede spektrene responderer til anti-cancermedikamenter. Ved å benytte den probabilistiske KNN klassifisereren, som beskrevet her ovenfor, kan klassifisereren bli generert. Resulterende fra klassifisereren kan det være tre potensielle klassemerker, "god", "dårlig" eller

"udefinert". Et klassemerke eller klassifisering av "god" betyr at klassifisereren, i prosessering av testspektrum, bestemmer testspektrumet til å være i den samme gruppen som den "gode" gruppen av klassemerkede spektre. Resultatene av en slik blindtest er vist i fig. 14, og bekrefter resultatene av utviklingsfasen.

Ved trinn 314 av fig. 3, og som tidligere beskrevet, kan visualisering bli utført, hvor visualiseringen kan inkludere verktøy for å utføre (i) beregne gjennomsnitt av spektre, (ii) spektral variering og (iii) lokalisering av trekk. Disse visualiseringsverktøyene kan være nyttige for diagnostiske formål.

Om det er bestemt ved klassifiseren at testspektrumet er nærmest relatert til den "gode" gruppen av spektre, så vil testspektrene bli klassifisert som "gode" og pasient kan bli preskribert anti-cancermedikamentet med et bestemt nivå av konfidens for at han eller hun vil respondere. Om det er bestemt ved klassifisereren at spektrumet er nærmest relatert til den "dårlige" gruppen av spektre, så vil testspektrene bli klassifisert som "dårlig" og pasienten vil ikke bli preskribert anti-cancermedikamentet. Om det ikke kan bli bestemt at testspektrumet er assosiert med enten den "gode" eller "dårlige" gruppen av klassemerkede spektre, så vil testspektrumet bli klassifisert "ubestemt" og pasienten vil ikke bli preskribert anti-cancermedikamentet.

Tabell VII presenterer et annet eksemplarisk sett av gjennomsnittlig differensieringstopp verdier, liknende dem av tabell V, som bestemt ved ekstraksjon av trekk og seleksjonsalgoritmer i trinn 308 og 310 av fig. 3. Disse spektrene er klassifisert og merket ved klassifisereren av trinn 312 av fig. 3 som "god", "dårlig" eller "udefinert". Som listet opp har de "dårlige" spektrene differensieringstopper som har store standardavvik, vanligvis større enn amplituden av toppen, slik at toppen ikke kan bli målt. Spektrene klassifisert som "gode" har differensieringstopper som tenderer til å ha mindre amplituder og standardavvik. De "udefinerte" spektrene er et eller annet sted i midten hvor amplitudene av differensieringstoppene er mindre over bestemte m/z lokaliseringer og høyere over andre.

Gruppe/MZ	**5794.38**	**5868.02**	**11483.44**	**11572.81**	**11729.95**	**12495.04**
Dårlig	190.83±207.43	232.88±301.35	798.03±964.81	1451.46±1541.45	1747.09±2208.33	97.96±109.81
God	8.74±3.69	6.40±4.34	6.06±6.21	15.34±17.77	20.15±19.91	2.68±4.50
Udefinert	17.62±6.76	16.62±7.94	37.84±20.51	89.82±47.87	105.52±53.71	8.18±6.08

TABELL VII. Eksemplariske differensieringstopper og standardavvik

Nivå av konfidens er basert på sannsynligheten av assosiasjon med treningssettet av spektre som er satt ved delta-p parameteren for den probabilistiske KNN algoritmen. Delta-p parameteren kan bli økt opp eller ned avhengig av nivået av konfidens ønsket for å assosiere testspektret med treningssettet. I en blind teststudie ble delta-parameteren satt til 0,2 og et beregningsresultat på 02% nøyaktighet resulterte.

Imens fig. 11 er nyttig i grafisk representering av spektret i todimensjonalt rom av trekk, resulterer virkelige spektre vanligvis i 8-12 dimensjonale rom av trekk, som ofte når 8-12 dimensjoner eller høyere. Høyere eller lavere dimensjonale rom av trekk kan bli bestemt til å være adekvate eller nødvendige for å bestemme hvorvidt en cancerpasient vil være mottakelig for et anti-cancer medikament. I bestemte utførelsesformer kan kanskje derfor en lege benytte bare en eller to differensielle topper, i andre utførelsesformer vil tre eller fire differensielle topper bli benyttet, i andre utførelsesformer vil fem eller seks differensielle topper bli benyttet, i andre utførelsesformer vil syv eller åtte differensielle topper bli benyttet, i andre utførelsesformer vil ni eller ti differensielle topper bli benyttet og i andre utførelsesformer vil elleve eller tolv differensielle topper bli benyttet. Å tilføye enda flere differensielle topper enn tolv er virkelig betraktet ved oppfinnelsen. Bestemmelsen av antallet av trekk som tilveiebringer nok informasjon til å være deterministisk kan bli basert på et antall av faktorer, inkluderende for eksempel amplitude av trekkene, klassifisering av spektrene og pasientrespons til anti-cancerbehandlingen.

Ved å fortsette med fig. 3 kan en database, slik som database 220 (fig. 2) bli benyttet for å motta og lagre differensieringstopper, massespektrometer diagnostikk og/eller andre resultatparametere fra klassifiseringen og en den diagnostiske prosessen som beskrevet. Disse parametrene kan bli lagret og benyttet for fremtidig klassifisering av nye spektre fra nye cancerpasienter. Eventuelt kan databasen bli utvidet til det omfanget at presishet og pålitelighet i klassifisering av testspektre hovedsakelig sikrer en høy sannsynlighet, slik som 98%, for at en cancerpasient vil respondere til anti-cancermedikamentet.

Fig. 12 er et Kaplan-Meyer plot 1200 av testdata som viser overlevelsesrater av grupper av pasienter som klassifisert i overensstemmelse med prinsippene av den foreliggende oppfinnelsen. Kaplan-Meyer plotet 1200 er et mortalitetsplot som indikerer overlevelsesratene over bestemte varigheter av tid. Som vist levde de cancerpasientene som ble kategorisert som "gode" lengst på grunn av mottak av anti-cancermedikamenter. De cancerpasientene som ble kategorisert som "dårlig" hadde et bratt frafall i de første få månedene. De cancerpasientene som ble kategorisert som

"udefinerte" gikk jevnt ned med en lav overlevelsesrate. Dette plottet ble oppnådd i oppdagelsesfasen med testing av en klassifiserer forsøk på italiensk 1 prøver på japansk 1 og 2 prøver.

Fig. 13 er et Kaplan-Meyer plot 1300 liknende fig. 12 hvor en klassifiserer forsøkt på de japanske prøvene 1 og 2 ble testet på italiensk 1 settet. Som vist ble pasienter hvis assosierte spektrum ble klassifisert som "god" beregnet til å ha utvidet levetid fra behandling med anti-cancermedikamenter. Pasienter klassifisert som "dårlige" ble beregnet til å ha en bratt mortalitetsrate med en mindre prosentandel som strekker seg utover et år. De pasientene klassifisert som "udefinerte" hadde en bratt nedgang og ingen ble beregnet til å leve utover seks måneder. Disse beregningene viste seg å være nøyaktige med de kliniske testene.

Fig. 14 er et Kaplan-Meyer plot 1400 liknende figur 12 og 13 oppnådd fra å benytte den validerte klassifiserer blindt på italiensk 2 prøvene. Ved tidspunktet av testen var det ingen kunnskap om overlevelsesdataene ettersom de ble opprettholdt konfidensielle. Etter at klassifiseringen ble utført ble overlevelsesdataene angitt, og kurvene i fig. 14 bekreftet resultatene fra utviklingstestingen. Som vist ble pasientene klassifisert som "gode" beregnet til å ha en forlenget overlevelsesrate, og de klassifisert som "dårlig" hadde et bratt frafall med en mer begrenset levetid. I dette bestemte tilfellet ble testen kjørt med en lav delta-p, så det var ingen pasienter klassifisert som "udefinerte". Igjen var resultatene samsvarende med den aktuelle kliniske testen.

Fig. 15 er et blokkdiagram av en eksemplarisk prosess 1500 for å bestemme hvorvidt en cancerpasient vil være mottakelig for et anti-cancermedikament i overensstemmelse med prinsippene av den foreliggende oppfinnelsen. Prosessen 1500 starter med et trinn 1502 hvor testspektrum produsert ved et massespektrometer fra et serum produsert fra en cancerpasient er oppnådd. Ved trinn 1504 er testspektrumet prosessert for å bestemme et forhold til en gruppe av klassemerkede spektre produsert fra respektivt serum fra andre cancerpasienter som har det samme eller liknende kliniske trinn av canceren og kjent for å har respondert eller ikke-respondert for et anti-cancermedikament. Forholdet betyr at testspektrumet mer sannsynlig vil være assosiert eller ha de samme eller liknende egenskapene som et eller et annet klassemerket spektrum. Anti-cancermedikamentet kan være et som behandler ikke-småcelle lungecancer. Ved trinn 1506 er en bestemmelse gjort, basert på forholdet av testspektrumet til gruppen av klassifisert spektrum, for hvorvidt pasienten vil være mottakelig for anti-cancermedikamentet. Å være mottakelig betyr at anti-

cancermedikamentet vil ha noen positiv fordel for cancerpasienten. Den positive responsen vil forhåpentligvis forlenge pasientens levetid, men andre positive fordeler kan resultere fra cancerpasienten som blir behandlet med anti-cancermedikamentet.

Biomarkørene målt ved den foreliggende oppfinnelsen kan være enhver type kvantifiserbare parametere som viser seg som en topp i et massespektroskopi spektrum. Parameteren som forårsaker massespektroskopitoppen kan være forårsaket ved enhver substans, inkluderende, men ikke begrenset til spesifikke enzymer, hormoner, mRNA, DNA, RNA, proteiner, lipider, vitaminer, mineraler, metabolitter og kjemiske forbindelser. Videre kan biomarkørene bli malt fra ethvert vev eller væske samlet fra pasienten, inkluderende, men ikke begrenset til, serum, røde blodceller, hvite blodceller, negl, hud, hår, vevsbiopsi, cerebral spinalvæske, benmarg, urin, avføring, sputum, galle, bronkooalveolar væske, pleural væske og periotoneal væske.

Biomarkører kan reflektere et utvalg av sykdomsegenskaper, inkluderende nivået av eksponering for en omgivelse eller genetisk utløser, et element av sykdomsprosessen i seg selv, et mellomliggende trinn mellom eksponering og start av sykdom, eller en uavhengig faktor assosiert med sykdomstilstanden, men ikke forårsakende av patogenese. På denne måten er det betraktet at prinsippene av den foreliggende oppfinnelsen også kan være anvendbare for å bestemme spesifikke trinn av sykdom og lidelser.

Selv om eksemplene av prinsippet av den foreliggende oppfinnelsen har blitt beskrevet med hensyn på ikke-småcelle lungecancer og behandling med enkelte anti-cancermedikamenter, burde det bli forstått at prinsippene kan bli benyttet til andre cancere og andre anti-cancermedikamenter tilgjengelig nå eller i fremtiden. Videre kan prinsippene og fremgangsmåtene bli benyttet for deteksjon av enhver sykdom eller lidelse, inkluderende, men ikke begrenset til, cancer, autoimmune sykdommer eller lidelser, diabetes, genetiske sykdommer eller lidelser, virale infeksjoner, bakterielle infeksjoner, parasititt infeksjoner, prion sykdommer, ernæringsmessige mangler, vitaminmangler, mineralmangler, mitokondrielle sykdommer eller lidelser, seksuelt overførbare sykdommer eller lidelser, fødselsmangler, seksuelle sykdommer eller lidelser, immunsykdom eller lidelser, balansesykdommer eller lidelser, smerte, systemiske sykdommer eller lidelser, blodsykdommer eller lidelser, blodkarsykdommer eller lidelser, nervesykdommer eller lidelser, muskulatursykdommer eller lidelser, hjertesykdommer eller lidelser, ryggmarkssykdommer eller lidelser, øyesykdommer eller lidelser, mental sykdommer eller lidelser, metaboliske sykdommer eller lidelser,

innvendige organsykdommer eller lidelser, lungesykdommer eller lidelser, leversykdommer eller lidelser, nyresykdommer eller lidelser, galleblæresykdommer eller lidelser, pankreasykdommer eller lidelser, gastrointestinale sykdommer eller lidelser, prostatasykdommer eller lidelser, gynekologiske sykdommer eller lidelser og hørselssykdommer eller lidelser. Videre kan prinsippene og fremgangsmåtene av den foreliggende oppfinnelsen også bli benyttet for å bestemme om en behandling vil virke for omgivelsesmessig eksponering og dens effekter, substansmisbruk og epidemiologiske studier.

Prinsippene og fremgangsmåtene kan bli benyttet for enhver medikamentbehandling, inkluderende, men ikke begrenset til generelle anestetiske medikamenter, angst og søvnlidelse medikamenter, medikamenter for psykiatrisk lidelse, antipsykotiske midler, medikamenter for affektiv lidelse, medikamenter for bevegelseslidelser, epileptiske og antiepileptiske medikamenter, medikamenter for å håndtere hjertesvikt, anti-iskemiske medikamenter, antiarrytmiske medikamenter, vaskulære medikamenter, kardiovaskulære og pulmonare medikamenter, opioid analgestika og antagonister, bronkiodilatorer, anti-inflammatoriske medikamenter, medikamenter for å håndtere bronkiospastisk sykdom, kromolynt natrium og relaterte medikamenter, respiratoriske stimulanter, hostestillende medikamenter, medikamenter som modulerer mukociliar transport, diuretika, antidiuretiske hormoner, syntetiske analoger og relaterte medikamenter, insulin, glukagon, orale hypoglykemiske midler, medikamenter for å behandle diabetes mellitus, paratyroid hormon medikamenter, bisfosfonater, calcitonin, adrenale kortikosteroider, kortikotropinfrigjørende hormon, adrenokortikotropin og antiadrenale medikamenter, tyroid hormoner, tyroidstimulerende hormon, tyrotropinfrigjørende hormone og antitryoid medikamenter, østrogener, antiøstrogener, progestiner, prevensjonsmidler, androgene og anabole midler og antagonister, gonadotropiner, antiprogestiner, aktiviner, inhibiner, gonadotropinfrigjørende hormon (GNRH), GNRH supragonister og antagonister, veksthormon, insulinliknende vekstfaktorer, prolaktin, medikamenter for å behandle en typerprolaktinemisk tilstand, fettløselige vitaminer, vannløselige vitaminer, makromineraler, mikromineraler, fluorider, avføringsmidler, medikamenter mot diaré, medikamenter som rammer gastrointestinal motilitet, antiemetika, medikamenter som virker på blod og blood-dannende organer, medikamenter som virker på immunsystemene, ikke-opiate analgetika, anti-inflammatoriske medikamenter, plasma lipid modifiserende midler, topikal kortikosteroider, tjærer, ditranol, sinkpreparater, retinoider, antimikrobielle forbindelser, medikamenter som behandler keratinisering, medikamenter for å behandle ectoparasitter, medikamenter for å behandle neoplastiske lidelser av hud, antihistaminer,

behandling av blæredannende lidelser av huden, sulfonamider, sulfoner, trimetoprin-sulfametoksazol, aminoglykosider, tetrasykliner, kloramfenikol, erytromiycin, protein synteseinhibitorer, fluorquinoloner, quinoloner, nitrofuraner, metenamin, β-laktam antibiotika, medikamenter for å behandle mykobakterielle infeksjoner, anti-soppmiddel, antivirale medikamenter, antiparasitiske medikamenter og cancer kjemoterapeutiske medikamenter.

I tillegg kan prinsippene bli benyttet for andre arter enn menneske. Imens beskrevet som å benytte serum for å utføre klassifiseringen og analysen, burde det bli forstått at ulike aspekter av prinsippene på liknende måte kan bli benyttet ved å benytte andre væsker eller vevsprøver for å generere spektre i stand til å ha differensieringstopper for å bestemme om en pasient har egenskaper av andre cancerpasienter som responderte til et anti-cancermedikament.

P a t e n t k r a v

1.

Fremgangsmåte for å bestemme hvorvidt det vil være sannsynlig at en pasient som lider av ikke-småcelle lungecancer vil ha fordel av behandling med gefitinib eller erlotinib som målsøker en epidermal vekstfaktor reseptor vei, eller hvorvidt det ikke vil være sannsynlig at nevnte pasient har fordel av behandling med gefitinib eller erlotinib k a r a k t e r i s e r t v e d at fremgangsmåten omfatter følgende trinn:

a) oppnå et massespektrum fra en blodbasert prøve fra pasienten;

b) utføre ett eller flere forhåndsdefinerte pre-prosesseringstrinn på massespektrumet oppnådd i trinn a);

c) oppnå integrerte intensitetsverdier for valgte trekk i nevnte spektrum på ett eller flere forhåndsdefinerte m/z områder etter at pre-prosesseringstrinnene av massespektrumet i trinn b) har blitt utført;

d) benytte verdiene oppnådd i trinn c) i en klassifiseringsalgoritme ved å bruke et treningssett som omfatter klassemerkede spektre oppnådd fra blodbaserte prøver fra andre pasienter med ikke-småcelle lungecancer for å bestemme om det er sannsynlig eller ikke sannsynlig at pasienten drar fordel av gefitinib eller erlotinib;

hvori nevnte ett eller flere forhåndsdefinerte m/z områder omfatter ett eller flere m/z områder valgt fra gruppen av m/z områder bestående av:

5732 til 5795
5811 til 5875
6398 til 6469
11376 til 11515
11459 til 11599
11614 til 11756
11687 til 11831
11830 til 11976
23183 til 23525
23279 til 23622
65902 til 67502.

2.

Fremgangsmåte ifølge krav 1, k a r a k t e r i s e r t v e d at trinn c) oppnår verdier fra nevnte spektrum ved minst 8 av nevnte m/z områder i gruppen.

3.
Fremgangsmåte ifølge krav 1, k a r a k t e r i s e r t v e d at trinn c) oppnår integrerte intnesitetsverdier fra nevnte spektrum ved alle m/z områdene i gruppen.

4.
Fremgangsmåte ifølge krav 1, k a r a k t e r i s e r t v e d at klassifiseringsalgoritmen omfatter en K-nærmeste naboer klassifiseringsalgoritme.

5.
Fremgangsmåte ifølge krav 4, k a r a k t e r i s e r t v e d at K-nærmeste naboer klassifiseringsalgoritmen omfatter en probabilistisk klassifiseringsalgoritme.

6.
Fremgangsmåte ifølge krav 1, k a r a k t e r i s e r t v e d at det ene eller flere pre-prosesseringstrinn inkluderer et trinn av å subtrahere bakgrunn inneholdt i spektrumet.

7.
Fremgangsmåte ifølge krav 6, k a r a k t e r i s e r t v e d at trinnet av å subtrahere bakgrunn er utført ved å benytte et robust asymmetrisk estimat for bakgrunnen inneholdt i spektrumet.

8.
Fremgangsmåte ifølge krav 6, k a r a k t e r i s e r t v e d at det ene eller flere pre-produseringstrinn inkluderer et trinn av å normalisere det bakgrunns-subtraherte spektrumet.

9.
Fremgangsmåte ifølge krav 8, k a r a k t e r i s e r t v e d at trinnet av å normalisere det bakgrunns-subtraherte spektrumet omfatter å utføre en partiell ionestrøm normalisering.

10.

Fremgangsmåte ifølge krav 8, k a r a k t e r i s e r t v e d at trinnet å normalisere det bakgrunns-subtraherte spektrumet omfatter å utføre en total ionestrøm normalisering.

11.

Fremgangsmåte ifølge krav 6, k a r a k t e r i s e r t v e d at det ene eller flere pre-prosesseringstrinn inkluderer trinn av å oppstille det bakgrunns-subtraherte spektrumet med en på forhånd definert masseskala.

12.

Fremgangsmåte ifølge krav 1, k a r a k t e r i s e r t v e d at massespektrumet er oppnådd fra et MALDI massespektrometer.

13.

Apparat konfigurert for å bestemme hvorvidt det vil være sannsynlig at en pasient som lider av ikke-småcelle lungecancer vil ha fordel av behandling med gefitinib eller erlotinib som målsøker en epidermal vekstfaktor reseptor vei, eller hvorvidt det ikke vil være sannsynlig at nevnte pasient har fordel av behandling med gefitinib eller erlotinib, k a r a k t e r i s e r t v e d at det omfatter:

en lagringsanordning som lagrer et massespektrum av en blodbasert prøve fra pasienten, og

en prosessor som utfører programvareinstruksjoner konfigurert for å:

a) oppnå integrerte intensitetsverdier av trekk i nevnte massespektrum ved et eller flere m/z områder, hvor m/z områdene er valgt fra gruppen av m/z områder bestående av:

5732 til 5795

5811 til 5875

6398 til 6469

11376 til 11515

11459 til 11599

11614 til 11756

11687 til 11831

12375 til 12529

23183 til 23525

23279 til 23622

65902 til 67502; og

b) benytte en klassifiseringsalgoritme som opererer på verdiene av trekkene i spektrumet ved det valgte ene eller flere m/z områder og ved å benytte et treningssett omfattende klassemerkede spektre oppnådd fra blodbaserte prøver fra andre pasienter med ikke-småcelle lungecancer for å bestemme om det er sannsynlig eller ikke sannsynlig at pasienten drar fordel av gefitinib eller erlotinib som målsøker en epitel vekstfaktor reseptorvei.

14.

Apparat ifølge krav 13, k a r a k t e r i s e r t v e d at instruksjonene oppnår integrerte intensitetsverdier ved minst åtte av m/z områdene i gruppen.

15.

Apparat ifølge krav 14, k a r a k t e r i s e r t v e d at instruksjonene oppnår integrerte intensitetsverdier ved alle m/z områdene i gruppen.

16.

Apparat ifølge krav 13, k a r a k t e r i s e r t v e d at klassifiseringsalgoritmen omfatter en K-nærmeste nabo klassifiseringsalgoritme.

17.

Apparat ifølge krav 16, k a r a k t e r i s e r t v e d at K-nærmeste nabo klassifiseringsalgoritmen omfatter en probabilistisk K-nærmeste nabo klassifiseringsalgoritme.

FIG. 1

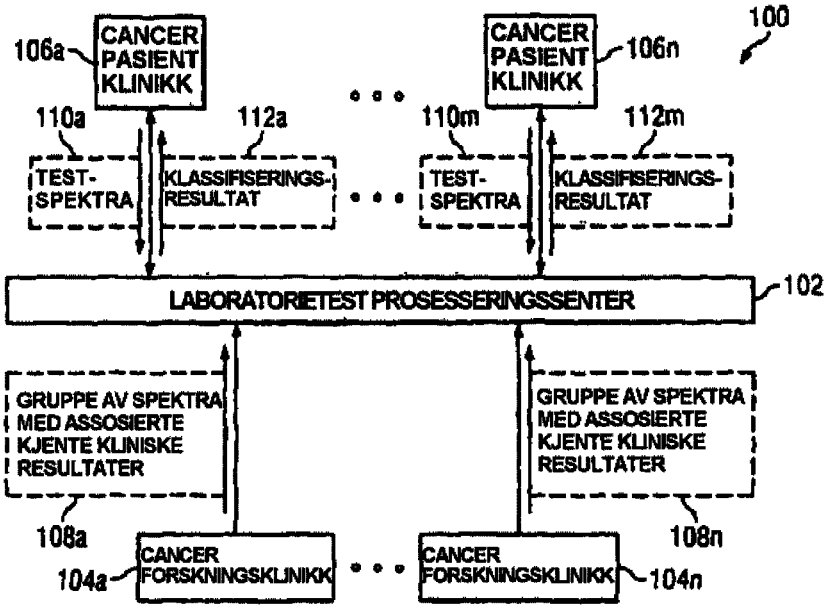

100
106a
CANCER PASIENT KLINIKK
106n
CANCER PASIENT KLINIKK
110a
112a
110m
112m
TEST-SPEKTRA
KLASSIFISERINGS-RESULTAT
TEST-SPEKTRA
KLASSIFISERINGS-RESULTAT
LABORATORIETEST PROSESSERINGSSENTER
102
GRUPPE AV SPEKTRA MED ASSOSIERTE KJENTE KLINISKE RESULTATER
GRUPPE AV SPEKTRA MED ASSOSIERTE KJENTE KLINISKE RESULTATER
108a
108n
CANCER FORSKNINGSKLINIKK
104a
CANCER FORSKNINGSKLINIKK
104n

FIG. 2

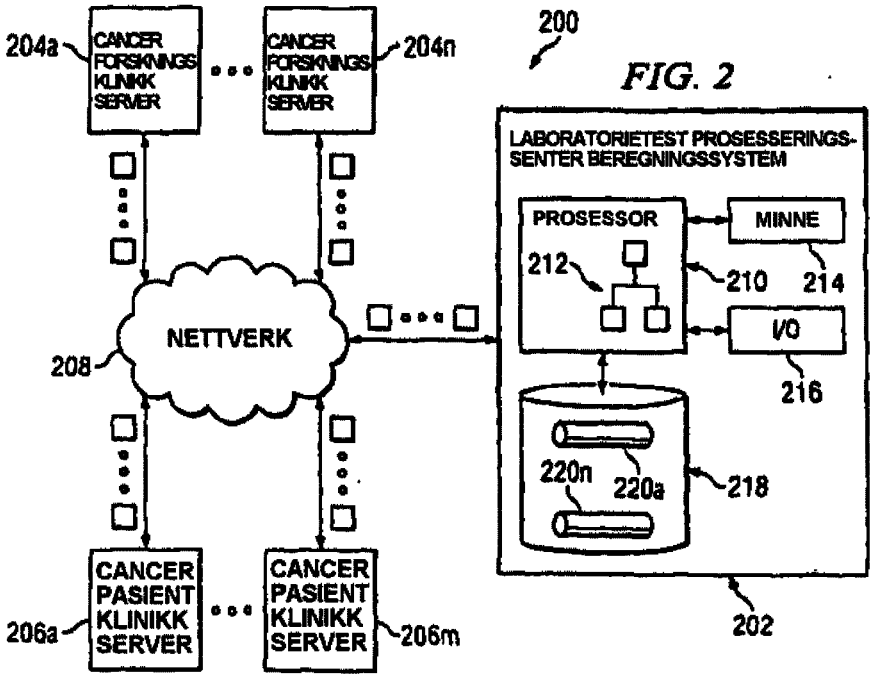
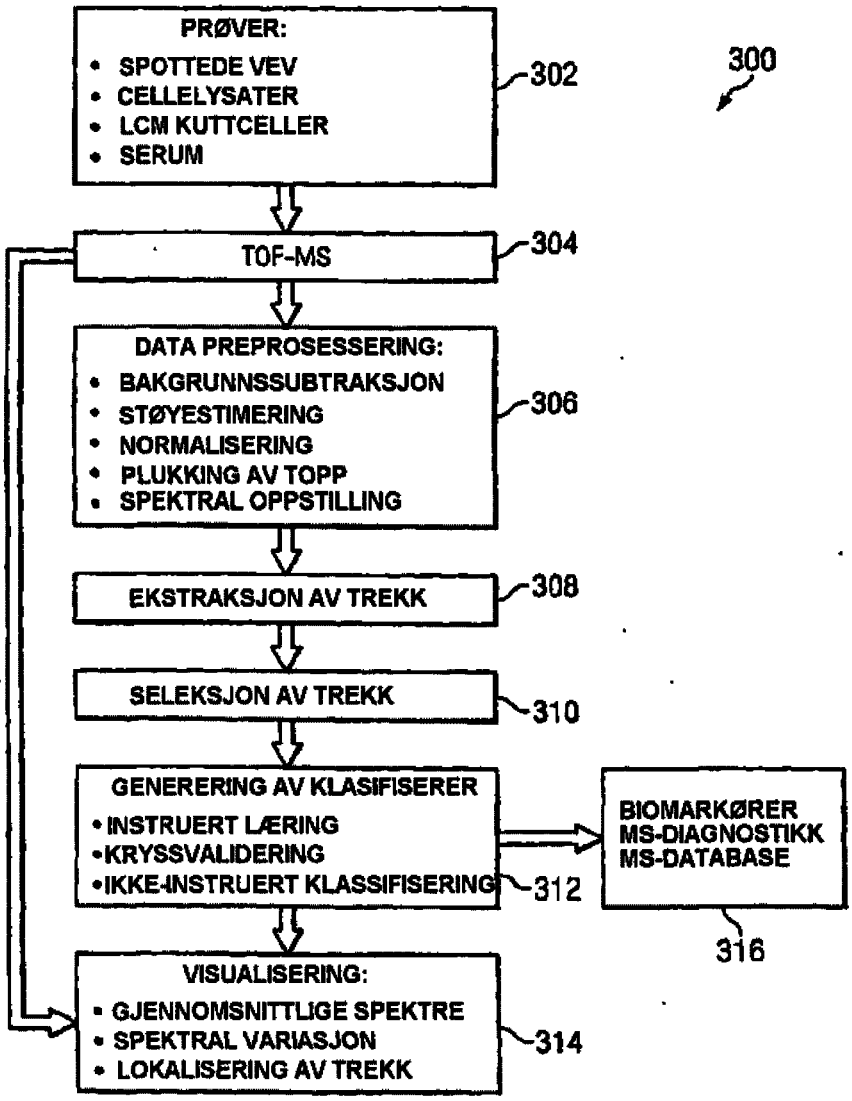

200
204a
CANCER FORSKNINGS KLINIKK SERVER
CANCER FORSKNINGS KLINIKK SERVER
204n
LABORATORIETEST PROSESSERINGS-SENTER BEREGNINGSSYSTEM
PROSESSOR
MINNE
212
210
214
I/O
216
NETTVERK
208
220n
220a
218
CANCER PASIENT KLINIKK SERVER
206a
CANCER PASIENT KLINIKK SERVER
206m
202

FIG. 3


PRØVER:
• SPOTTEDE VEV
• CELLELYSATER
• LCM KUTTCELLER
• SERUM
302
300
TOF-MS
304
DATA PREPROSESSERING:
• BAKGRUNNSSUBTRAKSJON
• STØYESTIMERING
• NORMALISERING
• PLUKKING AV TOPP
• SPEKTRAL OPPSTILLING
306
EKSTRAKSJON AV TREKK
308
SELEKSJON AV TREKK
310
GENERERING AV KLASIFISERER
• INSTRUERT LÆRING
• KRYSSVALIDERING
• IKKE-INSTRUERT KLASSIFISERING
312
BIOMARKØRER
MS-DIAGNOSTIKK
MS-DATABASE
316
VISUALISERING:
• GJENNOMSNITTLIGE SPEKTRE
• SPEKTRAL VARIASJON
• LOKALISERING AV TREKK
314

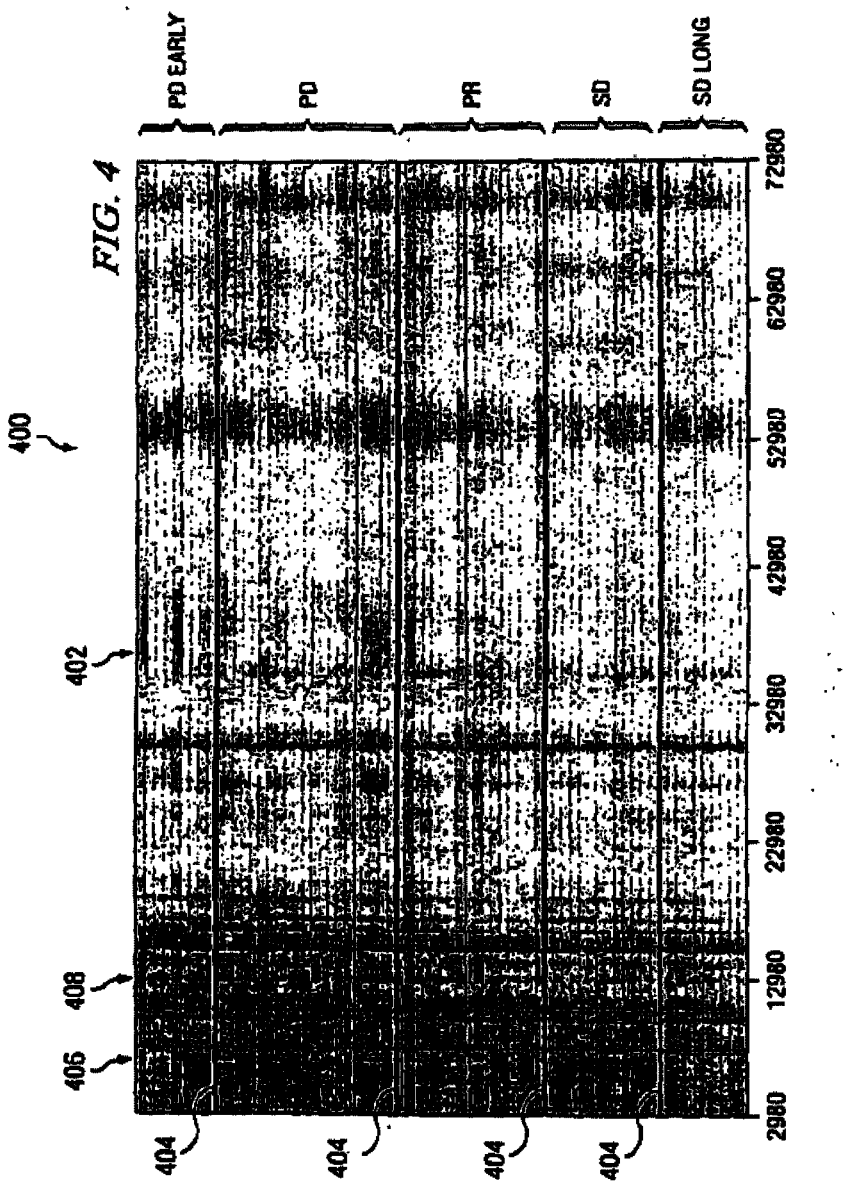
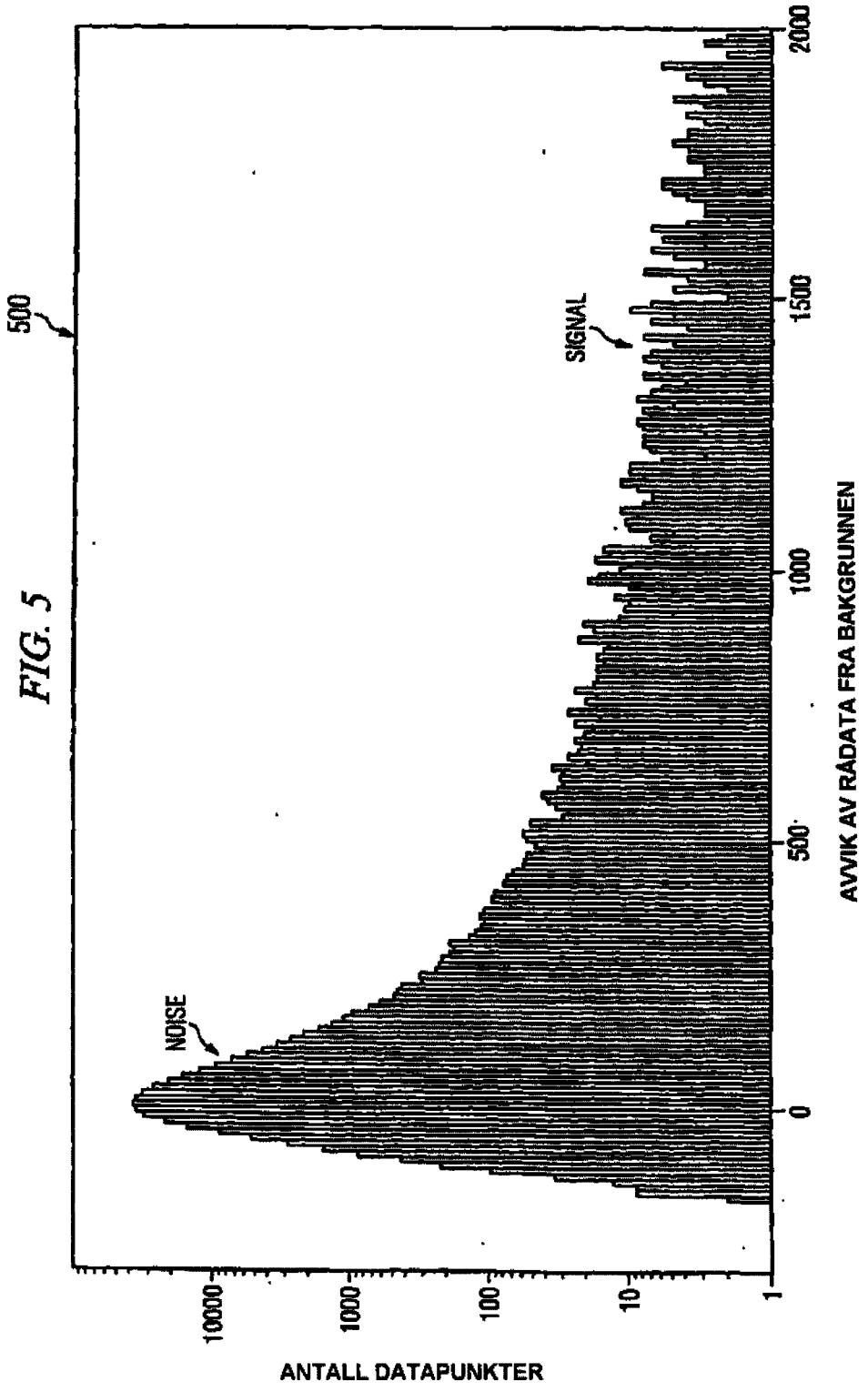

PD EARLY
PD
PR
SD
SD LONG
400
402
406
408
404
404
404
404
72980
62980
52980
42980
32980
22980
12980
2980

FIG. 4

FIG. 5


500
NOISE
SIGNAL
10000
1000
100
10
1
0
500
1000
1500
2000
ANTALL DATAPUNKTER
AVVIK AV RÅDATA FRA BAKGRUNNEN

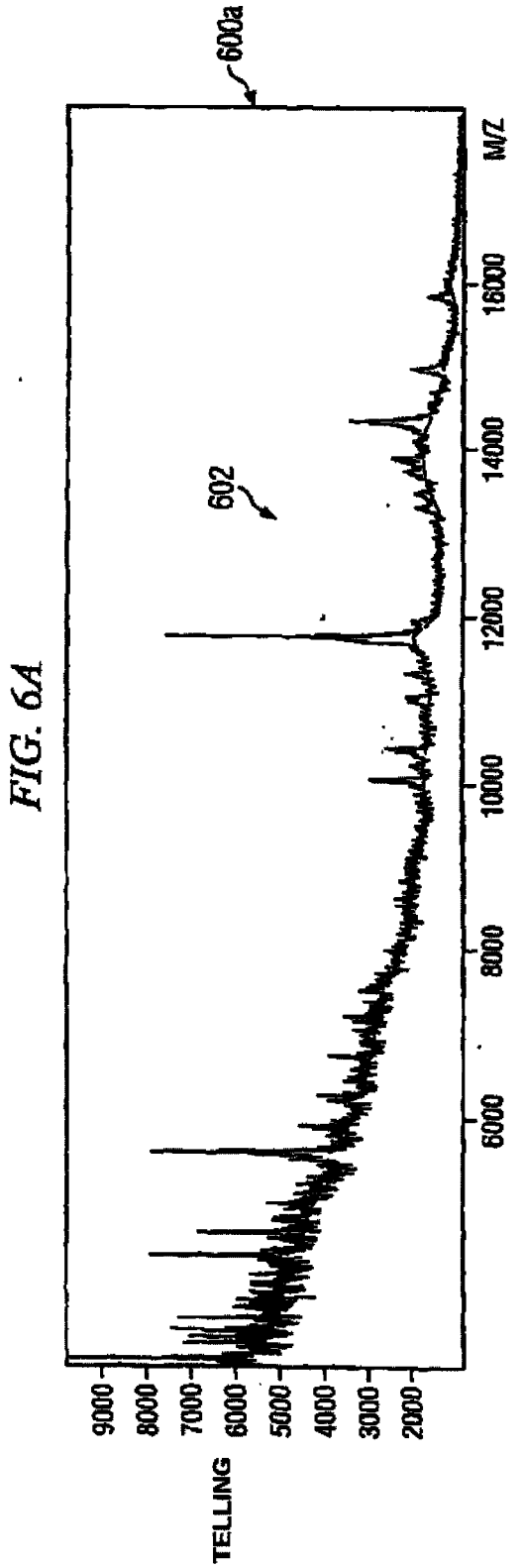

600a
602
TELLING
M/Z
9000
8000
7000
6000
5000
4000
3000
2000
6000
8000
10000
12000
14000
16000

FIG. 6A

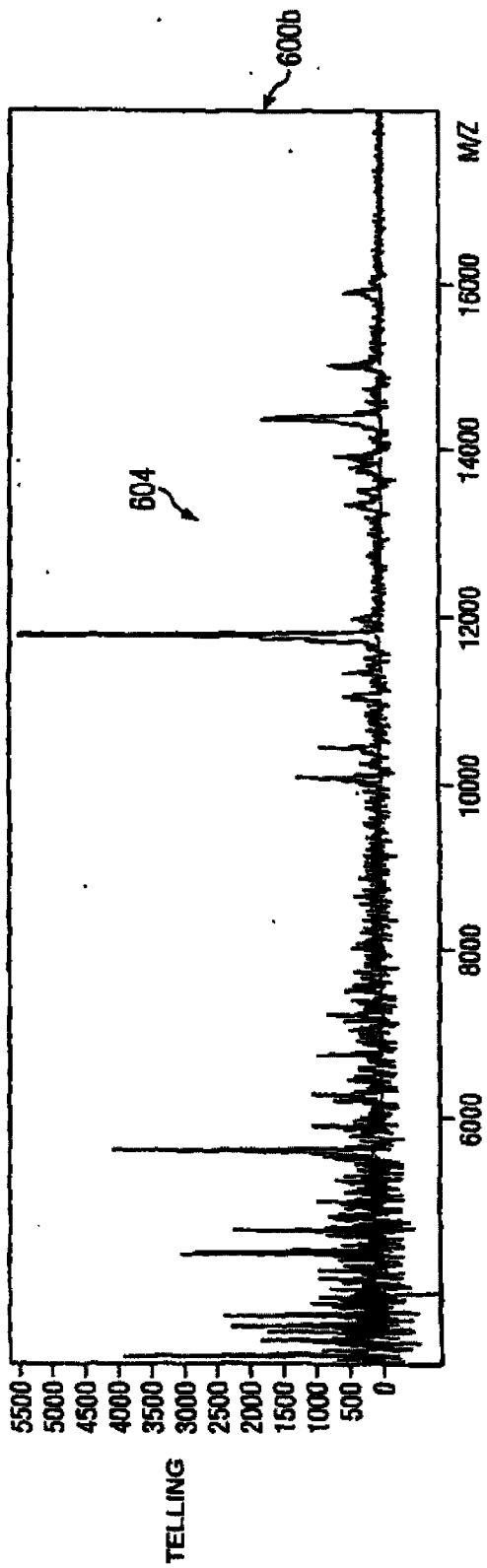

600b
604
TELLING
M/Z
5500
5000
4500
4000
3500
3000
2500
2000
1500
1000
500
0
6000
8000
10000
12000
14000
16000

FIG. 6B

FIG. 7A

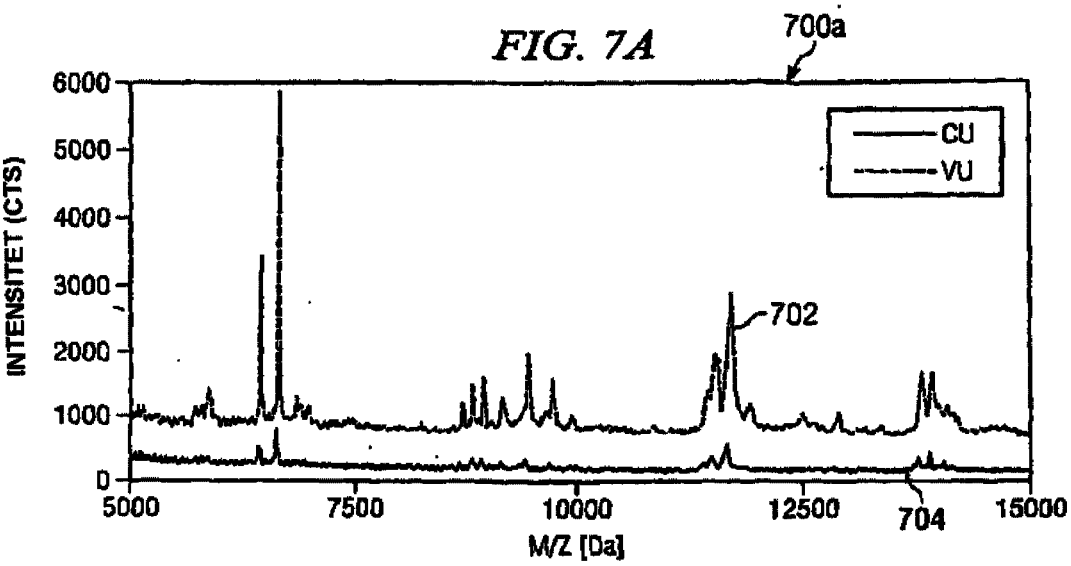
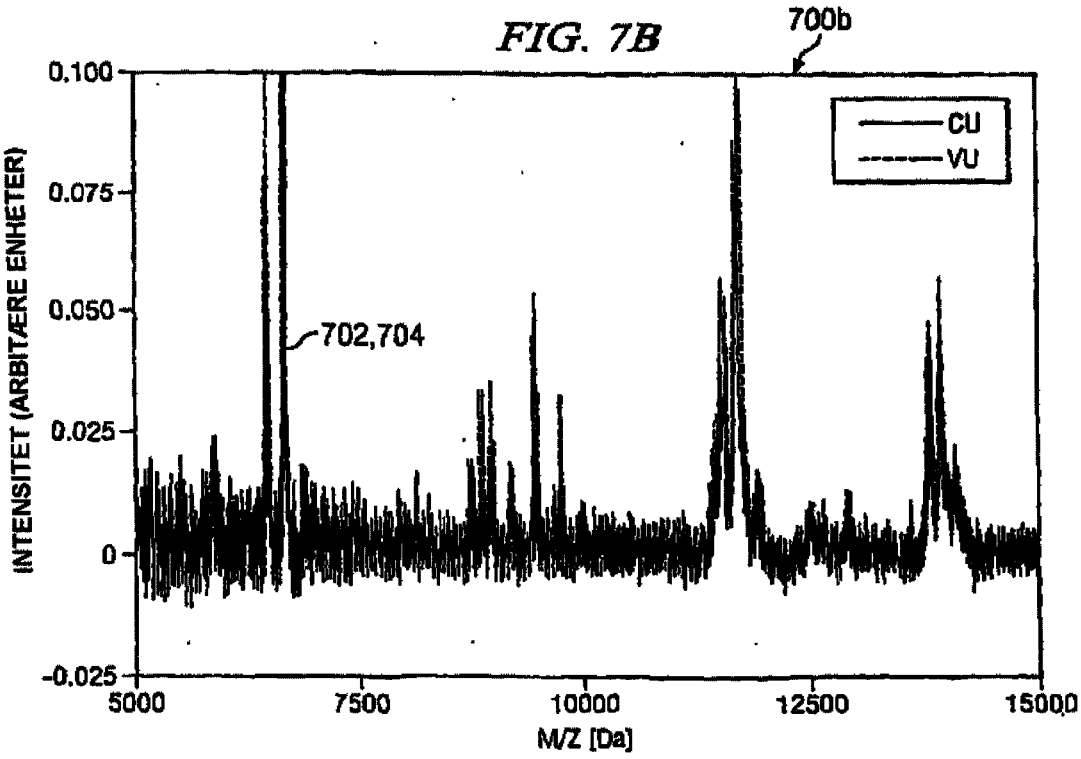

700a
6000
5000
4000
3000
2000
1000
0
INTENSITET (CTS)
CU
VU
702
704
5000
7500
10000
12500
15000
M/Z [Da]

FIG. 7B


700b
0.100
0.075
0.050
0.025
0
-0.025
INTENSITET (ARBITÆRE ENHETER)
CU
VU
702,704
5000
7500
10000
12500
15000
M/Z [Da]

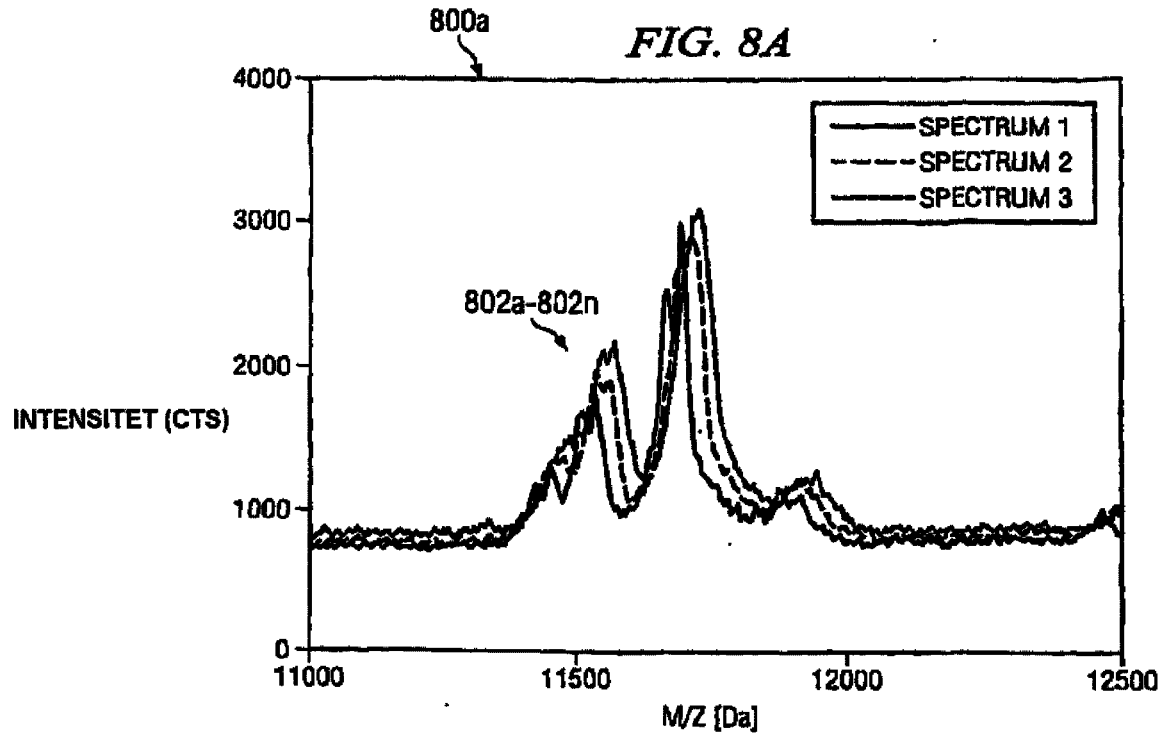

800a
4000
3000
2000
1000
0
INTENSITET (CTS)
802a-802n
SPECTRUM 1
SPECTRUM 2
SPECTRUM 3
11000
11500
12000
12500
M/Z [Da]

FIG. 8A

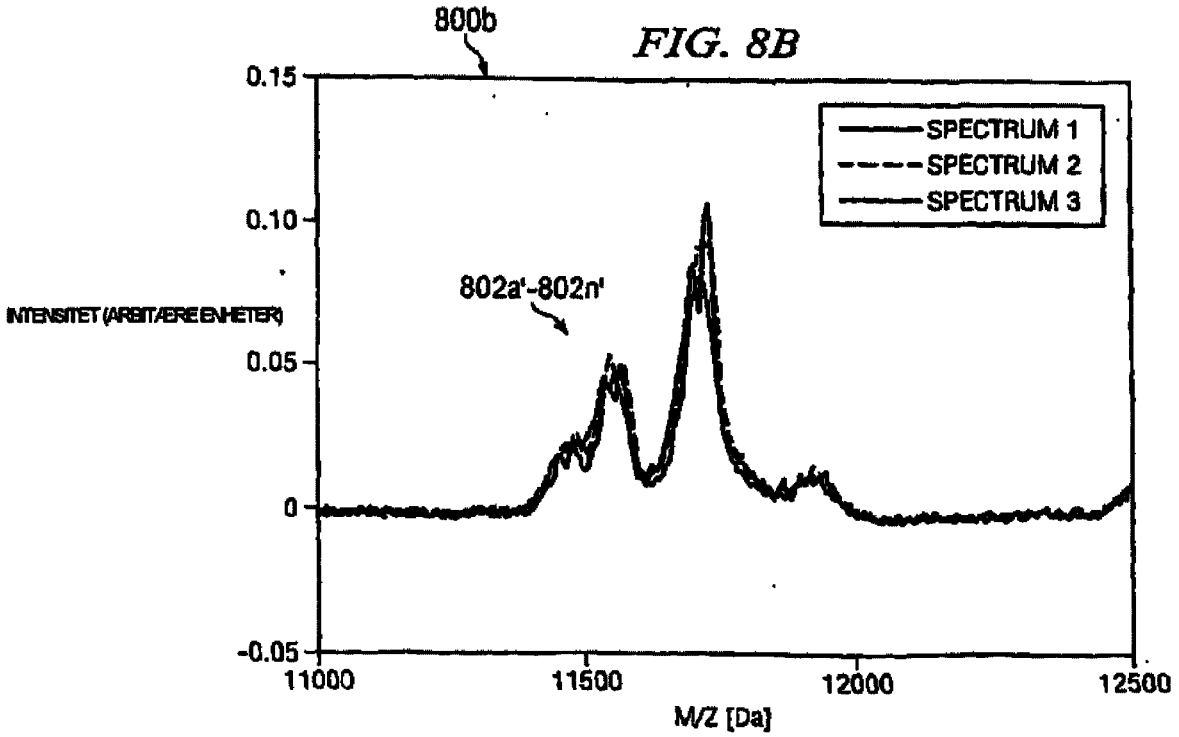

800b
0.15
0.10
0.05
0
-0.05
INTENSITET (ARBITÆRE ENHETER)
802a'-802n'
SPECTRUM 1
SPECTRUM 2
SPECTRUM 3
11000
11500
12000
12500
M/Z [Da]

FIG. 8B

FIG. 9

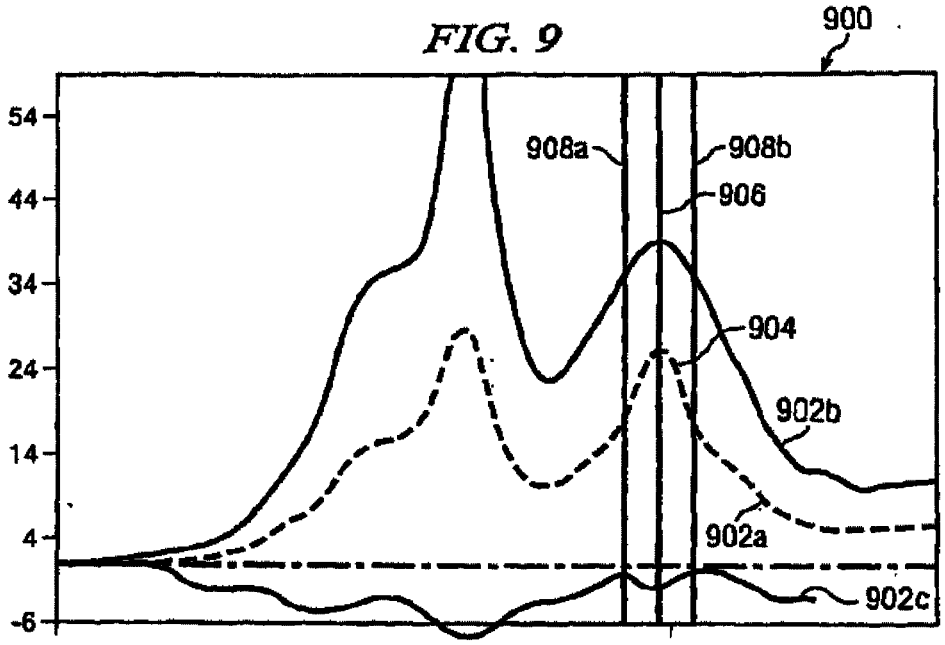

900
54
44
34
24
14
4
-6
908a
908b
906
904
902b
902a
902c

FIG. 10A

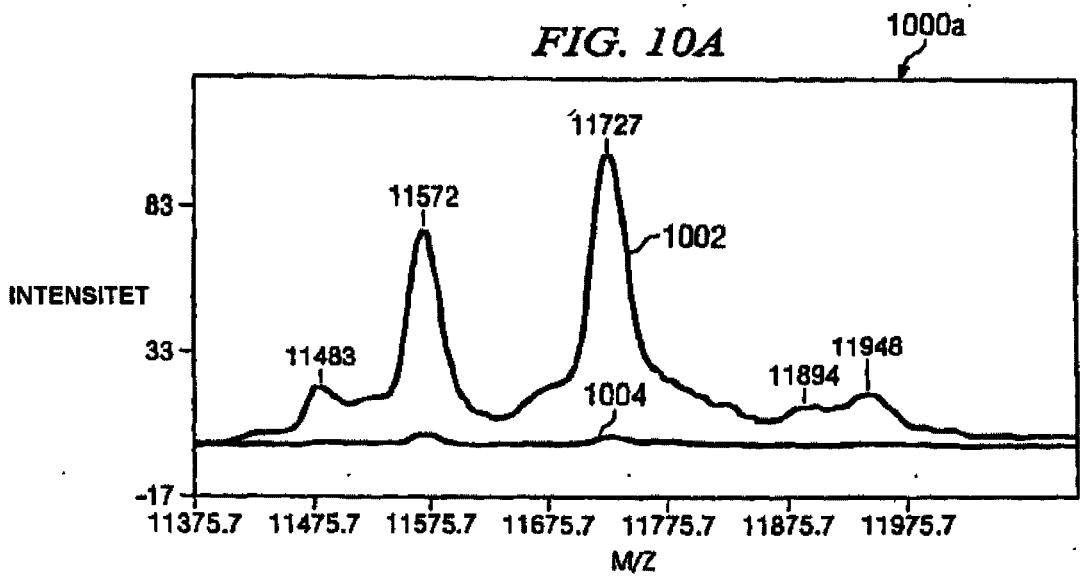

1000a
INTENSITET
83
33
-17
11727
11572
1002
11483
11894
11948
1004
11375.7
11475.7
11575.7
11675.7
11775.7
11875.7
11975.7
M/Z

FIG. 10B

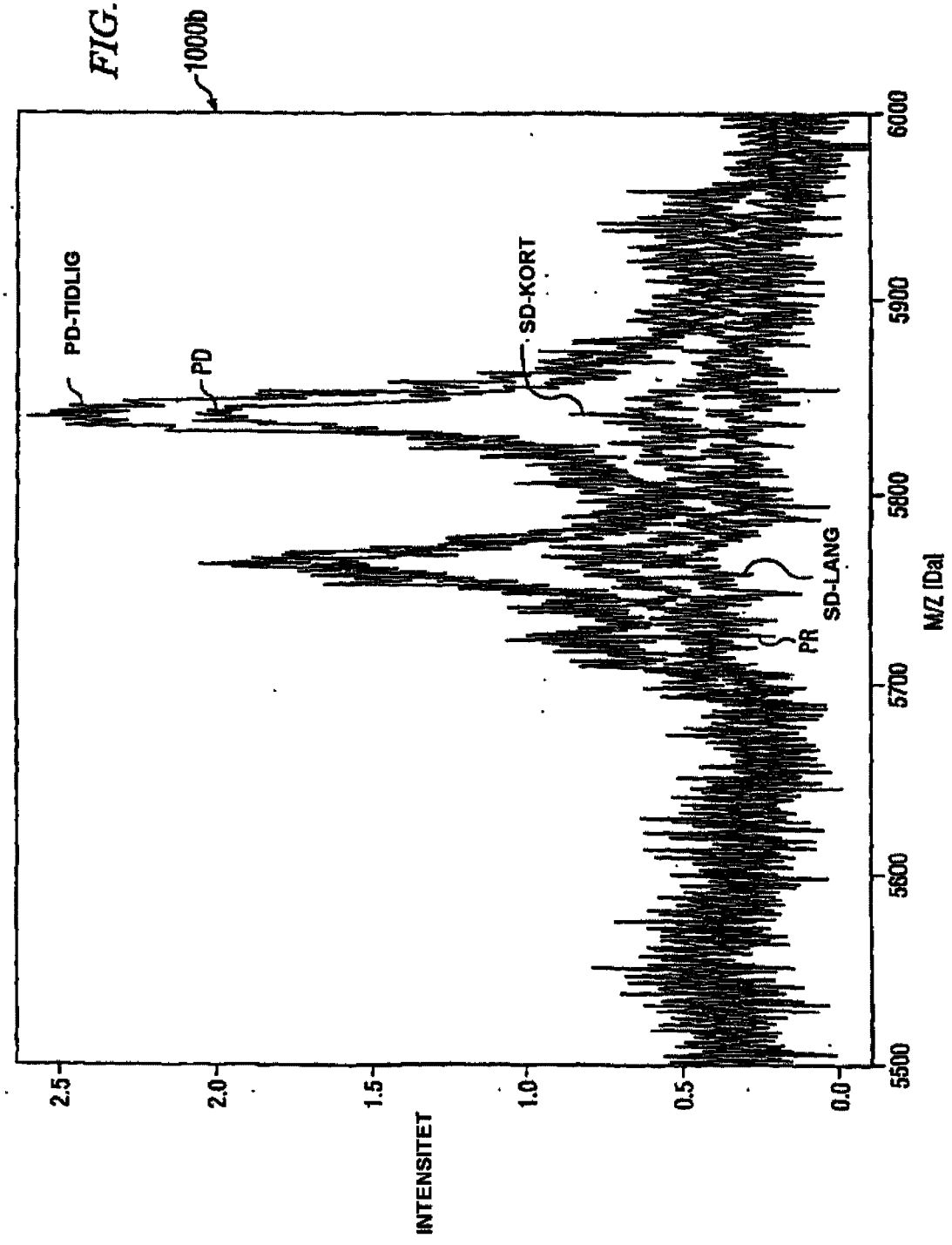

1000b
PD-TIDLIG
PD
SD-KORT
SD-LANG
PR
INTENSITET
M/Z [Dal
2.5
2.0
1.5
1.0
0.5
0.0
5500
5600
5700
5800
5900
6000

FIG. 10C

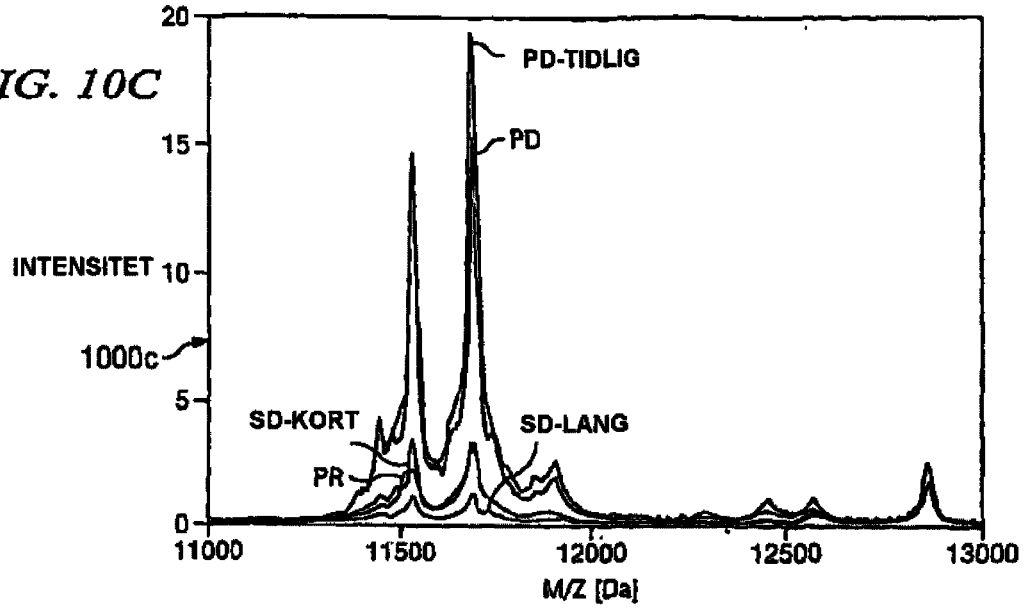


FIG. 11

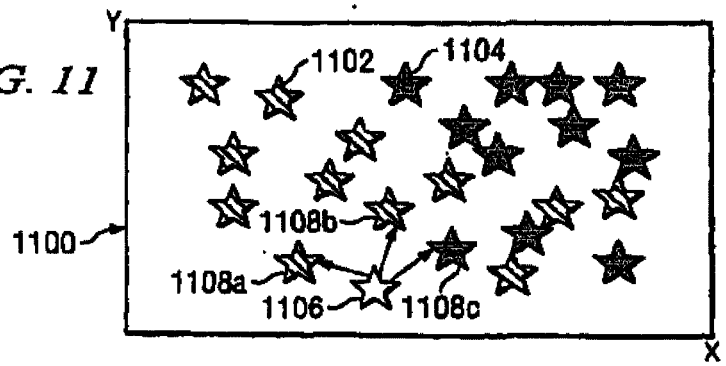
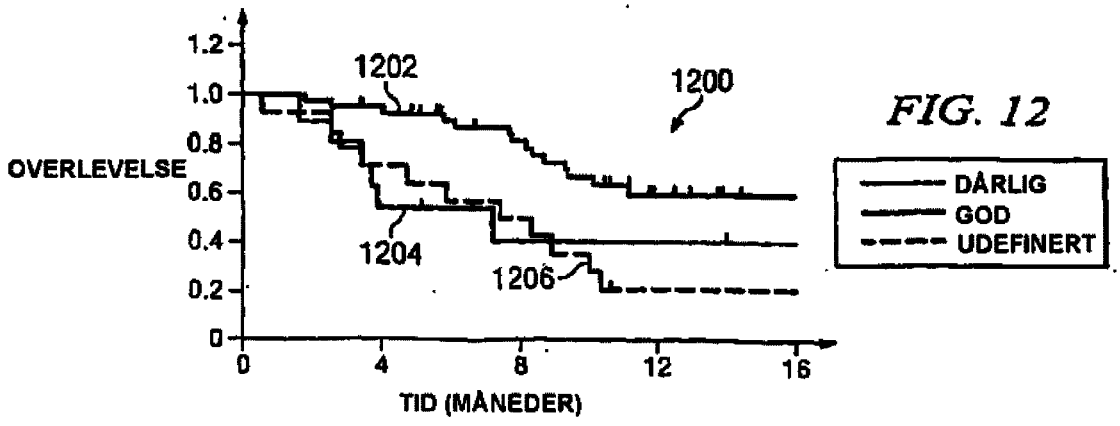


FIG. 12



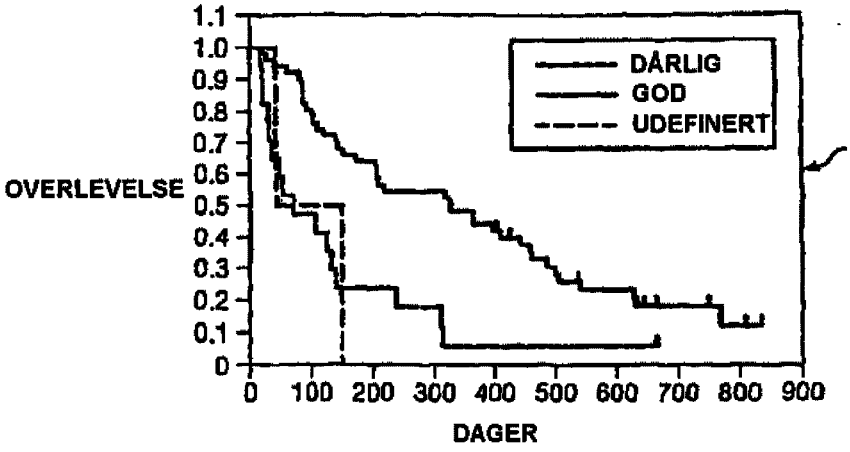


FIG. 13

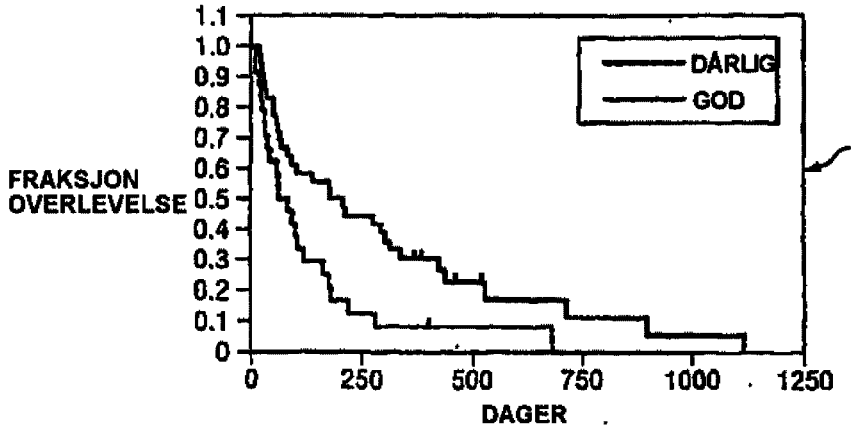


FIG. 14

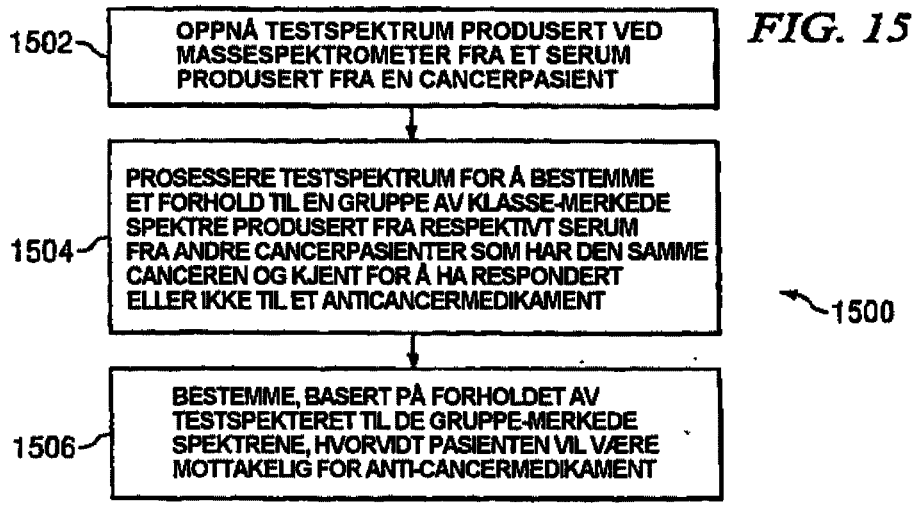


FIG. 15