BIOINFORMATIKK – ET INTERESSANT FORSKNINGSFELT FOR STATISTIKERE?

METTE LANGAAS, NORSK REGNESENTRAL
Mette.Langaas@nr.no

HVA ER BIOINFORMATIKK?

Bioinformatikk er et tverrfaglig forskningsfelt der matematikk, statistikk og informatikk anvendes til å analysere data som er produsert av eksperimentelt arbeid innen biokjemi, cellebiologi og genetikk[1].

I en lederartikkel i tidsskriftet *Bioinformatics* foreslår Professor Russ Altman at pensum for en utdanning innen bioinformatikk bør deles inn i fem områder; biologi, informatikk, statistikk, etikk og "bioinformatiske kjernefag". Innen biologi bør det legges vekt på molekylærbiologi, cellebiologi og genetikk. Den statistiske delen av utdanningen bør inneholde kurs i sannsynlighetsteori, statistisk forsøksplanlegging, statistisk analyse og stokastiske prosesser. Spesielt blir optimering, dynamisk programmering, søkealgoritmer, klusteranalyse, klassifikasjon, nevrale nettverk, genetiske algoritmer og Bayesiansk inferens nevnt som metoder som ofte benyttes innen bioinformatikk. Innen informatikk er programmering, datastrukturer, algoritmer og håndtering av databaser grunnleggende. Bioinformatiske kjernefag består av blant annet av design og implementasjon av biologiske databaser, annotasjon av biologiske sekvenser, metoder som støtter laboratorie-arbeid (derunder DNA arrays), modellering og prediksjon av proteinstruktur, metoder for sammenligning av biologiske sekvenser (sequence alignment), Hidden Markov Models og fylogeniske trær.

I Norge finnes hovedfagsutdanning innen bioinformatikk ved Gruppe for bioinformatikk ved Institutt for informatikk ved Universitetet i Bergen og Gruppe for bioinformatikk ved Institutt for informatikk ved Universitetet i Oslo. Et nytt hovedfagsstudium innen bioinformatikk er planlagt opprettet fra 2002 ved Norges Landbrukshøgskole i Ås (naturvitenskapelige fag). Ved Institutt for datateknikk og informasjonsvitenskap (faggruppe for kunnskapssystemer) ved Norges Teknisk-Naturvitenskapelige Universitet i Trondheim arbeider man med bioinformatiske problemstillinger. Ved Norsk Regnesentral arbeider vi med statistiske aspekter av bioinformatikk, og planlegger en økt satsning på feltet fremover. Vi samarbeider med "Mikromatriseprosjektet på Radiumhospitalet" og ønsker å opprette ytterlig samarbeid med akademiske og kommersielle aktører innen bioinformatikk i Norge og i utlandet.

GENOM OG GEN

Kroppen vår består av mer enn 10^{12} celler. Inne i hver celle finner vi organeller som er ansvarlige for energiproduksjon, proteinsyntese og lagring og videreføring av genetisk informasjon til nye celler. Den totale genetiske informasjonen kalles *genomet.* Alle cellene i kroppen vår inneholder den samme genetiske informasjonen, men kun deler av denne informasjonen er aktiv i en gitt celle. Deler av informasjonen som er aktiv i en hjernecelle er inaktiv (sovende) i en levercelle og vice versa. Hvilken del av den genetiske informasjonen som er aktiv i en celle bestemmer hvordan cellen fungerer.

Det humane genomet er fordelt på 46 kromosomer, der hvert kromosom inneholder et molekyl deoksyribonukleinsyre (DNA). Byggesteinene i DNA-molekylet er nukleotide-enheter (sukkeret deoksyribose, en fosfatgruppe og en base), der fire mulige baser inngår[2]. DNA er laget som en dobbel helix, der to kjeder av nukleotider er festet sammen med hydrogenbindinger mellom basene[3]. Man har anslått at det humane genom består av 3 milliarder basepar. Et gen er et segment av et DNA molekyl, lokalisert i en spesifikk posisjon på et spesifikt kromosom. De aller fleste genene inneholder oppskriften på et protein (enzymer er proteiner som er viktige i produksjonen av ulike molekyler). Et viktig mellomprodukt i syntesen av proteiner heter mRNA (messenger RNA). Hvis vi ønsker å se om

---

[1] Denne definisjonen er i bruk ved Chalmers Universitet. Mange ulike definisjoner av bioinformatikk finnes; en meget bred definisjon er "bioinformatikk er å bruke informasjonsteknologi til å forstå biologi". Et relatert begrep er "computational biology".

[2] Basene heter G=guanine, C=cytosine, T=thymine og A=adenine

[3] Bindingene dannes mellom basene G og C og mellom T og A. Dette kalles basepar.

et gen er aktivt i en celle (*genekspresjon*) kan vi kvantifisere mengden av proteinet som genet koder for i denne cellen. Det er mer komplekst å måle proteinmengden enn mRNA mengden i en celle, og målinger av mRNA mengde er i dag ofte brukt når man ønsker å måle genekspresjon.

DET HUMANE GENOM PROSJEKT

Kartlegging av genomet vil si å identifisere posisjonen og DNA-sekvensen til hvert gen. Den komplette DNA-sekvensen for mange genomer[4] er kjent, mens sekvensiering av genomet til mange planter og dyr pågår. "Det humane genom prosjekt" (HGP) ble startet i 1990 under ledelse av The National Center for Human Genome Research i USA og har som mål å kartlegge hele det humane genomet. Et konkurrerende privat selskap, Celera Genomics, er blitt dannet med det samme målet. De to selskapene har benyttet to ulike teknikker for å sekvensiere genomet. I februar 2001 publiserte HGP sine resultater i Nature og Celera Genomics sine resultater i Science. Antall gener ble der anslått til å være rundt 30 000 (et mye mindre tall enn tidligere antatt). Kart over plasseringen av de ulike genene på de ulike kromosomene, og informasjon om egenskapene til det ulike genene er tilgjengelige fra HGPs WWW-sider[5]. Informasjonen legges ut kontinuerlig og kan benyttes fritt.

ULIKE FORSKNINGSOMRÅDER INNENFOR BIOINFORMATIKK

Ved at genomet til mange organismer nå er ferdig kartlagt, har vi beveget oss fra den pre-genome æra til den post-genome æra. Den viktigste utfordringen i den post-genome æra er å utforske funksjonen til genomet, dvs. rollen til hvert gen. Dette kalles *funksjonell genomikk*, og vil være hovedtemaet for resten av denne presentasjonen.

Det finnes imidlertid mange andre forskningsområder innen bioinformatikk, og mye av det eksperimentelle arbeidet og metodegrunnlaget har felles komponenter innen de ulike områdene. Noen av de mest profilerte områdene er som følger. Innen *proteomikk* ser man på hvilke proteiner som er tilstede i en celle [6] og hvordan de ulike proteinene samvirker. I *strukturell genomikk* ønsker man å bestemme den tredimensjonale strukturen til proteiner som er kodet i genomet. Hvordan ulike genetiske forskjeller påvirker en pasients respons på medisiner er hovedspørsmålet innen *pharmacogenetikk*. I *komparativ genomikk* ser man på funksjonen til menneskets gener og andre områder av menneskets DNA ved å studere deres paralleller i andre organismer (f.eks. mus).

FUNKSJONELL GENOMIKK: GENEKSPRESJON OG DATA FRA DNA MIKROARRAYS

Hvordan en celle fungerer kan generelt ikke beskrives ved aktiviteten til ett gen, den oppstår som et resultat av samspill mellom mange gener. Måling av aktiviteten til flere tusen gener samtidig kan danne grunnlaget for økt forståelse av ulike biologiske prosesser. Før ble ett og ett gen studert separat. Nye teknikker, som DNA mikroarrays, gjør det mulig å studere rollen til tusenvis av gener samtidig og dermed kunne avdekke komplekse samspill mellom gener.

Det finnes to hovedformater DNA mikroarrays; oligonukleotide- og cDNA mikroarray. Det mest brukte DNA mikroarray formatet i den akademiske forskningen er cDNA mikroarray for analyse av genekspresjon (se figur 1). I et cDNA mikroarray eksperiment inngår en glassplate, der små dråper av gener (kalt prober) er avsatt ved hjelp av en mikroarray-robot[7], og to ulike celleprøver som er bearbeidet og merket[8]. Antall gener som avsettes er i størrelsesorden 2000-10000, men i fremtiden vil glassplaten kunne inneholde hele det humane genomet. Genekspresjonen til de to celleprøvene anslås ved at celleprøvene reagerer (hybridiserer) med genene på glassplaten. Ved hjelp av en laserscanner avbildes glassplaten ved to ulike bølgelengder (jmf. merkingen). Resultatet er to gråtonebilder. Ved hjelp av bildeanalyse gjøres de to bildene om til et datasett med en intensitet for hvert gen for hver

---

[4] Blant annet gjær, bakterien E.coli, ormen C.elegans, bananflue.

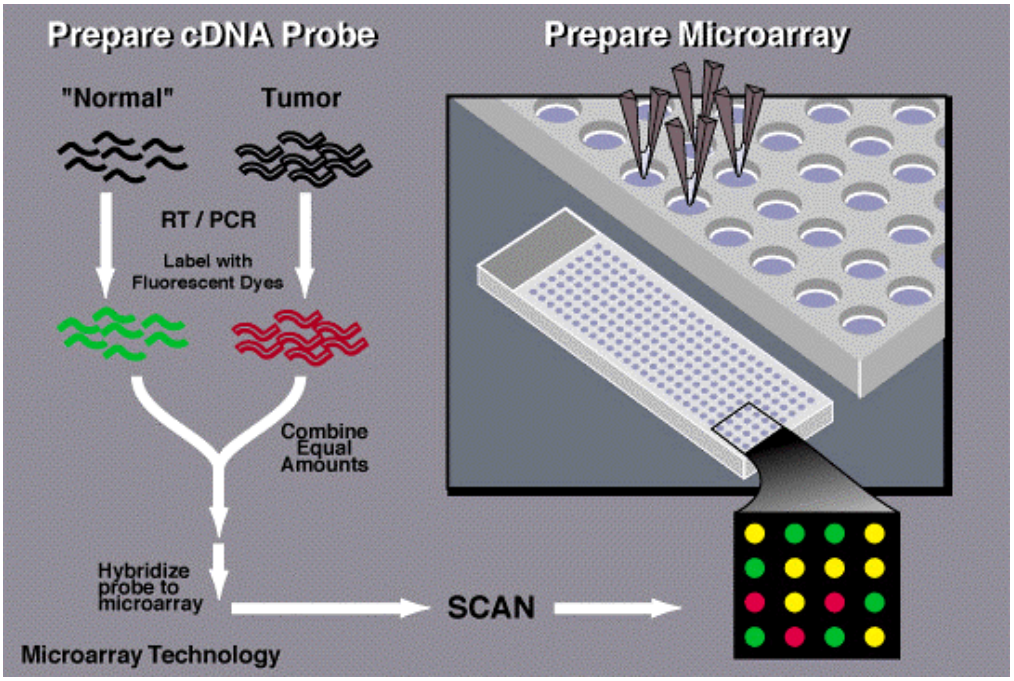
[5] http://www.ornl.gov/hgmis/

[6] Man har anslått at mennesket kan produsere i størrelsesorden 80000 ulike proteiner.

[7] I Norge finnes det to mikroarray-roboter; en ved NTNU i Trondheim og en ved Det Norske Radiumhospital. Minst to nye roboter er planlagt bygget.

[8] Vi tenker oss her at prøvene er merket med to typer fluorescent fargestoff (rødt og grønt). De to fargestoffene emitterer lys når de blir belyst med lys av to ulike bølgelengder. Farging med fluorescent fargestoff er den mest populære merketeknikken.

Figur 1: Et cDNA mikroarray eksperiment. Figuren er tatt fra HGPs sider http://www.ornl.gov/hgmis/



celleprøve. Intensitetene er anslag for mengden av mRNA assosiert med de ulike genene i de to celleprøvene (dvs. hvor mye et gen er "slått-på", aktivt, eller "slått-av", inaktivt").

Grunnen til at to celleprøver inngår i et cDNA mikroarray eksperiment er at det er vanskelig å kontrollere nøyaktig mengden av probemateriale som avsettes i hvert område på mikroarrayen (probeområdene er kalt spots). For hver spot studeres derfor vanligvis ratioen mellom intensitetene for de to celleprøvene. Ønsker man å undersøke mer enn to celleprøver er det mest brukte forsøksdesignet et såkalt referansedesign, der en av celleprøvene defineres som en referanse. Hvis man har n celleprøver, inngår en av celleprøvene i alle cDNA mikroarray eksperimentene, dvs. man bruker (n - 1) mikroarray eksperimenter til å undersøke n celleprøver. F.eks. i et forsøk der man vil undersøke effekten av ulike doser av radioaktiv stråling på en kreftsvulst, kan referansen være en celleprøve som ikke har vært utsatt for stråling. Et cDNA mikroarray datasett inneholder typisk 2000 til 10000 gener (variable) og 2 til 100 celleprøver (observasjoner).

STATISTISKE METODER FOR Å ANALYSERE GENEKSPRESJONS-DATA

Teknikker som DNA mikroarrays produserer store datamengder. I tabell 1 har vi identifisert noen problemstillinger der statistisk kompetanse er viktig for å utnytte funksjonell genomikk data maksimalt. Dette kommenteres kort nedenfor.

La oss tenke oss at vi ønsker å undersøke hvilken effekt ulike doser av radioaktiv stråling har på ulike gener i kreftsvulster. F.eks. kan vi ønske å finne grupper av gener som har reagerer likt på strålingen. Det første vi må bestemme er hvordan cDNA mikroarray eksperimentene skal utføres. Det vanligste valget er å bruke et referansedesign, men ved hjelp av statistisk forsøksplanlegging kan vi finne andre design som kan være mer effektive[9]. Etter at eksperimentene er utført brukes bildeanalyse til å bestemme intensiteten til genene fra gråtonebildene. I bildeanalysen inngår segmentering, identifikasjon av spots og lokal bakgrunn, og man beregner intensiteter korrigert for at celleprøvene har reagert utenfor probeområdene på glassplaten. Det er mange kilder tilfeldige og systematiske feil i

[9] Logikken ved å benytte en referanse-celleprøve er å kunne gjøre indirekte sammenligninger, men introduksjon av en referanse er ineffektivt fordi halvparten av dataene samles inn fra denne referansen. Et referansedesign vil være et ufullstendig blokk-design, da referanse-celleprøven merkes likt i alle eksperimentene. Men, det er praktiske biologiske fordeler ved å bruke et referansedesign og forsøksplanlegging krever tett samarbeid mellom statistiker og biolog.

et cDNA mikroarray eksperiment. Benyttes de to mest populære merkestoffene (grønn og rød fluorescent farge) inkorporeres disse ulikt i celleprøvene og kan føre til at man for den ene celleprøven systematisk observerer en høyere intensitet enn for den andre. Korreksjon av denne type feil kalles i DNA mikroarrays for "normalisering". Observasjon av intensiteten til noen gener kan være manglende i noen eksperimenter. Det kan f.eks. skyldes at en eller annen form for forurensning (f.eks. et hårstrå) har kommet på glassplaten. Metoder for imputering av manglende data kan være nyttig.

Den mest populære statistiske metoden brukt på DNA mikroarray data er klusteranalyse. Her ønsker man å finne grupper av gener og/eller grupper av celleprøver som har lignende egenskaper. Hierarkisk klustring har vært mye benyttet i publiserte analyser av DNA mikroarray data, og det har også blitt utviklet nye statistiske metoder for to-veis klustring (gener og celleprøver samtidig). Klustring av celleprøver har i tillegg vært benyttet for å finne mulige undergrupper av sykdommer, f.eks. for noen kreftpasienter vil en behandling virke dårlig og en forklaring kan være at det finnes undergrupper av sykdommen som responderer ulikt på en gitt behandling. Har man celleprøver fra pasienter med forskjellige sykdomsdiagnoser, kan man bruke DNA mikroarray data til å finne gener som har ulik ekspresjon for de ulike sykdomsdiagnosene (diskriminantanalyse). Man kan også bruke disse genene videre til å lage en regel for å diagnostisere nye pasienter (klassifikasjon). En relatert oppgave er å finne et sett av gener som er uttrykt forskjellig i to eller flere celleprøver. Her har man i litteraturen sett på ulike måter å utføre multiple tester på. Data fra slike eksperimenter kan også analyseres med grunnlag i en variansanalysemodell (ANOVA), der hovedeffekten av gen, merking (rød og grønn), glassplate, stråledose og blant annet samspill mellom glassplate og gen (spot) inngår.

Det ligger store utfordringer i å utnytte data fra DNA mikroarray-eksperimenter effektiv. Disse datasettene er såkalte "stor p liten n" datasett, dvs. antall gener er mye større enn antall celleprøver. En liste av artikler som tar opp problemstillingene skissert her, kan finnes hos Rockefeller University [10] og Aas (2001) gir en bred oversikt over metoder som er blitt brukt til analyse av DNA mikroarrays. Det er ikke mange *metodiske* statistiske artikler som er publisert innen analyse av DNA mikroarrays. Grunnleggende spørsmål om kvaliteten til og reproduserbarheten av DNA mikroarray eksperimenter er i liten grad behandlet statistisk.

PROBLEM	STATISTISK METODE
Planlegge design av eksperiment	Statistisk forsøksplanlegging
Beregne genekspresjon fra gråtonebilde	Bildeanalyse
Håndtere systematiske feil (normalisering)	Generell statistisk analyse, variansanalyse, glatting
Håndtere manglende data	Analyse av manglende data
Finne gener eller celleprøver med lignende egenskaper	Klusteranalyse
Diskriminere mellom to eller flere grupper av celleprøver	Diskriminantanalyse
Klassifisere en ny celleprøve til en av mange klasser (eller beregne klassesannsynligheter)	Klassifikasjon
Finne gener som er forskjellig uttrykt i ulike celleprøver	Multippel testing, egenskapsutvelgelse

Tabell 1: Problemstillinger og tilhørende statistiske metoder for å analysere genekspresjonsdata.

HVORDAN KAN STATISTIKERE BIDRA?

Med noen få unntak ligger Norge langt etter genforskningen i utlandet. Det finnes i Norge i dag meget begrenset med forskningsmidler til arbeid innen bioinformatikk og genforskning generelt. Et samlet forsknings-Norge har laget en nasjonal plan for funksjonell genomforskning, forkortet til FUGE. FUGE-planen[11] ble presentert i januar 2001. Forskningsmiljøene ber om minst 300 millioner kroner årlig i de neste fem til ti årene.

[10] http://linkage.rockefeller.edu/wli/microarray/

[11] Les mer om FUGE-planen hos Norges Forskningsråd http://www.forskningsradet.no/fag/andre/fuge/

For å oppnå de beste resultatene innen funksjonell genomikk og bioinformatikk er det nødvendig med et tett samarbeid mellom forskere med statistisk (og bioinformatisk) bakgrunn og forskere innen genetikk, medisin og biologi. Behovet for statistisk tenkning er tilstede i alle faser av funksjonell genomforskning; fra planlegging av eksperimenter til analyser av eksperimentelle resultater. Det er et stort behov for statistiske analyser basert på klassisk metodikk og for utvikling av nye statistiske metoder for å imøtekomme spesielle datatyper og store datamengder.

BIOINFORMATIKK – ET INTERESSANT FORSKNINGSFELT FOR STATISTIKERE?

For å oppsummere; bioinformatikk er et fagfelt som er bygget på viktige biologiske og medisinske problemstillinger. Fagfeltet er nytt og svært aktivt. Hyppige teknologiske nyvinninger introduserer stadig nye utfordringer innen statistisk dataanalyse (datakvalitet og datamengde). Innen fagområdene genetikk, molekylærbiologi, biokjemi og cellebiologi er det et stort behov for statistisk tenkning og statistisk ekspertise. Kort sagt, det er store og mange statistiske utfordringer innen bioinformatikk, og bioinformatikk er uten tvil et meget interessant forskningsfelt for statistikere.

REFERANSER OG KILDER TIL MER INFORMASJON:

FRENGEN OG PRYDZ (2000):"Det humane genom prosjekt og kreft." Tilgjengelig fra Den Norske Kreftforenings WWW-sider på http://www.kreft.no/share/osas/cache/artid200244.html

THE HUMAN GENOME PROJECT (1996): "To know ourselves". Tilgjengelig fra http://www.ornl.gov/hgmis/publicat/tko/index.html/

AAS (2001): "Microarray Data Mining; A Survey", NR notat SAMBA/02/01, tilgjengelig fra http://www.nr.no/samba/bioinformatics.html

OVERSIKT OVER ARTIKLER OM ANALYSE AV DATA FRA MIKROARRAYS VED ROCKEFELLER UNIVERSITY, http://linkage.rockefeller.edu/wli/microarray/

GRUPPE FOR BIOINFORMATIKK VED INSTITUTT FOR INFORMATIKK VED UNIVERSITETET I OSLO, http://www.ifi.uio.no/bioinf/

MIKROMATRISEPROSJEKTET PÅ RADIUMHOSPITALET http://www.med.uio.no/dnr/microarray/

EN SAMLING BIOINFORMATIKK-LENKER FRA FORSKNINGSGRUPPEN FOR BIOINFORMATIKK VED UNIVERSITETET I BERGEN http://www.ii.uib.no/forskningsgrupper/bio/lenker/index-eng.shtml

GENOMIC CLASSIFIER RESEARCH VED NTNU http://www.idi.ntnu.no/grupper/KS-grp/GCR/

TIDSSKRIFT FRA BIOTEKNOLOGINEMNDA, WWW-sider http://www.bion.no

STATISTICAL METHODS IN BIOINFORMATICS, WWW-sider hos Norsk Regnesentral http://www.nr.no/samba/bioinformatics.html