

3.10.6 GENOM OG GENOMFORSKING – KVA ER DETTE?

Genomforskning har gitt store framsteg innan molekylær forskning dei siste åra. Torsk og lakselus er to artar der ein nyttar denne tilnærminga ved Havforskningsinstituttet. Ein er i ferd med å etablere ei rad ulike metodar som vil gi marin forskning eit viktig løft i åra framover. Ein mikromatrise på torsk er snart ferdig, og mange nye genetiske markørar er under uttesting. Dei store datamengdene ein genererer i genomprosjekta vil gi forskinga heilt nye verktøy, som i framtida kan gjere oss i stand til å svare på viktige spørsmål for mellom anna fiskeriforvaltinga.

Frank Nilsen
frank.nilsen@imr.no

Genomforskning er studium av gen og genprodukt der ein ynskjer å studera flest moglege gen – helst alle. Genom er definert som alt DNA i ei celle (kjerne-DNA og mitokondrium-DNA og andre liknande organellar). I fleircella organismar er DNA likt i alle celler, og genomstorleiken vert gitt som den haploide delen (Boks 1). Ein deler gjerne genomforskning inn i strukturell og funksjonell genomforskning. Den strukturelle delen tar for seg samansetjing av arvestoffet (DNA) der DNA-sekvensering (dvs. å lese heile eller delar av DNA-sekvensen) er ein viktig del, medan den funksjonelle delen tar for seg funksjonen til gen og genprodukta (RNA, protein). Målet med artikkelen er å gi ein smakebit på kva eit genom er og litt om kva genomforskning er.

Strukturell genomforskning

Med sekvensering meiner vi å lese av basesamansetjinga til DNA. Metoden som ein i stor grad nyttar i dag vart publisert i 1977 av Fred Sanger, noko som han vart lønna med Nobelpris for. Sjølv om prinsippet i metodikken er den same no som for 30 år sidan, har mange mindre og større endringar før til ein langt lettare bruk av

teknologien, ikkje minst når det gjeld kor raskt og mykje DNA ein kan “lese”. No er heilt ny teknologi i ferd med å lanserast. Denne teknologien byggjer ikkje på “Sanger prinsippet” og vil truleg føre til ein ny revolusjon i genomforskning. Den nye sekvenseringsmåten er særskild rask og prisen for bruk er svært låg rekna i kostnader / base lest (1-10% av prisen idag).

Ei drivkraft i genomforskning har vore målet om å sekvensere heile genomet til mennesket. Ideen til dette særskilde omfattande prosjektet vart fødd tidleg på 1990-talet, og ein viktig del for å kunne lukkast var å utvikla teknologien som skulle til. I 2001 vart sekvens og analysar av heile mennesket sitt genom publisert. Seinare er det kome fleire “oppdateringar” av sekvensen, og ein har no ein sekvens som er meir enn 99,999% nøyaktig.

Ein komplett genomsekvens er det beste utgangspunktet for funksjonelle studiar, då ein vil skaffe sekvensinformasjon om “alle” gen. Men dette er kostbart, og når ein i tillegg veit at meir enn 50% av genomet vanlegvis er intergenisk sekvens (sjå boks 1), lyt ein sekvensere mykje for kvart gen ein finn. For mindre prosjekt, dvs. der ein ikkje har nok ressursar til å foreta ein komplett genomsekvensering, har ein

Kva er eit genom?

Genomet er arvematerialet til ein organisme. Dette er vanlegvis DNA, men kan i nokre virustypar vera RNA. I fleircella organismar er det DNA i ei celle som utgjer genomet, sidan alle cellene har lik DNA. Unntaket er kjønsceller som er såkalla haploide, dvs. at kvart kromosom berre finst i ei utgåve. Eukaryote organismar er kjenneteikna ved at dei har DNA i cellekjernen, men òg i organellar som mitokondrium og kloroplastar, t.d. t.d. planter og algar. (“Organellar” er cellene sine “energiverk” – ordet tyder eigentleg “småorgan”.) DNA i cellekjernen utgjer størstedelen både med omsyn til mengd og tal på gen.

Storleiken (mengd DNA per haploid celle) og samansetjinga til genomet varierer mykje mellom ulike organismar. Ei typisk bakterie som *E. coli* har 4.7 millionar basar (Mb), bananfluge 180 Mb, torsk 1 000 Mb og menneske 3 000 Mb. Talet på gen er frå 3 200 i *E. coli* til om lag 30 000 i mennesket, noko som gir ein faktor på om lag ti for talet på gen. Ser ein på mengd DNA er forskjellane mykje større med ein faktor på nær 1 000 mellom menneske og *E. coli*. Dette tyder at *E. coli* i snitt har 1 gen/1 400 basepar (bp), medan mennesket har 1 gen/100 000 bp. Sjølv om gena typisk er større hjå menneske enn *E. coli*, er det langt meir DNA mellom gena (intergenisk DNA) i eit menneske (sjå Figur 3.10.6.1a). Ein har stor skilnad i genomstorleiken hjå eukaryotar, langt større

enn mellom prokaryotar (bakteriar), der dei minste har ein storleik som bakteriane og dei største mange gonger større enn mennesket.

Genomet til ein “typisk” fleircella (metazoa) organisme er illustrert i Figur 3.10.6.1a. Dei svarte boksane representerer gen, medan området mellom (raude) er intergenisk DNA. Intergenisk DNA har vore kalla “junk”-DNA, men dette er ei særskild misvisande nemning. Repeterte element utgjer ein stor del av genomet til fleircella organismar (omlag 45% i pattedyr). Ein har mange typar repeterte element, frå dei heilt enkle (di, tri, tetra nukleotidar) til store samansette repeterte einingar. Dei enkle, repeterte elementa kallast ofte for mikrosatellittar. Lengdene på desse er svært variable, og dei er difor mykje nytta som genetiske markørar. Dei andre repeterte elementa er delt inn i mange klassar, og to av dei heiter transposomer og retrotransposomer. Dei har heilt særskilde eigenskapar og kan flytte seg rundt om i genomet. Transposoma gjer dette ved hjelp av ein “klipp & lim”-metode, medan retrotransposoma nyttar “kopier & lim”. Retrotransposoma vert kopierte på vanleg måte via mRNA, men vert sette inn att som DNA og nyttar då eit særskilt enzym (reverstranskriptase) for å lage DNA frå RNA. Dette enzymet vert ofte nytta i molekylærbiologien når ein arbeider med RNA, og det vert bruk til å kopiere RNA til DNA, noko som gir opphav til komplementær (complementary) DNA (cDNA).

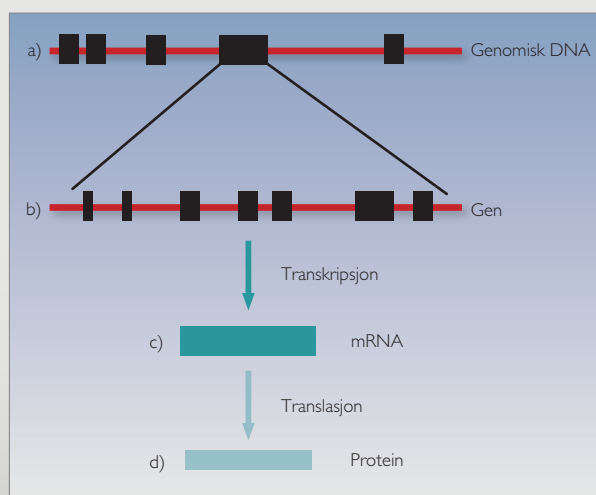
Eit typisk eukaryot gen er framstilt i Figur 3.10.6.1b, der dei svarte boksane er exon (den kodande delen) og mellomrommet mellom desse er intron (ikkje kodande). Heile eininga vert transkribert, men introna vert fjerna i ein prosess kalla spleising. mRNA består av dei ulike exonane. Den ferdige mRNA-en vert sendt ut av cellekjernen til cytoplasma der translasjon (dvs. informasjonsoverføring frå RNA til protein) føregår. Summen av alle mRNA i ein celletype vert ofte kalla transkriptomet, medan alle proteina vert kalla proteomet.

Ei av overraskingane då mennesket sitt genom var sekvensert var at det var langt færre gen enn mange hadde rekna med. Mennesket har berre om lag 30% flere gen enn rundmark! Når ein ser på kompleksiteten til organismane er dette verkeleg overraskande, og korleis kan ein forklare det? Eit viktig svar ligg i spleiseprosessen knytt til det ein kallar alternativ spleising. Genet i Figur 3.10.6.1b består av sju exon og seks intron og kan via alternativ spleising gje opphav til mange ulike transkript (og protein) ved at eit eller flere av exonane kan verta utelate. Dette betyr at til fleire exon eit gen har, til fleire ulike protein kan dette gje opphav til. I dette ligg det viktige skilnader mellom “enkle” dyr og t.d. pattedyr som typisk har komplekse gen (dvs. mange exon) med aktiv bruk av alternativ spleising som kan gje opphav til mange forskjellige protein.

Figur 3.10.6.1

Skjematisk framstilling av oppbygging av arvestoffet (genomet) og genomaktivitet. A) viser genomisk DNA der dei svarte boksane indikerar gen og raude områder syner intergenisk DNA. Hos dei fleste eukaryotar vil intergenisk DNA utgjera storparten av DNA-et. B) Skjematisk framstilling av eit typisk gen med exon (svarte boksar) og intron (raud). C) mRNA er budbringaren som gir opphav til protein. Transkripsjon av DNA gir opphav til mRNA. D) mRNA vert transportert til cytoplasma, og mRNA vert translert til protein i cytoplasma.

Overview of the genome and the genomic activity. A) shows genomic DNA where black indicates genes and red indicates intergenic regions. In most eukaryotic organisms the intergenic part is by far the largest. B) Drawing of a gene with exons (black) and introns (red). C) mRNA is the messenger giving rise to proteins. Transcription of DNA produces mRNA. D) mRNA is transported from the nucleus to the cytoplasm for translation into protein.



alternative strategiar for å skaffe sekvensdata frå mange gen. Ein metode er EST-sekvensering der ein startar med mRNA og lagar genbibliotek med utgangspunkt i aktive gen (ofte kalla cDNA-bibliotek). Ved å lage mange ulike cDNA-bibliotek frå ulike vev/organ og utviklingsstadium kan ein dekkje store delar av gena i genomet dersom ein produserer eit høgt tal med DNA-sekvensar.

Funksjonell genomforskning

Sjølv om sekvensering av eit genom er ei stor og krevjande oppgåve, er det ei langt større utfordring å finne funksjonen til alle gena og andre element i arvestoffet. Ein viktig del av den funksjonelle genomforskninga er å studere genaktivitet. Slik aktivitet er delt inn i tid og rom, og ein kan undersøke dette på fleire nivå (t.d. transkripsjonsnivå (mRNA) og proteomnivå (protein)). Nokre gen er aktive "alltid" og i "alle" celler, medan andre gen berre er aktive i nokre celletypar under særskilde tilhøve. Ein har metodar der ein kan ta føre seg aktiviteten til mange gen eller protein samstundes. Mikromatriser er ein slik metode der ein kan studere transkripsjonsaktiviteten til alle (mange) gen samstundes, men det føreset sjølvstakt at ein kjenner sekvensen til desse, sjå Boks 2. På same måte kan ein samanlikne proteina frå ulike prøvar for å finne kva for protein som er til stades ved ulike tilhøve.

Felles for desse og andre metodar i genomforskning, er at ein må ha tilgang på databasar med sekvensinformasjon og gode og effektive dataløysingar for å analysere sekvensinformasjonen. Bioinformatikk har difor ein sentral posisjon på alle nivå i genomforskning. Genbank (<http://www.ncbi.nlm.nih.gov/>) er døme på ei slik ressurs der store mengder sekvensdata er lagra og tilgjengeleg for søk. Datamengda i Genbank har auka eksponentielt dei siste

åra, og i 2006 var der over 160 milliardar basar med sekvens tilgjengeleg for søk. Utan effektiv og rask programvare vil det vera vanskeleg å finne relevante data i slike enorme databasar. Søk i Genbank med ein DNA- eller proteinsekvens vil raskt gi dei første indikasjonane på funksjon, og ved høg grad av sekvenslikskap kan ein vera ganske sikker på funksjonen til sekvensen ein har nytta i søket. Samanlikning av sekvensar mellom ulike artar kan gi viktig informasjon om funksjon. Ein ser at det ikkje berre er genprodukt som er konserverte mellom artar, men òg at organisering av gen kan vera konserverte mellom ulike artar.

Sekvenseringsprosjekt genererer store datamengder frå genomisk DNA og/eller cDNA (EST-ar). Sidan slike EST-ar er henta frå aktive gen, kan desse nyttast direkte som kandidatgen i funksjonelle studiar. Sekvensdata er òg startpunkt for å finne genetiske markørar som mikrosatellittar og SNP-ar. SNP (singel nucleotid polymorphisms) er enkeltbasar som er variable og som kan nyttast som genetiske markørar.

Genomforskning – kva så?

Ein viktig skilnad mellom tradisjonell molekylærbiologi og moderne genomforskning er at ein har gått frå ei gen-til-gen-tilnærming og over til ein strategi med alle/mange gen samstundes. Store framsteg i sekvenseringsteknologi i form av automatisering har gjort dette mogeleg. Ein har i dag store sekvenseringslaboratorium som kan produsere titusenvise av DNA-sekvensar per dag, noko som òg har ført til at kostnadane har gått kraftig ned. Sjølv små laboratorium kan i dag generere meir sekvensdata på ein dag enn store laboratorium kunne på eit år med "gamal" teknologi. Sidan eit stort tal med sekvensar vert generert, veljer ein ut eit stort tal med klonar for sekvensering, og etter at sekvenseringa er gjort ser ein kva ein har. Ved ei slik tilnærming vil noko av det ein finn vera venta og kjent, medan ein god del vil vera nytt og ukjent (avhengig av art).

Genomprosjekt vil generere store mengder med sekvensinformasjon, sjølv om ein berre sekvenserer EST-ar. Dette betyr at brukarar av dette vil ha tilgang til mange kandidatgen eller mange genetiske mar-

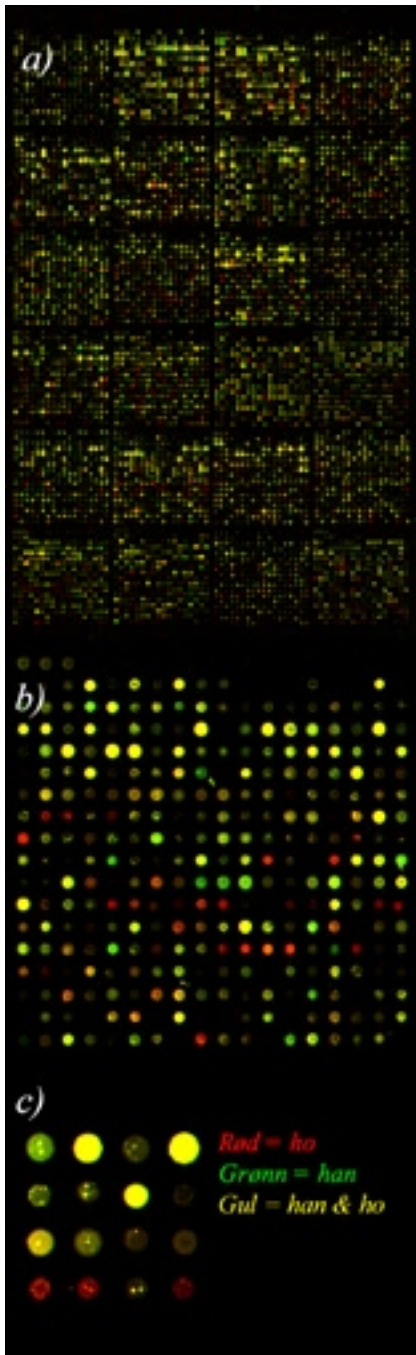
Mikromatriser eit sentralt verktøy i funksjonelle studiar

Mikromatriser er basert på DNA-hybridisering, der enkeltråda DNA som har lik sekvens kan gå saman og danna dobbeltråda DNA (slik ein typisk finn DNA i cellene). Enkeltråda DNA kan lett lagast ved at ein varmar opp DNA slik at hydrogenbindingane mellom basane vert brotne. Når ein så avkjølar dette, vil sekvensar som er like gå saman og danne dobbeltråda DNA. Mikromatrisestudier krev at ein har sekvens av heile eller deler av genomet til ein organisme. Basert på denne sekvensinformasjonen kan ein lage DNA-bitar (prober) som representerer eit gen eller delar av eit gen. Desse probene kan så enten prentast ut eller syntetiserast på små glasplater (storleik som objektglas), og

det er no mogeleg å ha langt over 1 million ulike prober på ei mikromatrise. Mikromatriser kan brukast på fleire ulike måtar, men den mest velkjente bruken er til å studere transkripsjon av gen. I slike studiar ynskjer ein å identifisere gen som er aktive i ulike situasjonar (t.d. på ulike tidspunkt i utviklinga, under ulike miljøtilhøve). Utgangspunktet vil vera RNA (mRNA) som ein nyttar til å lage cDNA. cDNA vert merka med raud og grøn fluoriserande farge, der to og to prøvar vert blanda og nytta på ei mikromatrise. Då vil aktive gen (som finnst på mikromatrisa) kome opp som raude, grønne eller gule (dei som er aktive i begge prøvane) signal (Figur 3.10.6.2).

Figur 3.10.6.2

Mikromatrise frå lakselus. A) viser mikromatrise frå lakselus som har over 7 000 prober. Desse 7 000 probene representerer 2 600 ulike gen. B) Forstørring av eit subarray (deltest-område) som viser ulike signal etter hybridisering. I dette dømet er cDNA frå hannlus (grøn) og holus (raud) samanlikna. Ein ser at ein har mange fargenyansar frå grøn til raud (B og C). Ein får informasjon om kva for gen som er aktive i hann og ho og tilhøvet mellom desse (dvs. kor mykje raud og grøn farge det er i kvar spott).
Microarray from the salmon louse. A) Shows a microarray from the salmon louse containing more than 7 000 probes. These probes represent 2 600 different genes. B) Enlargement of a subarray showing different signals after hybridisation. cDNA from male lice (green) and female lice (red) are compared. In this experiment it is possible to identify genes active in males and females, in addition to those active in both sexes (yellow).



kørar dersom ein er ute etter dette. I tillegg kan ein nytte verktøy som t.d. mikromatriser som gjer at ein kan studere aktiviteten til tusenvis av gen samstundes. Kva er så fordelane med dette? Ved at ein har tilgang til mykje meir data kan nye biologiske spørsmål stillast, og ein vil vera i stand til å svare på gamle spørsmål med betre datagrunnlag. Det er ikkje dermed sagt at biologisk forskning vert betre ved å nytta “genom-verktøykassa” men det kan vera mykje tid å spare ved rett bruk av ressursane. Ein stor EST-database frå ein art vil kunne nyttast til mange ulike føremål. Den vil mellom anna gi sekvensinformasjon om enkeltgen, den vil gje opphav til mange genetiske markørar (mikrosatellittar og SNP-ar), ein har eit utgangspunkt for å lage mikromatriser og avgjerande for proteinidentifisering frå proteomstudiar.

Marin funksjonell genomforskning

Funksjonell genomforskning er eit satsingsområde i Forskningsrådet via FUGE-programmet. Ei utfordring til den marine delen av denne satsinga er at ein har lite strukturelle data (dvs. sekvensdata) frå relevante artar for Norge. Per i dag er det stort sett berre EST-data som er tilgjengeleg frå eit fåtal artar. Laks er den arten ein har mest data tilgjengeleg for, men storparten av dette er data generert i utlandet. For mange viktige marine dyregrupper (t.d. kopepodar) finnast det svært lite data.

Lakselus

Ved Havforskningsinstituttet har ein etablert ressursar for funksjonell genomforskning på lakselus i løpet av dei siste åra. Ein har produsert EST-data og nytta dette til å etablere ei mikromatrise. Desse ressursane vert nytta i samband med prosjekt som arbeider med ei vaksine mot lakselus. Strategien baserar seg på å nytte funksjonell genomforskning for å finne kandidatgen som kan evaluerast som potensielle vaksine-antigen. I dette arbeidet er det mange utfordringar, særleg når ein ser at om lag 45 % av alle lakselusgena er nye (dvs. ein får ikkje signifikante treff ved søk i databasar).

Mikromatriseteknologien er i denne samanheng særst nyttig, då ein her kan knytte ukjente gen til biologiske prosessar

og slik “sile vekk” irrelevante gen. Sjølv om ein på denne måten kan knytte gen til biologiske prosessar vil ein ikkje ha funksjonen klarlagt.

Ein av dei mest effektive måtane å finne funksjonen til eit gen (dvs. funksjonen til proteinet det kodar for), er ved å “skru av” genet. Det kan ein gjere på fleire vis, og for modellorganismar (t.d. bananfluge) har ein standardiserte protokollar for dette. På lakselus har me fått det til ved å bruke RNA-interferens (RNAi). I virvellause dyr (evertebratar) kan ein “skru av” eit gen ved å injisere dobbeltrådig RNA (dsRNA) frå genet ein er interessert å stoppe. Det som skjer er at mRNA (med same sekvensen som dsRNA) vert brote ned og ein får ikkje laga protein. Ved å sjå kva for konsekvensar dette har, vil ein få data på funksjonen til proteinet. Metoden vil òg gi informasjon om proteinet er ein god vaksine kandidat eller ikkje. Dersom resultatet av å fjerne eit gitt protein er fatalt for lakselusa, vil dette vera ein god vaksine kandidat dersom ein kan få same effekten via vaksineringsring. RNAi vil difor vera eit særst nyttig hjelpemiddel til å velje ut kandidatgen til testvaksineringsring.

Fiskegenomforskning

Ein kjenner alle gena hos fleire fiskearter som fugu, tetraodon og sebrafisk, men enno ingen av dei kommersielt viktige artane. Havforskningsinstituttet har hatt ei sentral rolle i å byggje opp ressursar for funksjonell genomforskning på torsk. Arbeidet har fokusert på EST-sekvensering og produksjon av ei torskemikromatrise. I tillegg har me etablert eit BAC-bibliotek og ein protokoll for å nytte dette. Instituttet har ei viktig rolle i den nyleg etablerte Genofisk-plattformen i det nasjonale FUGE-programmet. Dette er ei ressurs- og teknologiplattform for funksjonell genomforskning på laks og torsk. Eit viktig mål med arbeidet er ressursetablering frå desse viktige kommersielle fiskeartene og å betre kompetansen på dette fagområdet i fiskeriforskningsmiljøa. Ressursane vil vera eit grunnlag for å auke den biologiske kunnskapen om desse artane og gjere ein betre rusta til å gi gode råd til styresmaktene i marine spørsmål.

Cod genomics

Recent development in genomics has been an important boost for development in molecular biology. Through sequencing of the human genome and several model species a new set of tools have been established, which can be used in other species. Comparative genomics has shown that the number of genes in different species are rather similar (less than tenfold variation

between bacteria and mammals) whereas the amount of intergenic DNA is much more variable. During the last few years IMR has started genome-related activity on cod and salmon louse. A main goal in these projects has been to establish functional genomics tools for these species. This includes EST-sequencing and establishment of microarrays.