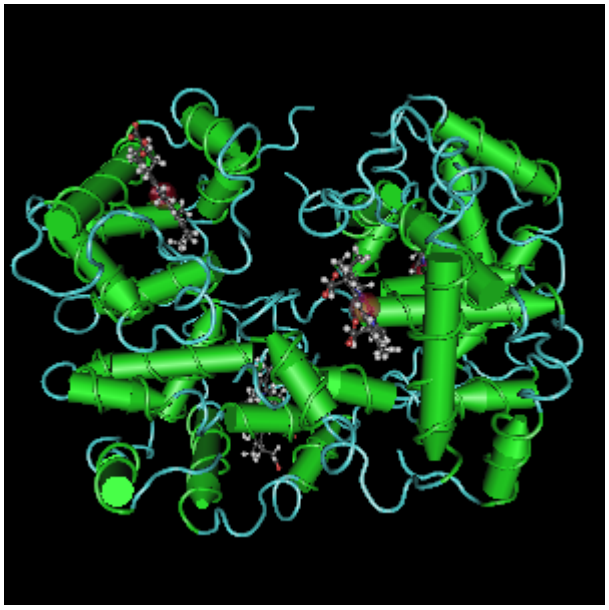
# Her går det rundt - cellebiologiens mysterier og ENTREZ sin underlige verden

**Av Regina Küfner Lein, Universitetsbibliotekar ved Bibliotek for medisinske fag, Universitetsbiblioteket i Bergen**

*Endelig skulle jeg lære meg alt om de molekylærbiologiske databasene som ligger i menylinjen til høyre for PubMed. Jeg deltok på "Bridging the Gap between PubMed and the Entrez Life Science Databases" som ble tilbudt som dagskurs før EAHIL-konferansen i Helsinki. David Herron fra Karolinska Instituttet i Stockholm holdt et informativt og morsomt kurs og viste oss det mest elementære innen molekylær cellebiologi før vi så nærmere på noen av databasene.*
*Det er nok ikke bygget noen bro mellom PubMed og Entrez for min del enda, men jeg hangler i det minste langsmed et tau og forstår deler av materien - tror jeg.*

Kanskje jeg skal starte neste undervisning for studentene med å vise en 3-dimensjonal modell av hemoglobin-molekylet som spinner rundt på PCen? Så enkelt kan det gjøres:



Jeg åpner PubMed. I den svarte menylinjen til høyre for PubMed ligger databasen **Structure**, og jeg søker på "human adult hemoglobin". I trefflisten heter det "2H35" men fullt navn står til høyre. Jeg klikker på spiralen ved siden av bildet, og vips kan jeg snurre rundt og rundt ved å bruke musetasten. Ved hjelp av et lite program, Cn3D, som IT-avdelingen meget raskt lastet inn på alle PCene på biblioteket kan alle brukerne gjøre det samme!

Hva er det vi ser? De viktigste strukturer i menneskekroppen består av proteiner, for eksempel hemoglobin, eller muskelfibre, enzymer, antistoffer og hormoner. Disse proteiner er som regel store og er kveilet opp til en 3-dimensjonal struktur som et nøste. Det er denne strukturen vi ser. For å laste ned programmet Cn3D 4.1 og mer info gå til http://www.ncbi.nlm.nih.gov/sites/entrez?db=structure
Og så kan du la de spinne automatisk (velg View, animation, spin etter å ha klikket på spiralen) eller merke av hver 5. aminosyre med navn og mye mer morsomt; det leste jeg nettopp om i Cn3D-tutorial som er lenket fra den blå margen.

Hittil har jeg vært best kjent med litteraturbasene med PubMed på topp. Jeg vet litt om OMIM (arvelige sykdommer), og jeg har brukt Books og PubMedCentral. Molekylærbasene derimot har vært ukjente for meg. De gir ikke først og fremst litteraturreferanser, men viser oppbyggingen av molekyler, og gjerne bilder av strukturene. Det er henvisninger fra den ene til den andre basen.

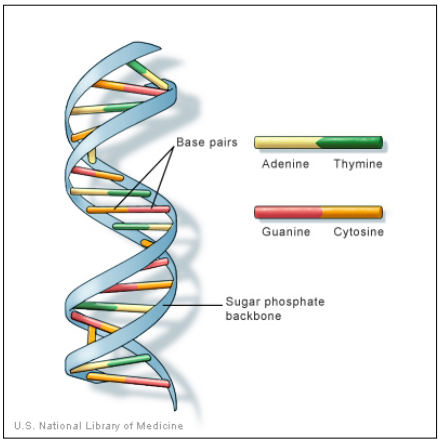
Databasen *Structure* er lenket til PubMed på bakgrunn av en publikasjon om denne strukturen av hemoglobin-molekylet. Bare klikk videre fra "2H35" og vi er snart på kjent grunn i PubMed. Sjekk også ut "Links" i PubMed som fører oss tilbake til Stucture!

Nå har jeg allerede lært så mye nytt at kurset godt kunne stoppet her. Ta gjerne en pause i lesingen av artikkelen i alle fall. På kurset lærte vi mer:

Entrez består for tiden av 67 databaser, for eksempel Nucleotide, Protein, Gene, Structure, Map Viewer og BLASTp. Og alt henger sammen på et vis. Se kartet over alle Entrez-baser i About Entrez og beveg museknappen over de enkelte baser for å se nettverket:
http://www.ncbi.nlm.nih.gov/Database/datamodel/index.html

Det menneskelige genom (=det samlete arvestoffet) var ferdig kartlagt i 2003, takket være flere store forskningsmiljøer. Også åpningsforedraget til konferansen dreide seg om genom-forskningen. Professor i humangenetikk og forskningsdirektør Leena Peltonen-Palotie påpekte at en så stor informasjonsmengde krever gjennomtenkt struktur for lagring, analyse og presentasjon av dataene. Det er vesentlig med umiddelbar tilgang til dataene for å kunne utnytte disse til lettere å forstå sykdom, men også for bedre diagnostikk og for å forstå individuell risiko og tilby adekvat behandling. Og mange av disse dataene har vi altså gratis tilgjengelig via NCBI (National Center for Biotechnology Information) og Entrez. Her ligger data fra de ulike laboratoriene (ofte lenket under "archived") og en gjeldende utgave ("curated") der det er ett treff per unike enhet.

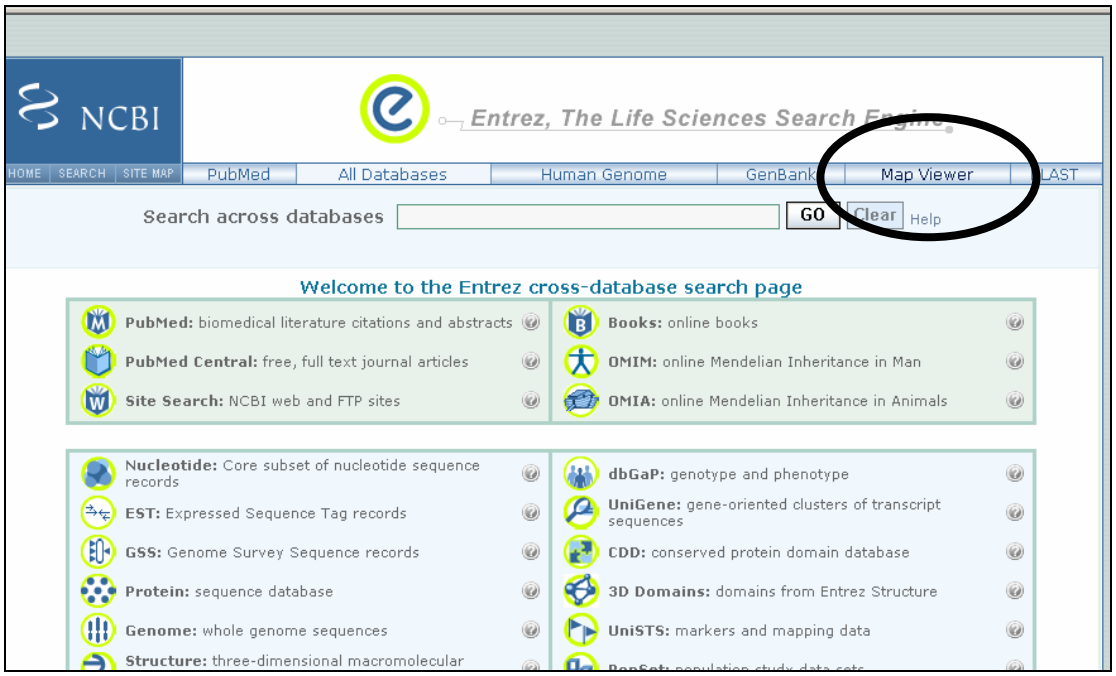
For bedre å forstå sammenhengen mellom gener og proteiner er det lurt å friske opp kunnskapene i cellebiologi. Kursdeltakerne ble oppfordret til å lese de første to kapitlene i læreboken Genetics Home Reference Handbook, Help Me Understand Genetics (http://ghr.nlm.nih.gov/handbook). Det kan anbefales.




Arvestoffet vårt er hovedsakelig lokalisert i cellekjernen, i form av 23 kromosomer. Hvert kromosom består av en lang rekke gener. Gener er avgrensete deler av et kromosom og laget av DNA. Vi husker at DNA er en dobbeltspiral. "Trinnene" i spiralen består av basepar av 4 nukleotider som forkortes med første bokstaven a,t,c,g. Dobbeltspiralen ble oppdaget av Watson, Crick, Wilkins og Rosalind Franklin i 1953. Tenk, på bare 50 år har hele genomet blitt kartlagt!

Vi skal se nærmere på noen databaser i Entrez:

Fra Entrez sin startside (klikk på "All databases" i den svarte menylinjen fra PubMed) kan vi søke på tvers av alle baser, både litteraturbasene oppe (som PubMed, OMIM) og "oppslagsverkene" for gener og proteiner som Nucleotide, Genome, Protein, med mer. Jeg vil starte med det største av de små og dukker videre inn i de mindre enhetene: Fra Genome til Gene og Nucleotide.




**Databasen Genome:** Her finner vi rekkefølgen av gener på alle menneskets 23 kromosomer, men også genmaterialet som er kartlagt for andre organismer. Det enkleste er å søke genom-basen via "Map viewer" fra Entrez-siden (oppe i menylinjen).

I Map Viewer listes organismene gruppevis med sitt vitenskapelige navn. Heldigvis står den engelske oversettelsen for de vanligste av disse, Homo sapiens står øverst. Ellers finner vi både mus og rotte, bananfluen, løk og tomat og mange flere. Vi velger Homo sapiens og får så vist alle 23 kromosomene.

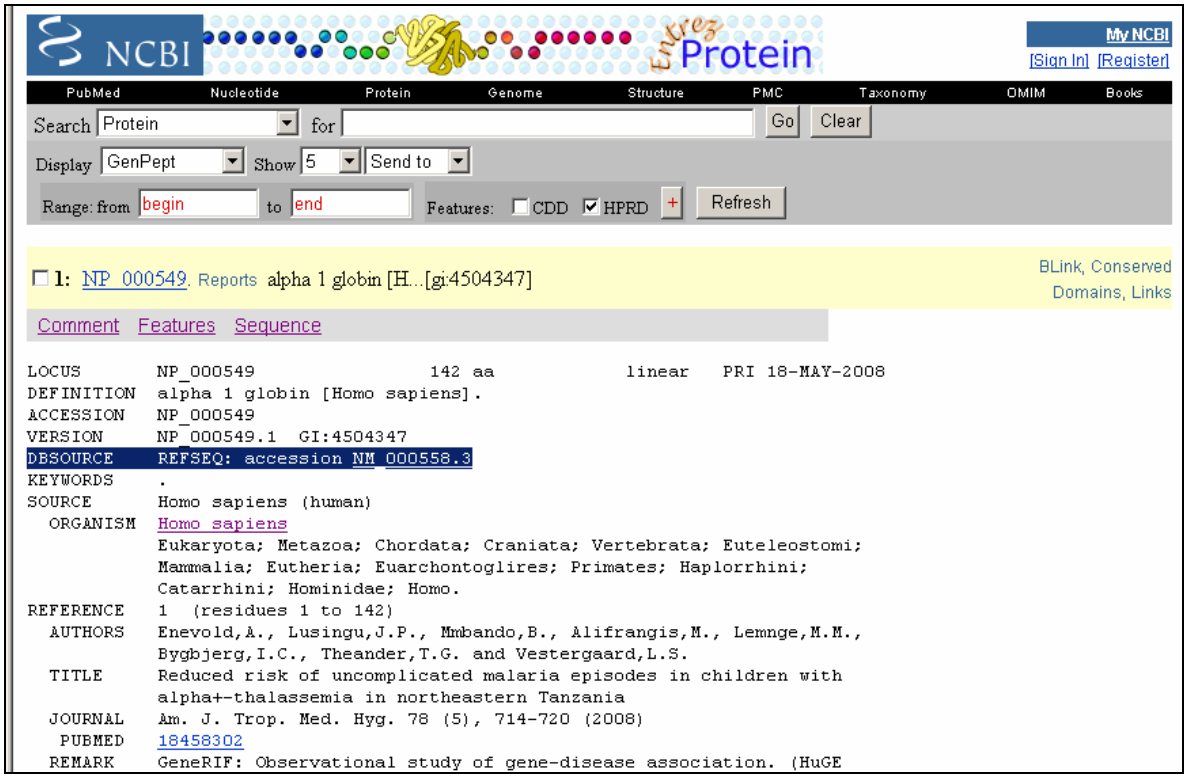
Her kan vi klikke på et enkelt kromosom og se alle gener på det eller søke på et protein og finne ut på hvilket chromosom de ansvarlige genene ligger. Søk for eksempel på hemoglobin. Vi får se at genene for hemoglobin ligger på flere ulike chromosomer. Lengre ned på siden er det det tre seksjoner som likner, den første fra samling "Reference", neste fra "Celera" og siste som heter "Huref". Det er flere miljø som har jobbet med kartleggingen, blant annet et firma som heter Celera. Alle data er nå i NCBI; den siste seksjonen Huref, er en oppsummering av alle data, og best å forholde seg til. For å forstå detaljene videre bør man være ekspert innen cellebiologi.

**Databasen Gene**: I Entrez Gene kan vi søke på gener som hører til de ferdig kartlagte genomer. Den inneholder "nomenclature, map location, gene products and their attributes, markers, phenotypes, and links to citations, sequences, variation details, maps, expression, homologs, protein domains and external databases" (Maglott et al. 2005 http://www.pubmedcentral.nih.gov/articlerender.fcgi?tool=pubmed&pubmedid=15608257). Det er heldigvis en grei punktvis liste på åpningssiden som viser ulike måter å søke på, f.eks. fritekst "human muscular dystrophy". Når vi så ser resultatene, forstår jeg for min del ikke så mye mer, men jeg vil gjerne vise dette til lånerne som jobber innen gen-feltet - jeg er overbevist om at de får noe nyttig ut av dette.

**Databasen Nucleotide**: Denne databasen viser sekvensen av nukleotider for ulike gener. Vi kan søke på sekvensnummer (accession number) - på kurset søkte vi på U46667. Den hører til bakterien Escherichia coli og enzymet citrate lyase. Nedert på siden står bokstavkodene til nukleotidene, de første 10 er: aaagcctcgg - og alt det andre håper jeg fagfolk forstår. Jeg regner også med at de kjenner sekvensnummeret til det de er interessert i, eller enzymnavnet, eller lignende.

Og hvor kommer så proteinene inn her? Hvordan et protein er bygget opp er kodet i de tilsvarende genene. Om kroppen skal lage mer hemoglobin, så skal informasjonen fra DNA i cellekjernen formidles til ribosomene der selve proteinsyntesen foregår. Ribosomene ligger utenfor cellekjernen, mens DNA er så stor at den ikke kan forlate cellekjernen. Derfor lages det en kopi av det aktuelle område på dobbeltspiralen som sendes ut av cellekjernen (mRNA) - akkurat som våre trykte tidsskrifter ikke skal ut av biblioteket, men en enkelt artikkel kan kopieres og taes med. Proteinet er bygget opp av aminosyrer, der rekkefølgen av aminosyrene avgjør hvilket protein det dreier seg om. Det finnes 20 ulike aminosyrer som forkortes med første bokstaven. I **databasen Protein** kan vi søke opp proteiner og se deres oppbygging, for eksempel *human adult hemoglobin*. Jeg får mange treff, men kan velge *Homo sapiens* fra høyre vindu og få treffmengden noe redusert. For hvert treff får jeg nederst på siden sekvensen av aminosyrer, dvs. rekkefølgen av aminosyrene, der hver bokstav står for én aminosyre:

```
ORIGIN
        1 mlsaqeraqi aqvwdliagh eaqfgaelll rlftvypstk vyfphlsacq datqllshgq
       61 rmlaavgaav qhvdnlraal spladlhalv lrvdpanfpl liqcfhvvla shlqdeftvq
      121 mqaawdkflt gvavvlteky r
```




Proteinene har gjerne betegnelsen NP…. og noen tall, feks. NP_000549. Kromosomer heter NC.. og noen tall, og mRNA kalles for NM.. og noen tall. Alle N-numre har med NCBI å gjøre og representerer de gjeldende ”curated” versjoner av proteinet (P), kromosomet (C) og m-RNA (M), og de er lenket til hverandre slik det vises i eksemplet i skjermbildet. Vi ser også i skjermbildet at det er lenke til litteraturreferanser i PubMed.

Fra ”Links” kan jeg velge ”Structure” eller ”Related structure” og ser igjen den tredimensjonale strukturen til dette proteinet. Selvfølgelig skal jeg la det snurre rundt og glede meg over farger og bevegelser.

Som en good repetisjon anbefalte David Herron artikkelen "Entrez: making use of its power" av Geer, R. & E. Sayers (Briefings in bioinformatics, 2003, 4(2):179-184). Dessuten finnes på NCBI sin hjemmeside greie, små artikler, samlet under "Coffee Break" i høyre margen, og en online håndbook om de ulike databasene.

Jeg synes fortsatt at dette er fryktelig vanskelig, og liker ikke at jeg ikke forstår alt. Forkortelser og rare termer, og lenker i alle retninger! Kurslederen David Herron overrasket oss med å si at han heller ikke visste alt, men at det han vet allikevel holder til å markedsføre disse flotte basene i de rette miljøene. Så da skal også jeg våge å bruke dem. Og for ikke å glemme hvordan det hele henger sammen, har jeg som forsett å lese denne artikkelen minst én gang i måneden.