Bioinformatik

Föreläsare:
Mats Nilsson
mats.nilsson.6o88@med.lu.se

= Sätt att samla info och göra den praktiskt användbar för ändamålet.

Definitioner

Göra en föreställning om molekylers fysiska kemi och tillämpa informationsteknik för att förstå och organisera informationen associerad med dessa molekyler. På stor skala.

Enligt Oxford ordbok.

Research, utveckling, tillämpning av datoriserade verktyg och inställningar för expansion av biologisk-, medicinsk-, beteendemässig- eller hälsomässig data. Denna data insamlas, lagras, organiseras, arkiveras, analyseras och visualiseras.

Enliglit Committee National Institute of Mental Health

Tvärvetenskaplig disciplin där algoritmer för analys av biologiska (ffa. Molekylärbiologiska) data utvecklas. Man använder också termen synonymt med engelskans computational biology.

Enligt wikipedia

(Man vill öka förståelsen för biologiska processer. Tar fram algoritmer osv. Metoder för att lokalisera gener inom gensekvenser. Skapa stora databaser med denna typ av information.)

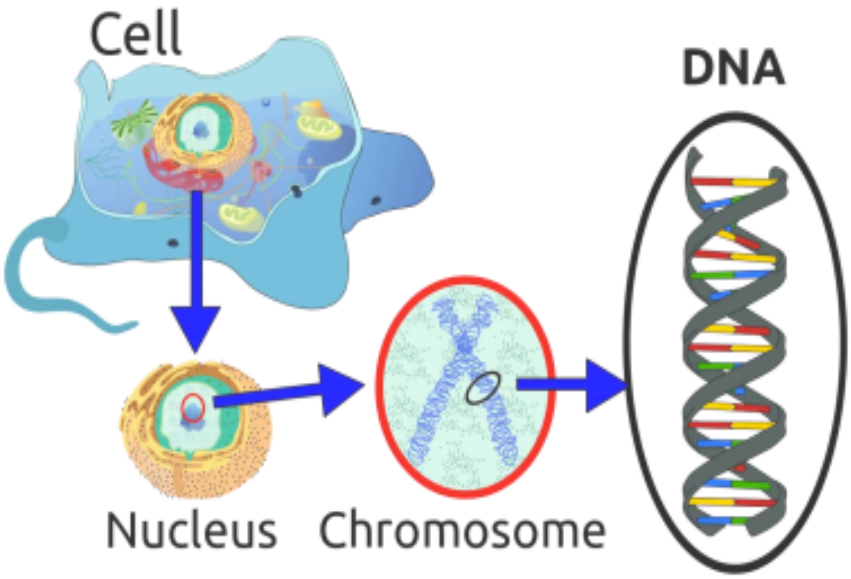
Primära målet är att öka förståelsen för biologiska processer.

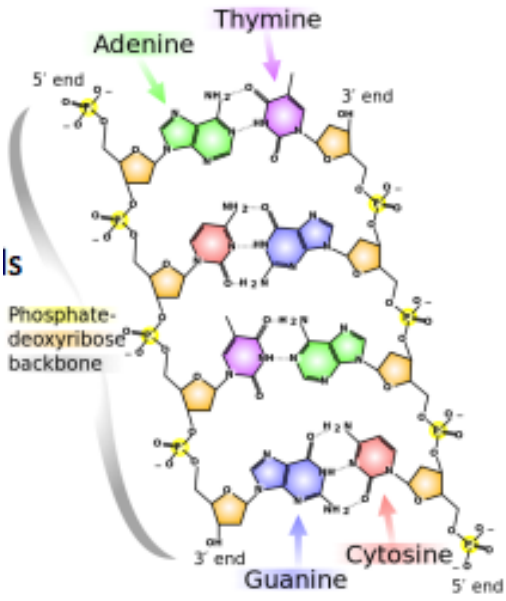
- Utveckling av nya algoritmer, statistiska mätmetoder och datorprogram för utvecklingen av stora databaser.
- Implementering av utvecklade algoritmer, program i datautvärdering och utvärdering av resultaten.
- Skapa och förbättra publika tillgängliga databaser.

Genetisk kod – DNA

C-G, T-A.

Strängar blir kopior av varandra. Om någonting blir fel kan enzym gå in och rätta till felet.





Om DNA

Självreplikerande material som finns i nästan alla levande organismer som huvudsaklig konstituent av kromosomer.
Bärare av genetisk information.
Består av två biopolymersträngar tvinnade kring varandra i en dubbelhelix. Baser paras A-T och C-G. Vätebindningar sammanlänkar kvävebaserna från de två polynukleotiderna → dubbelsträngat DNA.

Det Centrala dogmat

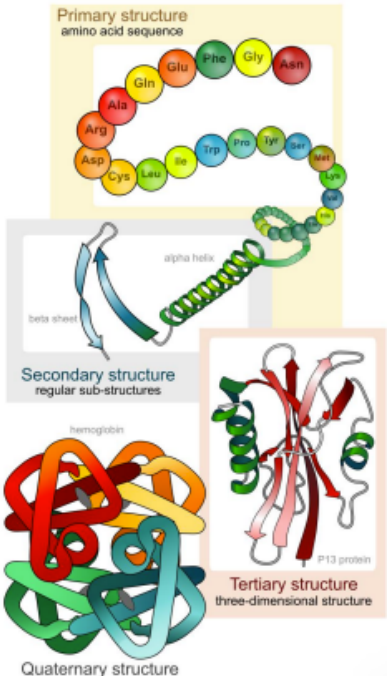
Består av codon. Vilka består av tre baser.
Innehåller G,A,U och C.
Varje codon motsvarar i sin tur en aminosyra vilken den ”omvandlas till” i translationen.
Finns även start- och stoppsekvenser.

Aminosyror som tillsammans blir protein.
Finns tre codon som endast kodar för start och stoppsignaler för de genetiska meddelandena.

Tittar på enskilda aminosyrors sidokedjor. Adiabatiska – Kväve och syre.
Finns syror, baser och polära. Kodas för att veta vad de kan ersättas med. En laddad ersätts med en annan motsvarande laddad.

Nivåer på proteinstrukturer

Nivå	Beskrivning	Stabiliseras av
primär	aminosyrasekvenser	peptidbindningar
sekundär	Alfa-helixar och beta-sheets i polypeptid	Vätebindningar mellan grupper längs peptidsträngen
tetriär	Tredimensionella formen av en polypeptid	Interaktion mellan olika R-grupper och peptidlängden.
kvartenär	Formationen av kombinerade polypeptider	Interaktion mellan R-grupper och peptidlängden på olika polypeptider.



Information sorteras i grupper baserat på dess likheter.

Homolog
Indikerar gener eller proteiner som är evolutionärt relaterade. Innebär gener med gemensamt evolutionärt ursprung. Antingen ortologa eller paraloga.

”Man kan aldrig vara säker på att två nästan identiska organ, proteiner eller nukleinsyrasekvenser verkligen har sitt ursprung hos en gemensam anfader, så termen används allmänt om organ etc. som är så lika varandra att det skulle vara mycket osannolikt att likheterna bara beror på slumpen. Märk att homologi är en absolut egenskap – antingen är proteinerna, organen etc. homologa eller så är de det inte. Det är därför inte korrekt att tala om olika grader av homologi, däremot kan man tala om olika grader av likhet. Två proteiner som är homologa kan alltså visa t.ex. 70 % sekvensidentitet, eller 96 % sekvensidentitet. Organ etc. som är homologa kan ha utvecklats i olika riktningar och se mycket olika ut, men förblir homologa, t.ex. våra armar och fåglarnas vingar som båda har sitt ursprung i framben”

Otrolog
En gen som är homolog till en gen hos en organism av en annan art.
Den gemensamma stamformen för ortologa gener existerade innan de utvecklingslinjer som ledde fram till respektive art skiljdes år. Exempelvis hemoglobin-B-genen hos människa och mus.

Paralog
En gen som är en homolog till en gen hos samma organism. Har uppkommit genom att ursprunglig gen har duplicerats så att de två kopiorna utvecklats oberoende av varandra och fått olika funktioner. Exempelvis Hemoglobiner och Myoglobiner hos människan.

KLASSIFIKATION

1. Mönsterigenkänning
   Igenkänning av särskild sekvens eller struktur som kan associeras med särskild karaktäristik.
2. Prediktion
   - Homologibaserad prediktion
     För okänt protein identifieras de med hjälp av att identifiera homologer med känd funktion eller struktur. Med hjälp av informationen om den kända homologen kan man förutspå strukturen/funktionen hos det okända proteinet.
   - Ab initio (de novo) prediktion
     ”from the beginning”. Prediktion görs genom användning av datoriserade modeller utan direkt jämförelse med existerande data
   - Sequence alignment
     Identifikation genom att identifiera liknande regioner.
     Lägger upp bredvid varandra. Hittat sekvenser som delvis liknar varandra.
     - Global alignment
       Arrangerar varje rest i varje sekvens. Bra för sekvenser med i grova drag samma storlek och likhet.
     - Local Alignment
       Användbar för olika sekvenser innehållande lika regioner ELLER lika sekvensteman inom de större sekvenskontexten.
   - Parwise sequence alignment methods
     - Dot-matrix method
       enkelt vis att evaluera likheter mellan två sekvenser. I en graf är ena sekvensen på Y-axeln och den andra på X-axeln. Markering finns där det finns en matchning. Om likhet finns bildas en linje.
     - Dynamisk programmering
       Likhet ger score. Straffpoäng för olikhet.
       Vanligtvis använder proteinalignments en substitutionsmatris för att tilldela aminosyran scores för dess matches och mismatches. Klyft-straff för att matcha en aminosyra i en sekvens med en klyfta en annan.

- Ordmetod
  Letar efter ganska exakt överensstämmelse i stor databas.

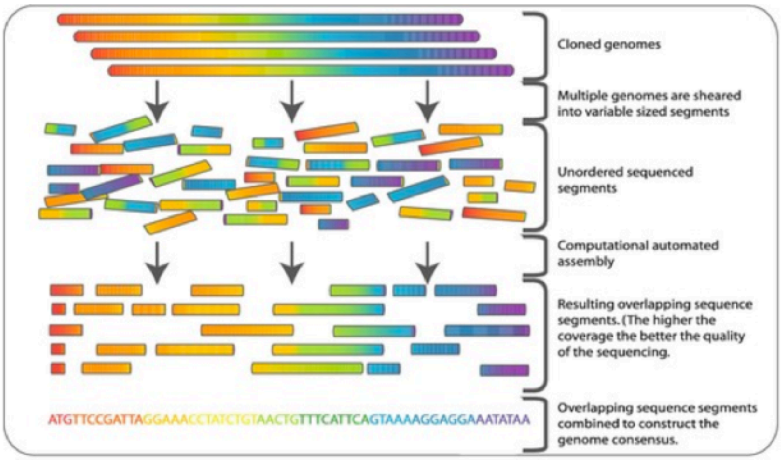
- Multipel sekvensalignment
  Utöka parvisa metoden. Alltså samma sak men i mycket större skala.
  Clustal program eller PSI-BLAST (position specific iterated basic local alignment search tool.

Användningsområden

- Genomik
- Genexpressionsanalys
- Proteomik
- Biologiska nätverk

1. Genomik

- **Shotgun – genome assembly**
  Man har ett antal exakt likadana genom, Sönderdelar dem I slumpvis stora delar. Sekvensbestämmer smådelarna.
  Hittar bitar som överlappar → talar om var de andra ska ligga.
  Är svår metod ty stora delar är identiska, upprepade alltså.
  Kallas för ”Repeats” och kan vara tusentals nukleotider långa och finnas på tudenrals olika platser. Särskilt i stora genom, såsom i plantor och människor.



- **Genprediktion**
  Man listar ut vilka delar i ett DNA som kodar för vilka gener.
  Man sorterar ut de proteiner som endast finns kvar av historiska skäl. Dvs filtrerar bort icke-kodande bitar. Detektera funktionella delar, mönster och ORF (open reading frame) med hjälp av

  - Empiriska metoder
    Letar efter sekvenser som liknar redan kända. Kan använda lokal alingment matchningar. Dessa kan vara totala eller partiella. Söks upp mha algoritmer för local alignment. T.ex BLAST. Utfall av metoden begränsas av Upplösning och innehållsmängd i sekvensdatabasen.

  - Ab initio
    Baserad på geninnehåll och signaldetektion. Startdekvens – mRNA
    Hel sekvens – open reading frame. Som ej har verklig funktion. Ligger mellan viktiga, kodande delar.
    Allt finns i OCF. ” grå zoner”.
    Det däremellan – intron (försvinner, tas bort med splicing)
    Det av intresse – exon.
    Sätts på en poly-A-svans. Mer komplext i eukaryot än prokaryot.
    CpG-öar – Gener som finns men ej kodar ty stängts av. Bland annat pga att det finns så mycket introner i Eukaryoter är det svårare att detektera periodicitet i eukaryoter.

- **Phylogenetiskt träd**
  Molekylär sekvensering och morfologiska datamatriser.
  Phylogenetiska trädet representerar en hypotes gällande det evolutionära förhållandet mellan olika biologiska arter. Se hur organismer är besläktade. Titta på sekvens.
  Nukleotidsekvenskodande gener är grunden för klassifikationen. Multipel sekvensalignment spelar viktig roll.
  Metoder:
  - multipel sekvens
  - avstånd mellan
  - antal mutationer mellan.

Sannolikhet för en mutation i förhållande till en annan.

- Distansmatrismetod
  Genetiska avståndet – andel som inte matchar
  Mellanrum ignoreras eller räknas som missmatch.
- Maximal parsimony method
  Snålaste. Titta på vad som mest ekonomiskt från en sekvens till en annan. Hur många mutationer som skett.
  Identifierar den potentiella phylogenetiska träd som kräver lägst totalt nummer av evolutionära händelser som förklaring till observerad sekvensdata.
  Användbar då alla event ej har samma sannolikhet.
- Max. Likelyhood
  Hur troligt att en sak hänt i förhållande till en annan.
  Fler mutationer = lägre sannolikhet

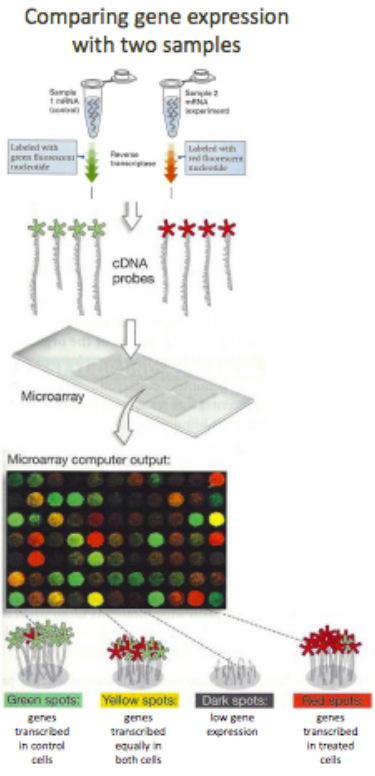
Genomspännande associationsstudier , GWAS

Tittar på antal individer. Tittar på enklaste nukleotiderna i DNA som bytts ut och givit skillnad mellan individerna. Se vad som ger upphov till visst särdrag. Beräknas statistiskt. Om den ena eller andra har såeciell benägenhet att utveckla särskild sjukdom eller dylikt. Får ut t-värden. Ska vara signigfikant skiljt från ±1.
Allel är då associerad med sjukdom.
C-DNA = komplementärt DNA.
Fokus ligger på att finna associationer mellan enkel polynukleotid polymorfos (SNPs) och drag såsom stora sjukdomar. Tittar på många vanliga genetiska varianter hos olika individer för att se om variationen kan associeras med ett visst mönster. Är ej en sökande metod ty den undersöker hela genomet.
odds ratio beräknas. Odds för sjukdomen om man har sen specifika allelens akt odds för sjukdom om man inte har allelen. P-värde för signifikansen hos oddset räknas sedan ut med tji-två-test.
Svaghet i dessa studier är att de endast fokuserar på vanliga genetiska variationer då dess antagande är att vanliga genetiska variationer förklarar ärftligheten i vanliga sjukdomar.

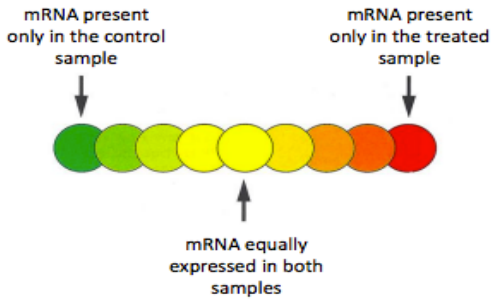
2. Genexpressionsanalys
Mätning av expression av tusentals gener simultant. Detta för att skapa en global bild av den cellulära funktionen. Sekvensen talar om vad cellen möjligen kan göra. Expressionsprofilen talar om vad den faktiskt gör vid viss punkt i tiden.

Tekniker

- DNA micro array
  Mäter den relativa aktiviteten av tidigare identifierade target-gener.
  Ett DNA-chip är en samling av mikroskopiska DNA-spots länkade till en fast yta. Varje DNA-spot innehåller en specifik DNA-sekvens, kallad prob. (kort gensektion) Proberna används för att hybridiseras med ett cDNAsampel under strikta förhållanden. Prob-targethybridiseringen detekteras och kvantifieras genom detektion av fluoroforer, silver och chemilumescence-märkta targets. Detta för att avgöra den relativa bindningen av nukleisyrasekvenser i target.
  Kan användas för att mäta förändringar i expressionsnivå, detektera enkla nukleotidpolymorfoser och för genotyp.



Dataanalys av mikroarray är viktig. Varje mikroarray ger tiotusentals datapunkter.



- Serial analysis of gene expression (SAGE)

Skapar snapshot av mRNApopulationen i samplet i form av små taggar som korresponderar till fragment av transkripten.

- RNA-seq (RNA-sequencing)
  Använder förmågan av next-generation sequencing för att visa snapshot av RNA-närvaro och kvantitet vid samma tidpunkt.

3. Proteomics

Struktur och prediktion
Varje protein består av en oveckad polypeptid eller random coil. Det vecklas sedan in till en karaktäristisk och funtionell tredimensionell struktur.
Korrekt tredimensionell struktur är essentiell för funktionen. Dock kan vissa delar hos funktionella protein förbli oveckade.

Somliga neurodegenerativa sjukdomar beror på en ackumulering av missveckning. dvs att de förblir oveckade.
Sekundär strukturprediktion. Lokalisera sekundära strukturer baserat på kunskap om dess aminosyrasekvens.

Tetriär strukturprediktion

- *Komparativ proteinmodellering.*
  Använder tidigare lösta strukturer som startpunkt.
  Homologisk modellering: Baserat på förutsatsen att två homologa proteiner delar lika strukturer.
  Används för att avgöra vilken del av en protein som är viktig i strukturformationen och interagerar med andra protein. Informationen används för att förutse strukturen av en protein från dess homolog.

Fold recognition. Metod för modellering av proteiner med samma vikning med kända strukturer, men som saknar kända homologer.

- *De novo physics-based modeling*
  Refererar till algoritmprocess genom vilken proteiners tetriära struktur kan förutspås från aminosyrasekvensen.(den primära strukturen)

4. Biologiska nätverk

Representerar och analyserar biologiska komplexa system.
Noder och förbindelselinjer. Står för interaktion. Linjer – edges.
Grad – antallinjer som förbinder noder.
Betweenness – Mäter hur central en nod är i ett nätverk. Noder med hög betweenness fungerar som broar mellan olika punkter i nätverket.

Interaktom: Fysiska interaktioner mellan molekyler i en cell.
Typer av interaktioner:

- Protein-protein nätverk
- Protein-DNA nätverk
- Metaboliska nätverg –
- Signalnätverk

Proteom: hela settet av proteiner uttryckta av ett genom, en cell, vävnad eller en organism vid ett särskilt tillfälle. ”Uttryckta protein av en given cell eller organism vid en given punkt i tiden under definierade förhållanden”