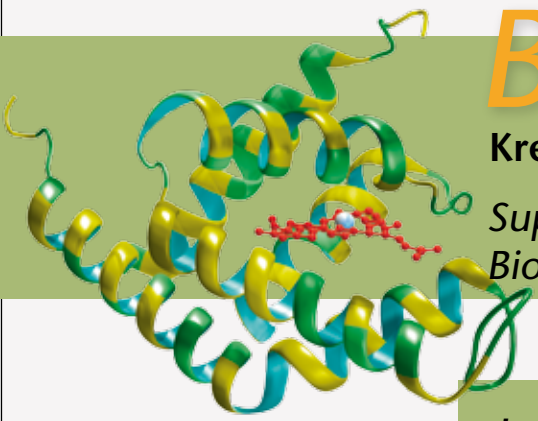

BIOINFORMATIK

Kresten Cæsar Torp

Supplerende materiale til
Biokemibogen – liv, funktion, molekyle

INDHOLD

DNA

DNA's struktur og metoderne til sekventering af DNA er beskrevet i **Biokemibogen** side 39-53. På side 51 opsummeres nogle af de vigtigste problemer man kan behandle vha. www-baserede databaser:

- Søge efter bestemte sekvenser eller gener.
- Sammenligne sekvenser af ønskede gener (alignment).
- Søge efter homologe sekvenser, dvs. sekvenser der går igen i gener fra forskellige arter.
- Søge efter hvor et bestemt gen kan klippes med bestemte enzymer.
- Søge på oversættelser af generne til protein.

Her er en praktisk vejledning til hvordan man kan prøve nogle af disse funktioner af.

Hvilke databaser skal man vælge?

DNA-sekvenser ligger på tre primære databaser, EMBL (Europa), GenBank (USA) og DDJB (Japan), som opdateres dagligt. Her ud over findes der et antal særlige databaser som letter specielle former for søgning. *Non-redundant (nr)*-databaser er omfattende databaser som har den fordel at man kan undgå flere udgaver af samme resultat. De mindre databaser, fx *expressed sequence tags (EST)* har flere referencer og kommentarer til søgeresultatet. Fra databaserne er der adgang til forskellige søgeprogrammer og hjælpeværktøjer.

Her starter vi på GenBank, www.ncbi.nlm.nih.gov
Det kan være en god ide først at orientere sig på siden. Under fanerne for oven på siden er der adgang til forskellige databaser og søgeprogrammer der går på tværs af disse. Der er også adgang til artikelsamlinger som dog kræver registrering. Se fx på kromosomkortene som du finder under *Entrez, Human Genome* og *Browse Genome*. Tryk på kromosomerne og se om du kan finde gener du kender fra **Biokemibogen**.

Søgning på en nukleotidsekvens

Vi har en nukleotidsekvens:

```
TTAATAATGTATTCCAGAGTGAAGCAGATATGTAGAGAAGATGGGCACACATATGTTGAAGTGAATGACTTAACTTTGAC
ATTGTCAAATCATTTGTCATTTCATGCTGCTTCAGAGTCTCTGAAGTTTTTGAAGGATATTGGTGTGGTGACATATGAGA
AGTCCTGTGTCTTCCCTTATGACCTTTACCATGCTGAAAGAGCCATGGCCTTTTCAATTTGTGACCTGATGAAGAAACCT
CCTTGGCATTTATGTGTCGATGTCGTAAAGGTGCTTGCCTCTATTCACACCACAAAACCTGAGAATTCAAGCGATGATGC
ATTGAATGAGAGCAAACCTGATGAATTAAGATTAGAAAATCCTGTGGATGTTGTGGACACACAGGACAATGGTGACCATA
TTTGGACTAATGGTGAAAAT
```

Hvor i det menneskelige genom hører den hjemme?

For at finde ud af det laver vi en homologisøgning med søgeprogrammet *BLAST* (Basic Local Alignment Search Tool). Homologi, enslydende, refererer egentlig til en antagelse om at to gener er evolutionært beslægtede fordi de stammer fra beslægtede organismer. Her anvendes det dog om i hvor høj grad nukleotiderne i to sekvenser stemmer overens, slægtskab eller ej.

1. Vælg *BLAST* i menuen. Herfra vil der være forskellige muligheder for søgning. Man kan vælge hvilken art man ønsker at søge på, om man vil søge på nukleotidsekvenser, proteiner eller oversættelser mellem nukleotider og aminosyrer. Søg i det menneskelige genom ved at vælge *human*.
2. Sæt sekvensen ind i søgefeltet (*Enter an accession*). Kopier ovenstående sekvens og sæt den ind.

3. Herefter er der flere muligheder:
    a. *Set subsequence*: Her kan man udvælge dele af sekvensen for søgningen.
    b. *Database*: Her kan man vælge mellem forskellige databaser. I dette tilfælde vælges genome (all assemblies) hvorved søgningen sker i forskelllige sekventerede genomer.
    c. *Program*: Her vælger man om man vil sammenligne med andre nukleotidsekvenser, med proteinsekvenser som kan være resultat af oversættelse af nukleotidsekvensen eller med andre arters genom. Vælg *megablast*.
    d. *Expect*: Dette tal angiver sandsynligheden for at den fundne sekvens stemmer overens med den indtastede ved et tilfælde. 0,01 angiver således at det vil kunne forventes i 1 ud af 100 gange. Ofte sættes værdien til 10 i første søgning. Resultater med en højere sandsynlighed end den angivne vises ikke.
    e. *Filter*: Her kan man aktivere et filter som frasorterer visse sekvenser som går igen mange steder på genomet.
    f. *Descriptions* og *alignments*: Her vælger man antallet af korte beskrivelser og sekvenser man ønsker at se. De sorteres efter hvilke der passer bedst med den indtastede, så selvom der vises mange, vil de først viste være dem der passer bedst

4. *Begin search* bringer næste side frem hvor man kan foretage yderligere valg for at begrænse søgningen. Her kan man bl.a. med flueben i *graphical overview* få en farveindikation for hvor sekvensen passer bedst med de fundne (*alignment score*).

5. Tryk på *view report* og se resultatet.
    a. Øverst vises i hvilken database matchene er fundet. Ved søgning i genomer kan man med *Genome View* se placeringen af sekvensen i genomet. På hvilket kromosom befinder sekvensen sig? Vælger du Genome view kan du blive nødt til at søge igen fordi siden forældes.
    b. Herefter følger en grafisk gengivelse af hvor godt de fundne sekvenser matcher. Hvad viser farven? Passer de fundne sekvenser godt med den indtastede?
    c. Referencen nedenunder virker som et direkte link til den database hvor sekvensen er fundet. Under referencen er evt. beskrevet *features*, dvs. egenskaber, for denne del af genomet. Hvilket gen stammer denne sekvens fra? Slå navnet efter i **Biokemibogen**, og find ud af hvilken funktion det har i cellen.
    d. Herefter følger en direkte sammenligning af baserne mellem *Query*, den indtastede forespørgsel, og *Sbjct*, den tilbageleverede sekvens. Er der fuld overensstemmelse? Er der baser som ikke stemmer overens? (De angives ved at der ikke er streger imellem dem).
6. Prøv nu at gå tilbage og lav en ny søgning på en mindre del af sekvensen. Gå tilbage og vælg fx *set subsequence* til 1-60.
    a. Hvor høj er alignmentscoren nu? Hvorfor kan man ikke have så stor tillid til resultatet når sekvensen bliver kortere?
7. For meget korte sekvenser skal man bruge andre søgeprogrammer. Prøv fx at sammenligne en sekvens på 30 nukleotider.
8. Lav mutationer i sekvensen. Byt baser ud, søg og se hvad der sker. Prøv også at indsætte eller fjerne baser. Hvordan gengives disse huller eller *gaps*? Hvad kalder man disse typer af mutationer? Se **Biokemibogen** side 61.

Har andre arter gener for helikase der minder om vores?
Gå tilbage til BLAST-startsiden, hvor du kan vælge at søge i andre arters genom. Ud over menneske, mus og rotte, anvendes arternes latinske navne. Vælg evt. søgeprogrammet *blastn* i stedet for *megaBLAST*, idet det tager flere sekvenser med og reguler evt. på *expect*.
Hvor finders tilsvarende sekvenser?

Proteiner

Søgning på aminosyresekvenser

Søgning på aminosyresekvenser kan også foretages vha. BLAST.
I databaser som disse gengives aminosyrernes navne normalt med ét-bogstav-forkortelser, i stedet for de tre som er brugt i **Biokemibogen**, side 72:

Aminosyre	**Forkortelser**	
Alanin	Ala	A
Arginin	Arg	R
Asparagin	Asn	N
Aspartat	Asp	D
Cystein	Cys	C
Glutamat	Glu	E
Glutamin	Gln	Q
Glycin	Gly	G
Histidin	His	H
Isoleucin	Ile	I
Leucin	Leu	L
Lysin	Lys	K
Methionin	Met	M
Phenylalanin	Phe	F
Prolin	Pro	P
Serin	Ser	S
Threonin	Thr	T
Tryptofan	Trp	W
Tyrosin	Tyr	Y
Valin	Val	V

Hvor stammer følgende aminosyresekvens fra?

TTSDVVVAGEFDQGSSSEKIQKLKIAKVFKNSKYNSLTINNDITLLKLSTAASFSQTV

Har du allerede prøvet at søge på nukleotidsekvenser, kan du også her bruge Genbank på adressen www.ncbi.nlm.nih.gov.

1. Vælg *BLAST* i menuen. Vælg derefter *protein blast* under *Basic BLAST*
2. Aminosyresekvensen kopieres til søgefeltet.
3. I databasen kan man vælge mellem flere muligheder. Både *pdb* (*Protein Data Bank*) og *Swissprot* er databaser hvor proteinerne er godt annoterede, og det er let at arbejde videre med deres tredimensionelle strukturer. Vælg *non-redundant protein sequences (nr)*.
4. Tryk *BLAST* og vurder resultatet.
    a. Først vises referencer til søgningen, og hvor mange sekvenser der er fundet.
    b. Dernæst vises med en farvekode hvor sandsynligt det er at de fundne sekvenser matcher den indtastede (Query). Sammenlign farverne med farveskalaen (alignment scores). Hvor godt stemmer de overens?
    c. Herefter følger referencer til databaser og artikler for de fundne sekvenser. Øverst vises de sekvenser der matcher bedst. Se de øverste titler igennem. Giver de et hint om hvilket protein der er tale om? Find proteinet i **Biokemibogen**. Hvad er det for et protein?
    d. Se listen over søgeresultater igennem. Hvilke arter kan du finde referencer til? Hvad fortæller det om proteinet at sekvenserne er forholdsvis ens? Sammenlign evt. afstanden mellem sekvenserne vha. *distance tree of results*.
    e. Søgningen finder også et andet protein, som du kan finde i **Biokemibogen** (se fx side 80-81). Er det overraskende at begge proteiner matcher sekvensen?

Det kan alternativt være en god ide at søge i en database med mere fyldestgørende beskrivelser af proteinerne. Gå fx tilbage, indsæt sekvensen og vælg i stedet Protein Data Bank, pdb. Hvordan ser søgeresultaterne nu ud?
Koden til venstre (fx pdb|1GCT|A) er molekylets PDB ID. 1GCT kan således bruges til at søge videre på proteinets struktur, bl.a. på den følgende database.

Søgning på proteiners 3-D-struktur

Nogle databaser indeholder lettilgængelig information om proteiners tredimensionelle struktur, både strukturer bestemt eksperimentelt og teoretisk beregnede modeller, se **Biokemibogen** side 78.
Her bruger vi www.rcsb.org/pdb. Orienter dig på hjemmesiden. Under *General education* og *Educational resources* er der fx gode animationer, posters, *flash tutorials* (vejledninger til databasen), vejledninger til at bygge fysiske modeller af proteinerne o.l.
Udskriv fx posteren *Molecular Machinery: A Tour of the Protein Data Bank* og hæng den op i klasselokalet.
Hvem kan forsyne flest af molekylerne med sidehenvisninger til bogen?

Søgning på baggrund af aminosyresekvens

I det følgende ser vi nærmere på et af bogens molekyler.
1. Vælg *search*.
2. Starter man med en aminosyresekvens, vælges *sequence*, og sekvensen kopiers til søgefeltet. Marker feltet *use sequence*.
3. Tryk *search* og vurder resultatet. Antallet af søgeresultater kan indsnævres ved at justere Expect-værdien (Her hedder den: *E cut off*) ned til fx 0.001.
4. Til venstre for hvert søgeresultat er en lille molekyltegning. Aktiveres linket kommer molekyltegningen frem i højre side. Under tegningen er der links til forskellige hjælpeprogrammer som kan bruges til at undersøge og arbejde videre med strukturen.

Søgning på baggrund af navn, art el. lign.

Da mange af molekylerne i databasen er modificerede i forbindelse med de forsøg de stammer fra, kan dette være vanskeligt.
1. Vælg *search* i venstre side Dette giver forskellige muligheder for søgninger.
2. Vælg *advanced search*.
3. Under *Choose a query type* vælges fx *molecule name*.
4. Skriv molekylets engelske navn i søgefeltet og tryk på *evaluate subquery*. Du kan nu se hvor mange resultater du får.
5. Vælg evt. at indsnævre søgningen ved at trykke på +.
6. Vælg fx hvilken organisme, *Source Organism* du søger indenfor.
7. Tryk *evaluate query*. Du får nu de fundne molekyler.

Søgning på baggrund af PDB ID

De molekyler som indrapporteres til databasen forsynes med et PDB ID. Med PDB ID kan man søge på molekylet, dels i søgebjælken for oven på hjemmesiden, dels via de øvrige søgemuligheder.
PDB ID for de vigtigste af Biokemibogens proteiner er gengivet i tabellen under 'Søg videre' nedenfor.
1. Skriv fx PDB ID for Carboxypeptidase, 1M4L i søgefeltet. Tryk *Site Search*.
2. Hvilke oplysninger får man om proteinet?

Undersøgelse af proteinets tredimensionelle struktur

Under tegningen til højre er der links til forskellige hjælpeprogrammer som kan bruges til at undersøge og arbejde videre med strukturen. Vælg fx *MBT ProteinWorkshop*. Flere af programmerne kræver at java er installeret på computeren. Det kan downloades gratis her.
1. Drej molekylet med musen ved at holde venstre museknap nede, og trække i molekylet. Ved at holde shift nede samtidig, kan der zoomes ud og ind. Sammenlign med **Biokemibogen** figur 87. Identificer zink-ionen, den hydrofobe grube og lukkearmen.
2. Programmet giver mulighed for at fremhæve egenskaber ved molekylet, og ændre i det.
    a. Vælg *shortcuts*. Giv proteinet farve efter hvor hydrofobe aminosyrerne er (*Hydrophobicity*). Sammenlign med **Biokemibogens** s. 76. Er der også en tendens til at hydrofobe aminosyrer vender indad og hydrofile vender udad på dette molekyle? Hvad betyder det for molekylet? Er den hydrofobe grube hydrofob?
    b. Farvelæg på baggrund af sekundærstruktur (*Conformation Type*). Hvor mange helix'er og foldeblade, *strands*, kan du tælle?
3. Hvor sidder de tre aktive aminosyrer i det aktive område?
    a. Vælg *Tools* og *Colors*.

   b. Vælg *Ribbons*. Dette giver mulighed for at ændre farverne på enkelte aminosyrer og større afsnit.
   c. Marker farvefeltet, og vælg en markant aktiv farve fra farveskemaet.
   d. Marker herefter de aktive aminosyrer, *145 arg*, *248 tyr* og *270 glu*. Åben først A-kæden, *Chain A* ved at trykke på +. Herved bliver de enkelte aminosyrer synlige. Derefter markeres de udvalgte aminosyrer.
4. De pågældende aminosyrers radikaler kan tilsvarende gøres synlige.
   a. Vælg *Tools* og *Visibility*.
   b. Marker *Atoms and Bonds*.
   c. Åben A-kæden som før, og marker aminosyrerne. Hvordan sidder de tre aminosyrers radikaler i forhold til hinanden?

Find tilsvarende forstadiet til carboxypeptidase, Procarboxypeptidase, 1AYE. Sammenlign de to molekylers tredimensionelle struktur.
1. Hvad er der sket ved aktiveringen af molekylet?
2. Hvilken type enzym skal der bruges for at aktivere enzymet på den måde? Hvilken type binding skal det kunne spalte? Sammenlign med **Biokemibogen** side 89.

Søg videre

Andre af **Biokemibogen**'s proteiner kan tilsvarende findes vha. deres PDB ID og undersøges nærmere. Suppler bogens informationer med oplysninger om molekylet fra andre kilder, fx ved en Google-søgning. Du kan finde molekylernes engelske stavemåde i bogens stikordsregister.
Forklar hvordan molekylet virker for klassen. Gengiv molekylets vinkel, farver osv., på den måde der bedst viser de vigtige pointer. Brug ideerne ovenfor.
Flere af molekylerne er symmetriske idet de består af flere ens underenheder. Ved flere af modellerne kan man skifte mellem hele komplekset og underenheden, den asymmetriske enhed. Nogle filer er store og kræver lidt tålmodighed.

Proteiner fra kapitel 3	**Side i Biokemibogen**	**PDB ID**
DNA-polymerase I – enzymkomplekset består af mange underenheder. Her ses den DNA-bindende del.	45	1D8Y
RNA-polymerase med en DNA-streng der deles.	54	1I6H
RISC's binding til dsRNA	69	1ytu
Proteiner fra kapitel 4		
Insulin	74	IB2F
Hæmoglobin /Hemoglobin	76-77	1A3N
Kymotrypsin /Chymotrypsin	80-81	1GCT
Kymotrypsinogen /Chymotrypsinogen	89	1CHG
Elastase (her med bundet kulhydrat). Sammenlign med de øvrige proteaser i Biokemibogen, særlig kymotrypsin og trypsin. Marker evt. den ladede asp 189 i gruben. Er denne grube hydrofob?	80-81	1BOF
Trypsin. Sammenlign med kymotrypsin og elastase. Marker evt. val 190 og val 216 som fylder op i den hydrofobe grube.	80-81	5PTP
Pepsin. Pepsin afsondres i maven og virker ikke i tolvfingertarmen hvor de øvrige proteaser tilsættes føden. Hvorfor kan vi ikke bruge det samme enzym begge steder? Læs evt. mere side 86-87 eller i en fysiologibog.	87	3PSG
P-loop ATP-ase	91-92	1I6I og 1I5S
Myosin (kan du finde P-loop ATP-asen?)	90, 92-93	1DFL
Aktin-figuren viser en af monomererne, underenhederne i aktinfilamentet. Sammenlign med gengivelsen i bogen.	90, 92-93	1ATN
Kinesin	94	1I6I
K^+-kanal	96-97	1K4C
Na^+/K^+-pumpe	97-98	1SV4
ABC-transporter /ATP Binding Casette-protein. Modellen viser Casettedelen, hvor ATP bindes og spaltes. Sammenlign med P-loop ATP-asen s. 91-92.	99	1F3O
Antistof. Identificer de lette og tunge kæder, de konstante og de variable dele. Hvor er det præcist at antigenet bindes? Sammenlign med figuren i bogen.	102	1IGT
Proteiner fra kapitel 5		
Fotosystem I	115	1JBO
Fotosystem II	114	1S5L
Cytokromer er oxidoreduktaser, dvs. enzymer som kan optage og afgive elektroner, de reduceres og oxideres. De indeholder oftest hæm-gruppen som en prostetisk gruppe. Her følger henvisninger for cytokrom bf fra kloroplasternes og elektrontransportkædens enzymkompleks III og IV (cytokrom bc og cytokrom c-oxidase). Sammenlign deres funktioner vha. **Biokemibogen** side 115 og 128.		

Cytokrom bf /cytochrome bf er et kompleks bestående bl.a. af tre cytokromer, af cytokrom b-typen og cytokrom f, som er et c-type kromosom. Her er cytokrom f. Bemærk hæm-gruppen, som optræder i alle cytokromerne.	115	1CTM
Cytokrom bc. Elektrontransportkædens enzymkompleks III, minder i opbygning om cytokrom bf, og består ligeledes af to asymmetriske enheder. Filen viser en asymmetrisk enhed. Farvelæg den efter polaritet (shortcuts og hydrophobicity). Molekylet er indlejret i mitokondriemembranen ved at midterafsnittet består af hydrofobe helixer som når gennem membranen. Passer det med farverne? Forklar princippet ud fra cellemembranens opbygning. Sammenlign med andre proteiner i cellemembranen, som optræder i **Biokemibogen**.	128 33-37	1BCC
Cytokrom c-oxidase, elektrontransportkædens enzymkompleks IV. Orienter dig i forhold til de parallelle helixer, som er indlejret i membranen. Elektroner overføres mellem kompleks III og IV af Cytokrom c, som kan bindes i gruben dannet af parallelle β-foldeblade.	128	2OCC
Cytokrom c. Molekylets struktur er stort set ens i alle organismer som respirerer.	128	3CYT
Nitratreduktase	135	1R27
Proteiner fra kapitel 6		
Nitrogenase består af to dele, reduktasen og nitrogenasen. Reduktasen. Find det kubiske Fe-S-kompleks. Her binder molekylet elektroner. Find også ATP-bindingsstederne. Disse bindingssteder er specialiserede P-loop ATP-aser. Når de spaltes ændrer reduktasen facon, FeS-komplekset bevæges ned til nitrogenasen og afleverer elektronerne. Processen er altså energikrævende. Nitrogenasen. Find cofaktoren i det aktive område, et kompleks som består af ni sulfider bundet til seks omkringliggende cystein-aminosyrer, syv jernatomer og et molybdænatom. Det er her N_2 modtager elektroner og reduceres.	153 91-92	1N2C 1M1N
Leghæmoglobin. Identificer hæmgruppen, og sammenlign molekylet med myoglobin og hæmoglobin.	154	1BIN
Proteiner fra kapitel 7		
Fosforylase a og b. Identificer de to asymmetriske enheder og ATP's bindingssteder. Sammenlign med bogens figur. Hvilke forskelle er der på de to former. Hvad er der sket med ATP?	178	1GPA og 1NOJ
HGH /human growth hormone /somatotropin. Sammenlign med figur 209, side 185 i **Biokemibogen**.	185	1HGU