nucleus
www.nucleus.dk

8

Databasesøgning på DNA-sekvenser

Relevant baggrundslæsning

- DNA's opbygning og nucleotidsekvenser: Bioteknologi 1, side 51-54.
- Den genetiske kode: Bioteknologi 1, side 56-57.
- Kromosomer og diversitet: Bioteknologi 1, side 60-62.
- DNA-sekventering: Bioteknologi 1, side 68-69.
- Isolering af gener: Bioteknologi 2, side 76-78.
- DNA-biblioteker: Bioteknologi 2, side 77.

På de internetbaserede databaser kan man fx:

- 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. Herudover findes der et antal særlige databaser som letter specielle former for søgning.

Non-redundant-databaser (nr) 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 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.
Forneden er der forskellige links. Vi arbejder videre fra GenBank:
http://www.ncbi.nlm.nih.gov/genbank/.

Søgning på en nucleotidsekvens

Vi har fundet en nucleotidsekvens:

ttctgcaccagtaaatctcaccagcaaattattttggttcataaaaacaggagtttctttttgaaggggaaattcatgtttct-
gttttttttttctttctttcttaaagaaagggttttatgtgtgttaagatctctgcacaaccaatcacctcaacaagtgtttggattactgttc-
caggagcagctgacagacgaagaaaagtcctggacaggaagggagaattctgacgccaacatgcagcgaagtaccatgtgag-
caccctcccttgcccctgagtctttccttctgcaagtcct

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 i betydningen af i hvor høj grad nucleotiderne i to sekvenser stemmer overens, slægtskab eller ej. Under BLAST-menuen er der forneden links til andre søgeværktøjer, fx Primer-BLAST som kan bruges til at finde egnede primere, og SNP-blast som kan være interessant, hvis man arbejder med genetisk fingeraftryk, se Bioteknologi 1, side 61.

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 nucleotidsekvenser, proteiner eller oversættelser mellem nucleotider og aminosyrer. Søg i det menneskelige genom ved at vælge *human*.
2. Skriv sekvensen ind i søgefeltet (*Enter an accession*).
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 forskellige sekventerede genomer.
   c. *Program*: Her vælger man om man vil sammenligne med andre nucleotidsekvenser, med proteinsekvenser som kan være resultat af oversættelse af nucleotidsekvensen 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 af hvor sekvensen passer bedst med de fundne (*alignment score*).
   a. Tryk på *View report* og se resultatet.
   b. Øverst vises i hvilken database matchene er fundet. Ved søgning i genomer kan man med *Human genome view* se placeringen af sekvensen i genomet. På hvilket kromosom befinder sekvensen sig? Vælger du *Human genome view,* kan du blive nødt til at søge igen fordi siden forældes.
   c. Herefter følger en grafisk gengivelse af hvor godt de fundne sekvenser matcher. Hvad viser farven? Passer de fundne sekvenser godt med den indtastede?
   d. Referencen (fx *ref|NM_000234.1|*) nedenunder virker som et direkte link til den database hvor sekvensen er fundet.
   e. Efter referencenummeret følger en kort beskrivelse af hvilket gen der er tale om, fra hvilken organisme det kommer, og evt. fra hvilket kromosom.
   f. Hvilket protein koder det gen for, som denne sekvens stammer fra? Slå navnet efter i Bioteknologi 1 eller 2, og find ud af hvilken funktion det har i cellen.

g. Herefter følger en direkte sammenligning af baserne mellem *Query*, den indtastede forespørgsel, og *Sbjct*, den sekvens som databasen leverer tilbage. Er der fuld overensstemmelse? Er der baser som ikke stemmer overens? (De angives ved at der ikke er streger imellem dem).

5. Prøv nu at gå tilbage til *BLAST Human Sequences* 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?

   b. For meget korte sekvenser skal man bruge andre søgeprogrammer.
   Prøv fx at sammenligne en sekvens på 30 nucleotider.

6. 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?