



Databasesøgning på aminosyresekvenser

Formål

At afprøve hvordan man kan bruge internetbaserede databaser til at søge efter og sammenligne proteiners aminosyresekvenser.

Relevant baggrundslæsning

- Proteinernes primærstruktur: Bioteknologi 2, side 51-52.
- Bestemmelse af aminosyresekvens: Bioteknologi 2, side 58-59.

På de internetbaserede databaser kan man fx:

- Søge efter bestemte aminosyresekvenser.
- Sammenligne sekvenser (alignment).
- Søge efter homologe sekvenser, dvs. sekvenser der går igen i gener fra forskellige arter.
- Søge på gener som kan være ophav til proteinet.

Overvej i klassen:

- Hvad kan man bruge sammenligninger af aminosyresekvenser til?
Lav en fælles liste over anvendelser, fordel dem mellem jer, og undersøg nærmere hvordan man gør dette i praksis. Det kan være beskrevet i bøger eller I kan finde eksempler.

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

Hvilke databaser skal man vælge?

Aminosyresekvenser 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.

Denne vejledning gælder GenBank: <http://www.ncbi.nlm.nih.gov/genbank/>.

Ved databasesøgning anvendes normalt aminosyrernes et-bogstav-forkortelser som kan findes i Bioteknologi 2, figur 3, side 52.

Hvor stammer følgende aminosyresekvens fra?

GTWTRPHKAQVDSWETVFGNEGWNWDNVAAYSLQAERARAPNAKQIAAGHYFNASCHG-VNGTVHAGPRD

Vælg *BLAST* (Basic Local Alignment Search Tool) i menuen.

1. Vælg derefter *protein blast* under *Basic BLAST*.
2. Skriv aminosyresekvensen i søgefeltet.

I databasen kan man vælge mellem flere muligheder. Både *pdb* (*Protein Data Bank*) og *Swissprot* er databaser hvor proteinerne er godt annoterede med beskrivelser og referencer, og det er let at arbejde videre med deres tredimensionelle strukturer. Vælg *non-redundant protein sequences (nr)*.

1. Tryk *BLAST* og vurder resultatet.
2. Først vises *Graphic Summary*, en grafisk opsummering af resultatet. Farvekode viser hvor sandsynligt det er at de fundne sekvenser matcher den indtastede (*Query*). I bjælken ovenover er vist hvilken score hver farve repræsenterer. Scoren er en indekssværdi. Databasen finder både sekvenser som er præcist som den indtastede, og sekvenser der næsten svarer til den; hvor godt de matcher angives med farvekoden. Den kan altså også finde sekvenser som stammer fra andre proteiner eller andre arter, sandsynligvis med en lavere score.
 - a. Sammenlign farverne med farveskalaen (*Alignment scores*). Hvor godt stemmer sekvenserne overens?
 - b. Kør musen ned over de farvede linier og se hvilke organismer de stammer fra i feltet foroven. Hvilket protein er der sandsynligvis tale om? Fra hvilken organisme?
2. Herefter følger *Description*, dvs. navne og referencer til databaser for de fundne sekvenser. Under *Query Coverage* kan du se hvor godt de enkelte aminosyresekvenser dækker den indtastede sekvens.
 - a. Hvilke arter kan du finde referencer til?
 - b. Hvad fortæller det om proteinet, at sekvenserne er forholdsvis ens?
3. Nederst følger *Alignment* hvor den indtastede sekvens (*Query*) sammenlignes direkte med databasens sekvenser (*subject (sbjct)*).
 - a. Scroll nedad og find sekvenser som ikke passer helt. Hvordan angives dette?
 - b. Få forskellene illustreret grafisk ved at trykke på fanebladet *distance tree of result*.
 - c. Er der sammenhæng imellem hvilke organismer sekvensen kommer fra, og hvor meget sekvenserne matcher?
 - d. Der er også navne på andre enzymer i træet. Hvad hedder de?
 - e. Hvordan kan det hænge sammen, at flere enzymer kan have matchende sekvenser?

Det kan alternativt være en god idé at søge i en database med mere fyldestgørende beskrivelser af proteinerne.

4. Gå fx tilbage, indsæt sekvensen og vælg i stedet *Protein Data Bank, pdb*. Hvordan ser søgeresultaterne nu ud? Koden til venstre under descriptions eller under alignment (fx [1CF3_A](#) eller [pdb|1CF3|A](#)) er molekylets PDB ID. Det kan bruges til at søge videre på proteinets struktur, se vejledning i Tema 4, Link 8.
5. Tryk på linkene i databasen og se hvad der kommer frem.