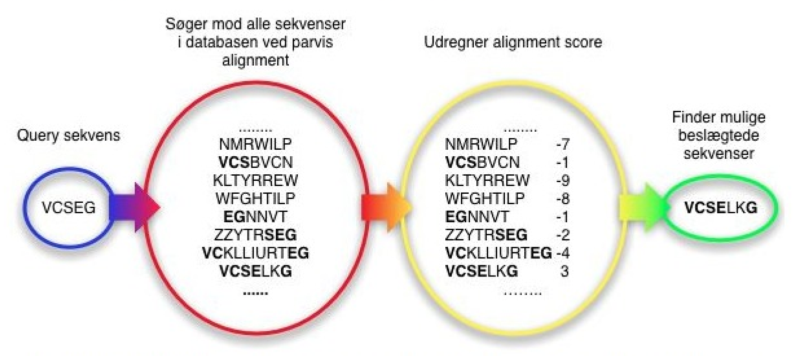
BLAST

« Back to Glossary Index

Basic Local Alignment Search Tool (BLAST) er en metode til at sammenligne DNA- eller proteinsekvenser (find det her). BLAST bliver også brugt i undervisningsforløbet: Bioinformatik – En introduktion.

Beskrivelse

Når man bruger BLAST, laver man en database-søgning, hvor man undersøger, om der i en database findes DNA- eller proteinsekvenser, der ligner den input-sekvens, man har givet BLAST. Den sekvens, man undersøger, betegnes en query-sekvens (se infoboks) og en undersøgelse i BLAST betegnes at blaste sekvensen (se infoboks). Disse to termer vil være gennemgående i det følgende materiale, og det er derfor vigtigt at blive bekendt med dem, se figur 1.

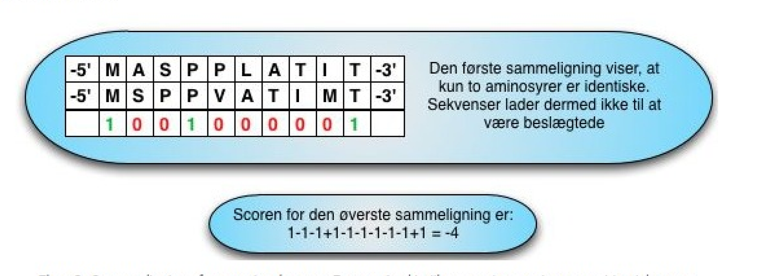


Figur 1: *Illustration af BLAST. Ved at BLASTe en sekvens mod en database laver man en parvis alignment mellem query-sekvensen og alle andre sekvenser i den pågældende database. Hver alignment får en score og tilslut kan man se om der er nogle sekvenser i databasen der er beslægtede med query-sekvensen.*

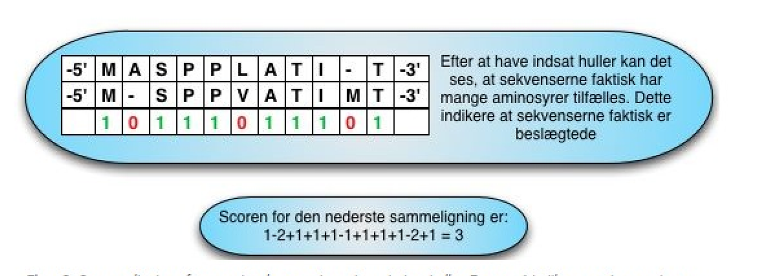
BLAST er opbygget omkring en matematisk model, der sammenligner DNA- eller proteinsekvenser ved brug af parvis alignment. Parvis alignment er en metode til at søge efter lokale ligheder (eng. local alignments) mellem query-sekvenser og de DNA- og proteinsekvenser, der findes i den biologiske database, man anvender. Et alignment er illustreret i figur 2. Man kan udregne en score for, hvor godt alignmentet er, ved at finde de positioner hvor sekvenserne er henholdsvis forskellige og identiske. På baggrund af scoren kan man konkludere, hvor tæt beslægtede sekvenserne er.

BLAST sorterer søgeresultaterne efter, hvor godt de matcher query-sekvensen. Resultatsekvenserne sorteres efter den føromtalte score, hvor en høj score er ensbetydende med et godt match og dermed en indikation af, at de to sekvenser er beslægtede.

De sekvenser, der alignes, er ikke altid lige lange. På grund af mutationer i DNA-sekvenserne kan der være fjernet eller indsat nukleotider i sekvensen, hvilket betyder, at der mangler eller er blevet indsat en (eller flere) ekstra aminosyre(r). Der tages højde for mutationer i alignmentet, da man kan indsætte huller (eng. gaps) i alignmentet for at få en bedre score og dermed en bedre sammenligning. Figur 3 viser, hvordan indsættelse af huller i query-sekvensen og i den sekvens man sammenligner med kan forbedre scoren. For at hele alignmentet ikke bliver fyldt med huller, bliver disse scoret med en negativ værdi, som er lavere end den for to forskellige aminosyre, der står over for hinanden. I figur 3 bliver match scoret med 1, et mis-match (to forskellige aminosyrer over for hinanden) bliver scoret med -1, og huller bliver scoret med -2. Ved sammenligning af figur 2 og 3 ses, at sekvenserne matcher bedre, efter at der er indsat huller i begge sekvenser.



Figur 2: *Sammenligning af to proteinsekvenser. Et grønt 1-tal indikerer, at de to aminosyre er identiske, og et rødt 0 indikerer, at de er forskellige. Til at udregne scoren bliver identiske aminosyrer tillagt værdien 1 og to forskellige aminosyrer får værdien -1.*



Figur 3: *Sammenligning af to proteinsekvenser, hvor der er indsat huller. Et grønt 1 indikerer, at de to aminosyrer er identiske, og et rødt 0 indikerer, at de er forskellige. Til at udregne scoren bliver identiske aminosyrer tillagt værdien 1, får -1, og et hul får -2. Gennem alignmentet kan det ses, at sekvenserne er mere beslægtede end først antaget.*

Når man laver en BLAST, undersøger man om en query-sekvens er beslægtet med andre DNA- eller proteinsekvenser i en database. Når man skal analysere resultaterne fundet gennem BLAST, er det derfor vigtigt at være kritisk, da man vil finde beslægtede sekvenser. I teorien kan ALLE sekvenser nemlig alignes, men blot fordi de kan det eller får en høj score, er det ikke nødvendigvis ensbetydende med, at resultat-sekvensen er beslægtet med query-sekvensen.

I BLAST-resultatet er der inkluderet metoder, der undersøger, om sekvenserne i alignmentet kan karakteriseres som værende beslægtede. Dette gøres blandt andet ved at finde resultat-sekvensernes query-dækning og e-værdi. Nedenfor gennemgås de tre forskellige værdier, som er vigtige at kigge på, når man fortolker et BLAST-resultat.

- ***Query-dækning*** (eng. *query coverage*) angiver, hvor mange procent af den pågældende resultatsekvens der er med i det parvise alignement, den danner med query-sekvensen. Hvis query-dækningen er høj, indikerer det, at en stor andel af resultatsekvensen matcher query-sekvensen. Det skal dog påpeges, at dækningsgraden ikke fortæller, om der er eventuelle huller i alignmentet (se infoboks 4), eller om alle aminosyrerne er identiske. Desuden kan den procentuelle dækning være misvisende, da en query-sekvens kan være en lille del af en større sekvens og den procentuelle dækning vil således ende med at være lille, da den kun angiver den del af resultatsekvensen, der indgår i det parvise alignment med query-sekvensen.
- ***Max+scoren*** er den score, som hver resultatsekvens får på baggrund af matchet med query-sekvensen. Max+scoren udregnes gennem den føromtalte matematiske model og tager ikke højde for, om sekvenserne egentlig er identiske. Query-sekvensen bliver nemlig sammenlignet med ALLE sekvenser i den pågældende database, og hver sammenligning vil få tildelt en score. I et generelt BLAST-output er det resultatsekvenserne med de højeste scorer, der bliver vist først (jo højere score, desto bedre match), men som nævnt er en høj score ikke ensbetydende med et identisk match. Grunden er, at sekvenser rent tilfældigt kan matche en query-sekvens uden egentlig at være beslægtet med den. Et mål for tilfædigheden er e-værdien, der beskrives nedenfor.
- ***E-værdien*** (eng. *e-value* eller *expected value*) angiver den forventede (tilfældige) tilstedeværelse af en resultat-sekvens i den database, man søgte mod. Grunden til, at man bruger e-værdier, er, at man af tilfældige årsager kan finde højt scorende resultat-sekvenser i en stor database. E-værdien for en given sekvens angiver, hvor mange resultater med den samme høje score man ville kunne forvente at få af tilfældige årsager.

Generel guide til BLAST

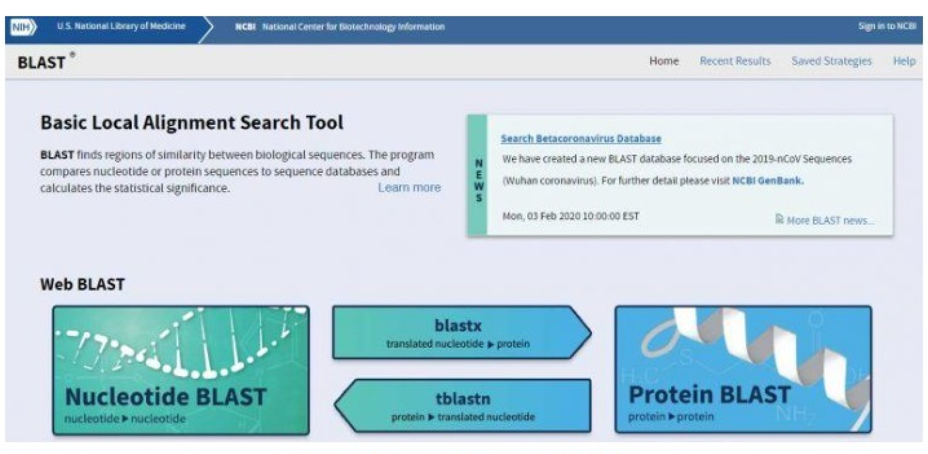
Når man skal foretage en BLAST skal følgende punkter gennemgås:

1. Type BLAST-søgning (protein eller nukleotid)
2. Valg af den organisme og/eller database man vil søge mod.
3. Valg af BLAST-type.
4. Indtastning af sekvens eller upload fil med sekvens.
5. Start af BLAST
6. Fortolkning af resultatet

De seks punkter vil blive gennemgået nedenfor.

Punkt 1

Vælg enten "Nucleotide BLAST" eller "Protein BLAST", afhængigt af din type data.

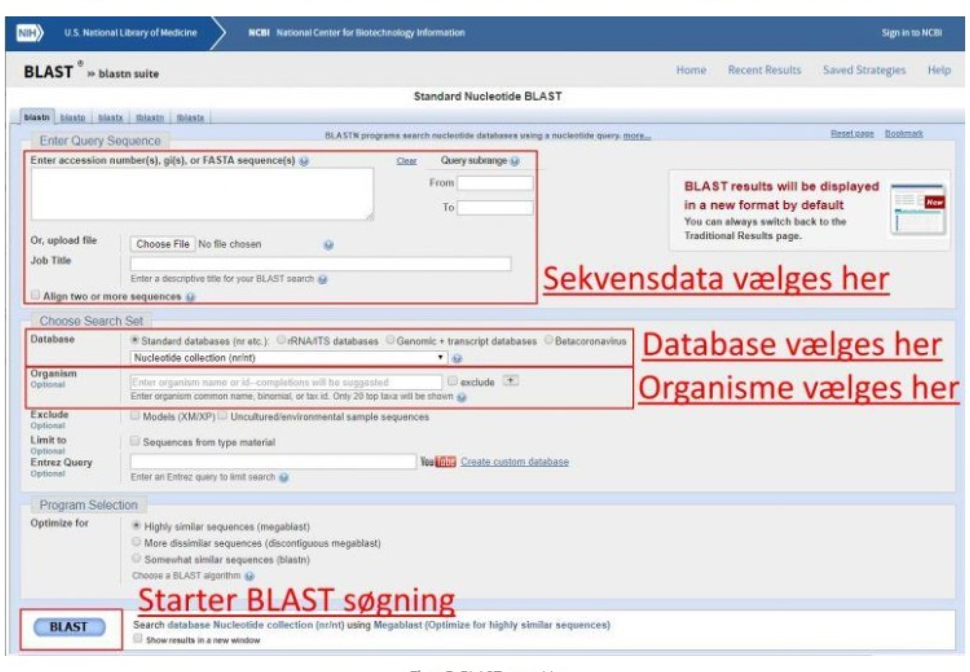


Figur 4: BLAST startsiden som den så ud i februar 2020

Punkt 2+3

På BLAST søgesiden, se figur 5, kan man blandt andet vælge, hvilken organisme man ønsker at blaste query-sekvensen mod. Ens sekvensdata kan enten indtastes manuelt eller i FASTA format, uploades som FASTA-fil, eller som et accession nummer.

Hvis der ikke er noget specifikt ønske om organisme, kan man BLASTe mod alle sekvenser i databasen ved at lade organisme-feltet stå tomt.



Figur 5: BLAST søgeside

Punkt 4+5

Afhængig af den query-sekvens man ønsker at blaste, og om man vil finde lignende DNA- eller proteinsekvenser, skal BLASTen specificeres. Dette gøres ved at vælge, hvilken type BLAST man vil foretage. Der findes følgende former:

- BLASTn: man har en nukleotidsekvens og søger i nukleotiddatabaser.
- BLASTp: man har en proteinsekvens og søger i proteindatabaser.
- BLASTx: man har en nukleotidsekvens og søger i proteindatabaser.
- tBLASTn: man har en proteinsekvens og søger i nukleotiddatabaser.
- tBLASTx: man har en nukleotidsekvens og søger i nukleotiddatabaser (denne er mere omfattende end BLASTn).

De typer af BLAST, der er mest benyttede, er BLASTn og BLASTp. I øvelserne, der omhandler BLAST, arbejdes der med BLASTp.

I figur 5 kan søgesiden for BLAST ses. Den øverste markering viser felterne, hvor der ens sekvensdata indtastes, den midterste markering viser, hvor man kan vælge, hvilken database man vil blaste imod, samt hvilken organisme der skal søges i. De mest benyttede databaser er dem, der indeholder flest sekvenser. Disse databaser er følgende:

- nucleotide collection nr/nt for BLASTn, tBLASTn og tBLASTx .
- non-redundant protein sequences for BLASTp og BLASTx.

Man kan begrænse sin søgning ved at vælge databaser, der indeholder et begrænset antal sekvenser.

Den nederste markering i figur 5 viser, hvor man starter sin BLAST.

Punkt 6

Outputtet af en BLAST er opdelt i fire dele: *Descriptions, Graphic Summary, Alignments og Taxonomy*. De tre førsnævnte dele er mest brugbare, og gennemgås derfor.

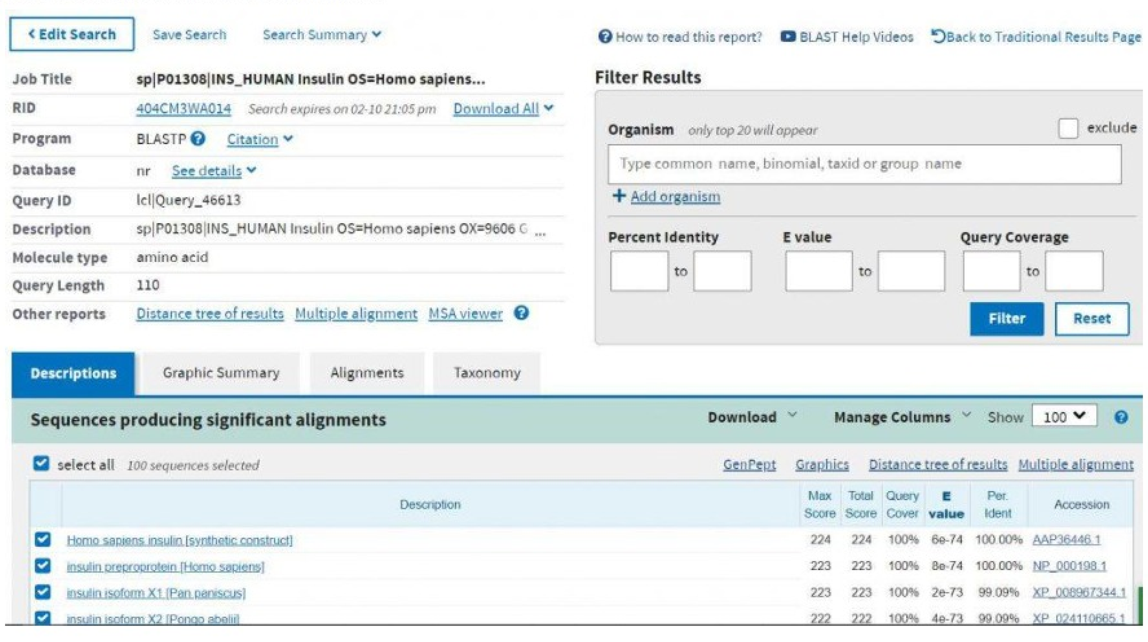
- ***Descriptions***
  Her er de bedste resultatsekvenser vist med accession-nummer, beskrivelse af proteinfunktion, hvor det stammer fra, max-score, total-score, procent lighed, e-værdi og links til databaser hvor sekvensen kan findes. Ved at trykke på accession-nummeret kan man komme til proteinets GenBank-side, hvor man kan finde flere informationer omkring proteinet. De vigtigste informationer om resultatsekvensen og matchet med query-sekvensen kan findes ved at se på e-værdien, max-scoren og procent ligheden, da disse tre størrelser beskriver, hvor identiske resultatsekvenserne er med query-sekvensen. En resultatsekvens kan antages at være et korrekt match til query-sekvensen, hvis e-værdien er under 10ˆ-4 – 10ˆ-5 (jo lavere, desto bedre).
- ***Graphic Summary***
  Øverst kan man se hele query-sekvensen og dens længde i aminosyrer (AA). Under query-sekvensen findes området med superfamilier (eng. *superfamilies*). Her angives, hvor på query-sekvensen der er, og om disse tilhører en bestemt superfamilie af proteiner. I den næste del af *Graphic Summary* kan man se, hvilke områder af resultatsekvenserne der matcher query-sekvensen, samt hvor de matchende områder er. Hver linje er en resultatsekvens, og de er listet i den rækkefølge, som sekvenserne også er listet i i Descriptions. Farverne for hver resultatsekvens (sort, blå, grøn, lilla og rød) indikerer, hvad max-scoren for resultatsekvensen er. Længden af linjerne indikerer, hvilke områder i query- og resultatsekvensen der er identiske.
- ***Alignments***
  Her vises sekvenssammenligningen mellem de enkelte resultatsekvenser og query-sekvensen.
  I hver sekvenssammenligning vises, hvilke steder sekvenserne er både identiske og forskellige. Linjen mellem input sekvensen (query) og hittet (subject) viser sammenligningen (alignment). Et "+" indikerer at aminosyren minder om hinanden mht. fysiske og kemiske egenskaber, f.eks. hvis de begge har sidekæder med carboxylsyrer, og et mellemrum indikerer, at sekvenserne er forskellige på denne position. Det anbefales altid at kigge på sekvenssammeligningerne, da man således kan se, hvor sekvenserne er identiske. En resultatsekvens kan endvidere undersøges nærmere ved at trykke på accession-nummeret, hvilket vil føre til en nærmere beskrivelse af sekvensen NCBI.

På baggrund af viden om en god e-værdi samt sammenligning af max-score og procentuel lighed, kan det bestemmes, hvilken/hvilke resultatsekvense(r), der matcher query-sekvensen bedst. Hvis man opnår resultater med en e-værdi, der er lavere en grænseværdien på 10ˆ-4 – 10ˆ-5, kan man konkludere, at query-sekvensen har samme funktion som resultatsekvensen med den lave e-værdi. Alle resultat-sekvenser vil altid kunne undersøges nærmere ved at trykke på det tilhørende accession-numme

Eksempel

Nedenfor vises resultatet for en BLASTp, hvor aminosyresekvensen for menneskeligt insulin er blevet brugt som query.

```
>sp|P01308|INS_HUMAN Insulin OS=Homo sapiens OX=9606 GN=INS PE=1 SV=1
MALWMRLLPLLALLALWGPDPAAAFVNQHLCGSHLVEALYLVCGERGFFYTPKTRREAED
LQVGQVELGGGPGAGSLQPLALEGSLQKRGIVEQCCTSICSLYQLENYCN
```



« Back to Glossary Index

Biotech Academy

Søltofts Plads, Bygning 227/041
2800 Kongens Lyngby
biotech@bio.dtu.dk

Sitemap

Eksterne ressourcer

Biosensor

Det Virtuelle Laboratorium

Om Biotech Academy

Biotech Academy tilbyder gratis undervisningsprojekter til gymnasier og grundskoler. Alle projekter involverer varierende undervisningsformer, og har et stort fokus på praktisk arbejde. Organisationen har et tæt samarbejde med forskere, lærere og elever under udviklingen af ethvert projekt.

Dansk