

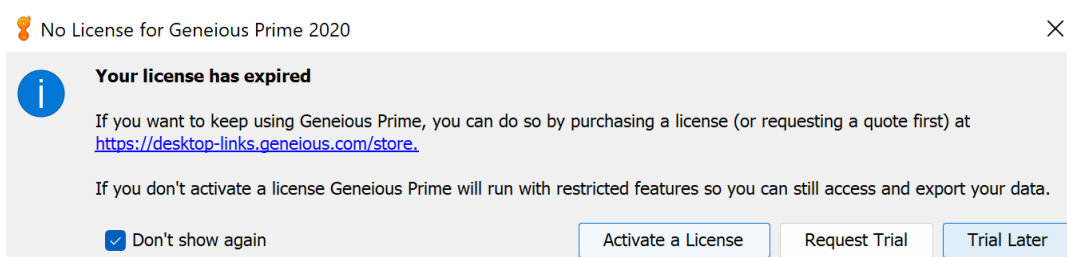


Geneious - en uddybende manual

– til bogen Bioinformatik i praksis

Installation

Programmet hentes ved at google Geneious Prime download eller ved at gå ind på adressen <https://www.geneious.com/updates>. Efter download og installation af Geneious Prime (bare vælg standardinstallation ved at klikke **Next** til alle bokse) skal man vælge at køre gratisversionen, og det gøres ved at klikke på **Trial Later** og sætte flueben i **Don't show again**, se figur 1. Så starter Geneious Prime op i en 'Restricted version', som fint dækker de behov skolebrug har.



Figur 1. Ved installation skal man vælge **Trial Later** og klikke på **Don't show again**.

Når installationen er gennemført, bliver man bedt om at angive sin mailadresse – det kan man gøre eller man kan skrive noget der ligner en e-mailadresse. Skriver man sin egen eller sin skolemail, får man beskeder om opdateringer af Geneious. Bliver man træt af at modtage beskeder, kan man blokere afsender.

Programmet

Fordelen ved at bruge Geneious Prime til bioinformatik og grunden til, at det anvendes gennem bogen **Bioinformatik i praksis** er, at man kan tilgå sekvenser som ligger på servere inde fra programmet og man kan nemt med en mappestruktur danne sig overblik over de bioinformatiske projekter, man arbejder med. Derudover indeholder programmet mange af de egenskaber man kan forvente af et bioinformatikprogram. Man kan:

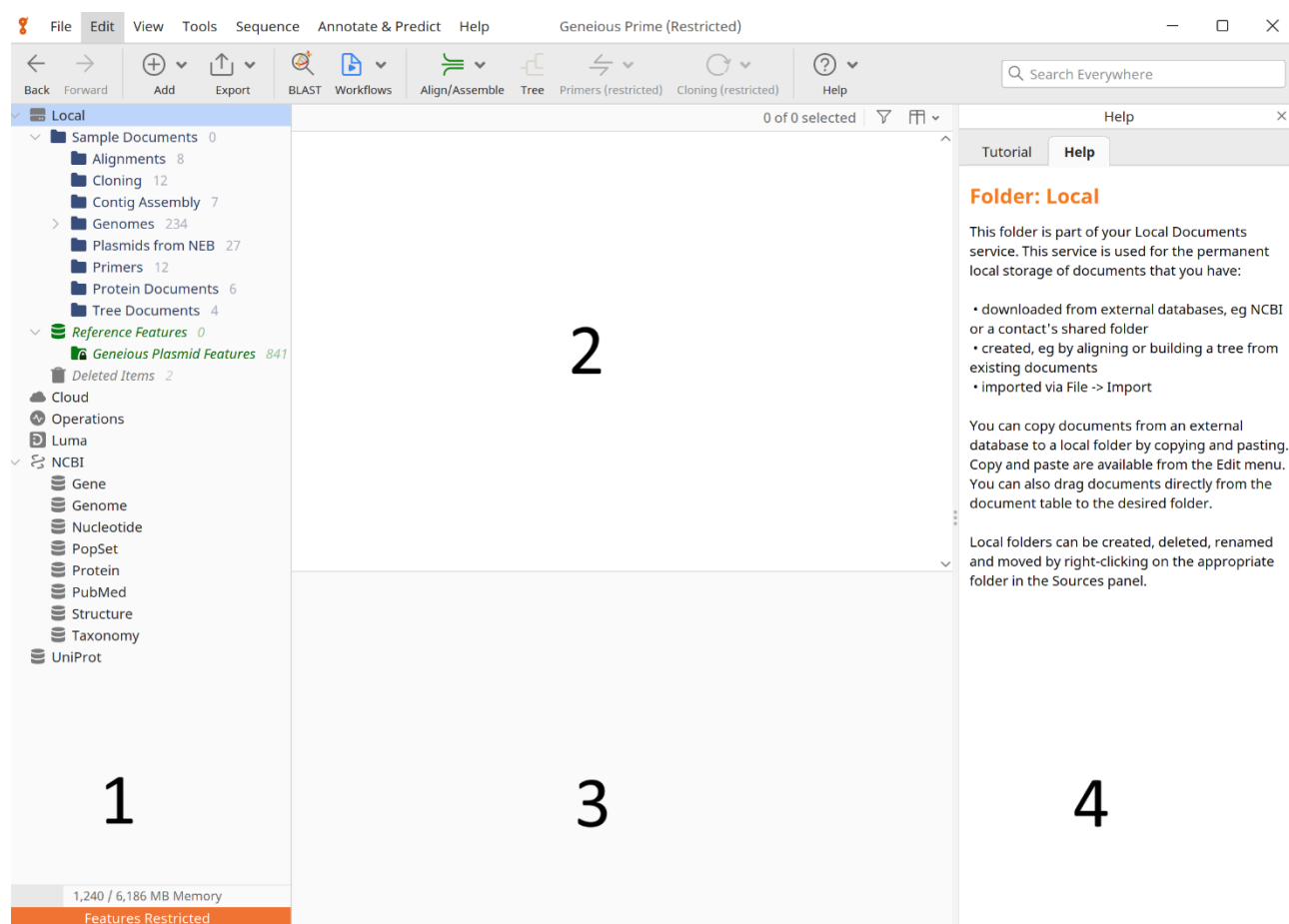
- fremstille alignments
- fremstille stamtræer
- vise strukturfiler af proteiner og arbejde med dem grafisk
- udføre BLAST fra programmet
- vise isoelektrisk punkt for proteiner
- føre almindelig, deskriptiv statistik over sekvenser
- have adgang til plugins (mest i en trial-version)
- kunne importere og eksportere filer fra og til andre programmer.

Når programmet startes op første gang, ser man en skærm der består af fire enheder. På figur 2 kan man se de fire felter. Felt 4 er et felt der viser hjælp og forklaringer til programmet. Det er lidt intelligent så det forklarer om der, hvor man har stillet sig ved at klikke på noget i programmet. I felt 1 har man øverst en mulighed for at lægge sekvenser og proteinstrukturer ind i mappen under **Local**. Man kan selv oprette



undermapper, så det svarer til hvad man arbejder med. Under ens egne filer er der nogle sekvenser og strukturer som programmet selv har installeret – såkaldte **Sample Documents**.

Nederst er der ikoner af servere. **NCBI** (National Center for Biotechnology Information) er en server, hvor man kan søge på mange forskellige typer sekvenser og strukturfiler – den er oftest anvendt som kilde til bioinformatik. **UniProt** er ligeledes en server, men den er specialiseret i proteinsekvenser og proteinstruktur.



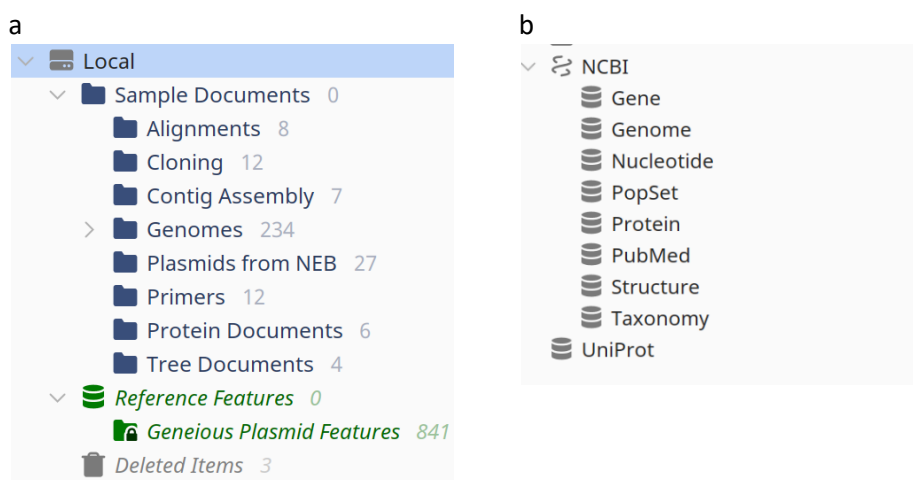
Figur 2. Programmet er opbygget af en række felter, der her er nummereret fra 1 til 4. Hvad de enkelte felter bruges til forklares i teksten.

Felt 2 i figur 2 har to funktioner. Hvis man klikker på en mappe under **Local** eller under **Sample Documents** i panel 1, vises indholdet af mappen i felt 2. Klikker man på en fil i felt 2 vises den i felt 3. Når man klikker på serverne NCBI og de tilhørende servere under NCBI eller på UniProt-serveren vil der i felt 2 komme en søgelinje øverst. Når man søger på noget ved hjælp af tekst og fx sekvenslængder kan man se en liste af søgeresultaterne i felt 2. For at få vist et søgeresultat skal man klikke på det, og så bliver det vist i felt 3. Når man har fundet det man søger efter, kan man fra felt 2 med musen trække filen over i mappen **Local** eller i en undermappe man selv laver (anbefales), og så har man sekvensen eller strukturen liggende på sin computer og kan arbejde med den.



Søgning

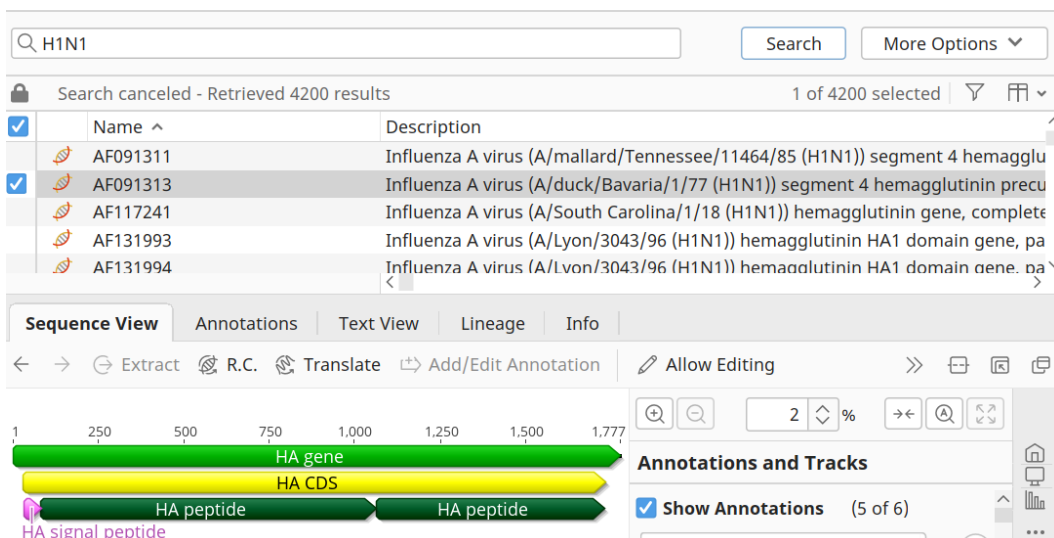
Ser man nærmere på venstre panel (1 i figur 2) er opdelingen vist i figur 3. Den øverste del (figur 3a) er den lokale del med de filer man har liggende på sin computer. Når man lukker programmet, vedbliver de lokale filer med at ligge der, hvor man har placeret dem. Man skal altså ikke trykke **Gem** eller lignende for at bevare det arbejde man har lavet. Programmet gemmer automatisk de lokale filer. Sletter man en lokal fil kan den gendannes fra mappen **Deleted Items**.



Figur 3.

- a. De lokale filer og de indlagte eksempelfiler i **Sample Documents**.
b. NCBI-serverens muligheder for download og nederst UniProt-serveren.

Figur 3b viser de søgemuligheder der er under NCBI-serveren og UniProt-serveren. Ved at klikke på en server under NCBI eller UniProt, kan man finde sekvenser eller strukturer. Hvordan søgningerne foregår, er vist i bogen **Bioinformatik i praksis**. Figur 4 viser en søgning på H1N1. Søgningen foregår ved, at man klikker på **Nucleotide** under NCBI og i søgefeltet skriver H1N1. Man kan se, at man øverst får vist søgeresultaterne, og nedenunder får man vist det valgte søgeresultat.



Figur 4. Ved søgning på H1N1 (influenzavirus) findes mange resultater. Det markerede resultat vises nedenfor søgeresultatet.



Nogle gange er de filer man har søgt på meget store. Når det er tilfældet, viser programmet ikke filens udseende, men en opsummering og en knap der gør det muligt at downloade filen, se figur 5. Det gør at man ikke skal vente så lang tid på søgningen og at man ud fra beskrivelsen af filen selv vælger, hvilken fil man gerne vil downloade.

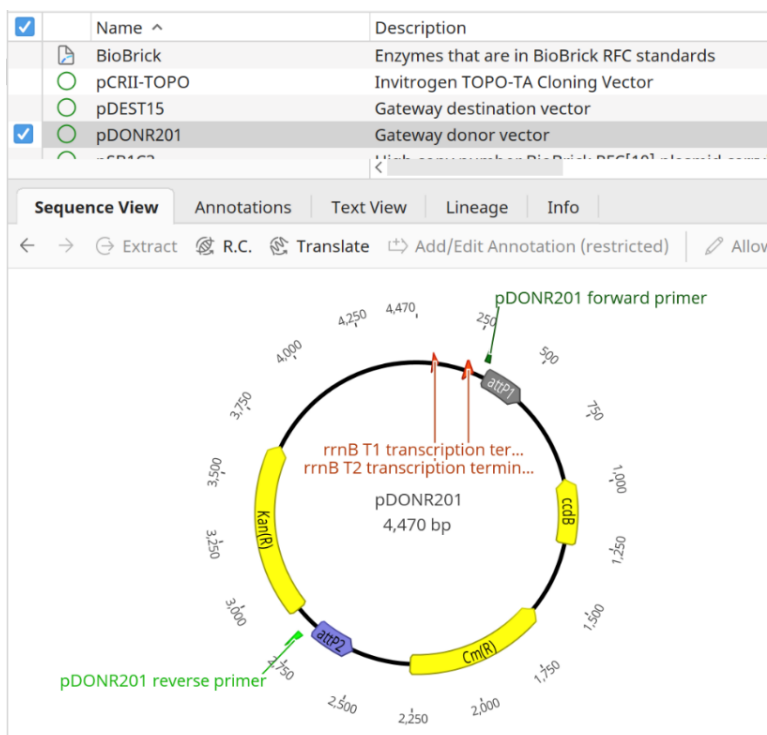


Figur 5. Download af store filer sker først efter et ekstra klik på **Download**.

Specielt søgninger på strukturfiler kan af forskellige årsager give anledning til problemer; det gennemgås til sidst i denne vejledning under overskriften Geneious og pdb-strukturer.

Visning af lokal fil

På samme måde som ved søgning, får man ved klik på en mappe under **Local** vist indholdet af mappen i felt 2 (figur 2) og klikker man på en fil i felt 2 vises den i felt 3, som vist på figur 2. Figur 6 viser hvordan man ved at klikke på **Sample Documents** → **Cloning** → **pDONR201** får vist et plasmid bestående af 4470 basepar.

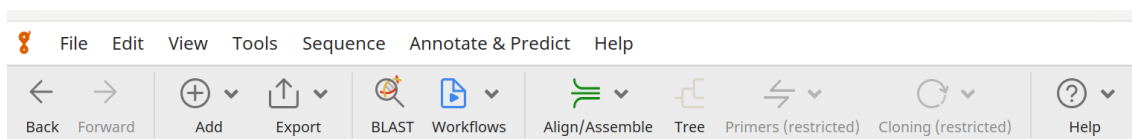


Figur 6. Visning af filer i Geneious Prime.



Funktioner i Geneious Prime

Geneious Prime har mange funktioner som kan tilgås fra enten menulinjen eller værktøjslinjen lige under menulinjen, se figur 7. Værktøjslinjen er egentlig bare en hurtig genvej til de mest anvendte funktioner som også findes i menulinjen. Den helt praksisnære del af funktionerne er beskrevet i [Bioinformatik i praksis](#), så de funktioner der beskrives her, bliver med fokus på at uddybe enkelte funktioner.



Figur 7. Menulinjen øverst og værktøjslinjen derunder.

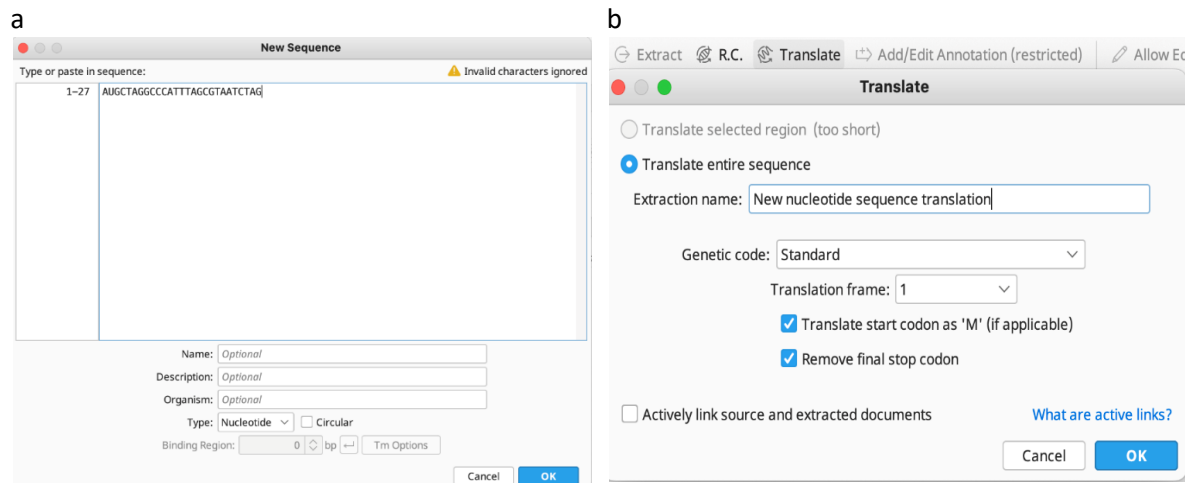
Alignments

Alignments bruges når man laver fylogenetiske stamtræer og til sammenligning af sekvenser. Alignments kan foretages når man har markeret to eller flere sekvenser. Det giver kun mening, at aligne homologe sekvenser. Når et alignment er foretaget, skal alle sekvenser der indgår i alignmentet overlappe hinanden – ellers er der sekvenser med der ikke er homologe. Alignments bruges til at sætte sekvenser op så de har mindst mulig forskel ved at sætte streger ind i sekvenserne, hvor der er sket deletioner eller insertioner af baser. Ved alignment af RNA- eller DNA-sekvenser tælles forskelle, og der hvor der er færrest forskelle, har man det bedste alignment. Det er en nem algoritme som en computer sagtens kan lave. Er man interesseret i at lave et alignment af et kortere stykke af sekvensen fx hen over et område med mutationer kan det være en god ide at udtrække det pågældende område så alignmentet ikke styres af, hvad der sker andre steder langt væk fra det interessante område i samme sekvens. Når man skal udtrække et bestemt område, gøres det ved at markere området og derefter vælge **Extract**.

Alignments af aminosyresekvenser gøres ud fra en BLOSUM62-matrix, se Bioinformatik i praksis side 13, hvor der tages hensyn til, om ændringerne i aminosyrerne er store, fx at en lille upolær aminosyre udskiftes med en stor polær aminosyre eller en negativt ladet aminosyre udskiftes med en uladet eller en positiv ladet aminosyre. Sådanne ændringer opfattes som mere usandsynlige end, hvis fx en lille upolær aminosyre udskiftes med en anden lille upolær aminosyre, fx alanin og valin, hvor der kun er en methylgruppe til forskel.

Oversætte sekvens til protein

I visse skriftlige opgaver kan man blive bedt om at oversætte en sekvens til protein. Man kan via **File → New → Sequence** skrive sekvensen ind og navngive den under **Name**. Man kan også bare lade alle felter stå tomme og arbejde videre med sekvensen. Når sekvensen er indtastet, se figur 8a kan man oversætte sekvensen til protein ved at klikke på sekvensen og derefter trykke på **Translate**, som vist i figur 8b.



Figur 8.

- a. Indtastning af sekvens manuelt.
b. Ved at klikke på **Translate** øverst i felt 3 (figur 2) kommer en oversætningsboks op. Der er forskellige kodemønstre. Standard er den der bruges til bl.a. mennesker og den der undervises i. Klik på **OK**.

Når man zoomer ud til 200 % eller mere kommer aminosyrerne til at stå med deres trebogstavsforkortelse. Da sidste codon er et stopcodon vises det ikke, se figur 9. Stopcodons vises i sekvenser ellers som en asterisk (*).



Figur 9. De otte aminosyrer som sekvensen koder for. Zoom til 200 % og de vises som trebogstavsforkortelse.

BLAST

BLAST er en søgemaskine, hvor man bruger en sekvens til at finde ud af hvor den kommer fra. Helt basalt vil det sige, at hvis man har en sekvens, kan man finde ud af, hvilken organisme eller virus den kommer fra, og hvor i genomet sekvensen stammer fra. Søgningen er en søgning i databaser, og det tager tid at søge i databaser. Modsat Google vises ikke kun topresultaterne andre har valgt var vigtige, men alle resultater. Resultaterne sorteres med en **Bit-score** som bør være høj, og en **E-value** som skal være lav. Det er også vigtigt at se på, hvor stor dækningsgraden er (**Grade**). Jo højere dækningsgrad, des flere baser er der med i søgeresultatet.

Når man foretager en BLAST-søgning kan det gøres på to måder fra Geneious. Den ene måde er inde i programmet, og den anden er via BLAST-websiden. Man kan søge på den sekvens man har markeret i Geneious Prime eller man kan søge på uformateret tekst, hvor man bare kopierer en sekvens ind som tekst i programmet.

I det følgende gennemgås, hvordan BLAST fungerer, for det kan drille i en undervisningssituation, hvor mange er på samtidig. Er BLAST-serveren svær at arbejde med, kan eleverne bruge deres mobiltelefon som mobil hotspot. Eksemplet der er valgt at vise, er hvor man via en aminosyresekvens kan finde en strukturfil via Protein Data Bank-sekvenserne (PDB). Man kan søge på andre sekvenstyper, og så er det andre Databaser man skal søge i.

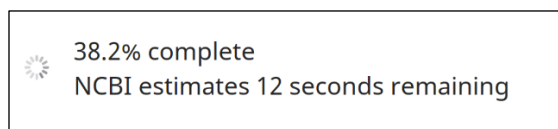


BLAST via Geneious Prime

Første eksempel er, at man søger på NCBI via Geneious. Det er det nemmeste, og fungerer det, er det fint og bør anvendes. Man klikker på en sekvensfil man har liggende. Her er valgt AAB53374. Nogle gange skal man dog ud at søge via BLAST-websiden, fordi det ikke umiddelbart virker inde fra programmet. Det forklares herefter. BLAST-søgning via Geneious sker ved ud for **BLAST Location** at klikke på **NCBI via Geneious**, som vist i figur 10. Læg mærke til, at databasen er sat til **Protein Data Bank (pdb)**. Herefter klikkes på **OK**.

Figur 10. BLAST-søgning via Geneious Prime.

Det viser sig, at søgningen tager tid. Der kommer et estimat op som kan være mere eller mindre korrekt. Hvis ikke søgningen når frem til et resultat i løbet af den estimerede tid fortsætter søgningen med et nyt tidsestimat, se figur 11.



Figur 11. At søgningen er i gang, indikeres med et tidsestimat.

Når resultatet kommer frem, ses det som tabel. I figur 12 er kun øverste resultat vist. Man kan på v-tegnet ud for Bit-Score se, at søgeresultaterne er vist med højest Bit-Score øverst. Det er vigtigt, at Bit-Scoren er høj og E-value lav – den er 0 i dette tilfælde, hvilket er flot. VIGTIGT: Man skal også tjekke beskrivelsen af proteinet for at se, om man rent faktisk har fat i det man regner med at finde. Det ses, at PDB-strukturfilen med navnet 7TLT er den afbildning af proteinstrukturen den passer bedst med den sekvens forespørgslen drejede sig om.

Hit Table				
Query Centric View				
Annotations				
Distances				
Info				
Hit Table				
✓	Name	E Value	Bit-Score ▾	Grade
✓	7TLT_A	0	736.873	98.6%

Figur 12. Tabel over søgeresultater i en BLAST-søgning.



BLAST via Webside

Man klikker igen på sekvensfilen – her AAB53374. BLAST-søgning via BLAST-webside sker ved ud for **BLAST Location** at klikke på **NCBI via Website**, se figur 10. Når man gør det, føres man over på en webside og den sekvens man har klikket på, skrives ind i websidens inputdataboks under **Enter accession number(s), gi(s), or FASTA sequence(s)**, som vist i figur 13. Der vælges også den rigtige BLAST-søgning, som i dette tilfælde er en proteinsøgning. Det er derfor **blastp** er valgt (en aminosyresekvens mod en strukturfil). **blastn** er for søgning på nucleotider. **blastx** oversætter nucleotidesekvensen til en aminosyresekvens og sammenligner den med protein databasen. Det der ikke sker ved søgningen er, at man ikke automatisk søger på Protein Data Bank (pdb) under database selvom, man har indtastet det i søgeboksen. Det skal man altså selv sætte korrekt op, som vist i figur 13. Når man har sat den rigtige database, trykkes på **BLAST**.

The screenshot shows the NCBI BLAST web interface. At the top, there are tabs for different BLAST programs: blastn, **blastp**, blastx, tblastn, and tblastx. The 'blastp' tab is active. Below the tabs, there's a section 'Enter Query Sequence'. It has a text input field for 'Enter accession number(s), gi(s), or FASTA sequence(s)' with a 'Clear' button. The input field contains the text '>AAB53374' followed by a multi-line FASTA sequence. To the right of this is a 'Query subrange' section with 'From' and 'To' input fields. Below the input field, there's a section 'Or, upload file' with a 'Vælg fil' button and a message 'Der er ikke valgt nogen fil'. There's also a 'Job Title' input field with a placeholder 'Enter a descriptive title for your BLAST search'. A checkbox 'Align two or more sequences' is also present. Below this is the 'Choose Search Set' section. It has a 'Databases' section with 'Standard databases (nr etc.):' selected and 'Experimental databases' unselected. There's a 'Compare' section with a checkbox 'Select to compare standard and experimental database'. Below this is a 'Standard' section with a 'Database' dropdown menu. The dropdown menu is open, showing a list of databases: 'Non-redundant protein sequences (nr)' (selected), 'RefSeq Select proteins (refseq_select)', 'Reference proteins (refseq_protein)', 'Model Organisms (landmark)', 'UniProtKB/Swiss-Prot (swissprot)', 'Patented protein sequences (patat)', 'Protein Data Bank proteins (pdb)', 'Metagenomic proteins (env_nr)', and 'Transcriptome Shotgun Assembly proteins (tsa_nr)'. Below the dropdown menu is a 'Program Selection' section with a 'Algorithm' dropdown menu. The dropdown menu is open, showing a list of algorithms: 'blastp (protein-protein BLAST)' (selected), 'PSI-BLAST (Position-Specific Iterated BLAST)', 'PHI-BLAST (Pattern Hit Initiated BLAST)', and 'DELTA-BLAST (Domain Enhanced Lookup Time Accelerated BLAST)'. Below the dropdown menu is a 'BLAST' button. To the right of the button is a checkbox 'Show results in a new window'.

Figur 13. BLAST-søgning via webide.

Ligesom når man har lavet en BLAST-søgning via Geneious kommer resultatet frem som en tabel. Denne gang på en webside, men resultatet er det samme. Det er stadigvæk 7TLT der er den mest veltilligende strukturfil, se figur 14.



Descriptions	Graphic Summary	Alignments	Taxonomy					
Sequences producing significant alignments								
		Download	Select columns					
		Show	100					
☒ select all 100 sequences selected								
GenPept Graphics Distance tree of results Multiple alignment MSA Viewer								
Description	Scientific Name	Max Score	Total Score	Query Cover	E value	Per. Ident	Acc. Len	Accession
☒ Chain A, HLA class I histocompatibility antigen, A alpha chain [Homo sapiens]	Homo sapiens	736	736	100%	0.0	97.26%	365	Z1LT_A
☒ Chain A, HLA class I antigen [Homo sapiens]	Homo sapiens	700	700	100%	0.0	92.88%	365	7RTD_A
☒ Chain A, HLA class I histocompatibility antigen [Homo sapiens]	Homo sapiens	677	677	100%	0.0	92.05%	365	8RBU_A

Figur 14. Resultatet af BLAST-søgningen ses som en tabel på websiden.

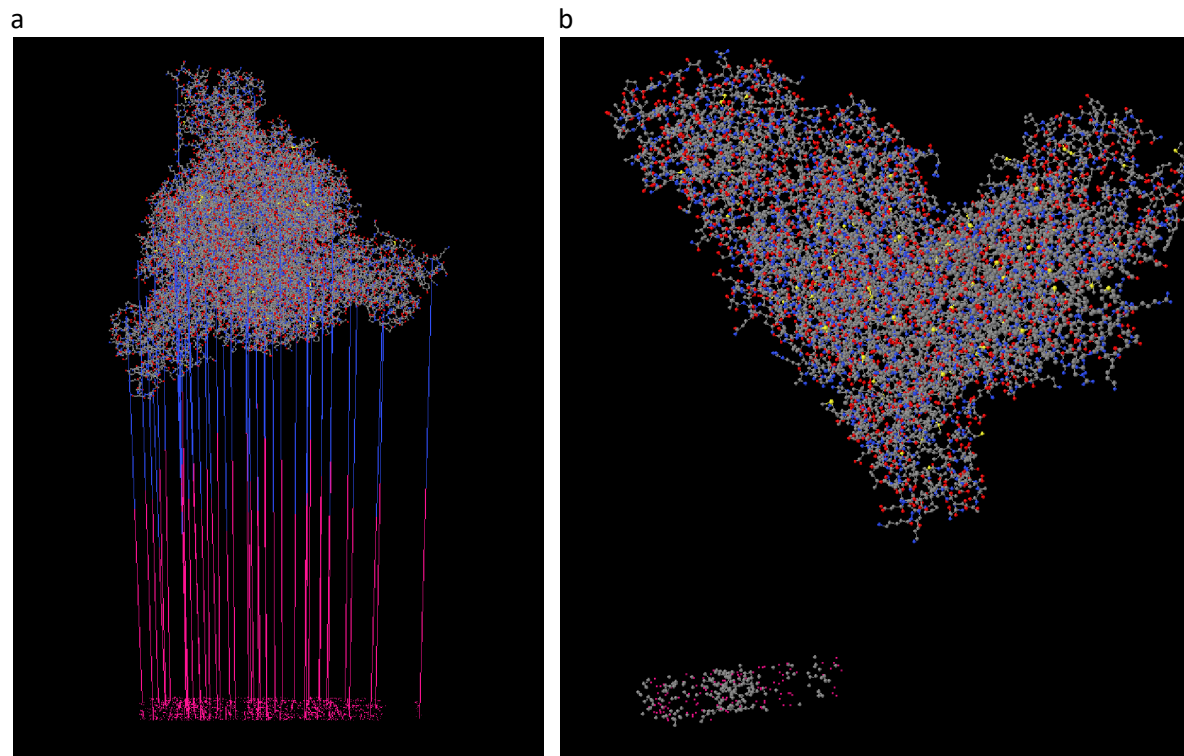
Geneious og pdb-strukturer

Strukturfiler og import af disse er ikke altid uden komplikationer. Det er muligt, at man kan få en fejl der gør, at man ikke inde fra Geneious kan downloade strukturfiler. Ofte kommer der en fejl op om, at der ikke er en slutning på sekvensen. Det er en fejl som normalt rettes ved næste opdatering og skyldes, at der er en ændring i den måde man skal forespørge om strukturfiler på. Derfor kan man direkte fra Protein Data Banks hjemmeside downloade filerne og trække dem ind i programmet. Protein Data Banks webside er <https://www.rcsb.org/>. Man søger med firetegnskombinationen oppe i søgefeltet til højre, som vist i figur 15.



Figur 15. Søgning på strukturfiler inde fra Protein Data Banks webside.

Derudover er der en fejl i visningen af visse strukturfiler i Geneious Prime. For visse strukturfiler har ligander og glykosileringer en fejl i deres koordinater ved indlæsning, hvorved de ikke placeres korrekt i forhold til proteinet. De kommer til at ligge som en bunke ved siden af proteinstrukturen, som vist i figur 16a, der viser SARS-CoV-2 spikeprotein 6XR8, som bruges i bogen **Bioinformatik i praksis**. De røde og blå streger er bindinger ned til glykosileringerne. Et andet eksempel ses i figur 16b der viser en μ -receptor med noget mærkeligt noget til venstre for nederen, men uden bindinger til strukturer.



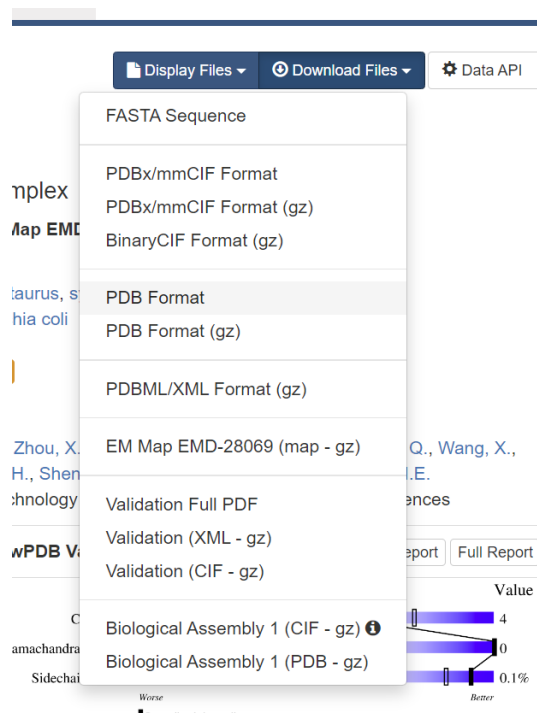
Figur 16. a. SARS-CoV-2 spikeproteinstruktur (6XR8). b. μ -receptor med noget neden for til venstre (8EF6).

Workaround

Der er to måder man kan undgå disse fejl på. Den ene er at man downloader via Protein Data Banks webside (pdb) på <https://www.rcsb.org/structure/xxxx>, hvor xxxx står for firetegnskoden, som pdb-filen har. For figur 16b er adressen til figuren <https://www.rcsb.org/structure/8EF6>. For SARS-CoV-2 spikeproteinstruktur (6XR8) er problemet, at filen er for stor til at downloade, og derfor skal den hentes via ChimeraX.

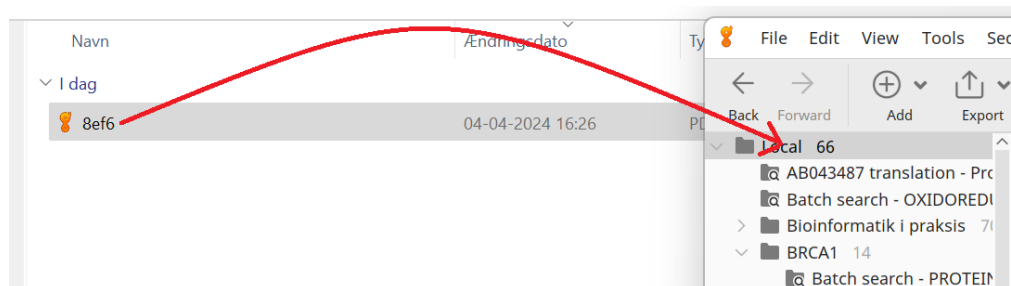
Hente fra websiden PDB

Til at begynde med kan man for de fleste filers vedkommende gøre følgende på websiden <https://www.rcsb.org/structure/> - her vist med 8EF6: Klik på **Download Files** og vælg **PDB-format**, som vist i figur 17.



Figur 17. Download af pdb-fil fra RCSB PDB's hjemmeside.

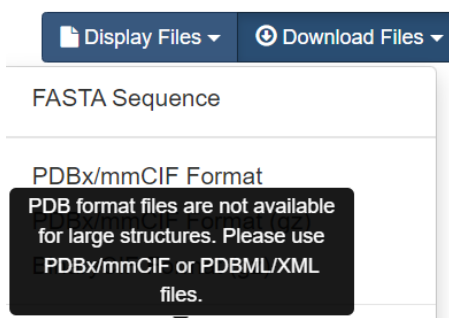
Filen gemmes i **Overførsler**. Herfra trækkes filen nu ind i Geneious Prime, som vist på figur 18. Man kan også trække filen over i anden mappe end **Local**. Den skal bare oprettes i Geneious Prime først.



Figur 18. Træk filen over i Geneious Prime.

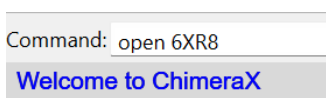
Brug ChimeraX til at hente filen

Hvis pdb-filen er for stor er det noget mere besværligt - også fordi man mister lidt information undervejs, hvis ikke man er opmærksom. Store filer kan nemlig ikke downloades fra RCSB PDB's hjemmeside. Der vil blive vist en fejl som i vist i figur 19. Eksemplet baserer sig på SARS-CoV-2 spikeprotein (6XR8). Løsningen er at hente filen via et andet program og gemme den som en pdb-fil, man kan importere i Geneious Prime.



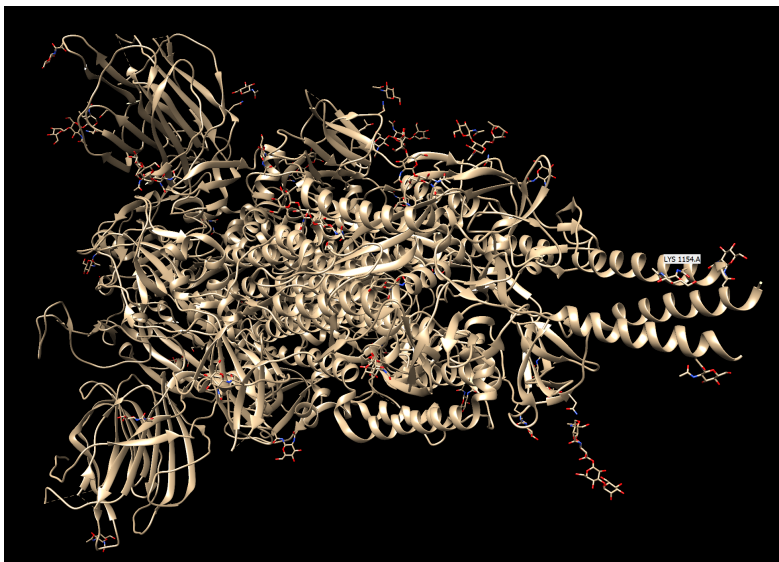
Figur 19. Når filen er for stor, kan den ikke downloades som en PDB-fil fra websiden. Denne fejl dukker op.

Åbn derfor ChimeraX version 1.8 eller højere (se vejledning om installation af ChimeraX på nucleus.dk under det supplerende materiale til Bioinformatik i praksis. Nederst i kommandolinjen skrives `open 6XR8` og der afsluttes med at trykke på Enter, se figur 20.



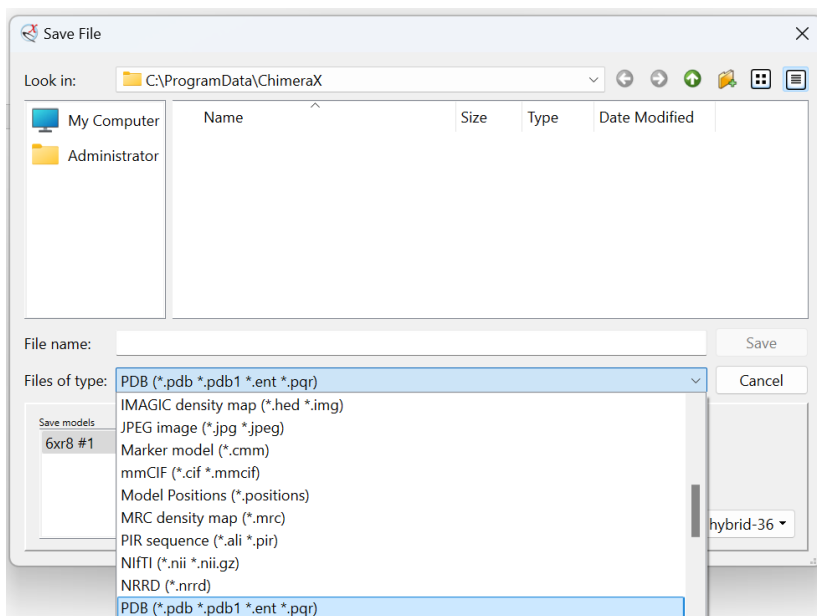
Figur 20. Åbn en strukturfil i ChimeraX via kommandolinjen.

Strukturfilen indlæses som vist på figur 21.



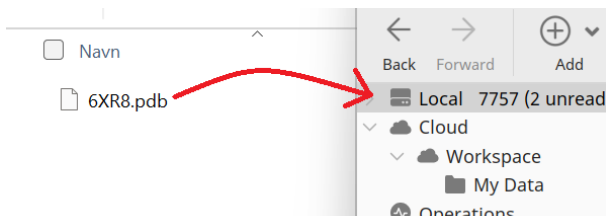
Figur 21. Strukturfil vist i ChimeraX

Nu skal filen gemmes. Vælg **File → Save**. En dialogboks, som vist på figur 22 dukker op og man kan gemme filen som en PDB-fil ved at vælge **Files of type** til at være PDB (*.pdb *.pdb1 *.ent *.pqr). Man kan ligeledes gemme filen et sted, hvor man kan finde den igen ved at dobbeltklikke på **My Computer** og finde mappen for fx Desktop/Skrivebord eller Downloads/Overførsler. Når man gemmer filen, bør man give den det navn pdb-serveren har angivet. I dette tilfælde 6XR8.



Figur 22. Gem PDB-fil i ChimeraX.

Åbn den mappe filen er i og træk den ind i Geneious Prime. Sørg for, at hverken mappen eller Geneious Prime fylder hele skærmen, se figur 23.



Figur 23. At flytte filer ind i Geneious prime sker nemmest ved bare at trække dem ind udefra.

Filen mangler nu et navn og en beskrivelse, som vist i figur 24.

☒	Name	Description ^
☒	-	-

Figur 24. Manglende navn og beskrivelse efter import af fil.

Vil man løse dette, kan man i Geneious Prime hente 6XR8 ind på sædvanlig vis **NCBI** → **Structure** og søge på 6XR8 og trække den ind i samme mappe, som den importerede fil ligger i. Det kunne se ud, som vist i figur 25.

☒	Name	Description ^	Modified	# Sequences	Max Seque...	Min Sequen...	PDB name	Path (Impor...	Filename (I...
☒	-	-	04 Apr 2024 6:45 pm	3	1,107	1,107	-	D:\Desktop	6XR8.pdb
☒	VIRAL PROTEIN (6XR8)	Distinct conformational states of SARS-CoV-2 spike prot...	11 Jul 2020	3	1,310	1,310	6XR8	-	-

Figur 25. Den importerede korrekte strukturfil uden navn og den forkerte strukturfil med navn.

Klik først på filen med det rigtige navn, men den forkerte struktur. Klik så igen på navnet (VIRAL PROTEIN (6XR8)), hvorefter det bliver markeret, så man kan kopiere det. Tryk på **Ctrl+C** eller **CMD+C**. Klik nu på filen uden navn. Klik igen på navnet og når strengen blinker så tryk på **Ctrl+V** eller **CMD+V** for at sætte navnet ind. Gentag for **Description** og også for **PDB name**. Slet derefter den forkerte strukturfil.