# HIV i retslokalet -sådan finder man data i Genbank

*Af: Kresten Cæsar Torp, Aalborghus Gymnasium Materialet inddrager dele af: Regassa, Laura; Naowarra Cheeptham og Michèle Shuster: Murder by HIV?, NSTA, Link: <u>Murder by HIV? Grades 9-12 Edition | NSTA</u>* 

# Oprettelse af bibliotek

Første skridt er at finde de sekvenser som skal sammenlignes. Dem samler man i en fil, en form for bibliotek. De skal have et format, som kan genkendes af programmet, når de skal sammenlignes. Selve sammenligningen kaldes et alignment, og man kan ud fra det konstruere et fylogentisk træ. Det kan du arbejde videre med i filen: *HIV i retsgenetik -alignment og fylogenetisk træ*.

- 1. Gå ind på NCBI's database Genbank. Link: <u>National Center for Biotechnology Information</u> (<u>nih.gov</u>)
- 2. Vælg "Nucleotide" i menuen til højre.

🗖 📔 Bioinformatik - Sysbio.dk	× National Center for Biotec	hnolog × +			- 0	$\times$
← C 🗘 https://www.n	cbi.nlm.nih.gov			^ & `	œ 🏽	
An official website of the United I	States government <u>Here's how you know s</u> ibrary of Medicine or Biotechnology Information	,		Lo Search	og in	Q +: 0
						+
NCBI Home	Welcome to NCBI			Popular Resources		
Resource List (A-Z)	The National Center for Biotechnol	ogy Information advances science and	d health by providing access to	PubMed		
All Resources	biomedical and genomic informatio	n.		Bookshelf		
Chemicals & Bioassays	About the NCBI   Mission   Organ	nization   NCBI News & Blog		PubMed Central		
Data & Software				BLAST		
DNA & RNA	Submit	Download	Learn	Nucleotide		
Domains & Structures	Deposit data or manuscripts	Transfer NCBI data to your	Find help documents, attend a	Genome		
Genes & Expression	into NCBI databases	computer	class or watch a tutorial	SNP		
Genetics & Medicine		_		Gene		
Genomes & Maps				Protein		
Homology	T			PubChem		
Literature						
Proteins				NCBI News & Blog		
Sequence Analysis				New annotations in RefSeq!		€
Taxonomy	Develop	Analyze	Research	14	Feb 2023	. 0.
https://www.ncbi.nlm.nih.gov/nucleotide/	Use NCBI APIs and code libraries to build applications	Identify an NCBI tool for your data analysis task	Explore NCBI research and collaborative projects	Eukaryotic Genome Annotation	·01	છુ
•	Q Søg	🔎 🔎 🖻 💽 🖬	i 📮 🖣 🧿 🐖 🦊	🔬 🔨 🔌 dan 👳 d	») <b>⊡</b> 17:38 ≫ 16-02-2023	8 3 4

- 3. Find den sekvens, vi vil sammenligne med de øvrige. Det gør man vha. en accession number, dvs. et ID, som sekvensen har i databasen. Kopier Accession Number: **AY156807**.
- 4. Indsæt nummeret i søgefeltet.

Bioinformatik - Sysbio.dk × N H	/nucleotide/	A* to		G   ·	ć @	٢	
An official website of the United States governmer	nt <u>Here's how you know</u> .∽					-	۹
NIH National Library of	f Medicine				Log in		+;
National Center for Biotechnolo	ogy information						0
Nucleotice AY	156807		Search		1		0
Adv	vanced				Hel	lp	
ACA CALCA -CAU	Nucleatide					100	т
TCTAGCACAC	ATTT The Nucleatide database is a solliestion	of acquiances from acused acurace, including Conf	Pank Doffa		and		
	ATTT CGCT PDB. Genome, gene and transcript sequ Nucleoside Table	of sequences from several sources, including Gent ence data provide the foundation for biomedical re	Bank, RefSe esearch and	q, TPA a discove	and ry.	ļ	
	ATTT CGCC PDB. Genome, gene and transcript sequences Nucleotide Tools	of sequences from several sources, including Gent lence data provide the foundation for biomedical re Other Resources	Bank, RefSe esearch and	q, TPA a discove	and ry.	ļ	
TGTAGCTTACCACAC GGTTTTGCTACCTCC Using Nucleotide Quick Start Guide	ATTT CGCC The Nucleotide database is a collection PDB. Genome, gene and transcript sequ Nucleotide Tools Submit to GenBank LiseCot	of sequences from several sources, including Gent lence data provide the foundation for biomedical re Other Resources GenBank Home DerSen Home	Bank, RefSe	q, TPA a	and ıry.	l	
TGTAGCTTACACAC GGTTTTGCTTACCTCC Using Nucleotide Auide Start Guide EAQ	The Nucleotide database is a collection PDB. Genome, gene and transcript sequ Nucleotide Tools Submit to GenBank LinkOut E-Hitlines	of sequences from several sources, including Gent lence data provide the foundation for biomedical re Other Resources GenBank Home RefSeq Homa Gene Homa	Bank, RefSe	q, TPA i discove	and ry.		
TGTAGCTTACCACAC GGTTTCCTTACCTCC Using Nucleotide Audick Start Guide EAQ Help GenBank FTP	The Nucleotide database is a collection PDB. Genome, gene and transcript sequence Nucleotide Tools Submit to GenBank LinkOut E-Utilities BLAST	of sequences from several sources, including Gent lence data provide the foundation for biomedical re Other Resources GenBank Home RefSeq Home Gene Home SRA Home	Bank, RefSe ssearch and	q, TPA a discove	and Iny.		
TGTAGCTTACACAC GGTTTGCTACCTCC Using Nucleotide Quick Start Guide EAQ Help GenBank FTP ReiSeo FTP	The Nucleotide database is a collection PDB. Genome, gene and transcript sequence Nucleotide Tools Submit to GenBank LinkOut E-Utilities BLAST Batch Entrez	of sequences from several sources, including Gent lence data provide the foundation for biomedical re Other Resources GenBank Home RefSeq Home Gene Home SRA Home INSDC	Bank, RefSe osearch and	ıq, TPA a	and ıry.		
Contraction of the second seco	The Nucleotide database is a collection PDB. Genome, gene and transcript sequences Nucleotide Tools Submit to GenBank LinkOut E-Utilities BLAST Batch Entrez	of sequences from several sources, including Gent lence data provide the foundation for biomedical re Other Resources GenBank Home Re/Saq Home Gene Home SRA.Home INSDC	Bank, RefSe	rq, TPA a discove	and ry.		*

## 5. Tryk "Search".

🖸   🗅 Bioinformatik - Sysbio.dk 🗙 🌓 HIV-1 clone V2.MIC.RT from USA 🗙 🕂		-	0	$\times$
← C	A" to	¢ f @		
An official website of the United States government Here's how you know.			-	0
NIH National Library of Medicine		Log in		+
				0
Nucleotide Nucleotide V	Search			
Advanced		Help	5	
CanBank Band In				+
Gendarik • Seria to: •	Change region shown	-		
HIV-1 clone V2.MIC.RT from USA reverse transcriptase (pol) gene, partial cds				
GenBank AY156807.1	Customize view	-	111	
FASTA graphics PopSet	Analyze this sequence		5	
<u>6010:</u> M	Run BLAST			
LOCUS AY156807 805 bp DNA linear VRL 22-0CT-2002	Pick Primers			
partial cds.	Highlight Sequence Feature	IS		
ACCESSION AY156807 VERSION AY156807.1	Find in this Sequence			
KEYWORDS . SOURCE Human immunodeficiency vicus 1 (HTV-1)				
ORGANISM Human immunodeficiency virus 1	Retrovirus Resource		Ĵ.	
viruses; kiboviria; Pararnavirae; Artverviricota; kevtraviricetes; Ortervirales; Retroviridae; Orthoretrovirinae; Lentivirus.	Retrieve, view, and downloa and protein sequences.	ad retroviral genomic		Ð
REFERENCE 1 (bases 1 to 805) AUTHORS Metzker,M.L., Mindell,D.P., Liu,X.M., Ptak,R.G., Gibbs,R.A. and			11	.0.
https://www.ncbi.nlm.nih.gov/nuccore/AY156807.1?report=fasta	Related information		<b>1</b> -	63
e 📑 🔁 🕵 🚛 🕞 💼	📕 🔬 🗛 🚺	DAN ବ⊄))⊡.	17:4	9 4
			0-02-202.	5

Nu får du en række oplysninger frem om sekvensen. Du skal imidlertid bruge selve DNA-sekvensen og du skal have den i et filformat, som programmet kan anvende senere. Filformatet er FASTA, som du finder lige under overskriften.

6. Tryk på "FASTA". Nu får du FASTA-filen frem. Den begynder med >, og du kan genkende nucleotid-sekvensen (beståede af A,T,G,C).



Første linje indeholder en række oplysninger om sekvensen. Dem vender vi lige med...

- 7. Marker kun selve DNA-sekvensen (GTAG......GTGC) og kopier den.
- 8. Gå tilbage til Genbanks hovedside.
- 9. Vælg BLAST (lige under "Nucleotide")



BLAST er et program som anvendes til at søge sekvenser i databasen, som ligner den sekvens, man indtaster (Query).

- 10. Tryk på Nucleotide BLAST.
- 11. Indsæt sekvensen fra før.

Computational Thinking på Biologiske Systemer

novo nordisk fonden

🔲   🕒 Bioinformat	ik - Sysbio.dk 🛛 🗙	S Nucleotide BL/	\ST: Search nucleo	× +								-	0	×
< C A	https://blast.ncbi.nlm	nih gov/Blast cgi?Pl	ROGRAM=blastn8	RPAGE TYPE=Blas	stSearch&UN	K LOC=blastbr	ome	D.	~		C2	= @		
	http://bidstifebi.ini		lo ora im- bidstire		CULINE DLA	<b>31</b>	onne	7.	10			- · U	•	1
blastn bla	astp blastx	tblastn	tblastx										^	٩
		BLA	STN programs sea	rch nucleotide data	abases using a	nucleotide que	ry. more				Reset page			+
Entra Origina	0										Bookmark			•
Enter Query	sequence	EASTA sequence(s	Clear	Query subran	ngo 🕗									
AGTGGGGATTTTT	CACACCAGACGAAA	ACACCAGAAAGAAG		rom										•
CGTTGGATGGGT CCATCCTGATAAA	FATGAACT TGGACAGTACAGCCT	ATAGTGC	-											+
				То										
Or, upload file	Vælg fil Der e	r ingen fil valgt	0											
Job Title	Estas a da aciatica		t. <b>A</b>											
Align two or m	ore sequences ?	TITLE FOR YOUR BLAST S	earch 😈											
Choose Sear	rch Set													
Database	Standard dat	abases (nr etc.): 〇	rRNA/ITS datab	ases 🔘 Genomi	iic + transcript	databases 〇	Betacoronavir	us						
	Nucleotide coll	ection (nr/nt)			✓ Ø									
Organism	Enter organism	name or idcomple	tions will be sug	gested	exclud	le Add organi	ism							
optional	Enter organism co	mmon name, binomial	, or tax id. Only 20	top taxa will be sho	own 😮									
Exclude	Models (XM/	XP) Uncultured/	environmental sa	mple sequences										
Limit to	Sequences f	rom type material												€
Entrez Query					You Tube	Create custom	database							ŝ
	Enter an Entrez qu	env to limit search ?	Q Søg	🔎 🛋	<b>e</b> 💼	<b>—</b>	o 🖷	<u>.</u>		~ 🔌	DAN 중	¢)) ⊡ 1	18:0! 6-02-202:	5 3 4

Under "Choose Search Set" vælg: "Standard database" og "Nucleotide collection (nr/nt)".
Under "Program Selection": Vælg "Somewhat similar"

Choose Sea	rch Sat
Onoose oca	
Database	● Standard databases (nr etc.): ○ rRNA/ITS databases ○ Genomic + transcript databases ○ Betacoronavirus
	Nucleotide collection (nr/nt)
Organism	
Optional	Enter organism name or idcompletions will be suggested <b>Add organism</b>
	Enter organism common name, binomial, or tax id. Only 20 top taxa will be shown 😮
Exclude	Models (XM/XP) Uncultured/environmental sample sequences
Optional	
Limit to	Sequences from type material
Optional	
Entrez Query	You Tube Create custom database
Optional	Enter an Entrez query to limit search 😧

- 14. Tryk: "BLAST". Det tager lidt tid af lave matchet i databasen. Husk at den indeholder millioner af sekvenser.
- 15. Scroll lidt ned. Her finder du en liste over de hits i databasen, som minder mest om den indtastede. Marker de første 8. De stammer fra retssagen:
  - Klon V1 og Klon V2 er virusvarianter fra offeret (victim)
  - Klonerne P1-P6 stammer fra patienten.
  - o "MIC.RT." angiver metoden: Revers Transkriptase
  - "(pol) gene" skyldes, at sekvensen kun er et udsnit af HIV-genomet, nemlig pol-genet.
- 16. Download filerne ved at trykke på "Download".
- 17. Vælg "FASTA (aligned sequence)".

🔲   🗅 Bioinformatik - Sysbio.dk 🛛 🗙 KBI Blast:Nucleotide Sequence 🗙 🕂	-	- 0	$\times$
← C	^` ☆ 📄 ৫   ৫ @	9	
Sequences producing significant alignments	Download ∽ Select colur ns ∽ Show 100 ✔ 3	^	۹
Select all 8 sequences selected	FASTA (complete sequence) FASTA (aligned sequences)	<u>er</u>	4
Description	ConPark (complete compare)     F     Per.     Acc.       Hit Table (text)     Ident     Len     Accession		0
HIV-1 clone V2.MIC.RT from USA reverse transcriptase (pol) gene, partial cds	Hit Table (CSV) 0.0 100.00% 805 AY156807.1		0
HIV-1 clone V1.MIC.RT from USA reverse transcriptase (pol).gene, partial cds	Ht Text 0.0 99.88% 805 AY156806.1		+
HIV-1 clone P6.MIC.RT from USA reverse transcriptase (pol).gene, partial cds	Lu Descriptions lable (CSV) 0.0 99.38% 805 AY156803.1		
HIV-1 clone P4.MIC.RT from USA reverse transcriptase (pol) gene, partial cds	ASN.1 0.0 99.13% 805 AY156801.1		
HIV-1 clone P5.MIC.RT from USA reverse transcriptase (pol) gene, partial cds	HL 0.0 98.26% 805 <u>AY156802.1</u>		
HIV-1 clone P3.MIC.RT from USA reverse transcriptase (pol) gene, partial cds	Human immun 1389 1389 100% 0.0 98.26% 805 AY156800.1		
HIV-1 clone P1.MIC.RT from USA reverse transcriptase (pol) gene, partial cds	Human immun 1380 1380 100% 0.0 98.01% 805 AY156797.1		
HIV-1 clone P2.MIC.RT from USA reverse transcriptase (pol) gene, partial cds	Human immun 1371 1371 100% 0.0 97.76% 805 AY156799.1		
HIV/1 isolate 5018-83 clone pbf4 from USA complete genome	Human.immun 1371 1371 100% 0.0 97.76% 9703 AY835777.1		
HIV-1 isolate PRRT_38 from USA pol protein gene, partial cds	Human immun 1370 1370 100% 0.0 97.39% 1497 KT167884.1		
HIV-1 isolate HIV1B_p6pol pol protein (pol) and gag.protein (gag).genes, partial cds	Human immun 1362 1362 100% 0.0 97.52% 1193 MF990830.1	*	
HIV-1 HAART patient isolate 180 from Italy pol protein (pol) gene, partial cds	Human immun 1362 1362 100% 0.0 97.52% 1129 AF251949.1	bacl	
HIV-1 isolate PRRT_273 from USA pol protein gene, partial cds	Human immun 1359 1359 99% 0.0 97.51% 1497 KT168119.1	Feed	
HIV-1 isolate SF20 from USA, partial genome	Human immun 1358 1358 100% 0.0 97.39% 8513 KJ704794.1		€
HIV-1 isolate NYC1 from USA, partial genome https://blast.ncbi.nlm.nih.gov/Blast.cgi#	Human immun 1358 1358 100% 0.0 97.39% 8505 <u>KJ704793.1</u>	<b>.</b>	ŝ
🕞 📑 🔍 Søg 📮 🖻 🖻 💽	i 📑 🐖 🌍 🐖 🦊 🤬 🛛 🗠 dan 👳 d) 🗗	18:18 16-02-2023	4

## 18. Gem filerne som en tekstfil (.txt) ved copy-paste. Word ol. Du'r ikke.

🗖 📔 Bioinformatik - Sysbio.dk 🛛 🗙	S NCBI Blast:Nucleotide Sequence X	+								$\times$
← C 🙃 https://blast.ncbi.nlm.	nih.gov/Blast.cgi#			A" to		G {	≡ 🕀	$\overline{1}$		
An official website of the United State	seqdump - Notesblok					- 0	×	0	-	
	Fil Rediger Vis						÷	~		q
NIIH National Libra										+
National Center for Biote	<pre>&gt;AY156807.1:1-805 HIV-1 clone GTAGGACCTACACCTGTCAACATAATTGGA</pre>	V2.MIC.RT from USA revers AAGAAATCTGTTGACTCAGATTGGTT	e transcriptase (p GCACTTTAAATTTTCCCA	ol) gene, TAAGTCC	partial	cds				6
	TATTGAAACTGTACCAGTAAAATTAAAGCC TAAAAGCATTAGTAGAAATTTGTACAGAA	CAGGAATGGATGGCCCAAAAGTTAAA ATAGAAAAGGAAGGAAAAATTTCAAA	CAATGGCCACTGACAGAA AATTGGGCCTGAAAATCC	GAAAAAA ATACAAT						
BLAST <sup>®</sup> » blastn suite » resul	ACTCCAGTATTTGCCATAAAGAAAAAAAAA	CAGTACTAGATGGAGAAAATTAGTAG	ATTTCAGAGAACTTAATA	AGAGAAC						o-
	TGGGTGATGCATATTTTTCAGTTCCCTTAC	GATAAAGAGTTCAGGAAGTATACTGC	ATTTACCATACCTAGTAT	AAACAAT						+
<b>Cedit Search</b> Save Search	GACAAAAATCTTAGAGCCTTTTAGAAAACA	AAAATCCAGACATAGTTATCTATCAA	TACATGGATGATCTGTAT	GTAGGAT						
Job Title	CTGACTTAGAAATAGGGCAGCATAGAATAA GACGAAAAAACACCAGAAAGAACCTCCATTO	AAAACAGAGGAACTAAGACAACATCT CCGTTGGATGGGTTATGAACTCCATC	GTTGAAGTGGGGATTTTT CTGATAAATGGACAGTAC	CACACCA AGCCTAT				df		
Nucleotide Sequence	AGTGC >AY156806.1:1-805 HTV-1 clone	V1.MTC.RT from USA revers	e transcriptase (n	ol) gene.	nartial	cds				
RID	GTAGGACCTACACCTGTCAACATAATTGGA	AAGAAATCTGTTGACTCAGATTGGTT	GCACTTTAAATTTTCCCA	TAAGTCC						
YWSS49H1016 Search expires on 02	TAAAAGCATTAGTAGAAATTTGTACAGAAA	ATAGAAAAGGAAGGAAAAAATTTCAAA	AATTGGGCCTGAAAATCC	ATACAAT						
Download All 🗸	ACTCCAGTATTTGCCATAAAGAAAAAAAAA TCAAGACTTCTGGGAAGTTCAATTAGGAA	CAGTACTAGATGGAGAAAATTAGTAG TACCACATCCTGCAGGGTTAAAAAAG	ATTTCAGAGAACTTAATA AAAAAATCAGTAACAGTG	AGAGAAC CTGGATG						
Program	TGGGTGATGCATATTTTTCAGTTCCCTTAG GAGACACCAGAGATTAGATATCAGTACAAT	GATAAAGAGTTCAGGAAGTATACTGC TGTGCTTCCACAGGGATGGAAAGGAT	ATTTACCATACCTAGTAT CACCAGCAATATTCCAAA	AAACAAT GTAGCAT						
BLASTN 2 Citation >	GACAAAAATCTTAGAGCCTTTTAGAAAACA	ΑΑΑΑΤCCAGACATAGTTATCTATCAA	TACATGGATGATCTGTAT	GTAGGAT						
Database	Ln 1, Col 1		100% Uni:	(LF)		UTF-8				
pt Soc details M						rtia			back	
						Fitter	Rese	τ	eed	
Query ID										€
lcl Query_62407									_	67
Description									-	43
•	📕 Q Søg 📕 💷	) 💼 💽 🗐 📄 !	💶 🧿 🐖 🦊		<u> </u>	🔪 🚵 DAI	N (\$ \$)	D 10	18:19 6-02-2023	4

- 19. For at undersøge, om virus fra offer og patient er tæt beslægtede, må man sammenligne med virus fra andre personer, der ikke har været involverede i sagen. Suppler filen med sekvenser fra tre andre personer fra USA og tre fra samme stat (Louisiana). Sekvensernes aquision number er:
  - AY835777 >USA\_1
  - o AY835778 >USA\_2

- o AY835769 >USA\_3
- o AY156793 >LA\_1
- o AY156789 >LA\_2
- o AY156788 >LA\_3

20. Nu skal filen rettes til, så sekvenserne får håndterbare navne (første linje inden sekvensen starter). De vigtigste regler er:

Sekvensen skal altid starte med ">"

Der må ikke være mellemrum. Brug "\_"

Du kan fx omdøbe sekvenserne til (men kan også vælge andre navne, som giver mening):

Acquisition number	Nyt navn
AY156807	<pre>&gt;victim_clone1</pre>
AY156806	<pre>&gt;victim_clone2</pre>
AY156803	>patient_clone1
AY156801	>patient_clone2
AY156802	>patient_clone3
AY156800	>patient_clone4
AY156797	>patient_clone5
AY156799	>patient_clone6
AY835777	>USA_1
AY835778	>USA_2
AY835769	>USA_3
AY156793	>LA_1
AY156789	>LA_2
AY156788	>LA_3

Du har nu en FASTA-fil du kan anvende til at lave det fylogenetiske træ. Gem din fil med et navn, som du kan genkende.