Biological Databases and file formats

T: +27 51 401 9111 | E: info@ufs.ac.za | www.ufs.ac.za **f** UFSUV | ♥ UFSweb | ♥ UFSweb | ♥ ufsuv

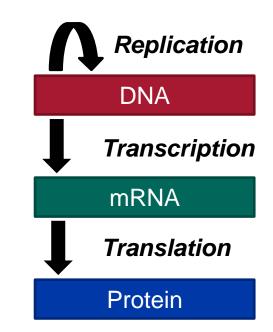
Inspiring excellence. Transforming lives.

UNIVERSITY OF THE FREE STATE UNIVERSITEIT VAN DIE VRYSTAAT YUNIVESITHI YA FREISTATA

UFS INFORMATION AND COMMUNICATION TECHNOLOGY SERVICES (ICT SERVICES)

Central Dogma of Molecular Biology

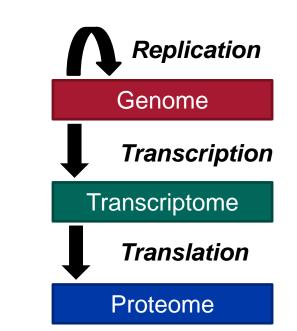
- Technological limitations→ single molecules
- Development of simple paradigm
- Last 20 years → Rapid technological development





Central Dogma of Molecular Biology 2.0

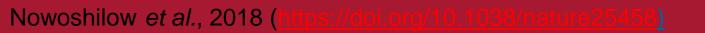
- New Technologies
 - Next Generation Sequencing (NGS)
 - Advances in Mass Spectrometry
 - Advances in NMR technologies
- Single molecule → Entire complement of a cell's biomolecules
- -Omes \rightarrow Birth of OMICS





Introduction – A problem of abundance

- Omics fields produce TONS of data
 - Mexican axolotl (Ambystoma mexicanum) (Salamander) genome
 - 32 giga-base-pairs (Gbp)
 - Human genome = 3272 Mbp
 - Genome + transcriptomics data = 8.5 GB (Gigabytes)
 - Largest genome sequenced thus far.









Databases

- Computer Science → Databases
- Shared, integrated computer structures that stores a collection of:
 - End-user data (raw facts e.g. Biological Sequences)
 - Metadata (data about data) → used to integrate and manage end-user data. E.g. Accession Number



Biological Databases

- Databases that store biological data
 - Publicly available
- Initially created to store sequence data (Sanger sequencing era)
- Evolved \rightarrow Integrated and interactive platforms for lab scientists
- Many biological databases \rightarrow specialized
 - Organisms
 - Diseases



Biological Databases - Examples

Name	Website	Description
GenBank (NCBI)	www.ncbi.nlm.nih.gov/genbank	International nucleotide sequence databases and repositories
ENA (EMBL-EBI)	www.ebi.ac.uk/ena	-
DDBJ	www.ddbj.nig.ac.jp	-
UniProt	www.uniport.org	Protein database, sequence and functional annotation
Ensembl	www.ensemble.org	Vertebrate and eukaryotic genomes
Ensembl genomes	www.ensemblegenomes.org	Genome-scale data for bacteria, protists, fungi, plants and invertebrate metazoa
InterPro	www.ebi.ac.uk/interpro	Functional analysis of protein sequences
Pfam	pfam.xfam.org	Manually curated collection of protein domain families
RCSB PDB	www.rscb.org	3D structure data for large biological molecules



Biological Databases – How do they grow?

- Incentives for data to be submitted/deposited:
 - Required for journal publication Data \rightarrow Publicly available
 - Required by funding agencies (especially public funders)
 - Data-exchange between databases



Types of Biological Databases – Primary DB

- Usually repositories of experimental data
- Good example \rightarrow Sequence databases
 - − Genbank, ENA, DDBJ \rightarrow nucleic acid sequences
 - Uniprot \rightarrow Protein sequences
- PDB → Structural data (X-ray Crystallography / NMR)



Types of Biological Databases – Secondary DB

- Layered on top of primary databases
- Provides additional information about records in the primary databases
- InterPro and Pfam → Classify protein structures into families based on structure/function

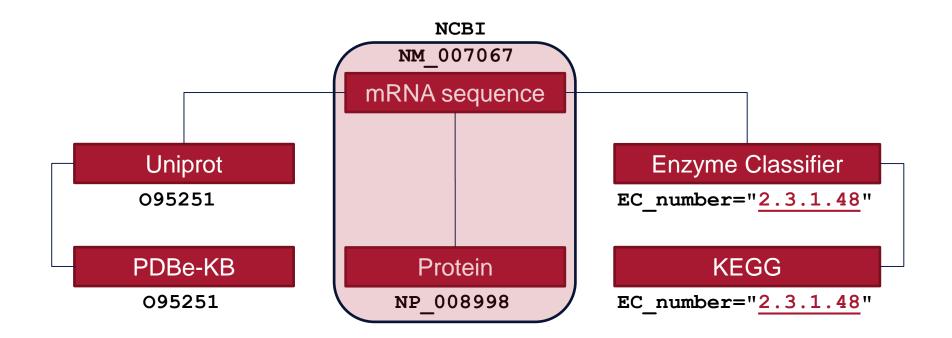


Types of Biological Databases – How are the connected

- Linked by accession numbers
 - Unique codes for records stored in a particular DB
- What would happen if the primary entry is corrupted / incorrect?



Example: Homo sapiens lysine acetyltransferase 7





Types of Biological Databases – Secondary DB

- Layered on top of primary databases
- Provides additional information about records in the primary databases
- InterPro and Pfam → Classify protein structures into families based on structure/function



Database Curation

- Accuracy of information in biological databases critical
 - Many downstream secondary databases rely on data to generate information used by lab scientists
- Curated in two ways:
 - Manually by humans \rightarrow Slow but high quality
 - Automatically \rightarrow Quick but less reliable
- Always double-check data entries manually in cases of unusual results



Biological Data Formats

- Downloading data from biological databases
 - Specific file formats. Ex. Sequences in FASTA format
 - Designed to be read by computer software (i.e. predictable)
 - Usually text files \rightarrow can be viewed in text editors
- Some important formats:
 - FASTA (sequence storage format)
 - GenBank (sequence and annotation format)
 - CLUSTAL (sequence alignment format)



FASTA File Format

- Common extensions → .fasta , .fa , .faa
- Single or Multiple sequences / records
- Records start with header:
 - ">" followed by a text description \rightarrow can contain accession number
- Next line(s) contains sequence
 - 70 characters per line until the end of sequence
- For multiple sequence files:
 - Start next sequence with a new header line



FASTA File Format Example

- This is only one line! (Look at the line-numbers)
- Good editors wrap the lines for you for easier reading
 - Thus it is a good idea to enable line numbers in your app of choice
- 1 NC_045512.2:21563-25384 S [organism=Severe acute respiratory syndrome coronavirus 2] [GeneID=43740568] [chromosome=]



GenBank File Format

- Common extensions → .gb , .genbank
- Contains annotation section and the sequence
 - Larger Files
 - More information than sequence
- Starts with the LOCUS keyword and record ends with "//" after sequence
- Goto <u>https://www.ncbi.nlm.nih.gov/Sitemap/samplerecord.html</u> for an interactive exploration of a genbank record



GenBank File Format Example

1	LOCUS	SCU49845 5028 bp DNA linear PLN 29-OCT-2018
2	DEFINITION	Saccharomyces cerevisiae TCP1-beta gene, partial cds; and Axl2p
3		(AXL2) and Rev7p (REV7) genes, complete cds.
4	ACCESSION	U49845
5	VERSION	U49845.1 GI:1293613
6	KEYWORDS	
7	SOURCE	Saccharomyces cerevisiae (baker's yeast)
8	ORGANISM	Saccharomyces cerevisiae
9		Eukaryota; Fungi; Dikarya; Ascomycota; Saccharomycotina;
10		Saccharomycetes; Saccharomycetales; Saccharomycetaceae;
11		Saccharomyces.
12	REFERENCE	1 (bases 1 to 5028)
13	AUTHORS	Roemer,T., Madden,K., Chang,J. and Snyder,M.
14	TITLE	Selection of axial growth sites in yeast requires Axl2p, a novel
15		plasma membrane glycoprotein
16	JOURNAL	Genes Dev. 10 (7), 777-793 (1996)
17	PUBMED	8846915
18	REFERENCE	2 (bases 1 to 5028)
19	AUTHORS	Roemer,T.
20	TITLE	Direct Submission
21	JOURNAL	Submitted (22-FEB-1996) Biology, Yale University, New Haven, CT
22		06520, USA
23	FEATURES	Location/Qualifiers
24	source	15028
25		/organism="Saccharomvces cerevisiae"



GenBank File Format Example

22	06	520, USA
23	FEATURES	Location/Qualifiers
24	source	15028
25		/organism="Saccharomyces cerevisiae"
26		/mol type="genomic DNA"
27		/db xref="taxon:4932"
28		/chromosome="IX"
29	mRNA	<1>206
30		/product="TCP1-beta"
31	CDS	<1206
32		/codon start=3
33		/product="TCP1-beta"
34		/protein id="AAA98665.1"
35		/db xref="GI:1293614"
36		/translation="SSIYNGISTSGLDLNNGTIADMROLGIVESYKLKRAVVSSASEA
37		AEVLLRVDNIIRARPRTANRQHM"
38	gene	<687>3158
39		/gene="AXL2"
40	mRNA	<687>3158
41		/gene="AXL2"
42		/product="Ax12p"
43	CDS	6873158
44		/gene="AXL2"
45		/note="plasma membrane glycoprotein"
46		/codon_start=1
47		/product="Ax12p"
48		/protein_id="AAA98666.1"
49		/db_xref="GI:1293615"
50		/translation="MTQLQISLLLTATISLLHLVVATPYEAYPIGKQYPPVARVNESF
51		TFQISNDTYKSSVDKTAQITYNCFDLPSWLSFDSSSRTFSGEPSSDLLSDANTTLYFN
52		VILEGTDSADSTSLNNTYQFVVTNRPSISLSSDFNLLALLKNYGYTNGKNALKLDPNE
53		VFNVTFDRSMFTNEESIVSYYGRSOLYNAPLPNWLFFDSGELKFTGTAPVINSAIAPE
54		TSYSFVIIATDIEGFSAVEVEFELVIGAHQLTTSIQNSLIINVTDIGNVSYDLPLNYV
55		YLDDDPISSDKLGSINLLDAPDWVALDNATISGSVPDELLGKNSNPANFSVSIYDTYG
56		DVIYFNFEVVSTTDLFAISSLPNINATRGEWFSYYFLPSQFTDYVNTNVSLEFTNSSQ
57 58		DHDWVKFQSSNLTLAGEVPKNFDKLSLGLKANQGSQSQELYFNIIGMDSKITHSNHSA NATSTRSSHHSTSTSSYTSSTYTAKISSTSAAATSSAPAALPAANKTSSHNKKAVAIA
59		CGVAIPLGVILVALICELIFWERBRENPDDENLPHAISGPDLNNPANKPNOENATPLN
59		CGVAIPLGVILVALICFLIFWRRRRENPDDENLPHAISGPDLNNPANKENQENAIPLN NPEDDDASSYDDTSIABBLAALNTLKLDNHSATESDISSUDEKEDSLSGMNTYNDOEO
61		SOSKEELLAKPPVOPPESPFFDPONRSSSVYMDSEPAVNKSWRYTCHLSPVSDIVRDS
62		YGSOKTUDTEKLFDLEAPEKEKRTSRDVTMSSLDPWNSNISPSPVRKSVTPSPYNVTK
63		HRNRHLONIODSOSGKNGITPTTMSTSSSDDFVPVKDGENFCWVHSMEPDRRPSKKRL
64		VDFSNKSNVNVGOVKDIHGRIPEML"
65	gene	complement (<3300>4037)
66		/gene="REV7"
67	mRNA	complement (<3300>4037)
68		/gene="REV7"
69		/product="Rev7p"
70	CDS	complement (33004037)
71		/gene="REV7"
72		/codon start=1
73		/product="Rev7p"
74		/protein_id="AAA98667.1"
75		/db_xref="GI:1293616"
76		/translation="MNRWVEKWLRVYLKCYINLILFYRNVYPPQSFDYTTYQSFNLPQ
77		FVPINRHPALIDYIEELILDVLSKLTHVYRFSICIINKKNDLCIEKYVLDFSELQHVD
78		KDDQIITETEVFDEFRSSLNSLIMHLEKLPKVNDDTITFEAVINAIELELGHKLDRNR
79		RVDSLEEKAEIERDSNWVKCQEDENLPDNNGFQPPKIKLTSLVGSDVGPLIIHQFSEK
80		LISGDDKILNGVYSQYEEGESIFGSLF"
81	ORIGIN	

82 1 gatectecat atacaacggt atetecaect caggtttaga teteaacaac ggaaccattg



GenBank File Format Example

getecteest stacsweggt stelesseet exggttings telesseese gas 61 cosecutors acceptage atoptogen of acceptor assecutors of additional 121 clgostchy angegiges gitterin gygtygeles esteriogi postgenes 181 gesengene tegeneset stytesent titegystet antegene testescog 241 consective stytestates tryansen gesegenese testescog 301 executions assessments according assectivity positions to consists 361 attingene character includes and an and a second secon 481 gagtogeont colligiogi glasititisi etittosisti gagacettat titelistie 541 titaeleten esteentytag tystigaese tyssaesgen amatoseta gaagacega 601 assattaett astagaassa titatette engassesga titeleget conservicta 661 catatatoss associtos citacostas osceptios astitustas tinciasos 721 checkstatic actacticus chaptaging consigned in spanishi inclutingas 781 ascatuces consignates apartesita antigities atticeasti terminati 841 cctataeatc gtotylagec asysceptic assistants castigottic ganthacras 901 golygottic gittigants atticage cylinteragy issecritic totyantian 961 histopic generacray tigiattica atginata cypegylacy generacray 1021 angestic titiyances and conset tigitylic esecritics toolchogo 1001 haspitegin titigenteks bernestent tigstigtink messigiste and 1001 haspitegin titigenteks terrestent ingesigine messigiste titigenigt 1141 angetetigen ærbagstett anligengist tissengiget titigenigt tonatigting 1201 changenige attospitigt tigstigtistenig gengthetiske gjoggitter. 1261 coastigat gitettogi telegongi inagita teggoogo oggipata 1371 acteggogi tigtooga acagetaa gititigtosi ostogias gantigang 1381 gittettoje gitettoga egitosis gititigtosi ostogias gantigang 1441 clattoness textituets atoaccits characters tescutics tatactter 1501 etchaweta tyttwicke gatgaegate chattette tyatwaity gytechiaa 1561 actaitega tyelessage tygegast bagabastge becomittee gygtechice 1621 cagatgastt actoggtasg aschonate exponentit thetgtgton attratgets 1681 ettatigitja tijtjäittää tiosaettej aagitijtete onnaegijat tijttigen 1741 tiagitetet teenaatatt aaegetana gyggijaatij giteteetae taittittige 1801 cttctcegtt tecegectec gtgestaces acgtttcatt egegtttect esttcsegcc 1921 agentitize congettes traggities angegenece angetteet setterage 1981 tatatttas catcattqc stgattcas squaectca clossocce: agtqcpastq 2041 comptone aspectic response optimized interaction of the second particular and the second pa 2221 clatomatt another classacte testimett estatate transcere 2281 gauggaoan trougocat gaoactic costects tagtgacet attracta 2341 acception tagtgacet capacity of the second tagtgacet attractant 2341 acceptions takenessed capacity accepting the second tractant 2401 atgetteete glacgalgat actionatag congregati ggetgetitg anceetitg 2461 aattgatas costicijo: actgasticij atatticosij cytypitgas asgagagatt 2521 ctotatosig tatgastaca taosatgato agticosato cosasgiasa gaspattat 2581 tecosesec: consistence octoneses accontect teconeces estagatett 2441 chickanali tangatagi gaacsagoag taashaalic chigogatat actigosacc 2701 titoccept chotastatt gloegeses gitactiaalic chigogatat actigosacc 2761 assaulttt casttages graccagege appearant; temptowage gatgtoects 2821 tytettoon operating accepted typesetti terrotary protocolar 2821 terrotary accepted typesetti terrotary accepted 2881 censtore attacepts accepted typesetti terrotary 2981 censessy terrotary accepted to the terrotary terrotary 2981 censessy terrotary accepted to the terrotary terrotary 2001 tildhogdy teenentyy microfild cancerigi accidents ingergen 3001 tigthogt teenentyy enentyti accident til gelygics tegespage comparingen 3061 gecompter geenengtte glagetilt coestenyng teetytoet gligtoeng 3121 theogenet teespegenge elocoseese tgelygicht arecgoeseg atellityte 3181 tastitutt thengitte attitute opping the opponent lattitutt 2011 agtitutat ethogons thattitut attended to be the second second 321 tastitutat ethogons thattitu attended to be this attitudes 321 tassesses attended to be the second 3361 attitutort connectest teattitics classified perturbations approximate 3421 tangances charged gatthat tanganget generation approximate 3421 tangances chargengt gatthat tanganget 3481 eestiticat citcinges tileaccos tilgentoc tilceatite typittics 3541 tonesectat operation gitteration sectiation chapteries theyeterate 3601 thesements concentry for the processing of the second 2021 optigitet tatisfierg organisati insensitia emigravit transport 301 emigravit tatisfierge organisati emissione emigravitet transport 301 engenetice globagtic ticitateg treating organisati engentitati 301 engenetice globagtic ticitateg treating organisati engentit 301 engenetice globagtic ticitategine globagticgi organisati engingette 301 engenetice globagtice ticitategine globagticgi organisati engingette 301 engenetice globagtice ticitategine globagticgi organisati engingette 301 engenetice globagtice ticitategine globagticgi organisati 301 engenetice globagtication globagticati enginetice enginetice 301 engenetice globagticati enginetice enginetice enginetice 301 engenetice enginetice enginetice enginetice enginetice 301 engenetice enginetice enginetice enginetice enginetice 301 engenetice enginetice enginetice 301 engenetice 301 eng 3961 acattletat assatasest constitueig tagostitta agtataccet cagocactte 4021 tctaccostc tettostasa gotgacgosa cgattactat tittittitc tictiggato 4081 tcagtogtog casaaacgta tacctictit ticcgaccti tittitagci ticiggasaa 4141 utiliatatta attasaccan utclastctt antatasaan chastauttt castinacta 4201 attraspes agtgenett antropic giptagengin antropic attraction attracts 4261 attraspes agtgenett antropic giptagengin attracting tetticing 4261 clyttlagt ticlengte: titlgette tegeneggig assegnets catectati 4321 titigitanag giganagcat antgianang clagestana atgacques tanagagagg 4381 citizgitung giputeres angoness angotesi composing 4381 citizgitung diputeres angoness angotesi composing 4441 contribute generators tigiging analysis to traceite cytigenti 4501 angogittig cytigitigin di contractor angotesi 4501 angogittig cytigittigi. 461 approprint topological interpolation integrated integrated characterized 4551 topological provides the standard of the second statistical second 4651 topological approximation of the second second second second 4651 topological topological second second second second second 4651 attractions applicable second second second second second 4651 attractions applicable second second second second second second 4651 attractions applicable second second second second second 4551 topologicable second second second second second second 4551 topologicable second second second second second second 4551 topologicable second second second second second second second 4551 topologicable second second second second second second 4551 topologicable second second second second second second second 4551 topologicable second second second second second second second 4551 topologicable second 4741 toectatett chegetatta tickegetee taatittet taatataatt eteettette 4801 getetcosett tettgagete threecoset techtigtet cepecastig acticicae 4861 tictosectt seciglogeg tigetcgtit thegegace sepatitiest ctoptitic 4921 titicagigi iagaligete taatteliig ageigitete tongeteete alaittitet 4981 igeoalgaet oogaltetaa titiaageta tioaattet ettigate



ClustalW File Format

- Common extensions → .aln
- For sequence alignments
- Format:
 - Start with "CLUSTALW" or "CLUSTAL W" (all other info is ignored)
 - One or more new lines (empty lines)
 - One or more blocks of sequence



ClustalW File Format – Sequence Block

- One line representing each sequence in alignment
- Formatted as follows:
 - Sequence Name
 - White space (usually a TAB)
 - Sequence symbols (max 60 per line). Gaps represented by "-"
 - Optional: White space followed by position of last sequence unit
- One line representing degree of conservation for columns in block
- One or more new lines



ClustalW File Format – Conservation Symbols

- * \rightarrow All residues are identical
- : \rightarrow Conserved substitutions observed
- . \rightarrow Semi-conserved substitutions observed
- "" \rightarrow No match



CLUSTALW File Format Example

1	CLUSTAL W (1.82) multiple sequence alignment	
2			
3			
4	FOSB_MOUSE	MFQAFPGDYDSGSRCSSSPSAESQYLSSVDSFGSPPTAAASQECAGLGEMPGSFVPTVTA	60
5	FOSB HUMAN	MFQAFPGDYDSGSRCSSSPSAESQYLSSVDSFGSPPTAAASQECAGLGEMPGSFVPTVTA	60
6 7		***************************************	
8	FOSB MOUSE	ITTSQDLQWLVQPTLISSMAQSQGQPLASQPPAVDPYDMPGTSYSTPGLSAYSTGGASGS	120
9	FOSB HUMAN	ITTSQDLQWLVQPTLISSMAQSQGQPLASQPPVVDPYDMPGTSYSTPGMSGYSSGGASGS	120
10 11		***************************************	
	ROAD MOULT		100
12	FOSB_MOUSE	GGPSTSTTTSGPVSARPARARPRRPREETLTPEEEEKRRVRRERNKLAAAKCRNRRELT	
13	FOSB_HUMAN		180
14		****** ***** •************************	
15			
16	FOSB_MOUSE	DRLQAETDQLEEEKAELESEIAELQKEKERLEFVLVAHKPGCKIPYEEGPGPGPLAEVRD	240
17	FOSB HUMAN	DRLQAETDQLEEEKAELESEIAELQKEKERLEFVLVAHKPGCKIPYEEGPGPGPLAEVRD	240
18		***************************************	
19			
20	FOSB MOUSE	LPGSTSAKEDGFGWLLPPPPPPPPPPPPSSRDAPPNLTASLFTHSEVQVLGDPFPVVSPSY	300
21	FOSB HUMAN	LPGSAPAKEDGFSWLLPPPPPPPPPPPPTPTSQDAPPNLTASLFTHSEVQVLGDPFPVVNPSY	300
2.2		*****	
2.3			
24	FOSB MOUSE	TSSFVLTCPEVSAFAGAORTSGSEOPSDPLNSPSLLAL 338	
25	FOSB HUMAN	TSSFVLTCPEVSAFAGAORTSGSDOPSDPLNSPSLLAL 338	
26		*****	



Take Home Message

- Get a good text editor
- Know your format of choice → will make understanding what software can do easier
- Using bioinformatics software \rightarrow Read input section
 - Parsing / input errors common
 - Some software applications may be strict or permissive with regards to format adherence
- Format converters
 - Understand what information may be discarded
 - Example → Coverting GenBank to FASTA

