

Einleitung

Was ist Bioinformatik?

Die Bioinformatik (englisch bioinformatics, auch computational biology) ist eine interdisziplinäre Wissenschaft, die Probleme aus den Lebenswissenschaften mit theoretischen computergestützten Methoden löst. Bioinformatik ist ein weitgefächertes Forschungsgebiet, sowohl bei Problemstellungen als auch den angewandten Methoden. Wesentliche Gebiete sind die Verwaltung und Integration biologischer Daten, die Sequenzanalyse, die Strukturbioinformatik und die Analyse von Daten aus Hochdurchsatzmethoden (~omics). Da Bioinformatik unentbehrlich ist, um Daten in großem Maßstab zu analysieren, bildet sie einen wesentlichen Pfeiler der Systembiologie. Im Folgenden werden einige grundlegende Begriffe erläutert:

Übliche Sequenztypen:

- DNA, RNA, Codons, Proteine

Repräsentation: 1 Letter code DNA/RNA

A	adenosine	C	cytidine	G	guanine
T	thymidine	N	A/G/C/T (any)	U	uridine
K	G/T (keto)	S	G/C (strong)	Y	T/C (pyrimidine)
M	A/C (amino)	W	A/T (weak)	R	G/A (purine)
B	G/T/C	D	G/A/T	H	A/C/T
V	G/C/A	-	gap of indeterminate length		

Repräsentation: 1 Letter code proteins

A	alanine	P	proline
B	aspartate/asparagine	Q	glutamine
C	cystine	R	arginine
D	aspartate	S	serine
E	glutamate	T	threonine
F	phenylalanine	U	selenocysteine
G	glycine	V	valine
H	histidine	W	tryptophan
I	isoleucine	Y	tyrosine
K	lysine	Z	glutamate/glutamine
L	leucine	X	any
M	methionine	*	translation stop
N	asparagine	-	gap of indeterminate length

Repräsentation: Codon Tabelle

Der genetische Standardcode ist in Abbildung 1 gezeigt.

		Second letter					
		U	C	A	G		
First letter	U	UUU } Phe UUC } UUA } Leu UUG }	UCU } UCC } Ser UCA } UCG }	UAU } Tyr UAC } UAA Stop UAG Stop	UGU } Cys UGC } UGA Stop UGG Trp	U C A G	Third letter
	C	CUU } CUC } Leu CUA } CUG }	CCU } CCC } Pro CCA } CCG }	CAU } His CAC } CAA } Gln CAG }	CGU } CGC } Arg CGA } CGG }	U C A G	
	A	AUU } AUC } Ile AUA } AUG Met	ACU } ACC } Thr ACA } ACG }	AAU } Asn AAC } AAA } Lys AAG }	AGU } Ser AGC } AGA } Arg AGG }	U C A G	
	G	GUU } GUC } Val GUA } GUG }	GCU } GCC } Ala GCA } GCG }	GAU } Asp GAC } GAA } Glu GAG }	GGU } GGC } Gly GGA } GGG }	U C A G	

Abbildung 1: Der genetische Standardcode

FASTA Format

Das FASTA-Format ist ein textbasiertes Format zur Darstellung und Speicherung der Primärstruktur von Nukleinsäuren (Nukleinsäuresequenz) und Proteinen (Proteinsequenz). Die Nukleinbasen bzw. Aminosäuren werden durch einen Ein-Buchstaben-Code dargestellt. Das Format erlaubt es, den Sequenzen einen Namen und Kommentare in einer Kopfzeile (geennzeichnet durch ein >) voranzustellen.

```
>gi|5524211|gb|AAD44166.1| cytochrome b [Elephas maximus maximus]
LCLYTHIGRNIYYGSYLYSETWNTGIMLLLITMATAFMGYVLPWGQMSFWGATVITNLFSAIPYIGTNLV
EWIWGGFSVDKATLNRFFAFHFILPFTMVALAGVHLTFLHETGSNNPLGLTSDSDKIPFHPYYTIKDFLG
LLILILLLLLLLALLSPDMLGDPDNHMPADPLNTPLHIKPEWYFLFAYAILRSVPNKLGGVLALFLSIVIL
GLMPFLHTSKHRSMMLRPLSQALFWTLTMDLLTLTWIGSQPVEYPYTIIGQMASILYFSIILAFPLIAGX
IENY
```

Identifiers

Identifiers sind normalerweise einfache Zifferncodes mit oder ohne Versionsnummer. Ein Identifier ist relativ flexibel und meist einfach (zum Beispiel ein einzelnes Wort oder Buchstaben/Zahlenkombination). Leerzeichen (sowohl davor als auch danach bzw. im Identifier selber führen meist zu Problemen), zum Beispiel:

```
ACCESSION    P01013
AAA68881. 1
gi| 129295
```

Korrekte Identifier dagegen sind

```
ACCESSION_P01013
AAA68881.1
gi|129295
```

Alignment

Sequenzalignment (von lateinisch sequentia, „Aufeinanderfolge“ und englisch alignment, „Abgleich, Anordnung, Ausrichtung“) bezeichnet den methodischen Vergleich zweier oder mehrerer Nukleotid- oder Aminosäuresequenzen in linearer Abfolge. Es wird u.a. in der molekularen Phylogenie verwendet, um die funktionelle oder evolutionäre Verwandtschaft (Homologie) von Nukleotidsequenzen, Aminosäuresequenzen oder Codonsequenzen zu untersuchen.

Vorbereitung Klausur

In der Klausur werden zahlreiche bioinformatische Onlinetools genutzt. Dazu ist es erforderlich, diese zunächst online zu testen um die Funktionalität während der Klausur zu gewährleisten. Im Folgenden sind einige beispielhafte Fragen für Datenbanken und Softwaretools genannt die sehr wahrscheinlich auch während der Klausur Anwendung finden werden.

WICHTIG (auch und insbesondere für die Klausur): Falls Dateien abgespeichert werden sollen, sind diese unter einem aussagekräftigem Dateinamen (achte auf die korrekte Endung) in dem *individuell* geteiltem Dropboxordner abzulegen. Dabei können auch Unterordner zur besseren Strukturierung angelegt werden.

(1) *Datenbank*: KEGG <https://www.kegg.jp/kegg/genome.html>

Suche nach dem Krebszyklus in *Pathways* (Identifier 00020) und markiere den Pathway für die Modellpflanze *Arabidopsis thaliana*. Speichere die erhaltene Abbildung unter einem geeignetem Dateinamen ab.

(2) *Datenbank*: GenBank <https://www.ncbi.nlm.nih.gov/genbank/>

Suche nach “Severe acute respiratory syndrome coronavirus 2 isolate Wuhan-Hu-1, complete genome” (Identifier NC_045512). Speichere alle codierenden DNA Abschnitte als Proteine (kann komplett über die Schaltfläche “Send to” und den geeigneten Einstellungen vorgenommen werden) unter einem geeigneten Dateinamen ab.

(3) *Datenbank*: <https://www.ncbi.nlm.nih.gov/datasets/genomes/>

Suche das Eisbären genom. Wie lautet der Assembly Identifier des Genoms? Wie viele Gene und Proteine wurden für das Eisbären genome vorhergesagt? Suche das Gene *PMCH* (Pro-melanin-concentrating hormone) des Eisbären. Speichere die CDS (codierende DNA Sequenz) und Proteinsequenz im FASTA Format unter zwei selbstgewählten Dateinamen ab.

(4) *Software*: <https://blast.ncbi.nlm.nih.gov/Blast.cgi>

Erstelle eine BLASTsuche von *PMCH* gegen die NCBI Refseq Datenbank. Ermittlen sie eine Spezies mit weniger als 95% Identität der Proteinsequenz.

(5) *Software*: Muscle <https://www.ebi.ac.uk/Tools/msa/muscle/>

Ermittle die Protein Sequenzen des PMCH Gens für den Braunbär und für die Weddellrobbe (*Leptonychotes weddellii*), und führe gemeinsam mit der oben ermittelten Eisbärsequenz ein Alignment mit muscle durch.

(6) *Software*: webPrank <https://www.ebi.ac.uk/goldman-srv/webprank/>

Führe mit den 3 CDS Sequenzen ein Alignment mit webPrank durch. Sind im Alignment die Codonlängen korrekt behandelt worden?

(7) *Software*: <https://kakscalculator.herokuapp.com/kaksform.jsp>

Benutze die vorgebenen Beispielsequenzen um paarweise Ka/Ks zu erhalten. Warum werden dir mehrere Werte ausgegeben?

(8) *Software*: Swiss model: <https://swissmodel.expasy.org/interactive#structure>

Erstelle ein 3D Modell des Eisbären PMCH Proteins und speichere es unter einem geeignetem Dateinamen ab. Wie viele Modelle hast du erhalten? Welche Templates wurden genutzt? Betrachte dein erstelltes 3D Proteinmodell mit dem integrierten Nglviewer und markiere eine selbstgewählte Folge von 5 Aminosäuren des Modells. Exportiere eine geeignete Abbildung, die deine Peptidmarkierung innerhalb des Modells verdeutlicht.