UniProt – Einführung in eine *freie Proteindatenbank C2BL

- **Das Hybrid-PDF besitzt Formularfunktion. Ausfüllen mit dem Okular oder LibreOffice möglich.**
- **Zu diesem Lernskript gibt es ein Dokument mit Musterlösungen:** http://laborberufe.de/c2bl/Lernskript_Einfuehrung_Proteindatenbank_UniProt_LOESUNG.pdf
- **Zu diesem Lernskript gibt es ein Video mit der Nachbesprechung:https://youtu.be/d4DdC-12T_Y**

UniProt (***uni***versal ***prot***ein database) ist die größte bioinformatische Datenbank für Proteine aller Lebewesen und Viren, und enthält Informationen über die Proteinfunktion und -struktur sowie Links zu anderen themenrelevanten Datenbanken. Gleich auf der Startseite (https://www.uniprot.org/), findet sich oben die Suchzeile:

Hier können Sie Suchbegriffe und Schlagworte eingeben, beispielsweise den englischen Namen des Proteins, die UniProt-Identifikationsnummer **(UniProtID,** z.B. über wikipedia.de ermittelt) oder die Spezies (Name der Tier- oder Pflanzenart). Unter „ADVANCED" können Sie weitere Suchkriterien eingeben, z.B. Länge, Masse, deutsche Bezeichnung.

Als Ergebnis der Suche erhalten Sie viele Informationen über das ausgewählte Protein, zu den einzelnen Abschnitten finden Sie hier Anmerkungen.

- ***1.1** Geben Sie die UniProtID **P02883** in die Suchzeile ein und scrollen Sie die Abschnitte ab und vergleichen Sie mit den Anmerkungen zu den entsprechenden Abschnitten hier unten. Stöbern Sie in Ruhe weiter, in dem Sie angegebene Links anklicken, Infoboxen anklicken (mit kleinem hoch stehenden [i] versehen), oder sich Hilfe holen („?" anklicken)."*

1. Abschnitt „Function": Molekulare und biologische Informationen

Unter *molekularen Funktionen* finden sie Angaben, wozu das Proteinmolekül auf Teilchenebene in der Lage ist, z.B. zur Bindung von Metallen, Transport von Teilchen. Unter *biologischen Funktionen* finden Sie Angaben, was das Protein für den Organismus insgesamt bedeutet. Thaumatin ist beispielsweise ein Süßstoff, der natürlicherweise in der Katamfe-Beere vorkommt.

2. Abschnitt „PTM/Processing": Informationen zur Prozessierung und zu posttranslationalen Modifikation (PTM)

Hier finden Sie Informationen zur Entstehung der Proteine ab dem Zeitpunkt der Transkription. Sagen Ihnen diese Grundlegenden Begriffe (noch) nichts, weil Sie diese noch nicht an der vorangegangenen Schule behandelt haben und an unserer Schule erst später (nach der Zwischenprüfung) unterrichtet werden, dann müssen Sie jetzt eigenständig etwas vorlernen. Hier hilft Ihnen youtube, z.B. https://youtu.be/HZ9yp38bq2Y Tipp: Nicht nur Lernvideos anschauen sondern mit mit Stift und Papier zusammenfassend mitschreiben. Es geht hier allerdings nur um einen Überblick, die genaue Behandlung erfolgt im Unterricht.

Initiator-Aminosäure. Das Startcodon codiert für die Aminosäure Methionin (M). Sie findet sich deshalb als erste AS an den meisten Eukaryotenproteinen. **2.1** *Prüfen Sie, dass die AS Nr. 1 hier auch Methionin (M) ist!*

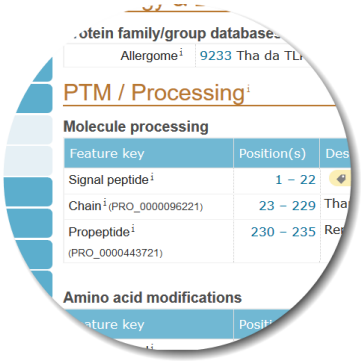


Abb. 2.1: Ausschnitt aus dem Datenbank Uniprot zu Thaumatin

Signalpeptid: Vor der eigentlichen Primärstruktur findet sich häufig noch ein *Signalpeptid (Signalsequenz)*. Diese AS-Sequenz entscheidet über den Bestimmungsort, den Transportweg des Proteins und die Sekretionseffizienz. Signalsequenzen finden sich typischerweise bei Proteinen, deren Bestimmungsort sich außerhalb der Zelle, in Biomembranen oder in Kompartimenten befindet. So ist für den Transport in das endoplasm. Retikulum, Chloroplasten, Mitochondrien oder den Zellkern und deren Membranen meist eine Signalsequenz erforderlich. Am Zielort wird diese Sequenz dann häufig durch bestimmte **Signalpeptidasen** abgespalten oder bleibt erhalten (z.B. als Transmembrandomäne).

Nach der Translation, also der Synthese an den Ribosomen werden die meisten Proteine noch weiter bearbeitet.

Knüpfen von Disulfidbrücken: Die Verknüpfung der Disulfidbrücken erfolgt durch bestimmte Enzyme nach der eigentlichen Translation.

Modifizierung von Aminosäuren: Aus Cystein wird häufig Selenocystein, als Serin bzw. Threonin entstehen häufig phosphorylierte Formen (Phosphoserin…). Sie finden auch eine Positionsangabe, also welche Nr. die AS im Angaben welche der Aminosäuren

Weitere Reifungsvorgänge: Allgemein werden häufig an bestimmten Stellen nach der Translation noch Aminosäurensequenzen abgespalten oder sogar der gesamte Proteinstrang zu zwei Untereinheiten gespalten. Hat die eine Untereinheit keine Funktion mehr, spricht man von einem **Propeptid**. *Thaumatin* enthält C-terminal ein 5 AS langes Propeptid.

3. „Structure“: Tertiär- und Quartärstruktur des Moleküls

Sie bekommen meist ein drehbares 3-D-Modell und können sehen, wo sich alpha-helicäre Bereiche etc. befinden. Durch Anklicken von Unterbereichen erhalten sie zusätzliche Informationen. Sie können verschiedene Datenbanken auswählen, die als Grundlage für die räumliche Darstellung gewählt werden. Die Analysemethoden die zur Ermittlung des räumlichen Baus gewählt werden, sind die **Kernresonanzspektroskopie (NMR)** und die **Röntgenkristallographie (X-RAY)**. Letztere setzt voraus, dass das Molekül kristallisierbar ist, also als hochregelmäßig aufgebauter Feststoff vorliegt.

- *Drehen Sie das Molekül etwas mit der Maus. Gehen Sie mit der Maus über einen Bereiche und lassen sich Position und dort liegende AS anzeigen (schwarzer Kasten links oben im Bildbereich).*
- *Klicken Sie in der Grafik einige Unterbereiche an, um sie mehr Infos zu holen und einen genaueren Ausschnitt anzeigen zu lassen.*
- *Vergleichen Sie die grafische Zusammenfassung der Sekundärstruktur. Überwiegen alpha-Helix oder beta-Faltblatt?*
- *Wählen Sie auch andere Datenbanken zur räumlichen Darstellung aus.*

Auf einem Balken finden Sie eine Übersicht, an welcher Stellen des Proteins welche Sekundrätstrukturen vorliegen. *„Beta strand“* = β-Faltblatt: *„Turn“* = Schleifenbereiche (Zuordnung zu α-Helix oder β-Faltblatt nicht mögglich)

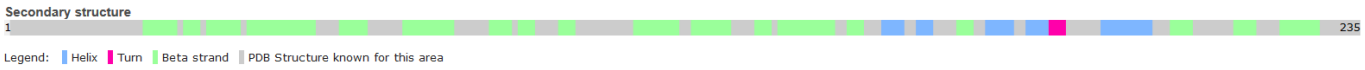


Abb. 3.1: Zusammenfassung der Sekundärstruktur von *Thaumatin* in UniProt

4. Proteinfamilie und Domänen

- Hier erhalten Sie Informationen darüber, in welchen weiteren Proteinen/Peptiden dieselben charakteristische Aminosäure-Sequenzen und charakteristische Domänen befinden.

5. Sequenz

- Hier finden Sie die Aminosäuresequenz als Einbuchstabencode. Sie finden diese Cods z.B. auf dem AS-Übersichtsblatt aus dem Unterricht. Rechts sind Gesamtlänge und Masse angegeben.
- Unter *"sequence databases"* finden sich auch Links die zur mRNA (Codons) bzw. den codogenen Sequenzen (Codogene) auf der DNA führen.
- *Klicken Sie an und lassen sie sich weiter anzeigen, welche Optionen diese weiteren Datenbanken bieten (z.B. ähnliche Sequenzen finden)*

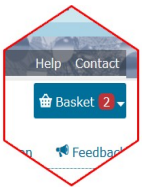
6. Ähnliche Proteine

- Hier finden Sie Verweise zu Proteinen die ähnliche Sequenzen haben. Sie sehen dass es noch Thaumatin II gibt, aus demselben Organismus (Katamfe-Pflanze). Weiterhin gibt es noch weitere Proteinsequenzen mit generischem Namen. Alle diese Proteine haben eine über 90%ige Übereinstimmung.

2. Proteinvergleich (Alignment): Vergleich mehrerer Einträge aus der Datenbank untereinander

*2.1 Legen Sie die Aminosäurensequenzen für das **humane** und das **bovine Serumalbumin** (selber in der Datenbank finden! Tipps: z.B. über die englischen Namen oder die UniProtID in den Korb (basket) und vergleichen Sie diese bezüglich der AS-Sequenz (Alignment).*

anklicken: "Add to basket"



rechts oben im Fenster Korb klicken

UniProtIDs („Entry“) und UniProt-Proteinbezecihnung („Entry name“):

humanes Serumalbumin: | bovines Serumalbiumin:

2.2 *Heben Sie durch Anklicken der entsprechenden Kästchen links, die Signalpeptide, die Bereiche mit Faltblatt etc. hervor.*

2.3 *Wie viel Prozent der Aminosäuren sind identisch?*

Bemerkung: Was genau diese Prozente bezeichnen, ist komplex: Da die verglichenen Proteine nicht exakt gleich lang sein müssen, muss man beim Vergleich auch erlauben, das zum Ausgleich die AS-Sequenzen auch auseinander geschnitten werden dürfen. Die Sequenzen werden dann so aneinander gelegt, dass es die größte Übereinstimmung gibt. Wen es näher interessiert: https://de.wikipedia.org/wiki/Sequenzalignment

3. BLAST: Durchsuchen einer Datenbank nach einer bestimmten Sequenz

BLAST (Abkürzung für englisch *Basic Local Alignment Search Tool*) ist der Überbegriff für eine Sammlung der weltweit am meisten genutzten Programme zur Analyse biologischer Sequenzdaten. BLAST wird dazu verwendet, experimentell ermittelte DNA- oder Protein-Sequenzen mit bereits in einer Datenbank vorhandenen Sequenzen zu vergleichen.
Als Ergebnis liefert das Programm eine Reihe lokaler Alignments, d.h. Gegenüberstellungen von Stücken der gesuchten Sequenz mit ähnlichen Stücken aus der Datenbank. Darüber hinaus gibt BLAST an, wie signifikant die gefundenen Treffer sind. Weitere Unterteilung der BLAST-Verfahren

blastp	• Vergleicht eine Aminosäuresequenz gegen eine Proteinsequenzdatenbank. *Beispielfrage: In welchen bekannten Proteinen gibt es die Aminosäuresequenz YGFIRTHRGT? Suche in einer Proteindatenbank (blastp)!* Man kann auch das gesamte Protein BLASTen. Dann bekommt man eine Übersicht über weitere Proteine, die die entsprechende Aminosäuresequenz enthalten. **vgl. auch Aufgabe 3.1 (unten)**
blastn	• Vergleicht eine Nukleotidsequenz gegen eine Nukleotidsequenzdatenbank
blastx	• Vergleicht eine Nukleotidsequenz (in ALLEN Leserastern translatiert) gegen eine Proteindatenbank • Man kann diese Möglichkeit nutzen, um eine mögliche Translation einer bekannten Nukleotidsequenz zu finden. Beispielsfrage: In wechen bekannten Proteinen steckt indirekt die DNA-Nucleotid-Sequenz: *agtgtcgctgtagagagtgcggagtgta*? Suche in einer Proteindatenbank. Mögliche Suchfilter: Beschränkung auf bestimmte Pflanzen- oder Tierart. **vgl. auch Aufgabe 3.2 (unten)**
tblastn	• Vergleicht eine AS-Sequenz gegen eine Nukleotiddatenbank (dynamisch in allen Leserastern translatiert)

3.1*. BLASTen Sie das gesamte humane Prolactin indem Sie auf BLAST drücken. Prüfen Sie auf der Basis Ihres Ergebnissen ob Menschen eher mit Bonobo-Schimpansen (Pan paniscus) oder eher mit Gorillas (Gorilla gorilla) oder eher mit Dromedaren (Camelus dromedarius) verwandt sind. Wie groß sind die Übereinstimmungen?*

3.2 *Führen Sie mit UniProt ein* ***blastx*** *für folgende DNA-Sequqenz durch: agtgtcgctgtagagagtgcggagtgta (Leseraster ist nicht bekannt!). Geben Sie das Protein mit UniProtID und Uniprot-Name an, zu dem die Sequenz passt.*

3.3 *Ermitteln Sie die proz. Übereinstimmung des Enzyms „**ATP-abhängige 6-Phosphofructokinase 1(ATP-PFK)“ (Enzym-nummer: 2.7.1.11)** im Vergleich zur humanen Variante. Existieren verschiedene **Isoformen** (gleiche Funktion, anderer Bau, meist in unterschiedlichem Gewebe zu finden), so vergleichen Sie die aus Muskelgewebe (PFKM oder PFKMA). Wenn kein oder mehrere passende Einträge, dann diejenige Form mit am ehesten vergleichbarer molarer Masse wählen:*

- *Möglichkeit 1: Protein Alignment. Alle Proteine in den Korb befördern. Dann jedes Protein einzeln mit dem humanen Protein vergleichen.*
- *Möglichkeit 2: blastp*

	Bonobo-Schimpanse (*Pan paniscus*)	**Wanderratte *(Rattus norvegicus)***	**Zebrafisch**	**Lanzettfischchen (*Branchiostoma sp.*)**	**ein Fadenwurm *(Caenorhabditis elegans)***	***Escherichia coli*** (Stamm K12)
letzter gemeinsamer Vorfahre lebte vor	5-6 Mio J	100-150 Mio J	200-300 Mio J	500 Mio J	800 – 1000 Mio J	unbekannt
UniProt ID	A0A2R9BRB6					
% Übereinstimmung	100,0%					

4. Weitere Werkzeuge zur Übersetzung DNA ↔ AS-Sequenz

Übersetzen einer Aminosäure-Sequenz in….

… die wahrscheinlichste Nucleotid-Sequenz: EMBOSS Backtranseq: https://www.ebi.ac.uk/Tools/st/emboss_backtranseq/

… alle denkbaren Nucleotidsequenzen: EMBOSS Backtranambig: https://www.ebi.ac.uk/Tools/st/emboss_backtranambig (Achtung: hier Abkürzungs-Verzeichnis beachten: Da eine gegebene AS durch mehrere Tripletts repräsentiert werden kann, sind für eine zusammenfassende Angabe in Form eines einzigen 3-Buchstaben-Codes noch weitere Abkürzungen erforderlich:

Übersetzen von Nucleotid-Abfolge in Aminosäuresequenz: Expasy translate: https://web.expasy.org/translate/

4.1. Aufgabe: Humanes Glucagon

a) Nehmen Sie den Aminosäuresequenz-Abschnitt zwischen der 15. und 21. Aminosäure aus dem humanen Peptidhormon Glucagon und übersetzen Sie es in die wahrscheinlichste Nucleotid-Sequenz.

b) Übersetzen Sie die AS-Sequenz mit einem anderem Werkzeug zurück in eine Aminosäuresequenz.

4.2. Ergänzen Sie die Tabelle unten. Existieren mehrere Möglichkeiten, so wählen Sie „REVIEWED“-Einträge, also solche die verifiziert wurden (goldfarben + Sternchen). Im Zweifelsfall wählen Sie den obersten passenden Eintrag.

Bezeichnung (DE/EN)	Wirts-organismus	UniProt ID	Gen-Name	erste 10 AS (N-Terminus)	letzte 10 AS(C-Terminus)	Länge (AS), Masse (kDa)	Bemerkung (z.B. wichtigste Eigenschaft/Biolog. Funktion)
Thaumatin I	*Thaumatococcus daniellii*						
		P02144					
		P02769					
	Human		ALB				*mit Advanced suchen!*
	Saccharomyces cerevisae						*enthält Fragment* MGRELGE **** mit“Peptide search“ suchen!***
	Human					ca. 26560 Da, 245-248 AS	*enthält Sequenz IAANS*
	Human					ca. 23 kDa	*Sequenzausschnitt* RYTHGRGFIT **** mit“Peptide search“ suchen!***
	Zellulärer oder subzellulärer unbekannter Erreger:			MFVFLVLLPL		ca. 141 kDa	Aus Sputum erkrankter Person isoliert. Ein weiteres nach Trypsin-Verdau gewonnenes Fragment besitzt die Sequenz: GIYQTSNFR

5. Werkzeuge zur Berechnung Fotometrischer Daten anhand der Aminosäuresequenz

Ist die AS-Sequenz eines Proteins oder die UniprotID bekannt, kann anhand der Zusammensetzung der theoretische Absorptionskoeffizient bei 280 nm und andere physikalischer Parameter, wie der theoretische Isoelektrische Punkt berechnet werden. Ein online-Werkzeug hierfür ist:

https://web.expasy.org/protparam/

5.1 Bestimmen Sie den isoelektrischen Punkt und die fotometrischen Daten für das menschliche Myoglobin (P02144)

- Molare Masse:
- molarer Absorptionskoeffizient ε_{280} in L/mol*cm:

  berechnen Sie anhand dieser Daten selbst des spezifischen Absorptionskoeffizienten: ε_{280} in L/g*cm:
- Absorbanz einer Lösung mit 1 g/L: (Prüfen Sie durch Rechnung auf Konsistenz!)
- theoretischer isoelektrischer Punkt:

5.2 Bestimmen Sie die fotometrischen Daten von Hühner-Ovalbum (Ovalbumin Chicken, P02012) durch Eingabe der Aminosäure-Sequenz (z.B. Copy and Paste aus Uniprot heraus):

MGSIGAASMEFCFDVFKELKVHHANENIFYCPIAIMSALAMVYLGAKDSTRTQINKVVRFDKLPGFGDSIEAQCGTSVNVHSSLRDILNQITKPNDVY
SFSLASRLYAEERYPILPEYLQCVKELYRGGLEPINFQTAADQARELINSWVESQTNGIIRNVLQPSSVDSQTAMVLVNAIVFKGLWEKAFKDEDTQAM
PFRVTEQESKPVQMMYQIGLFRVASMASEKMKILELPFASGTMSMLVLLPDEVSGLEQLESIINFEKLTEWTSSNVMEERKIKVYLPRMKMEEKYNL
TSVLMAMGITDVFSSSANLSGISSAESLKISQAVHAAHAEINEAGREVVGSAEAGVDAASVSEEFRADHPFLFCIKHIATNAVLFFGRCVSP

- Molare Masse:
- molarer Absorptionskoeffizient ε_{280} in L/mol*cm:

  berechnen Sie anhand dieser Daten selbst des spezifischen Absorptionskoeffizienten: ε_{280} in L/g*cm:
- Absorbanz einer Lösung mit 1 g/L: (Prüfen Sie durch Rechnung auf Konsistenz!)
- theoretischer isoelektrischer Punkt: