Fylogenetische analyse van hemoglobine

Studentnummer: 1128088 Instituut: Hogeschool Leiden

Auteur: Matthijs de Vries Module leider: Jeroen Pijpe

Datum: 06-12-2024 Module: BSEQAN

Versie: #2



Inhoudsopgave

[Inleiding 3](#_Toc182664651)

[Achtergrondinformatie 3](#_Toc182664652)

[Onderzoeksvraag 3](#_Toc182664653)

[Materiaal & Methoden 4](#_Toc182664654)

[Herkomst 4](#_Toc182664655)

[Achtergrondinformatie 5](#_Toc182664656)

[Scoringsmatrix 5](#_Toc182664657)

[Afstandsmatrix 5](#_Toc182664658)

[UPGMA 6](#_Toc182664659)

[Fylogenetische boom 7](#_Toc182664660)

[Resultaten 9](#_Toc182664661)

[Discussie & Conclusie 10](#_Toc182664662)

[Referenties 12](#_Toc182664663)

[Bijlagen 13](#_Toc182664664)

# Inleiding

## Achtergrondinformatie

Hemoglobine is het meest bestudeerde globuline eiwit, dat een heemgroep bevat. 1 Hemoglobine is de rode kleurstof van bloed, wat in 1840 door Hünefeld toevallig ontdekt was. 2 Hij zag dat gedroogd bloed kristallen vormde. 2 Deze kristallen werden later in 1864 door Hoppe-Seyler “Haemoglobin” genoemd. 2 Rond 1870 werd de rol van zuurstofvervoer ontdekt door Claude Bernard. 2

Cellen hebben zuurstof nodig om voedingsstoffen af te breken en ATP aan te maken. Dit proces wordt cellulaire respiratie genoemd, waarbij koolstofdioxide vrijkomt. ATP is essentieel voor het behouden van cellulaire homeostase.

Hemoglobine is een tetrameer dat bij volwassenen uit twee alfaketens en twee bètaketens bestaat. Hierbij bevat elke keten een heemgroep die een ijzer(II)ion omringt, waaraan zuurstof kan binden. Bij foetussen bestaat hemoglobine uit twee alfaketens en twee gammaketens. Gammaketens hebben een hogere zuurstofaffiniteit dan bètaketens, wat voor een efficiëntere zuurstofoverdracht van moeder naar foetus zorgt.

Hoe al deze verschillende ketens tot stand zijn gekomen, heeft met de duplicatie van het (voorouderlijk) hemoglobine te maken. Hierdoor ontstaat er een "genfamilie", een groep van homologe genen die (meestal) een vergelijkbare functie hebben (Fig. 1).

Afbeelding met tekst, schermopname, diagram, Plan

Automatisch gegenereerde beschrijving

Fig. 1. Voorgestelde "genfamilie" van het (voorouderlijk) globine dat voor globulines met heem codeert, afgeleid van sequentie analyses. Aan de linkerkant is de tijdlijn weergegeven. Bovenin zijn de chromosomen aangegeven waarin de genen zich bevinden. Duplicaties en translocaties hebben tot verschillende varianten van het (voorouderlijk) globine geleid. Uit hemoglobine zijn bijvoorbeeld alfaketens en bètaketens ontstaan. Vervolgens zijn gammaketens uit bètaketens ontstaan. 3

## Onderzoeksvraag

In dit onderzoek wordt een fylogenetische analyse gedaan naar hemoglobine. Hiervoor zijn 5 onbekende hemoglobine sequenties gebruikt van dieren. Het doel is om een fylogenetische boom te reconstrueren, waarbij de evolutionaire relatie van de 5 onbekende hemoglobine sequenties in kaart gebracht wordt.

# Materiaal & Methoden

## Herkomst

De 5 onbekende hemoglobine sequenties werden 1 voor 1 op de website van de NCBI vergeleken met de eiwitten van eiwitdatabases (GenBank, PDB, SwissProt, PIR en PRF). Dit werd gedaan door middel van de tool BlastP. Hiervoor waren de standaardinstellingen gebruikt (Tab. 1).

Tab. 1. BlastP resultaten van de 5 onbekende hemoglobine sequenties. Het nummer, identifier, soort, type en sequentie wordt respectievelijk voor elke sequentie in de tabel weergegeven.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| # | Identifier | Soort | Type | Sequentie |
| Sequentie 1 | XP\_034056698.1 | Gymnodraco acuticeps | subunit beta | MVNWTKTEKATITDIFSHLDYDDIGPKALSRCLIVYPWTQRYFSGFGNLYNAAAIIGNAKVAEHGIKVLH  GLDLGLKKMDNIEAAYADLSSLHSEKLHVDPDNFKLLSDCITIVLAAKLGSAFTAETQATFQKFLGAVMS  ALGKQYH |
| Sequentie 2 | XP\_034056697.1 | Gymnodraco acuticeps | subunit alpha | MSLSDKDKAAVRALWSTISKSSDAIGNDALSRMIVVYPQTKIYFSHWPEVIPGSIHIKEHGKKVMGGIEL  AVSKIDDLKTGLFELSEQHAFKLRVDPGNFKILNHCILVVIATMFPKEFTPEAHVSLDKFLSGVALALAE  RYR |
| Sequentie 3 | NP\_001257812.1 | Canis lupus familiaris | subunit beta-like | MVHLTAEEKSLVSGLWGKVNVDEVGGEALGRLLIVYPWTQRFFDSFGDLSTPDAVMSNAKVKAHGKKVLN  SFSDGLKNLDNLKGTFAKLSELHCDKLHVDPENFKLLGNVLVCVLAHHFGKEFTPQVQAAYQKVVAGVAN  ALAHKYH |
| Sequentie 4 | NP\_001257814.1 | Canis lupus familiaris | subunit alpha-like | MVLSPADKTNIKSTWDKIGGHAGDYGGEALDRTFQSFPTTKTYFPHFDLSPGSAQVKAHGKKVADALTTA  VAHLDDLPGALSALSDLHAYKLRVDPVNFKLLSHCLLVTLACHHPTEFTPAVHASLDKFFAAVSTVLTSK  YR |
| Sequentie 5 | P68068.2 | Ateles geoffroyi | subunit gamma | MSNFTAEDKAAITSLWGKVNVEDAGGETLGRLLVVYPWTQRFFDSFGSLSSPSAIMGNPKVKAHGVKVLT  SLGEAIKNLDDLKGTFGQLSELHCDKLHVDPENFRLLGNVLVTVLAILHGKEFTPEVQASWQKMVAGVAS |

## Achtergrondinformatie

De eerste en tweede sequenties zijn afkomstig van de Gymnodraco acuticeps die van de Antarctische drakenvissen familie afstamt. 4 De derde en vierde sequenties zijn afkomstig van de Canis lupus familiaris dat van de familie hondachtigen afstamt. De laatste sequentie is afkomstig van de Ateles geoffroyi dat van de grijpstaartapen familie afstamt. 5 Deze apensoort wordt beschouwd als een van de meest bedreigde primaten in de wereld. 5 Tot slot hadden alle blast hits een identiteit van 100%.

## Scoringsmatrix

Alle combinaties van de 5 sequenties werden lokaal tegen elkaar uitgelijnd door middel van de blast+ package op Ubuntu. Hierbij werd gebruikgemaakt van de tools makeblastdb en blastp. Met makeblastdb werd een eiwitdatabase aangemaakt, waardoor de eiwitten in het tekstbestand (FASTA-formaat) met de eiwitten in de eiwitdatabase paarsgewijs uitgelijnd konden worden:

makeblastdb -in eiwitten.fa -dbtype prot -out eiwitdatabase

BlastP werd gebruikt om elke combinatie uit te lijnen. Hierbij werd gebruikgemaakt van de BLOSUM45-matrix:

blastp -query eiwitten.fa -db eiwitdatabase -matrix BLOSUM45 -gapopen 12 -gapextend 2

De output van het bovenstaande commando werd op de terminal getoond. Bovenaan elke uitlijning stond de raw-score naast de bit-score tussen haken. Dit werd voor elke uitlijning genoteerd en in een tabel gezet (Tab. 2).

## Afstandsmatrix

Voor het omzetten van de scoringsmatrix naar een afstandsmatrix werd de volgende formule toegepast:

(1)

di, j = afstand tussen sequentie i en sequentie j

Si, j = raw-score van sequentie i tegen j

Si, i = raw-score van sequentie i tegen i

Sj, j = raw-score van sequentie j tegen j

Hierbij werden de berekende afstanden in een tabel geplaatst (Tab. 3).

## UPGMA

Tot slot werden de afstanden in de afstandsmatrix geütiliseerd om met het UPGMA-algoritme de sequenties te clusteren en de verwantschap te achterhalen.

Eerst werd de kortste afstand in de afstandsmatrix bepaald. Hierna werden de twee sequenties die deze afstand bevatten met elkaar geclusterd. Hierbij vormden deze sequenties samen één kolom en één rij. Bij het bepalen van de nieuwe afstanden voor de gevormde cluster werd de volgende formule benut:

(2)

d(i, j), k = afstand tussen een cluster (i, j) en sequentie k

di, k = afstand tussen geclusterde sequentie i en sequentie k

dj, k = afstand tussen geclusterde sequentie j en sequentie k

Hierbij is de cluster (i, j) de afstand van de twee sequenties die het dichtst bij elkaar liggen en k de sequentie die minder aan deze sequenties verwant is. Tot slot werd de bovenstaande formule voor elke sequentie k toegepast die niet geclusterd was.

Hieronder wordt weergegeven hoe het UPGMA-algoritme voor de 5 sequenties was toegepast (Fig. 2).

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | **A** | **B** | **C** | **D** | **E** |
| **A** | - |  |  |  |  |
| **B** | 0,463412 | - |  |  |  |
| **C** | **0,276874** | **0,428832** | - |  |  |
| **D** | 0,542360 | 0,281829 | **0,432451** | - |  |
| **E** | **0,276326** | **0,410357** | **0,062493** | **0,409931** | - |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **A** | **B** | **D** | **E**  **C** |
| **A** | - |  |  |  |
| **B** | 0,463412 | - |  |  |
| **D** | 0,542360 | 0,281829 | - |  |
| **E**  **C** | **0,276600** | **0,419595** | **0,421191** | - |

|  |  |  |  |
| --- | --- | --- | --- |
|  | **B** | **D** | **E**  **C**  **A** |
| **B** | - |  |  |
| **D** | 0,281829 | - |  |
| **E**  **C**  **A** | **0,441504** | **0,481776** | - |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **A** | **B** | **D** | **E**  **C** |
| **A** | - |  |  |  |
| **B** | **0,463412** | - |  |  |
| **D** | **0,542360** | 0,281829 | - |  |
| **E**  **C** | **0,276600** | **0,419595** | **0,421191** | - |

|  |  |  |  |
| --- | --- | --- | --- |
|  | **B** | **D** | **E**  **C**  **A** |
| **B** | - |  |  |
| **D** | **0,281829** | - |  |
| **E**  **C**  **A** | **0,441504** | **0,481776** | - |

|  |  |  |
| --- | --- | --- |
|  | **B**  **D** | **E**  **C**  **A** |
| **B**  **D** | - |  |
| **E**  **C A** | **0,461640** | - |

Fig. 2. Toepassing van het UPGMA-algoritme. Hierbij is de kortste afstand groen gekleurd. De gemiddelden van de blauwe, oranje en paarse afstanden bij de linkertabellen worden in de rechtertabellen ingevoerd.

**Legenda**

A = hemoglobin subunit beta [Gymnodraco acuticeps]

B = hemoglobin subunit alpha [Gymnodraco acuticeps]

C = hemoglobin subunit beta-like [Canis lupus familiaris]

D = hemoglobin subunit alpha-like [Canis lupus familiaris]

E = Hemoglobin subunit gamma [Ateles geoffroyi]

## Fylogenetische boom

Branchlengtes tussen twee sequenties werden bepaald door de kortste afstand door 2 te delen (Fig. 2). Naast de branchlengtes tussen twee sequenties werden ook de internal branchlengtes bepaald. Deze werden berekend door de lengtes van de lange branches minus de lengtes van de korte branches af te trekken:

Internal branch ECA = d(A) – d(EC) = 0,138300 – 0,031247 = 0,107053

Internal branch Ar = d(r) – d(A) = 0,230820 – 0,138300 = 0,092520

Internal branch BDr = d(r) - d(BD) = 0,230820 – 0,140915 = 0,089905

Hierbij is r de totale lengte van de boom.

# Resultaten

In de bijlage zijn de pairwise alignments weergegeven. Hierbij worden per pagina de alignments van 5 subjects met 1 query getoond (inclusief de query met zichzelf).

Hieronder is de scoringsmatrix weergegeven (Tab. 2):

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | **A** | **B** | **C** | **D** | **E** |
| **A** | 1075 | 337 | 506 | 275 | 508 |
| **B** | 337 | 1037 | 362 | 481 | 378 |
| **C** | 506 | 362 | 1061 | 355 | 798 |
| **D** | 275 | 481 | 355 | 1014 | 374 |
| **E** | 508 | 378 | 798 | 374 | 1067 |

Tab. 2. Scoringsmatrix van 5 hemoglobine sequenties.

Hieronder is de afstandsmatrix weergegeven (Tab. 3):

Tab. 3. Afstandsmatrix van 5 hemoglobine sequenties.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | **A** | **B** | **C** | **D** | **E** |
| **A** | - |  |  |  |  |
| **B** | 0,463412 | - |  |  |  |
| **C** | 0,276874 | 0,428832 | - |  |  |
| **D** | 0,542360 | 0,281829 | 0,432451 | - |  |
| **E** | 0,276326 | 0,410357 | 0,062493 | 0,409931 | - |

Door middel van het UPGMA-algoritme zijn de 5 sequenties geclusterd en is hun verwantschap achterhaald. Fylogenetische boom is hieronder weergegeven (Fig. 3):

Afbeelding met tekst, schermopname, diagram, lijn

Automatisch gegenereerde beschrijving

**Legenda**

A = hemoglobin subunit beta [Gymnodraco acuticeps]

B = hemoglobin subunit alpha [Gymnodraco acuticeps]

C = hemoglobin subunit beta-like [Canis lupus familiaris]

D = hemoglobin subunit alpha-like [Canis lupus familiaris]

E = Hemoglobin subunit gamma [Ateles geoffroyi]

Fig. 3.Fylogenetische boom van de 5

hemoglobine sequenties.

# Discussie & Conclusie

Fylogenetische boom is een rooted bifurcating tree. Het bevat 5 terminal nodes, 3 internal nodes en 1 root node. Verder heeft het 5 terminal branches en 3 internal branches. Kortste branch van C en E is een indicatie dat bij deze sequenties de minste evolutionaire veranderingen plaats hebben gevonden. Divergentie heeft bij deze eiwitten later afgespeeld, waardoor ze nauw verwant aan elkaar zijn: 71% identiteit. Het omgekeerde geldt voor B en D waarbij deze eiwitten (waarschijnlijk) de meeste mutaties ondergaan zijn: 46% identiteit.

Opvallend is dat A niet met C geclusterd is, want ze zijn beide bètaketens (E is een gammaketen). Een mogelijke verklaring is dat E en C van dieren afkomstig zijn die nauwer verwant aan elkaar zijn dan A. Dit lijkt aannemelijk aangezien E en C van zoogdieren afkomstig zijn en A van een Antarctische drakenvis. Verder stamt de gammaketen volgens de literatuur van de bètaketen af, waardoor ze erg vergelijkbaar zijn (Fig. 1). 3

Speciatie heeft plaatsgevonden tussen de sequenties C en A, en B en D. Hierbij zijn deze sequenties orthologen van elkaar. Aan de andere kant heeft genduplicatie plaatsgevonden tussen de sequenties A en B, en D en C. Hierbij zijn deze sequenties paralogen van elkaar. Hieronder worden deze "verwantschapstermen" visueel weergegeven (Fig. 4):

Afbeelding met tekst, schermopname, diagram, lijn

Automatisch gegenereerde beschrijving

Ortholoog

Homoloog Paraloog

Ortholoog

Fig. 4.Fylogenetische boom van de 5 hemoglobine sequenties met verwantschapstermen.

Het rekenwerk voor dit onderzoek is handmatig gedaan. Hierdoor is het mogelijk dat er fouten gemaakt zijn. Het onderzoek is daarom 3 keer herhaald, waardoor de kans hierop verkleind is. Verder hebben alle blast hits een identiteit van 100%, dus de kans dat hierbij iets fout is gegaan is nihil.

Verder is het mogelijk dat er een betere matrix is voor de pairwise alignment. Dit komt omdat de keuze voor de BLOSUM-matrices beperkt is. Er kan alleen gekozen worden tussen BLOSUM45, BLOSUM50, BLOSUM62, BLOSUM80 of BLOSUM90. Wellicht is een BLOSUM-matrix van 47 beter voor de pairwise alignment. Dit komt omdat de meeste alignments rond de ~47% liggen, waardoor deze matrix beter kan presteren dan de BLOSUM45- en BLOSUM50-matrix.

Vervolgens is het mogelijk dat er een betere combinatie van gapstraffen is. Dit komt ook doordat er een gelimiteerde keuze was aan gapopen- en gapextend-straffen. Specifieke keuze van gapopen 12 en gapextend 2 was gebaseerd op hoe goed de alignments waren. Aangezien deze waarden voor de beste alignments zorgden (hoogste raw-scores), werden deze geselecteerd.

Resultaten van dit onderzoek worden ondersteund door andere onderzoeken die gedaan zijn naar hemoglobine zoals (Schechter, 2008).3 Uit zijn onderzoek blijkt ook dat alfaketens en bètaketens zijn ontstaan uit hemoglobine, en dat gammaketens van bètaketens zijn afgesplitst.

Een vervolgonderzoek kan gedaan worden naar de onderlinge relatie tussen neuroglobine, cytoglobine, myoglobine en hemoglobine. Neem van elk eiwit 5 samples van verschillende dieren en maak daarvan een fylogenetische boom met een tijdlijn. Hierbij mag gebruikgemaakt worden van de 5 hemoglobine samples in dit onderzoek. Het vervolgonderzoek richt zich op het beantwoorden van de volgende onderzoeksvraag: wat is de evolutionaire relatie tussen neuroglobine, cytoglobine, myoglobine en hemoglobine? Relevantie van dit vervolgonderzoek ligt in het verduidelijken van de evolutie van het (voorouderlijk) globuline eiwit met een heemgroep.

Concluderend heeft hemoglobine zich eerst in alfaketens en bètaketens afgesplitst, waarna de gammaketen zich van de bètaketens heeft afgesplitst.

# Referenties

1. Ahmed MH, Ghatge MS, Safo MK. Hemoglobin: Structure, Function and Allostery. *Subcell Biochem*. 2020;94:345. doi:10.1007/978-3-030-41769-7\_14

2. Saha D, Patgaonkar M, Shroff A, Ayyar K, Bashir T, Reddy KVR. Hemoglobin Expression in Nonerythroid Cells: Novel or Ubiquitous? *Int J Inflamm*. 2014;2014:1-8. doi:10.1155/2014/803237

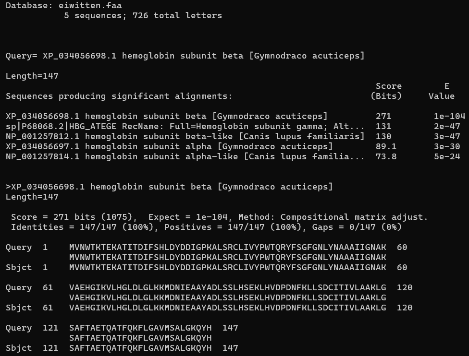
3. Schechter AN. Hemoglobin research and the origins of molecular medicine. *Blood*. 2008;112(10):3927-3938. doi:10.1182/blood-2008-04-078188

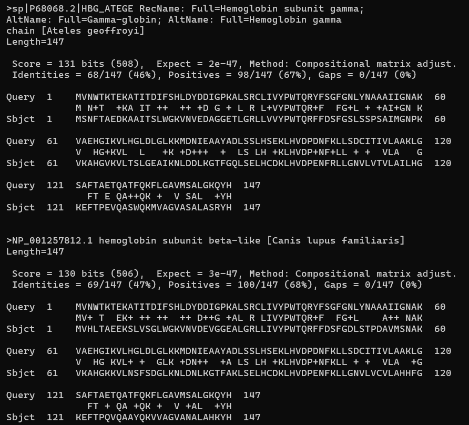
4. Song W, Li L, Huang H, et al. Isolation and characterization of the mitochondrial genome of *Gymnodraco acuticeps* (Perciformes: Bathydraconidae) with phylogenetic consideration. *Mitochondrial DNA Part B*. 2017;2(2):526-527. doi:10.1080/23802359.2017.1361361

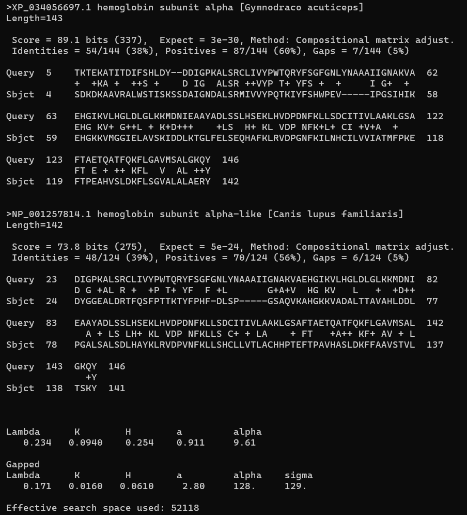
5. Girón-Galván LE, Zaldaña-Orantes K, Sanchéz-Trejo L, Castillo E, Rodríguez ME. Conservation and research efforts for Geoffroy’s spider monkey (*Ateles geoffroyi*) in El Salvador. *Neotropical Primates*. 2023;29(1):92-96. doi:10.62015/np.2023.v29.779

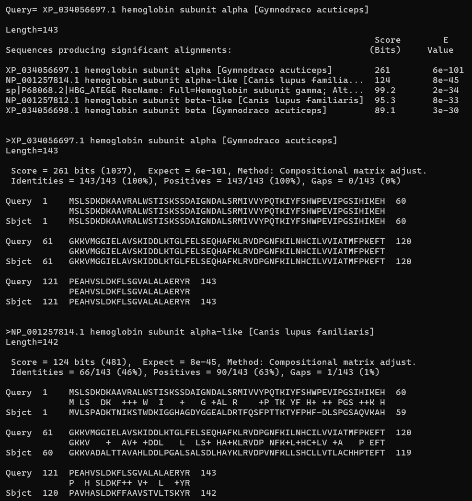
# Bijlagen

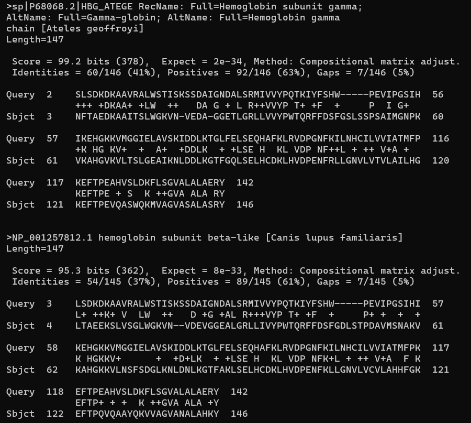
Pairwise alignments

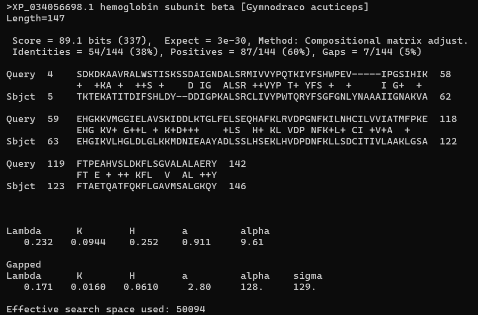


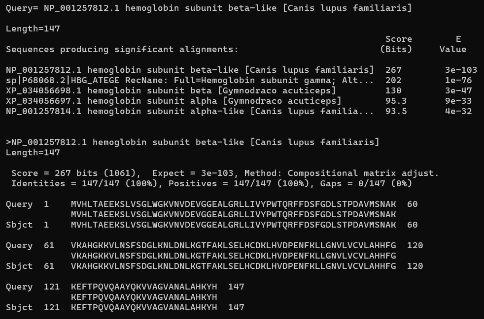
****

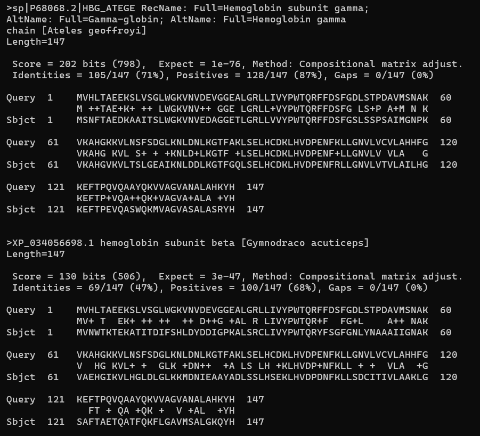


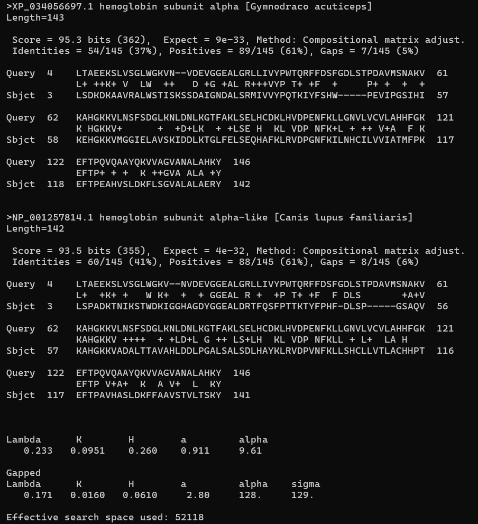








****

****

