Sprawozdanie nr 2

Analiza sekwencji i struktury drugorzędowej białek

Małgorzata Stęperska 151546, Adam Dachtera 147890

II. Algorytm Needlemana-Wunscha

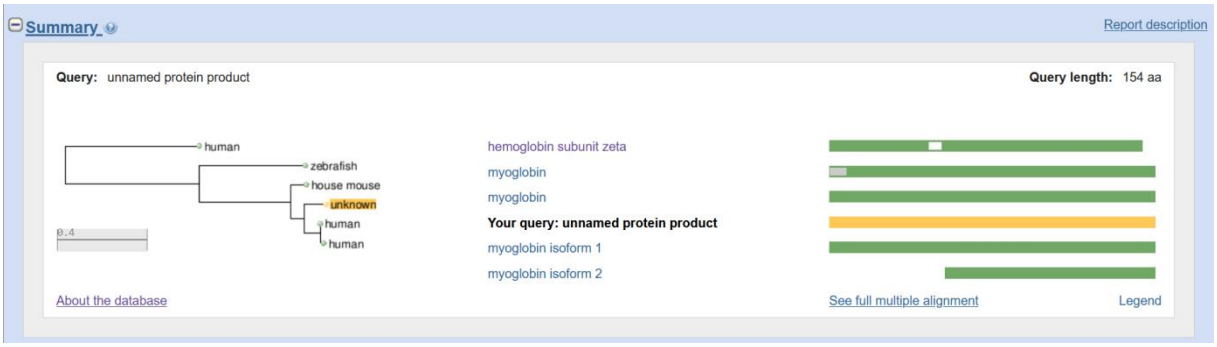
Link do repozytorium projektowego: GitHub - MSteperska/BIOS

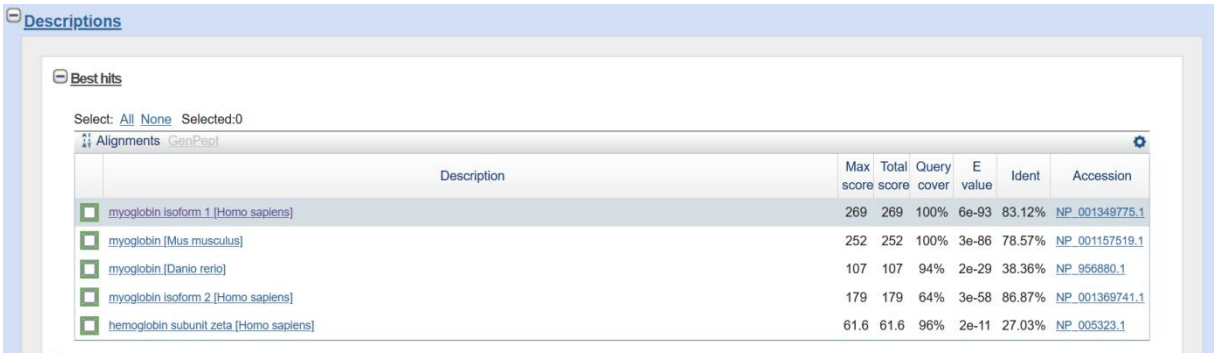
I. BLAST

W miejscu wyszukiwania wpisaliśmy „myoglobin”, dzięki czemu znaleźliśmy strukturę 1MCY (sperm whale myoglobin (mutant with initiator met and with his 64 replaced by gln, leu 29 replaced by phe)).

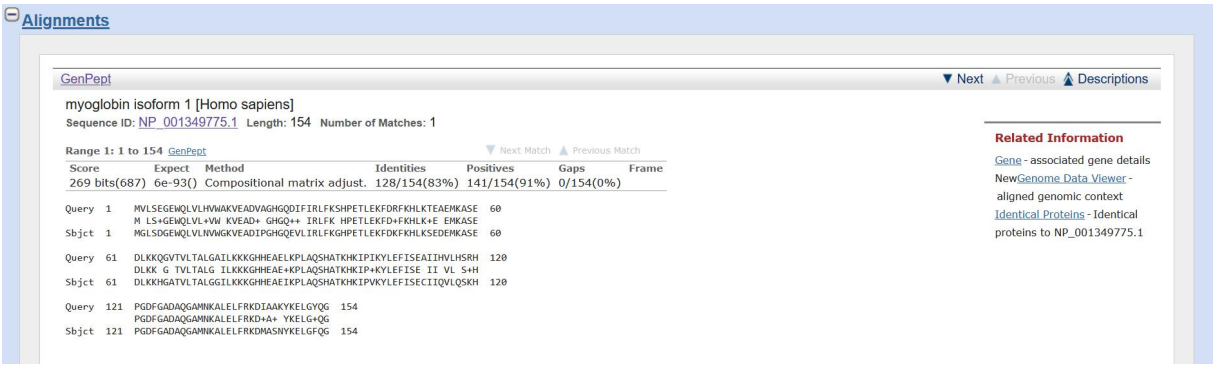
Następnie przy pomocy narzędzia blastp wyszukaliśmy sekwencji homologicznych do sekwencji z pliku .fasta dla struktury 1MCY pobranego z bazy PDB.

Wyniki i ich interpretacja





Description	Max score	Total score	Query cover	E value	Ident	Accession
myoglobin isoform 1 [Homo sapiens]	269	269	100%	6e-93	83.12%	NP_001349775.1
myoglobin [Mus musculus]	252	252	100%	3e-86	78.57%	NP_001157519.1
myoglobin [Danio rerio]	107	107	94%	2e-29	38.36%	NP_956880.1
myoglobin isoform 2 [Homo sapiens]	179	179	64%	3e-58	86.87%	NP_001369741.1
hemoglobin subunit zeta [Homo sapiens]	61.6	61.6	96%	2e-11	27.03%	NP_005323.1



Wynikiem działania narzędzia jest pięć najlepszych dopasowań. Jeśli jest taka możliwość, każde dopasowanie pochodzi z innego organizmu. Rezultaty przedstawione są w postaci drzewa filogenetycznego oraz graficznej, w postaci diagramu słupkowego (na schemacie po prawej stronie). Dodatkowo SmartBLAST pokazuje dopasowania z bazy konserwatywnych domen. Odniesienia w środkowej części przenoszą do sekcji *alignments,* gdzie możemy uzyskać informacje o dopasowaniu podanych sekwencji.
Ogólny graficzny zarys wyników (po prawej) należy odczytywać następująco: na zielono zaznaczone są landmark matches, dopasowania z nieredundantnej bazy białek określa kolor niebieski, kolorem szarym zaznaczone są niedopasowania, a kolorem białym przerwy. Sekwencja wprowadzona na wejściu zaznaczona jest kolorem żółtym.
W sekcji *descriptions* w formie tabelarycznej zebrane zostały najważniejsze informacje o dopasowaniach.

Max score – Najwyższy wynik dopasowania. Im wyższy wynik, tym lepsze dopasowanie.
Total score – Suma wyników dopasowania. Im wyższy wynik, tym lepsze dopasowanie.
Coverage – Określa procentową ilość sekwencji badanej, która pasuje/ pokrywa się z sekwencją z bazy NCBI. Wysoki coverage wskazuje, że duża część sekwencji badanej pasuje do sekwencji docelowej.
e-value – Oczekiwana liczba pozytywnie fałszywych wyników w porównaniu do bazy danych. Im wartość bliższa 0, tym wynik porównania sekwencji jest bardziej istotny.

ident – Procentowy udział zasad, które są identyczne z genomem referencyjnym.

Stwierdziliśmy pełne porycie podanej sekwencji z Myoglobin isoform 1 znalezionej w organiźmie ludzkim. Wiele z występujących różnic (na przykład zamiana izoleucyny na walinę) nie wpływa na funkcjonalność cząsteczki (w podanym przykładzie obie mają grupę hydrofobową). Podobnie stwierdziliśmy w przypadku mioglobiny pochodzącej z organizmu myszy, ale w jej przypadku procent identyczności był niższy.

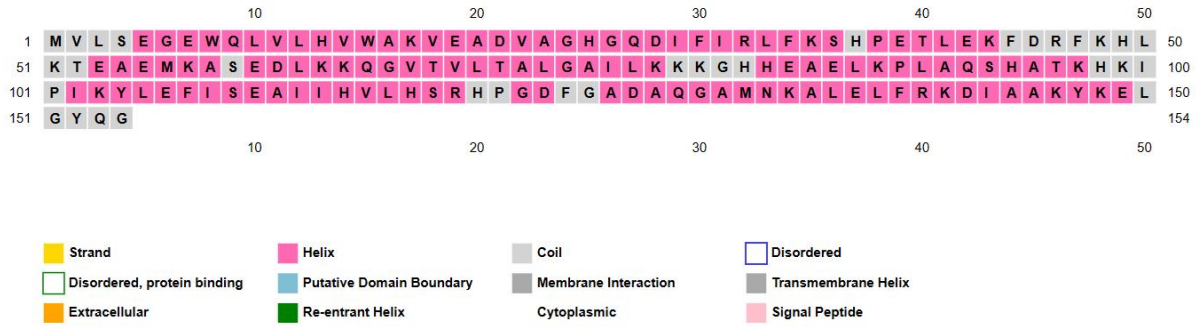
III. STRUKTURA DRUGORZĘDOWA BIAŁEK

```
# PSIPRED VFORMAT (PSIPRED V4.0)

  1 M C   0.999  0.001  0.000
  2 V C   0.975  0.008  0.001
  3 L C   0.984  0.007  0.001
  4 S C   0.977  0.017  0.001
  5 E H   0.024  0.980  0.000
  6 G H   0.018  0.985  0.000
  7 E H   0.008  0.993  0.000
  8 W H   0.004  0.997  0.000
  9 Q H   0.003  0.998  0.000
 10 L H   0.003  0.998  0.000
```

Pierwsza kolumna to indeks aminokwasu, kolejna to jednoliterowe oznaczenie aminokwasu. W kolumnie trzeciej znajduje się informacja o strukturze drugorzędowej. Ostatnie trzy kolumny to kolejno prawdopodobieństwo struktury: coil, helix, sheet.

Zgodnie z wynikami struktura drugorzędowa badanego białka składa się z



naprzemiennie występujących: helisy i coil. Prawdopodobieństwa tych struktur drugorzędowych w środku sekwencji aminokwasowej dość mocno spadają (momentami nawet <50%). Poza tym utrzymują się na dość wysokim poziomie (min. 89%).
Poniżej znajduje się schemat, który graficznie przedstawia strukturę drugorzędową.

6. Charakterystyka nazwa kolumn wykorzystywana do opisu aminokwasów wchodzących w skład analizowanej struktury

RESIDUE – dwukolumnowa wartość ilość reszt, pierwsza z nich tyczy się sekwencji razem z przerwami i służy do identyfikacji pozycji na danym łańcuchu, a druga wartość jest referencyjna dla całego dopasowania.

AA – jednoliterowy kod aminokwasowy, dla mostków disiarczkowych cysteiny używane są kolejne małe litery alfabetu, dla obu reszt.

STRUCTURE – informacje o strukturze wtórnej białek na podstawie analizy współrzędnych atomów. W wyniku można otrzymać oznaczenia takie jak: H dla α-helis, E dla β-kartek.

BP1, BP2 – numery pierwszej i drugiej reszty aminokwasowej pomiędzy którymi występują mostki wodorowe.

ACC – dostępność powierzchniowego obszaru dla danego aminokwasu w analizowanym fragmencie białka (mierzona w $Å^2$). Dla aminokwasów wewnętrznych wartość jest niska, ponieważ aminokwasy nie mają dostępu do otaczającej je wody ani innych czynników środowiskowych. Aminokwasy na powierzchni będą miały wyższą wartość ACC ze względu na większy dostęp do otoczenia.

N-H-->O – odległość między atomem azotu w jednym aminokwasie a atomem tlenu w innym (związane z mostkami wodorowymi).

O-->H-N – ta kolumna zawiera informacje o atomie tlenu w jednym aminokwasie, który połączony jest mostkiem wodorowym z atomem wodoru związanym z atomem azotu w innym aminokwasie.

N-H-->O – analogicznie do poprzednich dwóch kolumn; mostki wodorowe, w których atom azotu jest donorem wodoru a atom tlen - akceptorem

O-->H-N – atom azotu jest donorem wodoru, atom tlenu – akceptorem.

TCO – określa chwilowy stopień skręcenia α-helis w analizowanym fragmencie białka.

KAPPA – kąt pomiędzy trzema atomami Cα reszt I-2, I, I+2 (I – dowolna reszta). Używane do definiowania zakręcenia α-helisy.

ALPHA – kąt torsyjny zdefiniowany na czterech atomach Cα reszt (I-1, I, I+1, I+2). Używane do definiowania chiralności.

PHI, PSI – kąty torsyjne aminokwasów w danej sekwencji.

X-CA, Y-CA, Z-CA – współrzędne atomu Cα aminokwasu w danej sekwencji.

7. Znaczenie symboli opisujących strukturę drugorzędową białek:

C – coil
H – α-helix
T– β-turn
S – bend
G – 3_{10} -helix
I – π-helix
E – β-strand
B – β-bridge

C. Porównaj ze sobą powyższe struktury drugorzędowe (2D) białek, a mianowicie przewidzianą na podstawie znanej sekwencji (A) z wyekstrahowaną z referencyjnej struktury 3D (B) poprzez wyznaczenie wartości współczynnika identyczności (SSI). Czy porównywane struktury 2D są do siebie podobne? Odpowiedź uzasadnij

Seq – badana sekwencja aminokwasowa

SS – predykcja DSSP

PSI – predykcja PSIPRED

Na zielono zaznaczone zostały niedopasowania:

```
Seq: MVLSEGEWQL VLHVWAKVEA DVAGHGQDIF IRLFKSHPET LEKFDRFKHL
SS:  ccccHHHHHH HHHHHHHHGG GHHHHHHHHH HHHHHHcGGG GGGcTTTTTc
PSI: CCCCHHHHHH HHHHHHHHHH HHHHHHCHHH HHHCCCCCCC CCHHHHHHCC

Seq: KTEAEMKASE DLKKQGVTVL TALGAILKKK GHHEAELKPL AQSHATKHKI
SS:  cSHHHHHHcH HHHHHHHHHH HHHHHHHHTT TccHHHHHHH HHHHHHTScc
PSI: CCHHHHHHCH HHHHHHHHHH HHHHHHHHCC CCHHHHHHHH HHHHHHHCCC

Seq: PIKYLEFISE AIIHVLHSRH PGDFGADAQG AMNKALELFR KDIAAKYKEL
SS:  cHHHHHHHHH HHHHHHHHHc GGGcSHHHHH HHHHHHHHHH HHHHHHHHHH
PSI: CHHHHHHHHH HHHHHHHHHC CHHCCHHHHH HHHHHHHHHH HHHHHHHHHC

Seq: GYQG
SS:  Tccc
PSI: CCCC
```

Dopasowania: 154-32 = 122

$$SSI = \frac{122}{154} \approx 0{,}792 = 79{,}2\%$$

Wysoki współczynnik identyczności oznacza, że większość predykcji z obu metod pokrywa się, a więc są to mocno dopasowane i zbieżne sekwencje. Może to wskazywać na zbieżność działania obu narzędzi, ale większe zróżnicowanie i występowanie rzadszych stuktur może wskazywać na wyższą dokładność metody DSSP.