BIOINFORMATYKA - ZADANIA

BAZY DANYCH I SERWERY BIOINFORMATYCZNE

National Center for Biotechnology Information
https://www.ncbi.nlm.nih.gov/gquery/gquery.fcgi

The European Molecular Biology Open Software Suite:
http://emboss.sourceforge.net/

EMBL-EBI Services
http://www.ebi.ac.uk/services

Mobyle
http://mobyle.pasteur.fr/cgi-bin/portal.py#welcome

PUBMED

1. Wyszukaj publikacje w PubMed nt. dUTPase (www.ncbi.nlm.nih.gov/entrez). Ile wyników wyświetla się dla szukanej frazy?

2. Otwórz dowolną publikację w nowej karcie. Kiedy i gdzie została opublikowana? Czy jest dostęp do całego artykułu?

3. Zapisz artykuł (streszczenie) na dysku (send to file) i wyślij na swój adres e-mail (send to e-mail).

4. Dołącz artykuł (streszczenie) do „collections" (send to collections) – załóż konto NCBI.

5. Wróć do strony z wynikami o dUTPase i wyszukaj prace Abergela o dUTPase (napisz dwa hasła obok siebie).

6. Zapisz na lokalnym dysku darmowy artykuł Abergela pt. „Hidden ....." w formacie PDF

7. Zapoznaj się z organizacją bazy MEDLINE: Wyświetl streszczenie pracy Abergela i ustaw Display=MEDLINE. Zauważ, że tytuł publikacji znajduje się w polu TI. Jakie dwa pola zawierają adres laboratorium i nazwiska autorów?

8. Radzenie sobie z popularnymi hasłami jak np. Down, które może oznaczać syndrom, nazwisko lub ulicę. Wyszukaj hasło Down pojawiające się w tytule (wpisz Down [TI]), a następnie w adresie [AD] i wśród autorów [AU].

9. Czy ktoś z Polski pisał o dUTPase? (przeszukaj pola streszczeń [AB] i adresów [AD])

10. Wyszukiwanie nowych artykułów przeglądowych (Review) nt. dUTPazy: wyszukaj hasło dUTPase, w Filtrach ustaw Published in the Last =10 years, Languages=English, Type of Article=Review, Field Tag=Title/Abstract. Ile wyników się wyświetliło?

11. Wyszukaj artykuły o miRNA opublikowane od 01.01.2012. Sprawdź ile jest darmowo dostępnych całych artykułów. Czy ten filtr ma jakieś znaczenie?

12. NCBI udostępnia również darmowe książki. Zajrzyj do Bookshelf. Ile jest dostępnych książek z medycyny?

ŹRÓDŁA DANYCH O SEKWENCJACH NUKLEOTYDOWYCH

13. Szereg instytucji przyczynia się do rozwoju bioinformatycznych baz danych i narzędzi do analizy danych biologicznych. Korzystając z wyszukiwarki spróbuj rozszyfrować skróty nazw instytucji i odpowiedzieć na pytanie, gdzie mieszczą się i z jakich środków są finansowane: NCBI, EMBL-EBI, Wellcome Trust Sanger Institute.

14. NCBI prowadzi wiele baz danych. Korzystając z informacji zawartych pod adresem http://www.ncbi.nlm.nih.gov/guide odpowiedz na pytania: (a) Ile różnych baz nukleotydowych (DNA+RNA) jest prowadzonych przez NCBI? (b) Co zawiera baza GenBank? (c) Co zawiera baza RefSeq? (d) Co zawiera baza UniGene? (e) Co to jest PubMed?

15. Tradycyjnie, czasopismo Nucleic Acids Research poświęca swój pierwszy zeszyt w roku biologicznym bazom danych.
(a) Sprawdź, czy w najnowszym wydaniu znajdziesz nowe opracowania o GenBank.
(b) Czasopismo przygotowuje również listę on-line ważniejszych baz danych. Ile jest obecnie baz danych na tej liście?

16. Co to jest Ensembl i skąd czerpie środki na swoją misję?

17. Ensemble dostarcza baz danych dla wielu gatunków. Czy główny opis genomu odbywa się automatycznie czy ręcznie?

18. Ensembl dostarcza rozwiązań informatycznych dla zarządzania wieloma bazami danych. Strona 'Projects using Ensembl' wymienia aktualne' projekty wspomagane przez Ensembl. Jest wśród nich znany projekt "1000 genomes". Jaka jest misja tego projektu? Czy dane zebrane w ramach tego projektu są publicznie dostępne?

19. Połącz się z Entrez - systemem wyszukiwania informacji w NCBI. Wyszukaj informacje o sekwencji U49845. Przyjrzyj się wpisowi w formacie GenBank i odpowiedz na pytania:
(a) Jak długa jest sekwencja nukleotydowa?
(b) Od jakiego organizmu pochodzi? Czy jest to sekwencja genomowa czy też może RNA?
(c) Czym różni się numer ACCESSION od VERSION?
(d) Czy ta sekwencja była kiedykolwiek poprawiana?
(e) Kto opisał tę sekwencję i gdzie można znaleźć publikację towarzyszącą temu zgłoszeniu (sekwencji)?
(f) Jaki jeszcze numer pojawia się w polu VERSION?
(g) Jaką część sekwencji opisano jako gen (ang. gene)? Jak długi jest ten fragment? I jaką nazwę genu przypisano temu fragmentowi?
(h) Czy fragment genomu opisany jako gen AXL2 rozciąga się poza fragment 617..3158 oglądanej sekwencji?
(i) Ile w oglądanej sekwencji zidentyfikowano fragmentów kodujących? Podaj ich łączną długość.
(j) Który fragment kodujący jest odczytywany z komplementarnej nici?
(k) Czy pierwszy fragment kodujący zawiera kodon start (ATG)? Czy drugi również go zawiera?
(l) Czy ostatni fragment kodujący zawiera kodon STOP (TAG, TAA lub TGA)? Czy drugi również zawiera taki kodon?
(m) Jaką długość ma sekwencja aminokwasowa pochodząca z drugiego CDS? Jaki jest pierwszy aminokwas?
(n) Czy przypisana sekwencja białkowa została eksperymentalnie potwierdzona, czy też jest to tylko tłumaczenie komputerowe?
(o) Czy cecha sekwencji opisana jako mRNA rozciąga się poza cechę CDS? Dlaczego?
(p) Przełącz się na widok w formacie FASTA. Spróbuj rozszyfrować, co wpisano jako opis sekwencji.

20. Spróbuj utrwalić sekwencję z poprzedniego zadania w pliku tekstowym w formacie FASTA. Wskazówka: użyj linku SEND.

21. Wyszukaj informacji o bazie sekwencji dotyczącej psa. (a) Kiedy zbudowano bazę dla genomu psa? (b) Ile zawiera genów i transkryptów?

22. Warianty genu IGF1 (insulinopodoby czynnik wzrostu 1) decydują o wielkości psów. Znajdź informację o genie IGF1 psa w bazie Ensemble.
(a) Na ile jest on identyczny z genem człowieka?
(b) Jaki gen jest paralogiem genu IGF1?
(c) Czy są znane warianty strukturalne genu IGF1?

ANALIZA POJEDYNCZEJ SEKWENCJI DNA

Słowa

23. Pobierz sekwencję NT_007933.15. Uruchom program geecee (EMBOSS). (a) Jaka jest zawartość dinukleotydu CG w całej sekwencji?

24. Pobierz sekwencję NT_007933.15. Wejdź na stronę www.genomatix.de/cgi-bin/tools/tools.pl. Wybierz Create Sequence Statistics i wklej sekwencję w okno sequence input. Załaduj sekwencję i rozpocznij analizę. Odpowiedz na pytania (a) Jaka jest zawartość nukleotydów GC? (b) Ile jest par GC w sekwencji? (c) Która trójka nukleotydów występuje najczęściej?

25. Sekwencja Cross-over hotspot instigator (Chi) (5'-GCTGGTGG-3') licznie występują u E. coli i miejscowo sprzyja rekombinacjom. Zlicz słowa (Word size 8) w sekwencji U00096.2 z pomocą programu wordcount (EMBOSS).

26. Przeanalizuj sekwencję U00096.2 programem compseq. Jaki jest stosunek obserwowanej liczby słowa GCTGGTGG do oczekiwanej?

27. Analiza słów w dwóch podobnych genomach może ujawnić zadziwiające różnice. Genomy Lactococcus lactis i Stroptococcus pyogenes mają z grubsza podobną wielkość i zawartość GC. Przaanalizuj sekwencję genomową Lactococcus lactis i Stroptococcus pyogenes (NC_009004.1, NC_002737.1) pod kątem występowania słów o długości 13 nukleotydów i opisz wyniki.

Wyspy CpG

28. Pobierz sekwencję NT_011511. Uruchom program geecee z serwera (EMBOSS). Jaka jest zawartość dinukleotydu CG całej sekwencji?

29. Przy pomocy programu cgplot spróbuj zidentyfikować wyspy CpG w sekwencji NT_011511. (a) Ile wysp znaleziono i o jakich długościach? (b) Ile wysp można zidentyfikować przyjmując, że wyspa powinna mieć co najmniej 500pz.

Wykorzystanie kodonów

30. Odwiedź bazę Codon Usage Database. Jaki jest najczęściej wykorzystywany kodon dla izoleucyny przez E.coli?

31. Porównaj wartosci CAI (codon adaptation index) dla CDS kodujących białka P68919 i C4ZQ55 E.coli. Statystykę CAI liczy program codonw (mobyle.pasteur.fr). Kiedy wartość CAI osiągnie maksymalną wartość 1?

Powtórzenia w sekwencji DNA

32. Pobierz z pliku sekwencję „dot_plot". Wykorzystaj metodę DotPlot (program Dotlet). Co możesz powiedzieć o liczbie i wielkości powtórzeń w tej sekwencji?

33. Pobierz z GenBanku sekwencję NT_033778 z zakresu (Range) 11,760,176-11,763,000. Co możesz powiedzieć o liczbie i wielkości powtórzeń w tej sekwencji? Ile jest powtórzeń odwróconych?

34. Sekwencję NT_033778 przeanalizuj programem etandem (EMBOSS). Ile zidentyfikowałeś powtórzeń 6-nukletydowych?

35. Czy w sekwencji NW_001504437 istnieją jakieś powtórzenia skatalogowane w bazie RepBase. Skorzystaj z www.repeatmasker.org. (a) Ile w tej sekwencji zamaskowano nukleotydów? (b) Jakiego typu powtórzenia występowały najczęściej? (c) Z powodu jakich sekwencji zamaskowano najwięcej nukleotydów?

Enzymy restrykcyjne

36. Połącz się z http://rna.lundberg.gu.se/cutter2/. Dokonaj symulacji trawienia sekwencji z pliku VecScreen-sample1 enzymem HindIII. (a) Ile fragmentów powstanie i o jakiej długości? (b) Opisz dokładnie miejsce cięcia i sekwencję rozpoznawaną przez ten enzym.

37. Dla sekwencji VecScreen-sample1 dokonaj symulacji trawienia enzymem CfrI.(a) Ile fragmentów powstanie i o jakiej długości? (b) Opisz dokładnie miejsce cięcia i sekwencję rozpoznawaną przez ten enzym (c) Który enzym lepiej zastosować do genotypowania (CfrI czy HindIII) (d) Ile jest enzymów, które tną tę sekwencję dokładnie w trzech punktach? (e) Jaka będzie mapa restrykcyjna, jeżeli zastosujesz jednocześnie dwa enzymy Sse9I i BsrI?

38. Pobierz z bazy NCBI sekwencję 3 exonu leptyny (ang. leptin) człowieka. (a) Przeprowadź symulację cięcia enzymem MspI. Spróbuj narysować jaki obraz powstałby na żelu agarozowym pod wpływem cięcia tym enzymem. (b) Zapisz komputerową wizualizację rozkładu prążków w 2% żelu agarozowym po trawieniu exonu 3 genu LEP człowieka enzymem MspI. W tym celu połącz się z rebase.neb.com i wykorzystaj narzędzie RebSites.

PIERWOTNE REPOZYTORIA SEKWENCJI AMINOKWASOWYCH

39. Jedną z najbardziej wartościowych baz sekwencji aminokwasowych jest UniProtKB/Swiss-Prot. Baza jest kolekcją ręcznie opisanych i sprawdzonych sekwencji pochodzących z szerszej bazy UniProt Knowledgebase (UniProtKB). Cechuje ją wysoka jakość, brak nadreprezentacji sekwencji i dobry opis utworzony na podstawie wyników eksperymentalnych i metod obliczeniowych.
(a) Kiedy wydano najnowszą wersję bazy UniProtKB/Swiss-Prot?
(b) Ile zawiera wpisów o sekwencjach?
(c) Ile gatunków jest reprezentowanych w bazie?

40. Obok UniProtKB/Swiss-Prot istnieje UniProtKB/TrEMBL. Jest to kolekcja białek opisana jak dotąd tylko automatycznie. Z uwagi na pracochłonność 'ręcznego' opisu i weryfikacji sekwencji, nie wszystkie sekwencje mogą być reprezentowane w UniProtKB/Swiss-Prot.
(a) Ile wpisów zawiera baza UniProtKB/TrEMBL?
(b) Jakie są 3 najbardziej reprezentowane gatunki w bazie UniProtKB/TrEMBL?
(c) Czy w ostatnich latach tempo przyrostu bazy spada?

DOPASOWANIE PARY SEKWENCJI

Dopasowanie (ułożenie) pary sekwencji jest podstawową procedurą bioinformatyczną. Pozwala m.in. ocenić podobieństwo sekwencji, dociekać o spokrewnieniu (homologii) sekwencji i ich ewolucji, opisać zmienność sekwencji.

41. Znajdź najlepsze dopasowanie globalne dwóch sekwencji aminokwasowych, ortologów: IGF1 człowieka oraz IGF1 psa. Zastosuj program needle (EMBOSS). W jaki sposób program obliczył różne miary podobieństwa (alignement score, identity, similarity)? A teraz porównaj dwa paralogi: IGF1 człowieka z IGF2 człowieka. Która para jest bardziej podobna: ortologów czy paralogów? Jak wyjaśnisz różnice w podobieństwie?

42. Znajdź najlepsze miejscowe (lokalne) dopasowanie dwóch sekwencji np. AAF75024 oraz NP_040628. W zadaniu możesz się posłużyć programem LALIGN z pakietu FASTA lub water (EMBOSS). zastosuj macierz PAM120 oraz kary za wprowadzenie przerwy i jej poszerzenie 11 i 8, odpowiednio. Porównaj wynik z dopasowaniem globalnym.

43. Dwie spokrewnione, ale niezbyt podobne sekwencje mogą być trudne do optymalnego ułożenia, gdyż ułożenie może zależeć od drobnych zmian w systemie punktacji, np. od wielkości kary za przerwę. Białka RECA_ECOLI i RAD51_DROME mają podobną funkcję - łączą dwie pojedyncze homologiczne nici DNA. Mają też podobną strukturę trzeciorzędową. Dopasuj te dwie sekwencje przy zastosowaniu małej i dużej kary za przerwę (gap penalty). Zastosuj program LALIGN z pakietu FASTA. Wykonaj dwa porównania. W pierwszym zastosuj kary -12 (gap) i -2 (extend), a w drugim -5 i -1. Czy te sekwencje są łatwe do dopasowania?

44. Jedną z metod sprawdzenia, czy dwie sekwencje są statystycznie istotnie podobne jest wielokrotne porównanie jednej sekwencji z sekwencjami losowymi powstałymi przez przetasowanie aminokwasów drugiej sekwencji. W ten sposób uzyskujemy przybliżony rozkład prawdopodobieństwa punktacji podobieństwa między losowymi niespokrewnionymi białkami o składzie aminokwasów identycznym z sekwencjami, których podobieństwo rozpatrujemy. Posłuż się programem PRSS/PRFX (pakiet FASTA) by porównać sekwencje RECA_ECOLI i RAD51_DROME (Swiss-Prot). Oszacuj istotność dopasowania dwóch sekwencji (E-wartość).

45. Składanie sekwencji. Pobierz plik 6reads.txt z 6. odczytami z sekwenatora. Korzystając z programu CAP3 (http://pbil.univ-lyon1.fr/cap3.php) zbuduj contig. Jaka jest długość contigu?

HOMOLOGIA SEKWENCJI (BAZA HOMOLOGENE)

46. Ile grup sekwencji homologicznych wyróżniono w bazie HomoloGene dla człowieka?

47. Przeszukaj bazę HomoloGene pod kątem genu IGF2. (a) Na jakich chromosomach leżą geny Igf2a i Igf2b gatunku Danio reiro? (b) Jak nazwałbyś tę formę homologii (ksenologia/ortologia?)? (c) Porównaj sekwencje białkowe odpowiadające Igf2a i Igf2b. Jaki jest % identyczności?

48. Ile gatunków jest reprezentowanych w grupie HomoloGene 56594, a ile w grupie 115815? Ile par ortologicznych i ile paralogicznych zawiera grupa 88747?

49. Zapoznaj się z procedurą tworzenia bazy HomoloGene (build procedure). Czy punktem wyjścia do tworzenia bazy jest porównanie sekwencji DNA czy sekwencji białkowych?

50. Zapoznaj się z grupą HomoloGene 17097. Jakie podobieństwo sekwencji panuje w tej grupie? Czym to wyjaśnisz?

51. Czy elementy przenośne genomu człowieka TIGD1 i TIGD3 są homologami?

PRZESZUKIWANIE BAZ DANYCH O SEKWENCJACH

52. Pamiętaj, przeszukując duże bazy danych niemal zawsze znajdziesz sekwencje podobne do sekwencji w pytaniu. Statystyki związane z wyszukanymi sekwencjami pomogą rozstrzygnąć, czy dane ułożenie świadczy o homologii czy też jest przypadkowe. Korzystając z blastp przeszukaj nienadmiarową białkową bazę danych (nr) pod kątem sekwencji bez znaczenia biologicznego: CAPTAINKIRK. Zwróć uwagę na E-wartość.

53. Program blastp może posłużyć do zidentyfikowania białka. (a) Korzystając z bazy nr zidentyfikuj białko świni (Sus scrofa) na podstawie poniższej sekwencji. (b) Porównaj sekwencję do bazy PDB, aby dowiedzieć się czy białko to ma znaną strukturę 3D.

MLLLGAVLLLLALPSLGQETTEKPGALLPMPKGACAGWMAGIPGHPGHNGTPGRDGRDGV
PGEKGEKGDTGLTGPKGDTGESGVTGVEGPRGFPGIPGRKGEPGESAYVYRSAFSVGLET
RVTVPNNMPIRFTKIFYNQQNHYDVTTGKFHCNIPGLYYFSFHITVYLKDVKVSLYKKDKA
VLFTYDQYQDKNVDQASGSVLLYLEKGDQVWLQAYGDEENNGVYADNVNDSIFTGFLLYH
NIE

54. Poniżej podano nie opisaną sekwencję EST (znacznik sekwencji, która podlega ekspresji). Nie wiadomo, czy jest to sekwencja kodująca białko i jak przebiega ramka odczytu. Dla podanej sekwencji nukleotydowej spróbuj znaleźć w bazie sekwencji aminokwasowych białka homologiczne do białka, które może ta sekwencja kodować (zastosuj blastx). Z jakim znanym białkiem ta sekwencja może być powiązana?

CTTCACTCTCATTGCATGTAGTGAGGACCAAGGTGAAGGACCCTTCTCAAACCATGGTGA
TTAGGATCAATAAGGTAATCAACCAATCTATCAATCAGTCATCCAGTCAGTCAATCATTA
GTGTACTGATTTATGCAGTCTTTAGCTTGTTGATGAATTAAACAAACAAAAAAACAGCAT
CTATAGCCTACATTAATCGTATATTAATCACATTTATCTCTTCCTCCAGCTGGATATATC
TCCTAACCTAATCGAGGGCATGGACCCATTTCTCCCTGCGGCTGCAGTAGATCAGCATAT
AGAAAGTCTGCCTTCCATAATGGAGACCATGGGGTTCTACCAGGACCTAATGCTCTTCCT
AGACTGGGCAGACCTCAAACAGGTGGTGGAAGACACCTCCACTATGAGGGGTCTGCTGGA
GAACTGGATGATAT

55. Badając transkryptom organizmu uzyskujemy wiele różnych cDNA. Porównanie go z istniejącą bazą białek przy zastosowaniu blastx odpowiada na pytanie czy transkrypt koresponduje ze znanym białkiem. Czasami jednak nie znajdujemy żadnego znanego białka. Może transkrypt nie koduje białka lub białka nie ma jeszcze w bazie danych. Możemy wówczas zadać pytanie, czy nasz transkrypt koduje białko, które ma nie wykryte jeszcze homologi. W tym celu możemy zastosować program tblastx, aby przeszukać bazę danych est. Program tblastx porównuje sekwencje w pytaniu po jej tłumaczeniu wg 6. ramek odczytu z bazami nukleotydowymi również tłumaczonymi na bieżąco na potencjalne sekwencje białkowe. Porównaj transkrypt wyizolowany z pnia mózgu świni (GT916239.1) z bazą znanych białek (blastx) świni i potencjalnych białek (tblastx) świni (Sus scrofa).

56. Program psi-blast poszukuje sekwencji białkowych podobnych do sekwencji w pytaniu. Jednak w przeciwieństwie do klasycznego blastp, wyniki pierwszego przeszukiwania to tylko preludium. Psi-blast tworzy z nich macierz PSSM i stosuje ją do wyszukania sekwencji w następnych krokach. W ten sposób można w bazie danych znaleźć sekwencje homologiczne o mocno zatartym podobieństwie. (a) Użyj sekwencji HXK1_HUMAN jako sekwencji w pytaniu dla psi-blast. Po drugim kroku wyszukiwania zapisz macierz PSSM w formacie ASN1 i zapoznaj się z jej budową korzystając z narzędzia PSSM View. Za pomocą tej macierzy będą oceniane dopasowania w następnym etapie przeszukiwania bazy białek. (b) Uruchom trzeci etap przeszukiwania. Czy znalazłeś nowe sekwencje? Czy macierz PSSM została zmodyfikowana?

BIOMART

Ensembl (http://www.ensembl.org/) dostarcza narzędzie BioMart do zaawansowanego przeglądu i zestawiania danych (data mining).
Choose database – np. Ensembl Genes
Dataset – wybieramy gatunek
Filtry – ustawiamy ograniczenia
Atrybuty – ustawiamy jakie informacje chcemy uzyskać
Count – liczba genów spełniających nasze wymogi
Result – dla genów spełniających nasze wymogi podaje wszystkie atrybuty

57. Gen kodujący Myosin light chain kinase (MYLK_Bovin) zlokalizowany jest w chromosomie 1 genomu bydła.

1. Jakie inne geny zlokalizowane są na chromosomie 1. u bydła? Podaj ile ich jest.
2. Sprawdź ile z tych genów ma identyfikator w SwissProt
3. Sprawdź ile z nich ma homologa u człowieka
4. Ile z nich ma domenę transmembranową
5. Ile ma dodatkowo sekwencję sygnałową
6. Dla 2 przykładowych genów wypisz:
    i. Ensembl Gene ID
    ii. Gene Biotype
    iii. Gene Start (bp)
    iv. Gene end (bp)
    v. UniProt/SwissProt Accession
    vi. CDS Length
    vii. pobierz sekwencje 5' UTR
    viii. pobierz informacje o alternatywnych wariantach sekwencji: Reference ID i Variant Alleles. Jakiego rodzaju warianty znalazłeś/aś?

58. Skomponuj najprostsze zadanie dotyczące liczby genów microRNA w genomie człowieka i świni.

59. Otrzymaj strukturę genu myszy ENSMUSG00000042351 (podaj ile eksonów, jakie są ich pozycje, 5' i 3' UTR, długość CDS). Jaka jest nazwa genu – sprawdź w bazie Ensembl, czy korzystając z BioMartu dobrze określiłeś/aś strukturę genu.

60. Otrzymaj sekwencje białkowe genów ulokowanych na chromosomie 1 kury

61. Zespół Williamsa (zespół Williamsa-Beurena, ang. Williams' syndrome, Williams-Beuren syndrome) to rzadki, genetycznie uwarunkowany zespół wad wrodzonych. W 95% przypadków zespół Williamsa spowodowany jest delecją od 1,5 miliona do 1,8 miliona par zasad regionu q11.23 chromosomu 7. Sprawdź ile genów kodujących białka ulega delecji w tym przypadku. Pobierz ID i opisy tych genów

62. Wybierając bazę Ensembl Genes 70 stwórz własne zadanie (zastosuj kilka filtrów i atrybutów)

DOPASOWANIE SEKWENCJI WIELOKROTNYCH

Dopasowaniu wielu sekwencji pozwala uzyskać informację, która nie wynika z dopasowania sekwencji w parach. W ten sposób można uzyskać wiedzę o regionach konserwatywnych i ich charakterze (motywy). Dopasowanie sekwencji wielokrotnych jest także podstawą analizy filogenetycznej.

63. Pobierz z bazy SwissProt sekwencje o numerach P48125, P54050, P41199 i P96038. Jest to rodzina białek rybosomalnych l1, identycznych w 20-40%. Dopasuj sekwencje korzystając z programów ClustalW i T-Coffee. Porównaj wyniki programów z dopasowaniem w bazie Balibase, która zawiera najbardziej wiarygodne dopasowanie tych sekwencji (rodzina 1ad2). Który program spisał się najlepiej? Zapisz plik z dopasowaniem w swoim komputerze.

64. Zapoznaj się z bazą danych BLOCKs. (a) Co zawiera baza danych? (b) W jaki sposób powstają bloki? (c) Obejrzyj pojedynczy dowolny blok – opisz jego długość i polimorfizm.

65. Przeanalizuj dopasowanie wielosekwencyjne zapisane wcześniej w pliku - wyszukaj bloków w sekwencjach dopasowanych korzystając z serwera Blocks Multiple Alignment Processor. (a) Ile bloków zidentyfikowano? (b) Jaką mają długość? (c) Jaki wymiar ma PSSM dla pierwszego bloku? (d) Wyświetl logo bloków - co powiesz na ich podstawie. (e) Zapisz wyniki dla późniejszego porównania z MEME.

66. Istnieją sposoby na identyfikację motywów w sekwencjach nie dopasowanych. Zapoznaj się z MEME. (a) Co oznacza MEME? (b) W jaki sposób MEME reprezentuje motywy?

67. Dla tego samego zestawu sekwencji (bez dopasowywania) wyszukaj motywów korzystając z serwera MEME. (a) Ile różnych motywów znaleziono? (b) Porównaj wynik z wynikiem uzyskanym z BLOCKS.

SPOSOBY REPREZENTOWANIA MOTYWÓW KONSERWATYWNYCH - WZORCE SEKWENCYJNE

W bazach danych sekwencji obserwujemy krótkie konserwatywne odcinki. Nazywamy je motywami sekwencyjnymi. Poszczególne sekwencje reprezentujące motyw są do siebie dość podobne, ale nie muszą być identyczne. Motyw wykazuje więc zarówno pewien konserwatyzm (większy od sąsiednich odcinków) i jednocześnie umiarkowaną zmienność (mniejszą) niż sąsiadujące regiony. Istnieją różne sposoby by ten konserwatyzm i zmienność opisać. Taki reprezentacyjny opis motywu wykorzystywany jest następnie w analizie nie opisanych jeszcze sekwencji.

68. Wzorce sekwencyjne można zapisać za pomocą wyrażeń regularnych (regular expressions). Baza PROSITE stosuje zapis wzorców sekwencyjnych (patterns) w pewnym stopniu podobny do tego, jaki zastosowano w języku PERL. Jak zinterpretujesz wyrażenie C-x(2,4)-C-x(3)-[LIVMFYWC]-x(8)-H-x(3,5)-H?

69. Krótkie motywy z bazy PROSITE budowane są wokół funkcjonalnych regionów. Do takich regionów należą: (1) miejsca katalityczne enzymów, (2) miejsca wiązania grupy prostetycznej (struktury niebiałkowej uaktywniającej enzym, np. hem, biotyna), (3) miejsce wiązania jonu metalu, (4) miejsce wiązania ADP/ATP, GDP/GTP, wapnia, DNA, lub innego białka. (a) Zapoznaj się ze wzorcami PS00017, PS00080, PS00764 i PS00353. Jaką funkcję (1-4) przypisano każdemu z wzorców sekwencyjnych? (b) Ile wzorców (patterns) zawiera obecna wersja bazy PROSITE?

70. Korzystając z serwera Expasy i narzędzia ScanProsite odpowiedz na pytanie, ile sekwencji można znaleźć w UniProtKB/Swiss-Prot, które pasują do wzorca C-x(2,4)-C-x(3)-[LIVMFYWC]-x(8)-H-x(3,5)-H.

71. Baza PROSITE stanowi narzędzie opisu nowych białek. W pewnych badaniach wyizolowano i zsekwencjonowano transkrypt (mRNA) genu. Przetłumacz ten transkrypt na potencjalną sekwencję białkową (wg pierwszej ramki odczytu) i przeskanuj ją pod kątem występowania fragmentów pasujących do wzorców z bazy PROSITE. Do tłumaczenia sekwencji możesz zastosować program transeq (serwer Mobyle http://mobyle.pasteur.fr/cgi-bin/portal.py#welcome). Do skanowania sekwencji białkowej zastosuj ScanProsite (narzędzie bazy PROSITE). Jakim modyfikacjom potranslacyjnym to białko może podlegać?
Transkrypt:

atgaagcg caccccgact gccgaggaac gagagcgcga agctaagaaa ctgaggcttc ttgaagagct tgaagacact tggctccctt atctgacccc caaagatgat gaattctatc agcagtggca gctgaaatat cctaaactaa ttctccgaga agccagcagt gtatctgagg agctccataa agaggttcaa gaagcctttc tcacactgca caagcatggc tgcttatttc gggacctggt taggatccaa ggcaaagatc tgctcactcc ggtatctcgc atcctcattg gtaatccagg ctgcacctac aagtacctga acaccaggct ctttacggtc ccctggccag tgaaagggtc taatataaaa cacaccgagg ctgaaatagc cgctgcttgt gagaccttcc tcaagctcaa tgactacctg cagatagaaa ccatccaggc tttggaagaa cttgctgcca aagagaaggc taatgaggat gctgtgccat tgtgtatgtc tgcagatttc cccagggttg ggatgggttc atcctacaac ggacaagatg aagtggacat taagagcaga gcagcataca acgtaacttt gctgaatttc atggatcctc agaaaatgcc atacctgaaa gaggaacctt attttggcat ggggaaaatg gcagtgagct ggcatcatga tgaaaatctg gtggacaggt cagcggtggc agtgtacagt tatagctgtg aaggccctga agaggaaagt gaggatgact ctcatctcga aggcagggat cctgatattt ggcatgttgg ttttaagatc tcatgggaca tagagacacc tggtttggcg ataccccttc accaaggaga ctgctatttc atgcttgatg atctcaatgc cacccaccaa cactgtgttt tggccggttc acaacctcgg tttagttcca cccaccgagt ggcagagtgc tcaacaggaa ccttggatta tattttacaa cgctgtcagt tggctctgca gaatgtctgt gacgatgtgg acaatgatga tgtctctttg aaatcctttg agcctgcagt tttgaaacaa ggagaagaaa ttcataatga ggtcgagttt gagtggctga ggcagttttg gtttcaaggc aatcgataca gaaagtgcac tgactggtgg tgtcaaccca tggctcaact ggaagcactg tggaagaaga tggagggtgt gacaaatgct gtgcttcatg aagttaaaag agaggggctc cccgtggaac aaaggaatga aatcttgact gccatccttg cctcgctcac tgcacgccag aacctgagga gagaatggca tgccaggtgc cagtcacgaa ttgcccgaac attacctgct gatcagaagc cagaatgtcg gccatactgg gaaaaggatg atgcttcgat gcctctgccg tttgacctca cagacatcgt ttcagaactc agaggtcagc ttctggaagc aaaaccctag

72. Wzór sekwencyjny PROSITE ma charakteryzować się możliwie dużą czułością, co oznacza, że przeszukując bazę danych uzyskamy za jego pomocą możliwie największą liczbę sekwencji należących do grupy białek. Zapoznaj się z wpisem PS00236. Podano tam wyniki przeszukiwania bazy sekwencji UniProtKB/Swiss-Prot za pomocą opisywanego wzorca. Porównaj liczbę wyników prawdziwie pozytywnych i fałszywie negatywnych. Jaki odsetek sekwencji pozwala zidentyfikować opisany tu wzór sekwencyjny?

73. Wzór sekwencyjny powinien myć możliwie najbardziej specyficzny, czyli po przeszukaniu bazy danych powinniśmy uzyskać możliwie mało wyników fałszywie pozytywnych (sekwencji niepotrzebnie znalezionych). Zapoznaj się z wpisem PROSITE PS00045. Jaki odsetek wyszukanych sekwencji stanowią sekwencje fałszywie pozytywne?

74. W promotorach genów znaleziono 6 sekwencji:
TACGAT

TATAAT
TATAAT
GATACT
TATGAT
TATGTT
(a) Jaka jest sekwencja konsensusowa?
(b) Do ilu z tych sekwencji pasuje sekwencja konsensusowa?
(c) Czy sekwencja konsensusowa dobrze odzwierciedla zmienność w tym motywie?
(d) Na podstawie wielu sekwencji stworzono macierz wag (profil) dla motywu:
A -38 +19 +01 +12 +10 -48
C -15 -38 -08 -10 -03 -32
G -13 -48 -06 -07 -10 -48
T +17 -32 +08 -09 -06 +19
Z jakim wynikiem pasuje do profilu sekwencja konsensusowa, a z jakim sekwencja TACGTT?

75. Baza JASPAR (http://jaspar.genereg.net/) zawiera gotowe profile do identyfikacji w nieopisanych sekwencjach DNA możliwych miejsc wiązania czynników transkrypcyjnych. Znajdz profil GAL4 i zinterpretuj jego logo sekwencyjne.

76. Baza PROSITE była pierwotnie bazą motywów białkowych opisanych wyrażeniami regularnymi, ale obecnie jest także kolekcją dłuższych regionów czyli domen i odpowiadającym im profilom. Zapoznaj się z wpisami PS00155 i PS01031. Który z nich dotyczy motywu wyrażonego wzorcem sekwencyjnym, a który domeny opisanej macierzą wag (profilem)? Jaka jest długość motywu i domeny?

77. Zastosuj ScanProsite w celu przeskanowania dwóch poniższych sekwencji. Jakie domeny one zawierają?
Sekwencja A:
MGSLMLLFVETTRNSSACIFPVILNELSSTVETITHFPEVTDGECVFPFHYKNGTYYDCI
KSKARHKWCSLNKTYEGYWKFCSAEDFANCVFPFWYRRLIYWECTDDGEAFGKKWCSLTK
NFNKDRIWKYCE
Sekwencja B:
MSKGEELFTGVVPILVELDGDVNGHKFSVSGEGEGDATYGKLTLKFICTTGKLPVPWPTL
VTTFSYGVQCFSRYPDHMKQHDFFKSAMPEGYVQERTIFFKDDGNYKTRAEVKFEGDTLV
NRIELKGIDFKEDGNILGHKLEYNYNSHNVYIMADKQKNGIKVNFKIRHNIEDGSVQLAD
HYQQNTPIGDGPVLLPDNHYLSTQSALSKDPNEKRDHMVLLEFVTAAGITHGMDELYK

78. Bardziej zaawansowany opis motywu konserwatywnego można stworzyć za pomocą metody HMM. Program HMMbuild (serwer Mobyle http://mobyle.pasteur.fr/cgi-bin/portal.py#welcome) buduje profil HMM na podstawie wielu ułożonych sekwencji białkowych. Gotowy profil HMM może służyć do skanowania nieopisanej sekwencji pod kątem występowania wzorca konserwatywnego. Przykładowe ułożone sekwencje białkowe w formacie FASTA dostępne są w pliku inHMMbuild.txt. Zbuduj profil HMM na podstawie ułożonych sekwencji i zapisz go do pliku.

79. Wzorce sekwencyjne można opisać za pomocą profilowych ukrytych modeli Markowa. Analizując różne białka transbłonowe przygotowano profil, który charakteryzuje domeny transmembranowe. Profil ten może teraz służyć do poszukiwania domen transmembranowych w nieopisanych białkach. Przeanalizuj sekwencję białkową P04882 (Entrez) programem TMHMM (http://www.cbs.dtu.dk/services/TMHMM-2.0/). (a) Jaką metodę wykorzystuje ten program? (b) Porównaj wyniki z opisem sekwencji – jak oceniasz metodę?

ANALIZA POJEDYNCZEJ SEKWENCJI BIAŁKOWEJ

80. Agregacja białek koreluje z rozwojem wielu chorób neurodegeneracyjnych, np. choroba Alzheimera czy Parkinsona. Obserujemy w nich akumulację rekombinowanych białek w formie agregatów białkowych co prowadzi do wyniszczania mózgu. Dlatego na uwagę zasługuje rozwój metod przewidywania właściwości agregacji polipeptydów. AGGRESCAN to internetowe oprogramowanie do przewidywania podatnych fragmentów sekwencji białkowych do agregacji, w tym analizy wpływu mutacji na skłonności do agregacji białek i porównywania właściwości agregacji różnych białek lub zestawów białek. AGGRESCAN opiera się na skłonności naturalnych aminokwasów pochodzących z badań in vivo do agregacji, oraz przy założeniu, że krótkie i specyficzne obszary sekwencji modulują agregację białka. Program AGGRESCAN dostępny jest na stronie: http://bioinf.uab.es/aggrescan/ **a)** Znajdź sekwencje białkowe: Aβ42 peptide oraz synuclein w UniProt. Wprowadź je do programu AGGRESCAN w formacie FASTA. **b)** Porównaj oba białka występujące w chorobach neurodegenracyjnych pod względem takich parametrów jak: a3vSA (agregacja aminokwasów -częstotliwość występowania agregacji), nHS (liczba „hot spot"), NnHS (znormalizowana liczba „hot spot" dla 100 reszt), AAT (poziom agregacji powyżej progu „hot spot", THSA (całkowita ilość „hot spot"), TA (całkowity profil poziomu agregacji), AATr (AAT podzielona przez liczbę reszt w sekwencji aminokwasowej wejściowej), THSAr (THSA podzielona przez liczbę reszt w sekwencji aminokwasowej wejściowej), Na4vSS (a4vSS podzielona przez liczbę reszt w sekwencji aminokwasowej wejściowych i pomnożona przez 100) **c)** Kliknij w znak zapytania z lewej strony wartości aby zobaczyć dokładne wyjaśnienia dla powyższych parametrów. W podpunkcie amino-acid aggregation-propensity value. (a3v) kliknij w link do tabeli aby zobaczyć wartości a3v dla 20 prawidłowych aminokwasów. **d)** Kliknij w ikony P, A i A/N przy Graphics aby zobaczyć przedstawienie graficzne dla tych białek. Więcej informacji o programie AGGRESCAN możesz uzyskać z artykułu: „AGGRESCAN: a server for the prediction and evaluation of "hot spots" of aggregation in polypeptides", Oscar Conchillo-Solé, Natalia S de Groot, Francesc X Avilés, Josep Vendrell, Xavier Daura, Salvador Ventura (zadanie przygotowane przez Julię Rosiak).

81. Wewnętrznie nieuporządkowane białka (IDP) są białkami, w których brak jest stałej i uporządkowanej struktury trzeciorzędowej. Ta klasa białek obejmuje szereg białek, od w pełni nieuporządkowanych do częściowo nieuporządkowanych i zawierających w sobie m.in. przypadkowe pętle i białka złożone z wielu domen połączonych elastycznymi łącznikami, zbudowanych z wielu struktur pozostających z sobą w równowadze. DisCons jest narzędziem pozwalającym badać ilościowo nieuporządkowania w budowie białek na poziomie aminokwasów i klasyfikować je pod względem specjalnie dobranych kategorii, na podstawie sekwencji i skłonności fragmentu białka do występowania zaburzeń. Klasyfikacja taka pozwala na wskazanie czy dany nieuporządkowany segment jest funkcjonalnie ważny i może dać wskazówki dotyczące jego funkcji, np. regiony elastyczne pod względem struktury, mogą dotyczyć regionów, w których zachodzi wiązanie do receptorów i potranslacyjne modyfikacje. **a)** Skorzystaj ze strony: http://pedb.vib.be/discons/ i dowiedz się na jakie jeszcze kategorie program dzieli segmenty białek. **b)** Wyszukaj w bazie UniProt sekwencję aminokwasową białka p53 człowieka (Homo sapiens) i zastosuj szybki wariant narzędzia DisCons. Ile program zastosował dopasowań? Jaką część stanowi każda z grup sklasyfikowanych segmentów? Których z nich jest najwięcej? **c)** Wyszukaj teraz sekwencję lizozymu człowieka i zastosuj narzędzie DisCons. Jak wyglądają wyniki w tym przypadku? Które z białek jest bardziej „nieuporządkowane" i o czym to może świadczyć? **d)** Wykorzystaj sekwencję białka p53 i zastosuj zaawansowany wariant programu. (autorstwo zadania Anna Kotowska, Biotechnologia) Zastosuj różne parametry. Czy umożliwia to uzyskanie bardziej precyzyjnych wyników?

82. Oceń sekwencję białkową pod kątem obecności domen transmembranowych. Połącz się z http://www.expasy.org/tools/#proteome. Z pomocą metody 'sliding window' oblicz statystykę Kyte-Doolittle dla peptydu P04882 (VGLG_VSNJO) w poszukiwaniu domen transmembranowych. (a) Jaką długość okna dobierzesz? (b) Porównaj wyniki z opisem tego białka – jak oceniasz działanie metody?

83. Tę samą sekwencję przeanalizuj programem TMHMM (http://www.cbs.dtu.dk/services/TMHMM-2.0/). (a) Jaką metodę wykorzystuje ten program? (b) Porównaj wyniki z opisem sekwencji – jak oceniasz metodę?

84. Dokonaj analizy sekwencji GCN4_YEAST programem COILS. (a) Czy są dowody na istnienie struktur typu coiled-coil? W jakiej pozycji? (b) Sprawdź w opisie sekwencji jak jest określony ten region.

85. Suwak leucynowy ma strukturę coiled-coil oraz charakterystyczne powtórzenia leucyny. Wykorzystaj program 2ZIP do analizy sekwencji GCN4_YEAST. (a) Jaka jest dokładna lokalizacja struktury suwaka leucynowego? (b) Ile razy występuje leucyna w tej strukturze i co ile pozycji?

86. W genie KIF1A człowieka zidentyfikowano mutacje zmieniającą w sekwencji aminokwasowej alaninę na walinę w pozycji 255 (A255V). Przeanalizuj sekwencje aminokwasową odpowiadającą transkryptowi ENST00000320389. Dokonaj predykcji skutków tej mutacji programami SIFT i PolyPhen. Czy mutacja może prowadzić do zmiany fenotypu?

KLASYFIKACJA BIAŁEK I BAZY RODZIN BIAŁKOWYCH (BAZY DANYCH WZORCÓW SEKWENCJI)

Bazy białek różnią się tym, co rozumiemy pod pojęciem domena, motyw, rodzina. Do identyfikacji wzorców sekwencyjnych stosuje się różne definicje i algorytmy. Ostatecznie, powstało wiele metod klasyfikacji białek. Różne metody klasyfikacji białek posłużyły wygenerowaniu różnych baz rodzin białkowych. Pełnią one bardzo ważną rolę w opisie nowych białek. Stąd, warto zapoznać się z najważniejszymi bazami rodzin białek.

87. Zapoznaj się z bazą Pfam. (a) Jaki identyfikator ma największa rodzina białkowa i ile zawiera sekwencji? (b) Ile sekwencji zalążkowych (seed) zastosowano w celu zbudowania pierwszego wzorca identyfikującego tę rodzinę? (c) Ile gatunków reprezentują sekwencję z tej rodziny? (d) Do jakiego klanu należy ta rodzina? (e) Zapisz na dysku model HMM dla tej rodziny.

88. Zobacz jak wygląda logo sekwencyjne rodziny Pfam PF12729. Podaj interpretację logo dla rodziny Pfam PF12729: Jak długi jest motyw sekwencyjny? Na których pozycjach występują miejsca szczególnie konserwatywne? Czy logo budowane jest dla motywu czy domeny?

89. Pfam stosuje model HMM do reprezentowania wzorca sekwencji. Wzorce te mogą służyć do przeszukiwania baz danych. Zapoznaj się z rodziną GFP. (a) Zapisz na dysku model HMM. (b) Wykorzystaj HMMSearch (http://mobyle.pasteur.fr). Przeszukaj bazę Swiss-Prot korzystając z zapisanego modelu HMM. Ile sekwencji zidentyfikowałeś?

90. Pfam przy definicji rodzin stosuje obok podobieństwa sekwencji kryterium podobieństwa struktury. Czasem rodzinę tworzą białka o małym podobieństwie sekwencji. Zwróć uwagę na rodzinę białek histonowych PF00125. Zbadaj odsetek identyczności (a) między przedstawicielami jednej podrodziny (HIST1H3A i HIST1H3B człowieka), (b) przedstawicielami dwóch podrodzin jednej rodziny (HIST1H3A i HIST3H3) i (c) przedstawicielami różnych rodzin stanowiących jedną superrodzinę (HIST1H3a i HIST1H4A).

91. Na przykładzie wpisu PF02026 zapoznaj się z graficznym sposobem reprezentowania wzorców w bazie Pfam. (a) Jaki kształtem zaznaczone są domeny, a jakim motywy? Czy motywy Pfam występują w domenach? (b) Jaka jest długość modelu HMM dla domeny RyR. (c) Ile różne architektur zdefiniowano dla domeny RyR w bazie Pfam? (d) Czy w każdej architekturze sekwencje pasują do wzorca HMM na całej jego długości?

92. Czasami domeny zdefiniowane w bazie Pfam mogą być zagnieżdżone w innych domenach i nie zakłócać funkcji domeny -gospodarza. Zapoznaj się z domeną IMPDH (a) Jaka domena występuje często wewnątrz domeny IMPDH?

93. Baza PRINTS zbiera krótkie motywy tworząc "ślad rodziny" (fingerprint). Ślad lepiej reprezentuje rodzinę białkową, niż pojedynczy motyw. Białka współdzielące ślad należą do jednej rodziny, a współdzielące tylko część z zestawu motywów należą do podrodzin (subfamilies). **(a)** Ile minimalnie i ile maksymalnie motywów znajduje się w pojedynczym śladzie. **(b)** Czy w pojedynczym motywie dozwolone są przerwy w dopasowaniu?

94. Baza SMART to baza poświęcona szczególnie domenom zewnątrzkomórkowym, sygnałowym i powiązanym z chromatyną. **(a)** Ile domen zdefiniowano w basie SMART? **(b)** Znajdź białka człowieka mające dwie domeny CBS. Ile ich jest?

95. Baza COG definiuje rodzinę białkową jako klaster białek o tym samym pochodzeniu ewolucyjnym. Każdy klaster COG (cluster of orthologous groups) gromadzi pojedyncze białka (lub grupy paralogów) obecne w co najmniej trzech głównych liniach rozwojowych. Białka obecne w głównych liniach rozwojowych muszą pochodzić od wspólnego przodka bardzo odległego w historii ewolucyjnej. Białka w każdym COG mają pochodzić od wspólnego przodka i być produktem serii specjacji i duplikacji w genomach. Taki sposób klasyfikacji białek jest możliwy jedynie dla organizmów, dla których znane są całe genomy. Genomy porównywano parami, każdy gen z każdym, a najlepiej pasujące geny połączono linią. Jeżeli, dany gen ma najlepiej pasujące geny w dwóch innych genomach, a z kolei te

geny są najlepiej dopasowane, jeżeli porównać genomy z których pochodzą, to prawdopodobnie wszystkie 3 geny są ortologami. **(a)** Jakie gatunki porównywano w przypadku eukariota? **(b)** Jakie gatunki wchodzą w skład KOG1019? **(c)** Obejrzyj graf dla KOG1019. Czy graf zawiera trójkąt? Które z 3 białek są wzajemnie najlepiej do siebie pasujące po porównaniu wszystkich w genomach? **(d)** Czy baza COG gromadzi motywy sekwencyjne, czy też grupuje białka bez odwoływania się do motywów sekwencyjnych? (**e)** Sposób tworzenia COG, w którym identyfikujemy tylko najlepsze dopasowania po porównaniu całych genomów, jest niezależny od ogólnego poziomu podobieństwa między białkami, i stąd pozwala na detekcję ortologów zarówno między wolno jak i szybko ewoluującymi genami. Użyj narzędzia KOGNITOR by sprawdzić do jakiego klastra należy poniższa sekwencja i jakie gatunki obejmuje ten klaster.

MSAGGPCPAAAGGGPGGASCSVGAPGGVSMFRWLEVLEKEFDKAFVDVDLLLGEIDPDQADITYEGRQKM
TSLSSCFAQLCHKAQSVSQINHKLEAQLVDLKSELTETQAEKVVLEKEVHDQLLQLHSIQLQLHAKTGQS
ADSGTIKAKLSGPSVEELERELEANKKEKMKEAQLEAEVKLLRKENEALRRHIAVLQAEVYGARLAAKYL
DKELAGRVQQIQLLGRDMKGPAHDKLWNQLEAEIHLHRHKTVIRACRGRNDLKRPMQAPPGHDQDSLKKS
QGVGPIRKVLLLKEDHEGLGISITGGKEHGVPILISEIHPGQPADRCGGLHVGDAILAVNGVNLRDTKHK
EAVTILSQQRGEIEFEVVYVAPEVDSDDENVEYEDESGHRYRLYLDELEGGGNPGASCKDTSGEIKVLQG
FNKKAVTDTHENGDLGTASETPLDDGASKLDDLHTLYHKKSY

96. Porównaj wyniki skanowania dwóch sekwencji wg. ScanProsite i Pfam. Dla której sekwencji wyniki są zbieżne, a dla której rozbieżne?

Sekwencja A:

MSKGEELFTGVVPILVELDGDVNGHKFSVSGEGEGDATYGKLTLKFICTTGKLPVPWPTL
VTTFSYGVQCFSRYPDHMKQHDFFKSAMPEGYVQERTIFFKDDGNYKTRAEVKFEGDTLV
NRIELKGIDFKEDGNILGHKLEYNYNSHNVYIMADKQKNGIKVNFKIRHNIEDGSVQLAD
HYQQNTPIGDGPVLLPDNHYLSTQSALSKDPNEKRDHMVLLEFVTAAGITHGMDELYK

Sekwencja B:

MGSLMLLFVETTRNSSACIFPVILNELSSTVETITHFPEVTDGECVFPFHYKNGTYYDCI
KSKARHKWCSLNKTYEGYWKFCSAEDFANCVFPFWYRRLIYWECTDDGEAFGKKWCSLTK
NFNKDRIWKYCE

97. PHI-BLAST to metoda przeszukiwania bazy sekwencji białkowych pod kątem obecności wzorca i dobrego dopasowania sekwencji flankujących wzorzec do sekwencji w pytaniu. Metoda ułatwia znalezienie homologów sekwencji w pytaniu współdzielących dany motyw. Zastosuj PHI-BLAST mając podaną sekwencję i wzorzec, i porównaj wyniki z klasycznym programem BLASTP.

MPRGWAAPLLLLLLQGGWGCPDLVCYTDYLQTVICILEMWNLHPSTLTLTWQDQYEELKD
EATSCSLHRSAHNATHATYTCHMDVFHFMADDIFSVNITDQSGNYSQECGSFLLAESIKP
APPFNVTVTFSGQYNISWRSDYEDPAFYMLKGKLQYELQYRNRGDPWAVSPRRKLISVDS
RSVSLLPLEFRKDSSYELQVRAGPMPGSSYQGTWSEWSDPVIFQTQSEELKEGWNPHLLL
LLLLVIVFIPAFWSLKTHPLWRLWKKIWAVPSPERFFMPLYKGCSGDFKKWVGAPFTGSS
LELGPWSPEVPSTLEVYSCHPPRSPAKRLQLTELQEPAELVESDGVPKPSFWPTAQNSGG
SAYSEERDRPYGLVSIDTVTVLDAEGPCTWPCSCEDDGYPALDLDAGLEPSPGLEDPLLD
AGTTVLSCGCVSAGSPGLGGPLGSLLDRLKPPLADGEDWAGGLPWGGRSPGGVSESEAGS
PLAGLDMDTFDSGFVGSDCSSPVECDFTSPGDEGPPRSYLRQWVVIPPPLSSPGPQAS

[LIVF]-x(9)-[LIV]-[RK]-x(9,20)-W-S-x-W-S-x(4)-[FYW]

98. StemChecker to program umożliwiający sprawdzenie, czy analizowane geny mogą być powiązane z funkcjonowaniem i charakterystycznymi właściwościami różnych typów komórek macierzystych. W tym celu wybrane przez użytkownika geny są porównywane z genami zgromadzonymi w baize danych programu. Program ocenia czy wskazane geny mają związek ze specyficznymi cechami różnych typów komórek macierzystych („stemness signatures") na podstawie takich kryteriów jak profil ekspresji, dane literaturowe, interferencja RNA. Oceniane jest też czy ich aktywność jest regulowana przez czynniki transkrypcyjne charakterystyczne dla komórek macierzystych. Program StemCheker jest dostępny na stronie http://stemchecker.sysbiolab.eu/ **a)** W bazie GenBank (NCBI, zakładka „gene") znajdź geny o numerach ID: 6657, 701, 22, 5460 oraz 10370. **b)** Przejdź na stronę programu StemCheker, wybierz zakładkę „Analysis" i wprowadź w odpowiednie pole wszystkie pięć oficjalnych symboli znalezionych wcześniej genów. Uruchom analizę (submit). **c)** Jakiego typu informacje można uzyskać dzięki użyciu programu? W jakich typach komórek macierzystych wybrane geny wykazują działanie? Na podstawie jakich kryteriów wybrane geny zostały ocenione jako powiązane z funkcjonowaniem komórek macierzystych („stemness")? Sprawdź otrzymane wyniki w zakładce „Stemness signatures". **d)** Czy aktywność analizowanych genów jest regulowana przez czynniki transkrypcyjne występujące w komórkach macierzystych? Jakie to czynniki? (autorstwo zadania Monika Drobna, Natalia Mazurkiewicz)

PREDYKCJA GENÓW

99. Statystyka testcode odzwierciedla charakterystyczną własność sekwencji kodujących. Obserwujemy w nich w pewnym stopniu powtarzalność każdego trzeciego nukleotydu. Zjawisko to jest niezależne od gatunku. Program tcode wyznacza statystykę testcode dla wybranego obszaru genomu. Dokonaj analizy sekwencji NM_000018. (a) Czy tę sekwencję można uznać za kodującą czy nie? (b) Wypisz 3 regiony o najwyższej statystyce testcode i porównaj z adnotacją tej sekwencji w bazie danych. Czy Twoje potencjalne regiony kodujące pokrywają się z dostępnym opisem genomu?

100. Przy predykcji genów ważną informacją jest charakterystyka wykorzystania różnych kodonów w kodowaniu sekwencji aminokwasowej. Przykładowo, potencjalną ramkę odczytu można scharakteryzować pod kątem użycia kodonów i porównać obliczone częstości kodonów w potencjalnym ORF do znanych frekwencji. U różnych gatunków obserwowane są różne frekwencje wykorzystywanych kodonów. (a) Porównaj CDS genu MC4R człowieka, świni i myszy. Ile razy występuje w tych genach leucyna (L). Czy te gatunki używają kodonów w ten sam sposób? Teraz porównaj wykorzystanie kodonów dla dwóch genów człowieka: wykorzystaj CDS genów LEP i MC4R. Czy wykorzystanie kodonów leucyny jest podobne w obu genach człowieka? Czy częstości zgadzają się z informacjami w Codon Usage Database.

101. Predykcja otwartych ramek odczytu (ORF) jest dobrą metodą predykcji genów u prokariotów. Pobierz z GenBanku sekwencję E. coli X08087. (a) Wykorzystując plotorf znajdź potencjalne otwarte ramki odczytu. Która ramka jest najbardziej prawdopodobna? Jakiej może być długości? (b) Tę samą sekwencję zbadaj programem getorf i uzyskaj szczegółowe informacje o najdłuższym ORF. Porównaj wynik z opisem tej sekwencji w GenBanku.

102. Program MZEF (GeneFinder) odnajduje potencjalne eksony w nie opisanych sekwencjach genomowych. Zapoznaj się z opisem program. (a) Jaką metodę stosuje MZEF - liniową czy kwadratową analizę dyskryminacji? (b) Jakie statystyki (score) ocenia MZEF przy odnajdowaniu eksonów? (c) Do analizy jakich gatunków metoda została przyuczona? (d) Oceń jakość predykcji eksonów uzyskanej programem MZEF w opisanej sekwencji genu LEP człowieka (NG_007450, gi:169808406) porównując wyniki swojej analizy z opisem tej sekwencji w GenBanku: Ile gen LEP ma eksonów, a ile udało się zidentyfikować programem MZEF?

103. HMMGene to jeden z programów do predykcji genów. (a) Zapoznaj się krótko z dokumentacją. (a) Jaką metodę wykorzystuje program? (b) Przeanalizuj sekwencję genu LEP człowieka i porównaj wyniki z opisem tej sekwencji (NG_007450). Jak oceniasz prawidłowość predykcji dokonanej przez HMMgene? GrailEXP to znany program do predykcji genów. (a) Jaką metodę wykorzystuje program? (b) Przeanalizuj ponownie sekwencję genu LEP i porównaj z opisem tej sekwencji w GenBanku.

104. Ensembl dostarcza bazy danych i narzędzia do adnotacji genomów. Zapoznaj sie z bazą Ensembl. (a) Porównaj statystyki dotyczące opisu genomu człowieka i świni (Sus scrofa). Co możesz powiedzieć o stopniu opisu genomu świni? (b) Na podstawie dokumentacji opisz, w jaki sposób odbywa sie adnotacja genomu świni (genebuild) i jakie informacje wykorzystuje Ensembl w tym procesie?

105. Ensembl dostarcza narzędzie BioMart do zaawansowanego przeglądu i zestawiania danych (data mining). Skomponuj najprostsze zadanie dotyczące liczby genów microRNA w genomie człowieka i świni.

106. Dokonaj analizy sekwencji GCN4_YEAST programem COILS. (a) Czy są dowody na istnienie struktur typu coiled-coil? W jakiej pozycji? (b) Sprawdź w opisie sekwencji jak jest określony ten region.

107. Suwak leucynowy ma strukturę coiled-coil oraz charakterystyczne powtórzenia leucyny. Wykorzystaj program 2ZIP do analizy sekwencji GCN4_YEAST. (a) Jaka jest dokładna lokalizacja struktury suwaka leucynowego? (b) Ile razy występuje leucyna w tej strukturze i co ile pozycji?

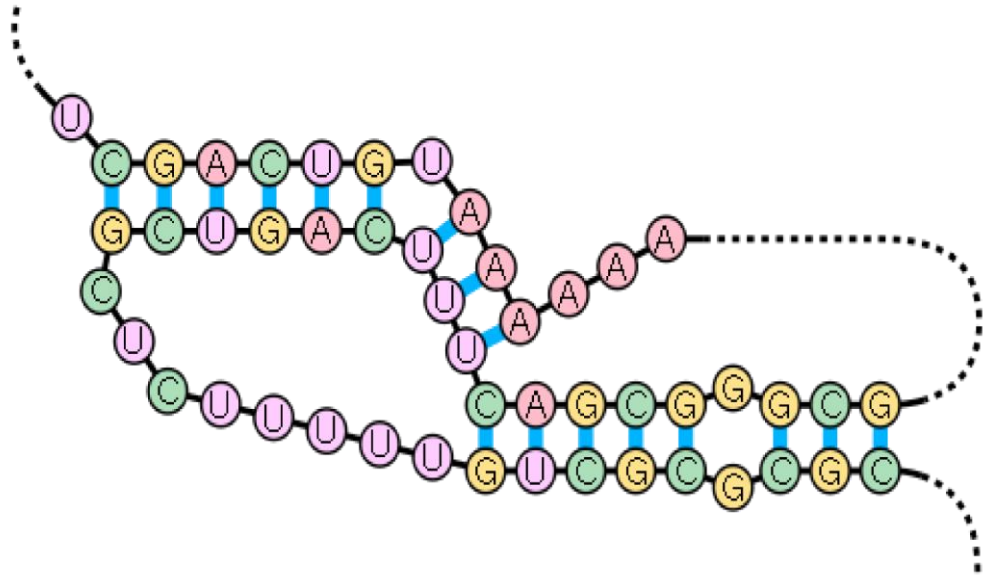
108. Rozważ możliwą strukturę krótkiej nici RNA(a) Znajdź w pojedynczej nici AGCCAUUUUUUGGCU fragmenty komplementarne (A/U i G/C, dla uproszczenia zignoruj G/U). (b) Narysuj na papierze, w jaki nić może się zwijać tworząc strukturę spinki do włosów. Korzystając z tabel, oceń jaka będzie energia takiej struktury? Uwaga, dinukleotyd 5'-CU-3' sparowany z 5'-AG-3' analizujemy jako parę U/A, która następuje po parze C/G.

			Para druga			
	A/U	**C/G**	**G/C**	**U/A**	**G/U**	**U/G**
A/U	-0,9	-1,8	-2,3	-1,1	-1,1	-0.8
C/G	-1,7	-2,9	-3,4	-2,3	-2,1	-1,4
G/C	-2,1	-2,0	-2,9	-1,8	-1,9	-1,2
U/A	-0,9	-1,7	-2,1	-0,9	-1,0	-0,5
G/U	-0,5	-1,2	-1,4	-0,8	-0,4	-0,2

Długość	**1**	**5**	**10**	**20**	**30**
Pętla wewnętrzna	-	5,3	6,6	7,0	7,4
Wybrzuszenie	3,9	4,8	5,5	6,3	6,7
Szpilka	-	4,4	5,3	6,1	6,5

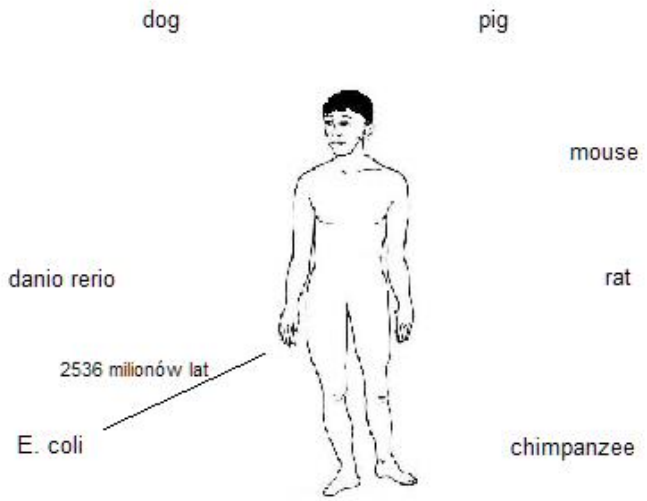
109. Z bazy tRNAdb pobierz sekwencję RD0260 i przeanalizuj ją programem RNAfold. (a) Jakie charakterystyczne elementy rozpoznajesz w wypredykowanej strukturze? (b) Jaka jest energia swobodna tej struktury. (c) Zazwyczaj dla jednej sekwencji można znaleźć wiele struktur, które mają energię swobodną blisko wartości minimalnej. Jeżeli dane wiązanie pojawia się często wśród suboptymalnych struktur, wówczas prawdopodobieństwo tego wiązania jest wysokie. Czy dla wszystkich elementów struktury optymalnej prawdopodobieństwo wiązań jest podobne? (d) Struktura centroidowa to struktura najmniej różniąca się od wszelkich prawdopodobnych struktur, inaczej mówiąc jest to struktura jakby uśredniona. Zazwyczaj jest ona inna niż struktura MFE (minimum free energy), ale często lepiej odzwierciedla prawdziwą strukturę RNA. Jaka jest enegia struktury centroidowej? Czy fragmenty 'pewne' w strukturze centroidowej, są również oznaczone jako wiarygodne w strukturze MFE? (e) Entropię można interpretować jako niepewność. Czy dla poszczególnych wiązanń entropia idzie w parze z prawdopodobieństwem? (f) Zapoznaj się ze sposobem zapisu struktury w formacie dot-bracket. Spróbuj zapisać strukturę szpilki z poprzedniego zadania.

110. Pseudowęzły tworzą strukturę, w której poszczególne sparowane nukleotydy nie są zagnieżdżone. Algorytm programowania dynamicznego nie rozpatruje takich struktur, stąd program RNAfold nigdy nie zaproponuje struktury z pseudowęzłem. Na rysunku przedstawiono strukturę RNA teleomerazy człowieka. Narysuj dla tej struktury wykres kołowy (circle plot). Co uwidacznia taki wykres?



111. IPknot jest programem służącym do przewidywania drugorzędowych struktur RNA, które zawierają pseudowęzły. Pseudowęzły występują w szeregu struktur funkcjonalnych RNA, stąd potrzeba tworzenia programów, które będą projektować takie struktury wystarczająco szybko i dokładnie. Metoda przewidywania struktury bazuje na MEA (maximum expected accuracy – wyszukująca statystycznie najtrafniejszą strukturę z szeregu możliwych) oraz na algorytmie heurystycznym (stosującym pewne założenia upraszczające proces, bez utraty dokładności przewidywania). Program IPknot dostępny jest na stronie: http://rna.naist.jp/ipknot/
(a) Znajdź sekwencję nukleotydową 5.8S ribosomal RNA u Sus scrofa. Po ręcznym przekształceniu jej w sekwencję RNA wypróbuj różne poziomy przewidywań dla tej sekwencji. Zaobserwuj, jak zmienia się struktura RNA oraz jej przedstawienie graficzne. **(b)** Sprawdź, jak działa dopasowanie wielosekwencyjne w programie IPknot klikając na przykład czwarty. **(c)** Porównaj wyniki predykcji otrzymywane w progamach RNAfold i IPknot na podstawie sekwencji użytej w zadaniu pierwszym. Jak (zasadniczo) różnią się oba programy? (pytanie przygotowane przez Alinę Dudzic, Weronikę Rybarczyk i Dorotę Kuczek)

112. Uzupełnij schemat informacjami o odległości ewolucyjnej (w milionach lat) między gatunkami. Skorzystaj z www.timetree.net



.

113. Uzupełnij tekst:

Filogenetyka bada historię ..... Pozwala odtworzyć kolejność ..... Wyniki analiz filogenetycznych są najczęściej przedstawiane w postaci..... Najstarszy wspólny przodek rozważanych gatunków to ..... Długość gałęzi może być wprost proporcjonalna do ..... lub ..... Topologia to sposób ..... Drzewo można przedstawić w dowolnej formie o ile będzie miało taką samą ..... Miejsca rozgałęzień to ..... Na drzewie ..... każdy węzeł oznacza punkt rozejścia się dokładnie dwóch linii ewolucyjnych. Drzewo nieukorzenione jest ..... informatywne od ukorzenionego. Drzewo można ukorzenić przez dodanie do zbioru danych ..... Grupa monofiletyczna nazywana jest ..... i obejmuje wszystkie gatunki wywodzące się od .....

Pierwszym etapem etap analizy filogenetycznej jest dobór ..... Sekwencje układamy wykorzystując dopasowanie ..... Następnie budujemy drzewo wybraną metodą. Jest wiele metod konstrukcji drzewa, niektóre oparte są na ..... a inne na ..... Metoda przyłączania najbliższego sąsiada (NJ) jest przykładem metody opartej na ..... Metoda parsymonii to przykład metody opartej na ..... W metodzie parsymonii przestrzegana jest zasada, że najlepsze drzewo jest .....

Uzyskane drzewo zazwyczaj jest jednym z wielu prawie równie możliwych. Stopień niepewności co do topologii drzewa można zmierzyć metodą .....

114. Strona www.onezoom.org oferuje interaktywny sposób patrzenia na drzewo życia. Zapoznaj się z ewolucją ssaków (mammals). Ile gatunków obejmuje ta linia ewolucyjna? Jaka część gatunków jest zagrożona wyginięciem? Kiedy wyodrębniła się ta linia ewolucyjna? Skąd pochodzą wyświetlane informacje?

115. Wszy to organizmy bardzo wyspecjalizowane - pasożytują z reguły na jednym gatunku. Zauważono, że wszy przechodzą ewolucję niejako równoległą do ewolucji ich żywicieli - gdy następuje wyodrębnienie nowego gatunku żywiciela, to powstaje również nowy gatunek wszy. Wszy stanowią więc dodatkową informację o filogenezie ptaków i ssaków. Wiadomo, że wszy nękały już pierwsze dinozaury. W tym zadaniu nie będziemy sięgać tak daleko wstecz, ale zajmiemy się filogenezą wesz z podrzędu Anoplura pasożytujących na ssakach.

Analizowanymi sekwencjami będzie fragment rybosomalnej jednostki 18S oraz sekwencja kodująca fragment podjednostki pierwszej oksydazy cytochromowej (COI). Te geny są często stosowane w analize filogenetycznej z uwagi na ich zmienność międzygatunkową. Poniżej podano numery sekwencji 18S (kolumna lewa) oraz COI (kolumna prawa) przynależne do sześciu różnych gatunków wszy oraz (gatunek żywiciela).

1. HM171379 EU375756 (Sus scrofa)
2. HM171380 EU375757 (Bubalus bubalis)
3. HM171397 EU375760 (Capra hircus)
4. HM171398 EU375761 (Ovis aires)
5. AY077775 HQ124316 (Homo sapiens)
6. AY077776 AY696000 (Homo sapiens)

(a) Wykorzystaj program muscle by ułożyć sekwencje (wskazówka: najpierw połącz sekwencję 18S oraz COI w jedną dłuższą sekwencję). Skorzystaj z serwera www.trex.uqam.ca. Zapisz ułożenie do pliku. **(b)** Zbuduj topologię drzewa filogenetycznego metodą największej wiarygodności. Zapisz plik z topologią do pliku. **(c)** Na podstawie zapisanej topologii stwórz plik graficzny. Dokonaj interpretacji drzewa filogenetycznego. **(d)** Nadaj swojemu drzewku ulubioną formę **(e)** Dokonaj oceny wiarogodności topologii drzewa za pomocą metody bootstrap.

PRIORYTYZACJA GENÓW

Genomowa analiza asocjacyjna oraz wielkoskalowa analiza ekspresji genów może doprowadzić do zidentyfikowania wielu genów potencjalnie powiązanych/odpowiedzialnych za badaną chorobę. Jeżeli przyjmiemy, że choroba jest monogenowa, wówczas warto zdecydować, który z genów kandydujących poddać szczegółowym badaniom w pierwszej kolejności, co potencjalnie może oszczędzić czas i środki przeznaczone na zidentyfikowanie mutacji sprawczej. Powstało wiele metod priorytyzacji genów. Przykładowo, program **Endeavour** priorytyzuje geny kandydujące na podstawie ich wielorakiego podobieństwa do wybranej grupy innych genów.

116. W pewnym badaniu nad przyczyną schorzenia HSP (heretidary spastic paraparesis) wytypowano 13 genów kandydujących, potencjalnie niosących mutację odpowiedzialną za przypadki tego schorzenia w pewnej rodzinie: *D2HGDH, HDLBP, ING5, DTYMK, OR6B3, AQP12A, THAP4, KIF1A, ATG4B, PASK, SNED1, OR6B3, AQP12B*. Znane są mutacje w innych genach, które również prowadzą do schorzenia HSP. Możesz posłużyć się tymi genami, jako zbiorem uczącym i wskazać, który z genów kandydujących jest najbardziej podobny do znanych już genów HSP. Znane geny HSP to: *ATL1, BSCL2, CYP7B1, GJC2, HSPD1, KIAA0196, KIAA1840, KIF5A, L1CAM, NIPA1, SPG7, PLP1, PNPLA6, REEP1, SLC16A2, SLC33A1, SPAST, SPG20, SPG21, ZFYVE26.* Wskaż, który z genów kandydujących warto zbadać jako pierwszy. Posłuż się programem **Endeavour.**
http://homes.esat.kuleuven.be/~bioiuser/endeavour/tool/endeavourweb.php

117. DisGeNET to kompleksowa platforma, zaprojektowana w celu dostarczania informacji o podłożu genetycznym chorób złożonych człowieka. Baza obejmuje ponad 16000 genów i 13000 sprzężonych z nimi chorób. Korzystając z DisGeNET (http://www.disgenet.org/) wykonaj następujące polecenia: **a)** Otyłość to przewlekła choroba, charakteryzująca się nadmiernym nagromadzeniem tkanki tłuszczowej, prowadzącym do zaburzeń stanu zdrowia. Coraz częściej nazywana jest pandemią XXI wieku. Czy przyczynę otyłości mogą stanowić czynniki genetyczne? **b)** Podaj symbol i pełną nazwę genu, którego mutacja najczęściej leży u podłoża tej choroby. Jaką funkcję pełni białko kodowane przez ten gen? Ile aminokwasów zawiera to białko? (wykorzystaj UniProt) **c)** Geny zaangażowane w rozwój otyłości są przyczyną również innych schorzeń. Jakich? **d)** Wyszukaj gen CFTR. Podaj nazwę choroby, z której występowaniem jest związany. **e)** Posługując się bazą danych OMIM (http://www.ncbi.nlm.nih.gov/, zakładka OMIM), określ dokładną lokalizację chromosomową tego genu. **f)** O czym jeszcze informuje baza OMIM? Analizując informacje zawarte pod hasłem „Inheritance" określ wzór dziedziczenia mukowiscydozy (dominujący lub recesywny). **g)** Korzystając z zakładki „Clinical Resources", a następnie „Clinical Trials" określ możliwości leczenia mukowiscydozy. Wskaż w których państwach prowadzi się najwięcej badań nad tym schorzeniem? W tym celu posłuż się mapą. (autorsto zadania Julita Matecka).