

mgr Katarzyna Stefańska

Katedra i Zakład Informatyki i Statystyki

Zalecenia dotyczące wykonywania analizy bioinformatycznej danych biologicznych

data przygotowania: 2024-04-22

Czym jest bioinformatyka?

Bioinformatyka jest interdyscyplinarną dziedziną nauki, obejmującą wykorzystanie metod obliczeniowych do badania danych biologicznych. Dane biologiczne zdobywane są na bazie przeprowadzanych w laboratorium eksperymentów i mogą dotyczyć kwasów nukleinowych, białek, lipidów, węglowodanów i innych cząsteczek. Informatyka, matematyka i statystyka dostarczają narzędzi, metod, algorytmów służących do analizy tych danych.

Odpowiednia analiza danych biologicznych jest niezwykle istotna, gdyż surowe dane (np. z sekwencjonowania) nie są informatywne. Dopiero przy wykorzystaniu narzędzi informatycznych takie dane nabierają wartości i stanowią punkt wyjścia do wnioskowania na ich podstawie.

Znaczenie bioinformatyki w naukach medycznych

Bioinformatyka umożliwia analizę sekwencji DNA, RNA i białek, co pozwala na pełniejsze zrozumienie procesów biologicznych na poziomie molekularnym (np. replikacji DNA, transkrypcji genów, translacji białek oraz regulacji ekspresji genów). Ponadto, dzięki modelowaniu homologicznemu możliwe jest przewidywanie struktury trzeciorzędowej białek na podstawie ich sekwencji aminokwasów. Jest to kluczowe dla zrozumienia funkcji biologicznych białek.

Współczesne technologie generują ogromne ilości danych biologicznych, takich jak sekwencje genomów, dane dotyczące ekspresji genów, czy dane z obrazowania medycznego. Bioinformatyka oferuje narzędzia i techniki analizy tych danych, co pozwala na wydobycie istotnych informacji oraz interpretację biologicznych zjawisk na wielu skalach, od molekularnej po populacyjną.

Niezwykle istotna jest też rola bioinformatyki w procesie odkrywania i opracowywania leków. Narzędzia bioinformatyczne umożliwiają identyfikację potencjalnych celów terapeutycznych, analizę struktury białek docelowych oraz projektowanie związków chemicznych

o pożądanymi właściwościami farmakologicznymi. Dzięki temu bioinformatyka przyspiesza proces badań nad nowymi lekami i może prowadzić do opracowania bardziej skutecznych i bezpiecznych terapii.

Narzędzia i analizy bioinformatyczne mogą odgrywać także istotną rolę w rozwoju medycyny spersonalizowanej poprzez umożliwienie analizy danych genetycznych pacjentów, co może prowadzić do bardziej precyzyjnych decyzji terapeutycznych dostosowanych do indywidualnych cech pacjentów.

Przykładowe biologiczne bazy danych oraz narzędzia do analizy bioinformatycznej

Istnieje wiele biologicznych baz danych, które są ogólnodostępne i darmowe. W poniższej tabeli zamieszczono kilka przykładowych baz danych wraz z linkami.

| Nazwa bazy | Zawartość | Link |
|-------------------------|--|---|
| GenBank | Informacje o sekwencjach DNA i RNA | https://www.ncbi.nlm.nih.gov/genbank/ |
| OMIM | Informacje o genach i chorobach genetycznych człowieka | https://www.omim.org/ |
| KEGG pathway | Informacje o szlakach metabolicznych | https://www.genome.jp/kegg/pathway.html |
| Gene Cards | Informacje o genach człowieka | https://www.genecards.org/ |
| Uniprot | Informacje o białkach i sekwencjach białkowych | https://www.uniprot.org/ |
| PDB (Protein Data Bank) | Informacje o trójwymiarowych strukturach białek | https://www.rcsb.org/?ref=nav_home |

Dodatkowo, istnieje wiele przydatnych narzędzi do przeprowadzania analiz bioinformatycznych, np.:

- narzędzia do dopasowywania sekwencji (np. BLAST, Clustal Omega)
- przeglądarki genomu (np. UCSC Genome Browser, Ensembl)
- oprogramowanie do przewidywania struktury białek (np. SWISS-MODEL, Phyre2)
- narzędzia do modelowania molekularnego (np. PyMOL, VMD)

O czym należy pamiętać przed rozpoczęciem analizy?

Przed przeprowadzeniem analizy bioinformatycznej konieczna jest znajomość podstawowych pojęć z zakresu biologii molekularnej i genetyki, m.in. zasady komplementarności zasad azotowych czy centralnego dogmatu biologii molekularnej.

Zasada komplementarności zasad azotowych

Zasada komplementarności zasad azotowych opisuje sposób, w jaki zasady azotowe (adenina, tymina, cytozyna i guanina) w niciach DNA łączą się ze sobą. Zasada ta mówi, że zasady azotowe łączą się w parze w taki sposób, że adenina (A) zawsze łączy się z tyminą (T) poprzez dwie wiązania wodorowe, podczas gdy cytozyna (C) zawsze łączy się z guaniną (G) poprzez trzy wiązania wodorowe.

To znaczy, że w podwójnej helisie DNA, jeśli w jednej nici występuje adenina, w przeciwnej nici będzie znajdować się tymina, a jeśli występuje cytozyna, w przeciwnej nici będzie guanina. Ta zasada jest kluczowa dla replikacji DNA, ponieważ umożliwia precyzyjne kopiowanie informacji genetycznej podczas podziału komórkowego, oraz dla procesu transkrypcji, podczas którego DNA jest przepisywane na RNA.

Centralny dogmat biologii molekularnej

Centralny dogmat biologii molekularnej to fundamentalna zasada opisująca przepływ informacji genetycznej w komórce. Składa się z trzech głównych etapów: replikacji, transkrypcji i translacji. Podczas replikacji DNA jest precyzyjnie kopiowane przed podziałem komórki, aby każda nowo powstała komórka miała identyczną kopię materiału genetycznego. Podczas replikacji podwójna helisa DNA jest rozwijana, a enzymy replikacyjne syntezują nowe komplementarne nici na wzór istniejących nici matrycowych.

Podczas transkrypcji informacja genetyczna z DNA jest przepisywana na cząsteczkę mRNA (RNA informacyjny). Ten proces zachodzi w jądrze komórkowym u eukariontów lub w cytoplazmie u prokariotów. Podczas transkrypcji enzym polimeraza RNA syntezuje cząsteczkę mRNA na wzór jednej z nici DNA, przy czym zasady azotowe RNA tworzą komplementarne pary z zasadami na matrycowej nici DNA (A-T, T-A, G-C, C-G). Dodatkowo, w cząsteczce mRNA zamiast tyminy (T) występuje uracyl (U).

Ostatnim etapem jest translacja, gdzie informacja zawarta w cząsteczce mRNA jest odczytywana przez rybosomy i przekształcana w sekwencję aminokwasów, tworzących łańcuch

polipeptydowy. Rybosomy przemieszczają się wzdłuż mRNA, odczytując kodony (trójki nukleotydów) i łącząc odpowiednie aminokwasy w procesie tzw. elongacji.

GenBank

GenBank to publiczna baza danych zawierająca sekwencje genomów, transkryptów oraz białek. Jest to jeden z największych i najważniejszych zasobów informacji genetycznej na świecie, stanowiący cenny zasób dla naukowców pracujących w dziedzinach biologii molekularnej, genetyki, biotechnologii oraz pokrewnych dziedzinach. GenBank przechowuje sekwencje nukleotydowe genomów różnych organizmów, zarówno eukariontów, jak i prokariontów. Sekwencje te obejmują zarówno sekwencje kodujące jak i sekwencje niekodujące. Baza danych zawiera również sekwencje RNA, w tym mRNA, rRNA, tRNA oraz inne rodzaje RNA. Ponadto, oprócz sekwencji nukleotydowych, GenBank przechowuje także sekwencje białek, które zostały przetłumaczone z sekwencji kodujących DNA lub RNA.

Każda sekwencja w GenBanku jest opatrzona metadanymi, takimi jak nazwa genów, nazwa organizmu, autorzy, referencje bibliograficzne oraz informacje dotyczące struktury, funkcji i ewolucji sekwencji. Co ważne, GenBank jest dostępny publicznie dla naukowców z całego świata. Dostępność tej bazy danych umożliwia badaczom dostęp do istniejących danych genetycznych, co przyspiesza postęp w badaniach naukowych.

Każdy rekord w GenBanku ma usystematyzowaną strukturę, składającą się z trzech głównych elementów: nagłówka, cech i sekwencji. W nagłówku znajdują się informacje takie jak numer dostępu („accession number”, czyli unikalny identyfikator danej sekwencji), opis sekwencji wraz z jej długością, słowa kluczowe i informacje o organizmie, z jakiego sekwencja pochodzi. Sekcja „cechy” zawiera szczegółowe informacje o cechach sekwencji, takie jak lokalizacja sekwencji kodującej oraz intronów i egzonów. W ostatniej części rekordu znajduje się sekwencja w postaci kolejności nukleotydów lub aminokwasów, najczęściej w formacie FASTA.

Human growth hormone gene (HGH-N), complete cds

GenBank: [M13438.1](#)

[FASTA](#) [Graphics](#)

[Go to:](#)

LOCUS HUMGHN 2657 bp DNA linear PRI 29-APR-1996
DEFINITION Human growth hormone gene (HGH-N), complete cds.
ACCESSION M13438
VERSION M13438.1
KEYWORDS Alu repeat; growth hormone; hormone; repeat region.
SOURCE Homo sapiens (human)
ORGANISM [Homo sapiens](#)
Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
Mammalia; Eutheria; Euarchontoglires; Primates; Haplorrhini;
Catarrhini; Hominidae; Homo.
REFERENCE 1 (bases 1 to 2657)
AUTHORS Seeburg,P.H.
TITLE The human growth hormone gene family: nucleotide sequences show
recent divergence and predict a new polypeptide hormone
JOURNAL DNA 1 (3), 239-249 (1982)
PUBMED [7169009](#)
COMMENT Original source text: Homo sapiens (clone: HGH-N.) (tissue library:
Lawn et al.) DNA.
The Alu family sequences are known to be transcribed by RNA
polymerase III and in [1] is inserted such that transcription would
be from the opposite strand to that of the growth hormone genes.

nagłówek

```
FEATURES             Location/Qualifiers
     source            1..2657
                     /organism="Homo sapiens"
                     /mol_type="genomic DNA"
                     /db_xref="taxon:9606"
                     /map="17q21-qter"
                     /clone="HGH-N."
                     /tissue_lib="Lawn et al."
     gene              497..2129
                     /gene="GH-N"
     prim_transcript   497..2129
                     /gene="GH-N"
                     /note="HGH-N mRNA"
     exon              497..568
                     /gene="GH-N"
                     /number=1
     CDS               join(559..568,828..988,1196..1315,1409..1573,1827..2024)
                     /gene="GH-N"
                     /note="precursor"
                     /codon_start=1
                     /product="growth hormone"
                     /protein_id="AA98618.1"
                     /translation="MATGSRTSLLAFGLLCLPWLQEGSAFPTIPLSRLLFDNAMLRLHQLAFDITYQEFEEAYIPKEQKYSFLQNPQTSLCFSESIPTPSNREETQKSNLEIRISLLLIQSWLEPVQFLRSVFNLSLVYASDQSNVYDLKKDLEEGIQLMGRLEDGSI  
TGQIFRQTYSKFDTNSHNDALLKNYGLLYCFRKMMDKVETFLRIVQCRSVEGSGI  
join(559..568,828..895)
     sig_peptide       497..568
                     /gene="GH-N"
     mat_peptide       join(896..988,1196..1315,1409..1573,1827..2021)
                     /gene="GH-N"
                     /product="growth hormone"
     intron            569..827
                     /gene="GH-N"
                     /number=1
     exon              828..988
                     /gene="GH-N"
                     /number=2
     intron            989..1195
                     /gene="GH-N"
                     /number=2
     exon              1196..1315
                     /gene="GH-N"
                     /number=3
     intron            1316..1408
                     /gene="GH-N"
                     /number=3
     exon              1409..1573
                     /gene="GH-N"
                     /number=4
     intron            1574..1826
                     /gene="GH-N"
                     /number=4
     exon              1827..2129
                     /gene="GH-N"
                     /number=5
     regulatory        2111..2116
                     /regulatory_class="polyA_signal_sequence"
                     /gene="GH-N"
     repeat_region     complement(2228..2501)
```

cechy

```

ORIGIN      1 bp upstream of EcoRI site.
1 gaattcagga ctgaatcgtg ctcaacaacc ccacaatcta ttggctgtgc ttggcccttt
61 ttccaacac acacattctg tctgggtgggt ggaggttaaa catgctgggga ggaggaaaag
121 gatagatag agaatgggat gtggctcgta gggggtctca aggactggcc tatcctgaca
181 tccttcgcc gctgacaggt tggccacat ggctgcggc cagagggcac ccactgacc
241 cttaaagaga ggacaagtgt ggtggtatct ctggctgaca ctctgtcac aacctcaca
301 acactggtga cgtggggaag ggaagaatga caagccaggg ggcatgatcc cagcatgtgt
361 gggaggagct tctaattat ccattagcac aagcccgtca gtggcccat gcataaatgt
421 agcacagaaa caggtggggt caacagtggg agagaagggg ccaggggata aaaagggccc
481 acaagagacc agctcaagga tccaagggc caactcccg aacctcag ggtcctgtgg
541 acagctacc tagctgcaat ggctacaggt aagcggcctt aaaatcctt tggcacaatg
601 tgcctgagg gggaggagcag cgacctgtag atgggacggg ggcaataacc ctcaagggtt
661 ggggttctga atgtgagtat cgccatctaa gccagattt tggccaatct cagaaagctc
721 ctggctcctt gggagatgga gagagaaaaa caaacagctc ctggagcagg gagagtgtg
781 gcctctgtct ctccggctcc ctctgttgc ctctggttc tccccaggct cccggaagtc
841 cctgctcctg gcttttggcc tgcctgect gccctggctt caagagggca gtgccttccc
901 aaccattccc ttatccaggc tttttgcaa cgctatgctc cgcccatct gtctgacca
961 gctggccttt gacacctacc aggagtgtt aagcttggg ggaatgggtg cgcctcaggg
1021 gtggcaggaa ggggtgactt tccccgctg gaaataagag gaggagacta aggagctcag
1081 ggtttttccc gaccgcaaaa atgcaggcag atgagcacac gctgagctag gttcccagaa
1141 aagtaaaatg ggagcaggtc tcagctcaga ccttggggg cggtccttct cctaggaaga
1201 agcctatctc ccaaaggaac agaagtattc attcctcag aacccccaga cctcctctg
1261 tttctcagag tctattcga caccctcaa caggagggaa acacaacaga aatcctgtag
1321 tggatgcctt ctccccagc ggggatggg gagacctgta gtcagagccc ccgggagca
1381 cagccaatgc cgtccttgc cctgcagaa cctagagctg ctcccatct ccctgctgct
1441 catccagctg tggctggagc cgtgacagtt cctcagaggt gtcttcgcca acagcctggt
1501 gtacggcgc tctgacagca acgtctatga cctcctaaag gacctagagg aaggctacca
1561 aacgctgatg ggggtgaggg tggcggcagg ggtcccaat cctggagccc cactgacttt
1621 gagagactgt gttagagaaa cactggctgc cctctttta gcagtcaggc cctgaccaa
1681 gagaactcac cttattctc atttcccctc gtgaatcctc caggctcttc tctacactga
1741 aggggaggga ggaaaatgaa tgaatgagaa agggagggaa cagtacccaa gccttggcc
1801 tctccttctc ttccttact ttgcagagc tggaaatgg cagccccgg actgggcaga
1861 tcttcaagca gactcacagc aagttcgaca caaactcaca caagatgac gcaactacta
1921 agaactacgg gctgctctac tgcctcagga aggacatgga caaggtcag acattctgc
1981 gcatctgca gtgcccctct gtggagggca gctgtggctt ctatgccc ggggtggctc
2041 cctgtgacc ctccccagtc ctctcctgg ccctggaagt tgccaactca gtgccacca
2101 gcctgtctct aataaaatta agttgcatca tttgtctga ctagggtgct ttctataata
2161 ttatgggggt ggggggggtg gtatggagca agggcccaa gttgggaaga caacctgtag
2221 ggcctgcggg gtctattcgg gaaccaagct gggagtgcagt ggcacaatct tggctcactg
2281 caatctcgc ctctgggtt caagcgattc tctgcctca gcctccgag ttgttgggat
2341 tccaggcatg catgaccagg ctacagtaat tttgtttt ttggtagaga cggggttca
2401 ccataattgc caggctggtc tccaactct aatctcaggt gatctacca ccttgcccct
2461 ccaaaattgt gggattacag gcgtgaacca ctgctccctt ccctgtctt ctgattttaa
2521 aataactata ccagcaggag gacgtccaga cacagcatag gctacctgcc atggcccaac
2581 cgggtggaca tttgagttg ttgcttgca ctgtcctctc atgcgttgg tccactcagt
2641 agatgcctgt tgaattc

```

sekwencja

Format FASTA

Format FASTA jest szeroko stosowany w bioinformatyce do przechowywania i wymiany danych sekwencji genetycznych, ponieważ jest czytelny dla ludzi i łatwy do przetwarzania przez programy komputerowe. Składa się z dwóch głównych elementów: nagłówka i sekwencji. Nagłówek zawsze rozpoczyna się od znaku większości „>”.

