Wprowadzenie do Informatyki Biomedycznej

Wykład 1:
Podstawy bioinformatyki

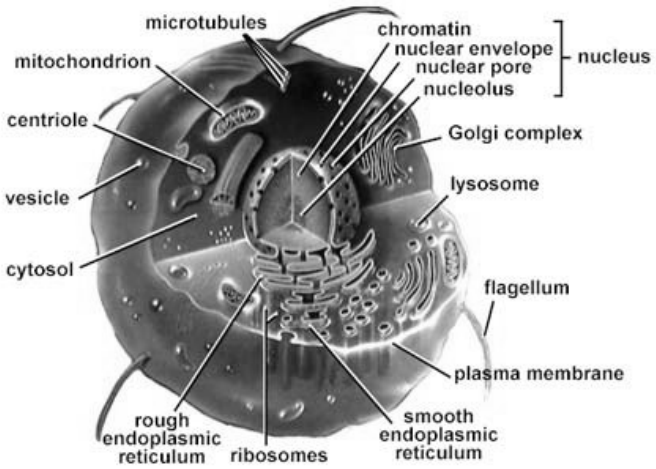
Marek Krętowski
pokój 206
e-mail: m.kretowski@pb.edu.pl
http://aragorn.pb.bialystok.pl/~mkret

Wersja 1.11

Podstawy biologiczne - komórki

- Wszystkie organizmy zbudowane są z komórek
  - komórka jest skomplikowanym systemem składającym się z wielu elementów otoczonych membraną
  - organizmy jednokomórkowe (np. bakterie czy drożdże piekarskie) lub wielokomórkowe
  - szacuje się, że w organizmie człowieka jest $6x10^{13}$ komórek 320 różnych typów (np. komórki skóry, mięśni, mózgu - neurony); rozmiar ich może się znacznie różnić
- Komórki eukariotyczne posiadają jądro (ang. *nucleus*), oddzielone od reszty komórki membraną
  - jądro zawiera chromosomy, które są nośnikami materiału genetycznego
- Kluczową cechą większości żywych komórek jest ich umiejętności wzrostu i podziału w odpowiednim środowisku (cykl komórkowy - ang. *cell cycle*)
- Komórki składają się z cząsteczek (ang. *molecules*)

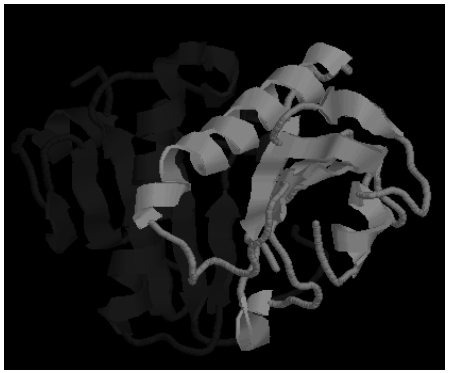
Model komórki eukariotycznej

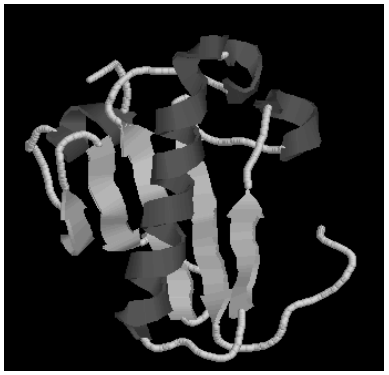


Podstawy biologiczne - cząsteczki

- Wyróżniamy 4 podstawowe cząsteczki istotne dla życia:
  - małe cząsteczki
  - białka, DNA i RNA (określane jako biologiczne makrocząsteczki)
- Małe cząsteczki mogą budować makrocząsteczki lub mogą mieć niezależne role(np. transmisja sygnałów, źródło energii, ...)
  - woda, cukry, kwasy tłuszczowe, aminokwasy, nukleotydy
  - istnieje 20 różnych cząsteczek aminokwasów, z których zbudowane są białka
- Białka są najistotniejszym budulcem i funkcjonalnymi cząsteczkami komórki (ok. 20% wagi komórek eukariotycznych; 70% to woda)
  - białka strukturalne (np. kolagen do budowy tkanki łącznej i kości)
  - enzymy katalizujące reakcje biochemiczne =>metabolizm
  - białka transmembranowe - regulatory komórkowe
- Białka posiadają złożoną strukturę trójwymiarową

Podstawy biologiczne - struktura białek (1)

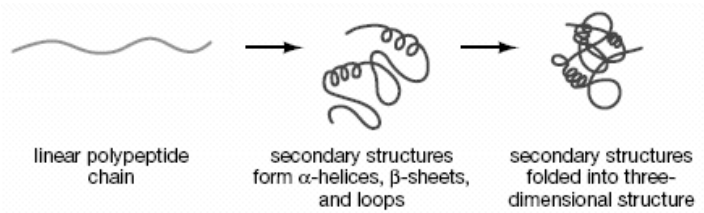


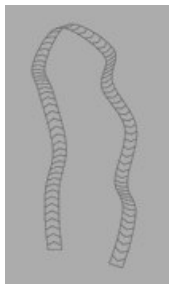


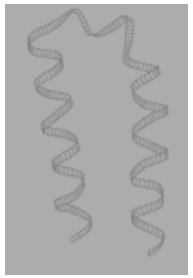
Podstawy biologiczne - struktura białek (2)

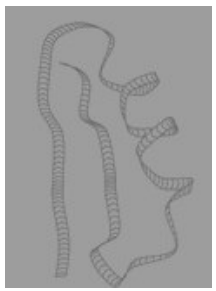
- Rozróżniane są 4 poziomy struktury białek:
  - liniowa sekwencja 20 różnych typów aminokwasów (ang. *poly-peptide chains*); podstawowa (ang. *primary*) struktura reprezentowana jako łańcuch liter odpowiadających aminokwasom; długość łańcucha od kilku do wielu tysięcy znaków (np. insulina - 51 aminokwasów, ale *titin* - 28.000)
  - struktura drugorzędna (ang. *secondary*): zaginanie się i zawijanie łańcucha; typowe podstruktury to *alfa-helices* czy *beta-strands*, połączone zwykle przez mniej regularne struktury zwane *loops*
  - struktura trzeciorzędowa (ang. *tertiary*)w rezultacie zawijania łańcuchy zbliżają się do siebie co powoduje pojawianie się różnych sił przyciągających i odpychających, dzięki którym tworzy się ustalona i relatywnie stabilna struktura trójwymiarowa
  - białko może być uformowane z większej liczby łańcuchów - struktura czwartorzędowa (ang. quaternary), np. hemoglobina zbudowana jest z 4 łańcuchów
- Przyjmuje się, że zasadniczo struktura wyższych rzędów jest uzależniona od struktury pierwszorzędowej

Podstawy biologiczne - struktura białek (3)









DNA (kwas dezoksyrybonukleinowy)

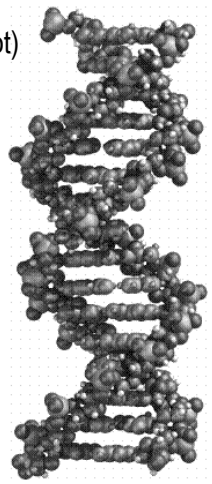
- DNA to podstawowy nośnik informacji w komórce, zorganizowany w postaci pojedynczego lub podwójnego łańcucha (ang. *strand*)
  - pojedyncza nić - polinukleotyd (ang. *polynucleotid*)
- Zbudowany z nukleotydów (4 rodzaje):
  - adenina (**A**), tymina (**T**), guanina (**G**) i cytozyna (**C**) określane mianem baz
  - nić może mieć dowolną długość i kodować dowolną sekwencję
  - końce nici są chemicznie rozróżnialne (sekwencja ma kierunek); końce nici oznaczone są przez 5' i 3', zgodnie z konwencją 5' zawsze po lewej stronie i łańcuch kodujący u góry;
  - dwie nici określane są mianem komplementarnych, gdy otrzymane są poprzez wymianę **A** z **T** oraz **C** z **G** i odwrócenie kierunku

DNA (2)

- Dwa komplementarne łańcuchy tworzą stabilną strukturę przypominającą heliks(10 par baz pozwala na pełen obrót)

```
5′ C-G-A-T-T-G-C-A-A  3′
   | | | | | | | | | |
3′ G-C-T-A-A-C-G-T-T 5′
```

- "Gęstość zapisu" informacji DNA ~75 GB na cm
- Komplementarność dwu nici jest wykorzystywana podczas powielania informacji genetycznej w procesie tworzenia nowych cząsteczek DNA (replikacja DNA)



RNA

- RNA podobnie do DNA zbudowana jest z nukleotydów, przy czym zamiast tyminy występuje uracyl (**U**)
- Skutkuje to tym, że RNA występuje zwykle w postaci pojedynczego łańcucha, przy czym może mieć skomplikowaną strukturę przestrzenną wynikającą z powiązań pomiędzy częściami tego samego łańcucha
- RNA ma wiele różnych funkcji w komórce, i w zależności od pełnionej funkcji rozróżniane są rodzaje np. mRNA, tRNA (oba odgrywają kluczową rolę w syntezie białek)
- RNA może dowiązać się do pojedynczej nici DNA, przy czym **T** jest zastąpione przez **U**

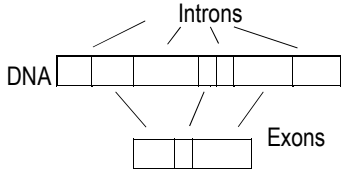
```
C-G-A-T-T-G-C-A-A  DNA
| | | | | | | | |
G-C-U-A-A-C-G-U-U  RNA
```


Chromosomy i genom

- W typowej komórce znajduje się wiele długich podwójnych nici -cząsteczek DNA zorganizowanych w postaci chromosomów
- Człowiek posiada 23 pary chromosomów; długość DNA w pojedynczej komórce człowieka po wyprostowaniu rzędu 1 m
- Genom organizmu jest tworzony poprzez DNA zawarte w chromosomach i mitochondriach (śladowe ilości w porównaniu do chromosomów)
- Wszystkie komórki organizmu zawierają identyczny genom
  - poza pewnymi specyficznymi wyjątkami, np. czerwone krwinki, które w stanie pełnego zróżnicowania nie mają jądra
- Rozmiar genomu różni w zależności od organizmu
  - bakteria (1 chromosom): od 400.000 do 10.000.000
  - drożdże (12 chromosomów): 14.000.000
  - mucha (4): 300.000.000 - robak (6): 100.000.000
  - człowiek (23): ~ 2.850.000.000 (Nature 2004)

Geny i synteza białek (1)

- **Gen** jest to ciągły obszar (odcinek) cząsteczki DNA, na podstawie którego złożony mechanizm molekularny może odczytać informację genetyczną (zakodowaną jako łańcuch **A**, **T**, **G** i **C**) i na jej podstawie utworzyć szczególny rodzaj białka lub kilka różnych białek
- Synteza białek:
  1. Transkrypcja : kopiowane jest pre mRNA, m-messenger)
  2. Sklejanie (ang. splicing) : eliminowane są introny a eksony są łączone tworząc mRNA
  3. Translacja : złożony proces powstawania białek poprzez łączenie aminokwasów w kolejności zakodowanej w mRNA. Każda kolejna trójka nukleotydów (kodon) koduje 1 aminokwas (kod jest redundantny, 64 kodony - 20 aminokwasów)



Początkowo uważano, że jeden gen pozwala tworzyć jedno białko; w tej chwili wiadomo, że tak nie jest (w wyniku alternatywnego sklejania oraz post-modyfikacji)

Geny i synteza białek (2)

- Według najnowszych szacunków (stan na rok 2004) człowiek ma 20-25 tys. genów (znacznie mniejsza liczba niż wcześniej sądzono)
  - ponad 19 tys. lokalizacji już potwierdzono, 2 tys w trakcie
- Człowiek ma ponad 1 tysiąc genów, które pojawiły się dopiero 3 mln lat temu (okres występowania australopiteka)
  - w tych najmłodszych genach zapisane są informacje m.in. o białkach związanych ze zmysłem powonienia, układem odpornościowym i rozmnażaniem się
- Znaleziono też 32 geny "umierające", czyli takie które już nie funkcjonują

Bioinformatyka - definicja dziedziny

- Bioinformatyka jest nauką o tym w jaki sposób informacja jest reprezentowana i przekazywana w systemach biologicznych, poczynając od poziomu molekularnego;
- Bioinfomatyka jest aktualnie w okresie szybkiego rozwoju, gdyż gwałtownie rosną potrzeby związane z przechowywaniem, wyszukiwaniem i analizą informacji biologicznej (w szczególności biologia molekularna, np. Human Genome Project)

- Genetyka a genomika
  - genetyka zajmuje się pojedynczymi genami i ich funkcjonowaniem
  - genomika zajmuje się nie pojedynczymi genami, ale funkcjonowaniem i zależnościami wszystkich genów w genomie, a także interakcją genów z czynnikami środowiskowymi

Dopasowywanie sekwencji
(ang. sequence alignment)

- Najbardziej podstawowe zadanie polegające na porównywaniu pary (lub większej liczby) sekwencji w celu stwierdzenia czy są one podobne i jak ustawić je względem siebie
- Mutacje (zamiana, dodanie lub usunięcie znaków)
- Poszukiwanie optymalnego dopasowania
  - dokładne dopasowanie
  - częściowe dopasowanie zachowujące całościowe własności
  - przerwy
- Wykorzystanie programowania dynamicznego (wysoka złożoność)
- Metody heurystyczne (FASTA, BLAST - Basic Linear Alignment and Search Technique)

Dopasowywanie struktury
(ang. structural alignment)

- Znając trójwymiarowe współrzędne poszczególnych elementów tworzonych złożoną strukturę 3D jak je wzajemnie ustawić aby wychwycić podobieństwa i różnice
- Istnieje wiele algorytmów dopasowywania struktury i wyznaczania stopnia podobieństwa, bazując na nich możliwa jest np. klasyfikacja białek (SCOP -Systematic Classification of Proteins, bazuje na kształcie i funkcjach)
- Zbliżonym problemem jest wykorzystanie struktury dużych biocząsteczek i struktury małych organicznych cząstek (np. leku) w celu zbadania ich interakcji
- Szczególnie istotne podczas tworzenia nowych leków, pozwalające znacząco obniżać koszty badań eksperymentalnych

Przewidywanie struktury i funkcji na podstawie sekwencji

- Jednym z najbardziej ważnych wyzwań bioinformatyki jest przewidywanie na podstawie nowo uzyskanej sekwencji DNA (lub sekwencji aminokwasów w białku) struktury cząsteczek jak również ich funkcji
- Należy przy tym zdawać sobie jasno sprawę z zagrożeń związanych wnioskowaniem bez przeprowadzenia badań eksperymentalnych; tym niemniej nawet aktualnie zebrane sekwencje pozwalają na dobrą predykcję w określonych przypadkach (np. w przypadku wystarczająco dużego podobieństwa makrocząsteczek białek - pow. 40%)
- Dzięki opracowaniu podstawowych algorytmów (np. dopasowania sekwencji i struktury) pojawia się szansa na bardziej zintegrowaną analizę procesów w których poszczególnych cząsteczki grają rolę oraz na odkrywanie sposobów manipulowania cząsteczkami w celach leczniczych

Bioinformatyczne bazy danych

- Pierwotne bazy danych sekwencji nukleotydowych (INSD)
- Pierwotne bazy danych sekwencji białek (PIR, MIPS, Swiss-Prot, ...)
  - Swiss-Prot – minimalny poziom redundancji, liczne powiązania z rekordami innych baz oraz wysoka jakość adnotacji do sekwencji; format rekordów EMBL
- Bazy danych rodzin białek (PROSITE, PRINTS, Pfam...)
  - tzw. wtórne; często na podstawie Swiss-Prot i TrEMBL
- Złożone bazy danych wzorców sekwencji białek
- Bazy danych struktur białek, np:
  - PDB (Protein Data Bank) – informacje o strukturze przestrzennej makrocząsteczek (białek, peptydów, wirusów, ...); wyznaczone za pośrednictwem dyfrakcji rentgenowskiej, spektroskopii jądrowej rezonansu magnetycznego i modelowania
  - SCOP (Structural Classification of Proteins) – wyniki klasyfikacji białek przeprowadzonej na podstawie badania zależności ewolucyjnych i strukturalnych
- i wiele innych (często mocno wyspecjalizowanych)

Bazy sekwencji nukleotydowych

- Podstawowe bazy danych sekwencji nukleotydowych (INSD):
  - European Molecular Biology Laboratory (EMBL) – Anglia, zarządzana przez European Bioinfomatics Institute (EBI)
  - GenBank, rozwijana przez National Center for Biotechnology Information (NCBI); zawiera (stan na sierpień 2009, tylko tradycyjne, bez danych uzyskanych z wysokoprzepustowych technik sekwencjonowania) 106,533,156,756 baz w 108,431,692 rekordów sekwencji
  - DNA Data Bank of Japan (DDBJ), Japonia
- Wymieniają się informacjami – niemal identyczne
- Nadrzędny cel: zapewnienie publicznego i nieograniczonego dostępu do informacji zawartych w sekwencjach DNA; dla potrzeb badań
- Większość czasopism wymaga aby powołując się na nowo zidentyfikowaną sekwencję zdeponować ją w jakiejś publicznie dostępnej bazie