BIOINFORMATYKA

edycja 2019 / 2020

wykład 1

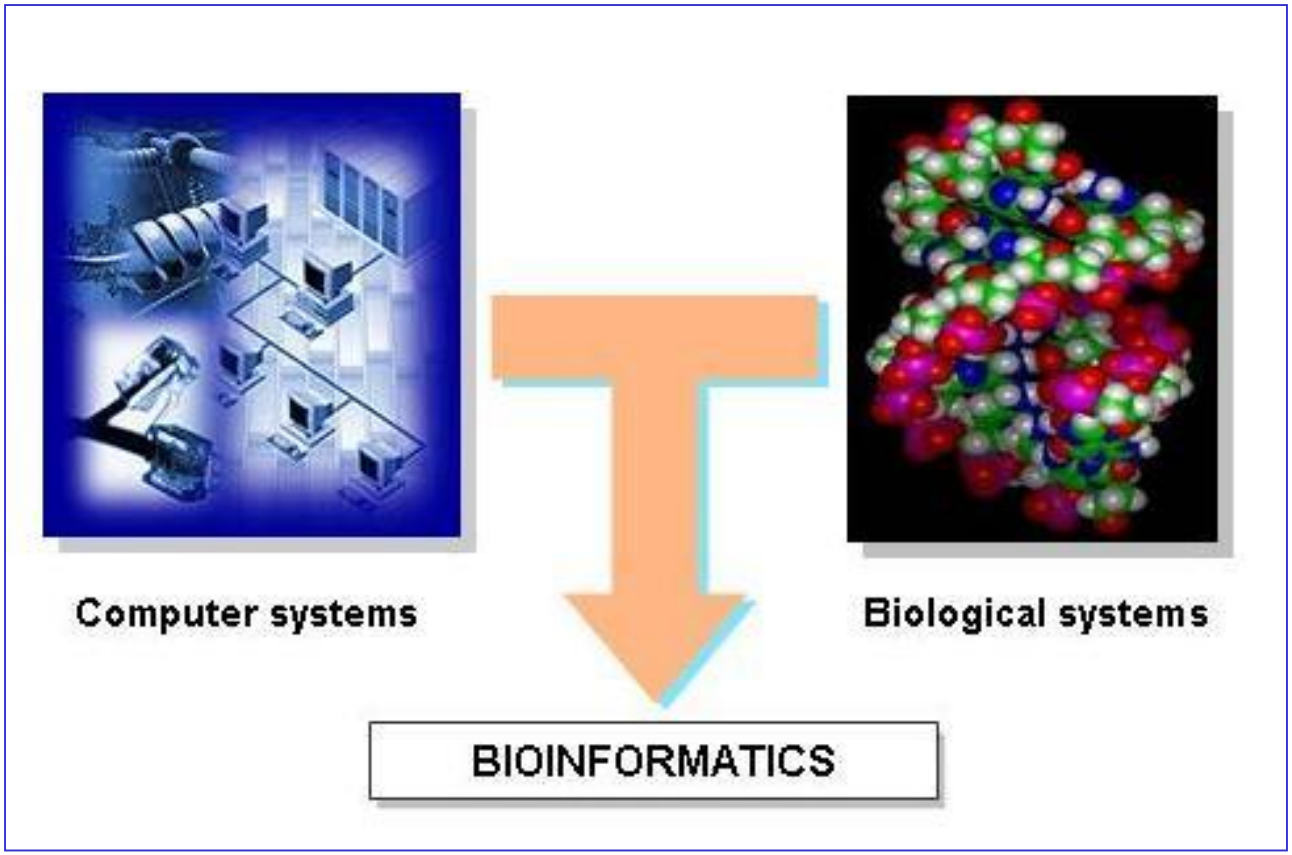
Zadania bioinformatyki

dr Jacek Śmietański
jacek.smietanski@ii.uj.edu.pl
http://jaceksmietanski.net

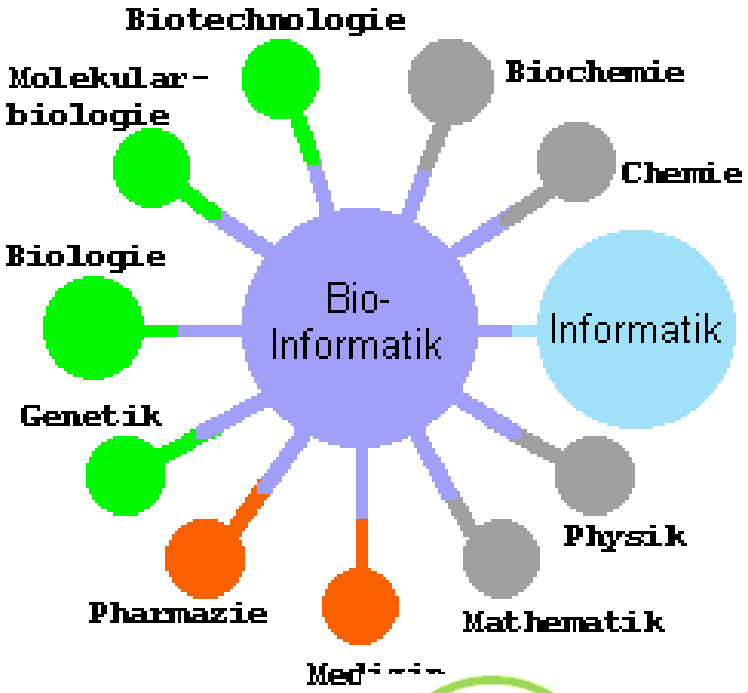
Bioinformatyka wśród innych nauk

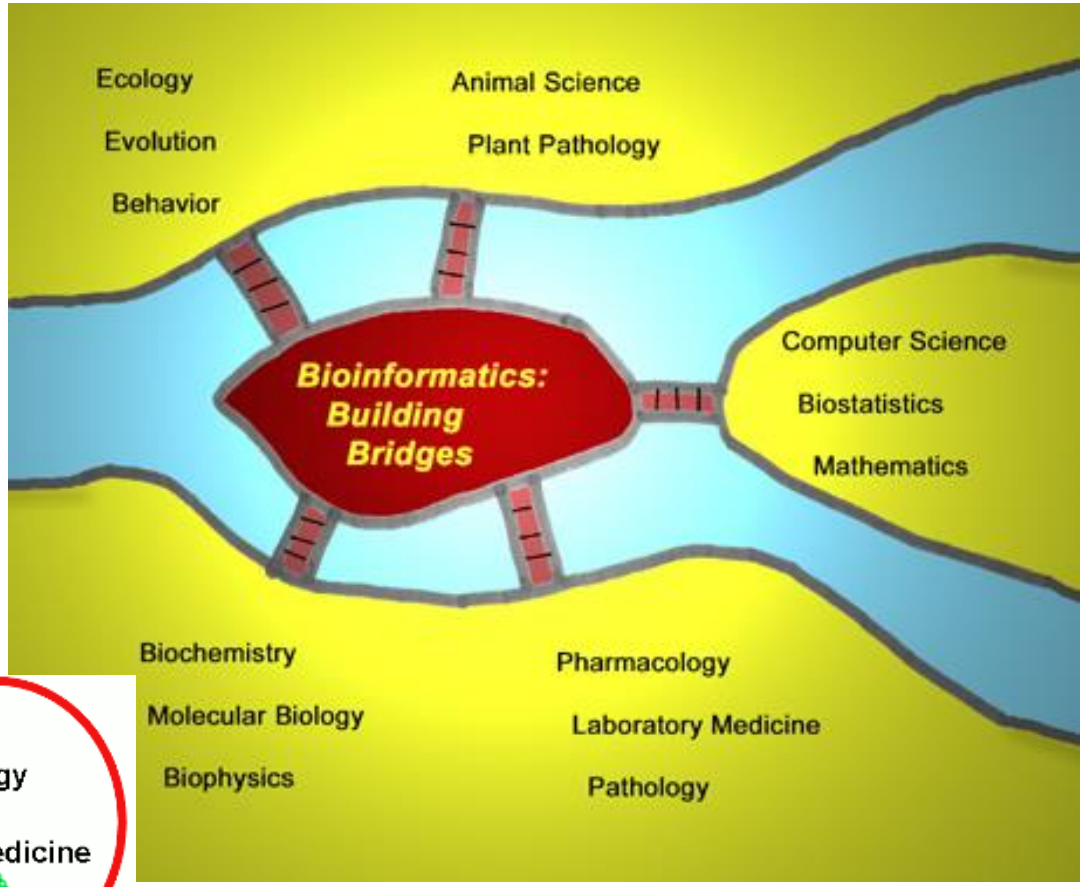
Czym zajmuje się bioinformatyka?

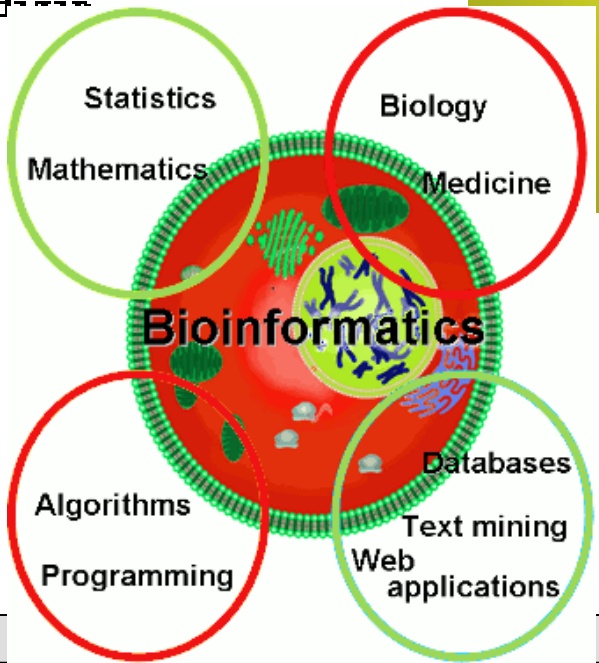
Intuicja: wykorzystanie komputerów w badaniach biologicznych.



Różne ujęcia







Formalna definicja NIH *(Narodowego Instytutu Zdrowia, USA)*

„Research, development, or application of computational tools and approaches for expanding the use of biological, medical, behavioral or health data, including those to acquire, store, organize, archive, analyze, or visualize such data.”

Definicja bardzo obszerna (obejmuje praktycznie wszystkie nauki o życiu).

Nie jest to jedyna istniejąca definicja.
Nie ma jednoznacznego, precyzyjnego określenia zakresu bioinformatyki.
Pamiętajmy też, że jest to nauka bardzo dynamicznie się rozwijająca, co za tym idzie, zakres badań też może się zmieniać.

Wielu badaczy, mówiąc o bioinformatyce, ma na myśli głównie aspekty związane z biologią na poziomie molekularnym (DNA, RNA, białko).

Osobiście uznaję definicję NIH, ale ten przedmiot koncentrował się będzie wyłącznie na aspektach molekularnych.

Często pojęcia te są utożsamiane ze sobą.
Z kolei źródła dokonujące rozróżnienia często robią to w zgoła odmienny sposób.

Np. wg „*Harper's Illustrated Biochemistry*”:
„Bioinformatyka to zbieranie i wykorzystywanie istniejących danych, natomiast istotą biologii obliczeniowej jest wykorzystanie mocy obliczeniowej w eksperymentach biologicznych.”

W podręczniku Xionga:
„Bioinformatyka różni się od powiązanej z nią dziedziny zwanej biologią obliczeniową, gdyż ogranicza się do analizy sekwencji, struktury oraz funkcji genów i genomów oraz odpowiadających im produktów ekspresji. Dlatego często określa się ją mianem molekularnej biologii obliczeniowej. Biologia obliczeniowa natomiast obejmuje wszystkie obszary biologii, które wymagają obliczeń. Na przykład w modelowaniu matematycznym ekosystemów i dynamiki populacji, w zastosowaniu teorii gier do analiz behawioralnych i rekonstrukcjach filogenetycznych wykorzystujących dane kopalne stosuje się narzędzia obliczeniowe, które nie muszą mieć związku z makrocząsteczkami biologicznymi”.

Bioinformatyka a biologia obliczeniowa (2)

Biologia obliczeniowa

Przetwarzanie danych wcale nie musi być trudne pojęciowo i algorytmicznie skomplikowane – wymaga jednak wykonania wielu obliczeń (dlatego przymiotnik „obliczeniowa”). Zajęcie mało twórcze, wręcz mechaniczne.

Bioinformatyka

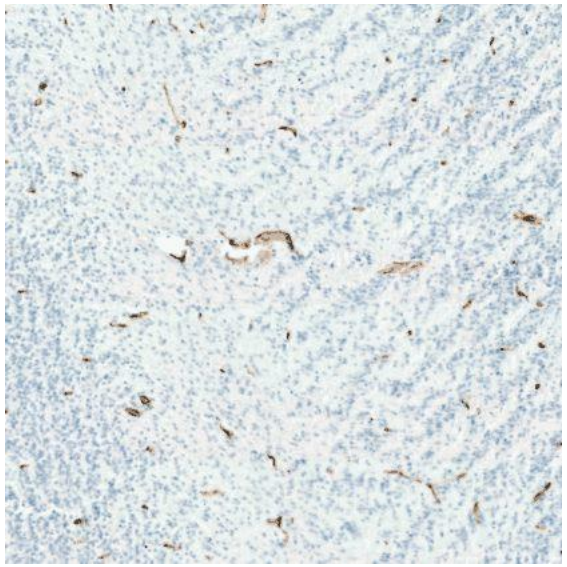
Wykorzystuje często zaawansowane techniki i algorytmy opracowane w ramach rozwoju informatyki. W wielu zadaniach wymaga indywidualnego podejścia do problemu i dedykowanych algorytmów.

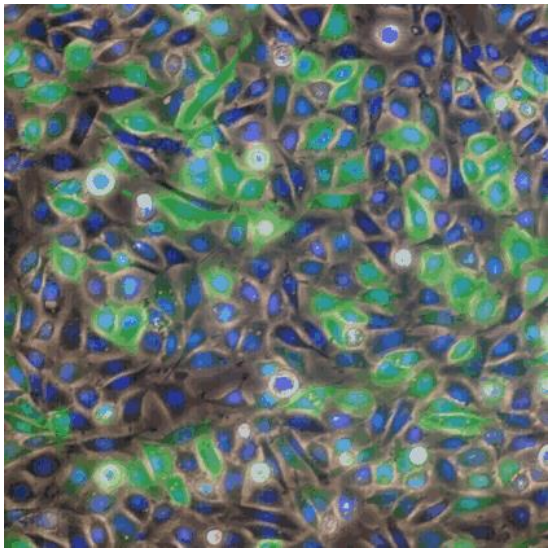
Jeszcze jedna opinia:

„biologia obliczeniowa zajmuje się poznaniem tego co jest, natomiast bioinformatyka – tworzeniem tego, czego jeszcze nie ma”

W temacie istoty bioinformatyki polecam też wykład prof Jacka Błażewicza pt. „Bioinformatyka i jej perspektywy”: http://www2.cs.put.poznan.pl/wp-content/uploads/2011/11/wyklad_inauguracyjny_2011.pdf

Rozpoznawanie obrazów?





a) obraz medyczny na poziomie tkankowym (tu: tomografia)
b) obraz medyczny na poziomie komórkowym (mikroskopowy)
c) obraz biologiczny (mikroskopowy)

Zgodnie definicją NIH – **tak**, to wchodzi w zakres bioinformatyki.
Ale wielu bioinformatyków nie uwzględnia tego obszaru.
Na tym wykładzie zagadnienia związane z analizą obrazów zostaną* pominięte.

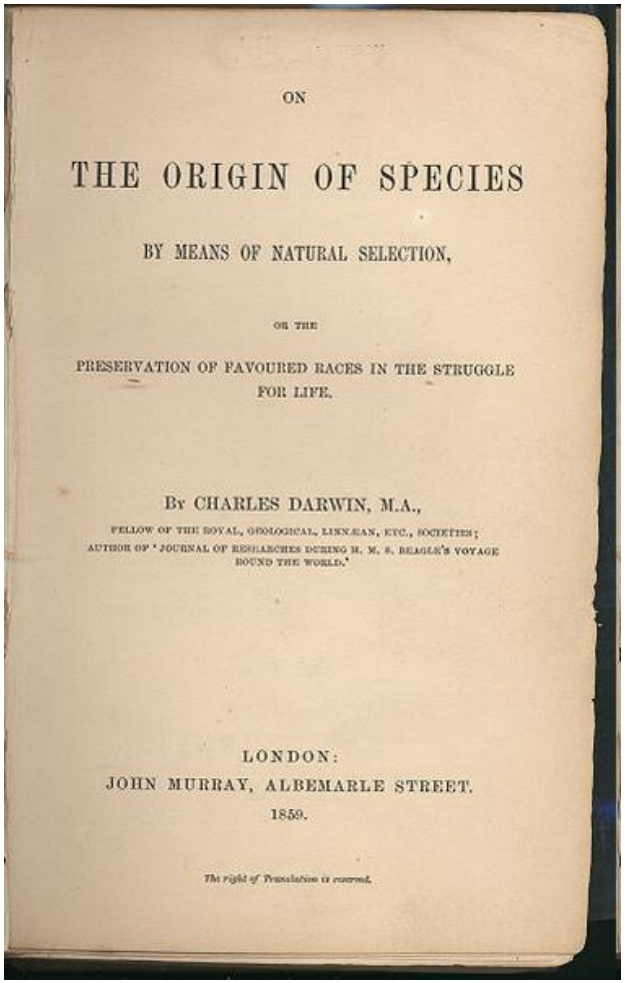
* Do analizy obrazów wrócimy na chwilę podczas omawiania metod analizy mikromacierzy, są to jednak stosunkowo proste zagadnienia (w porównaniu z przykładami powyżej), a z drugiej strony nie będziemy wnikali w szczegóły stosowanych tam algorytmów.

Rozwój bioinformatyki (1)

1859 – Charles Darwin

Podstawy teorii ewolucji:
publikacja pracy „*O powstawaniu gatunków drogą naturalnego doboru czyli o utrzymywaniu się doskonalszych ras w walce o byt*”

(„On the Origin of Species by Means of Natural Selection, or the Preservation of Favoured Races in the Struggle for Life”)


ON

THE ORIGIN OF SPECIES

BY MEANS OF NATURAL SELECTION,

OR THE

PRESERVATION OF FAVOURED RACES IN THE STRUGGLE FOR LIFE.

By CHARLES DARWIN, M.A.,

FELLOW OF THE ROYAL, GEOLOGICAL, LINNÆAN, ETC., SOCIETIES;
AUTHOR OF 'JOURNAL OF RESEARCHES DURING H. M. S. BEAGLE'S VOYAGE ROUND THE WORLD.'

LONDON:
JOHN MURRAY, ALBEMARLE STREET.
1859.

Rozwój bioinformatyki (2)

1865 – Mendel eksperymentując z grochem, wykazuje, że cechy dziedziczą się w odrębnych jednostkach;

1869 – Meischer wyizolował DNA;

1895 – Rőntgen odkrywa promienie X;

1902 – Sutton proponuje chromosomową teorię dziedziczności;

1911 – Morgan z współpracownikami stabilizuje tą teorię, badając muszkę owocówkę;

1943 – Astbury obserwuje wzór DNA przy użyciu promieni X;

1944 - Avery, MacLeod i McCarty wykazują, że DNA przenosi cechy dziedziczne (nie białka!)

Rozwój bioinformatyki (3)

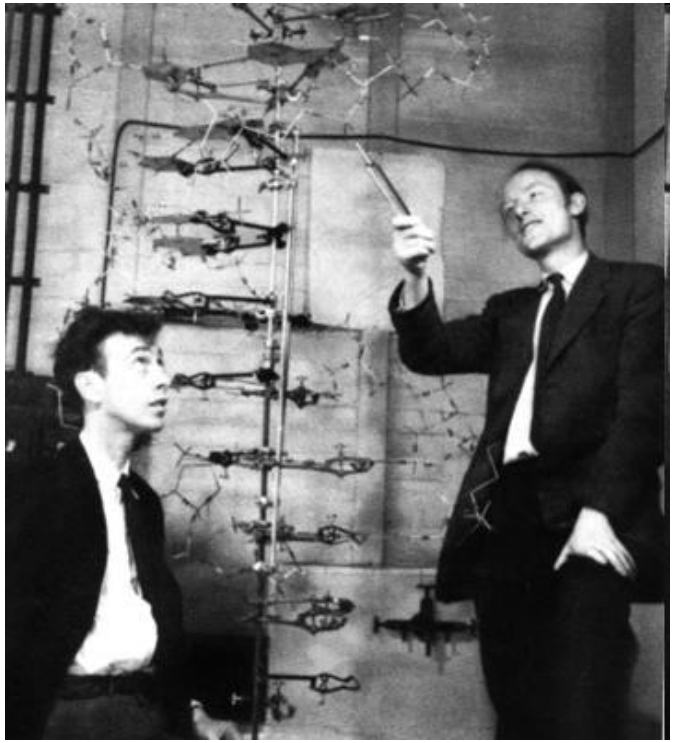
1951 - Pauling and Corey przewidują strukturę II-rzędową białek (α-helisę i β-kartkę) (Proc. Natl. Acad. Sci. USA, 27: 205-211, 1951; Proc. Natl. Acad. Sci. USA, 37: 729-740, 1951);

1953 – Watson i Crick proponują model podwójnej helisy DNA, bazując na badaniach krystalograficznych Franklin i Wilkins (Nature, 171: 737-738, 1953);

1955 – Sanger przedstawia pierwszą sekwencję białkową (insulina bydlęca);

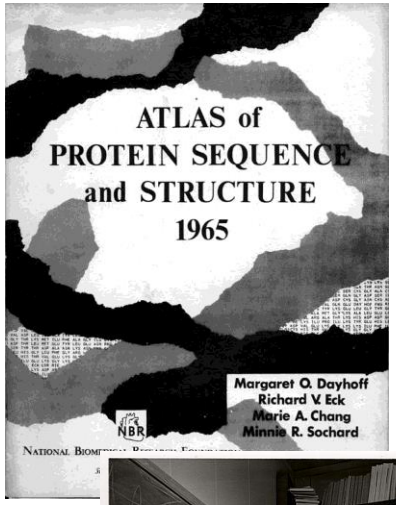
1955 – Kornberg izoluje enzym polimerazę DNA;

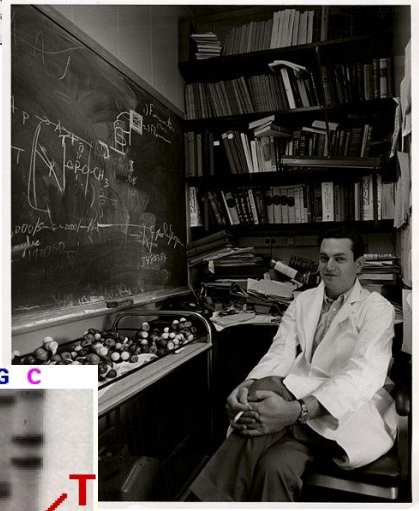
1958 – powstaje pierwszy układ scalony w korporacji Texas Instruments;

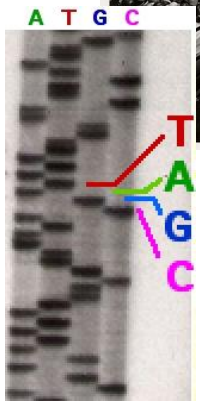


Rozwój bioinformatyki (4)

- 1959 - Perutz i Kendrew otrzymują pierwszą strukturę krystalograficzną białka (hemoglobina i mioglobina);
- 1961 – Brenner, Jacob i Meselson odkrywają mRNA przekazujące informację z DNA jądra do cytoplazmy;
- 1965 – Dayhoff – atlas sekwencji i struktur białkowych;
- 1965 – Nirenberg, Khorana, Ochoa i inni łamią kod genetyczny;
- 1970 – powstaje algorytm do porównywania sekwencji (Needleman-Wunsch);
- 1972 – Berg ze współpracownikami tworzą pierwszą rekombinowaną molekułę DNA;
- 1973 – Cohen odkrywa klonowanie DNA;
- 1975 – Sanger i inni (Maxam, Gilbert) opracowują metody sekwencjonowania;







Rozwój bioinformatyki (5)

1977 - pierwsza kompletna sekwencja genu (bakteriofag FX174) – 5386 zasad;
1981 – algorytm Smith-Waterman;
1981 – IBM wprowadza komputer osobisty na rynek;
1982 – powstaje baza danych GenBank;
1982 – zsekwencjonowano genom faga lambda;
1983 – algorytm poszukiwania sekwencji (Wilbur-Lipman);
1983 – Mullins odkrywa reakcję PCR;
1985 - Lipman i Pearson odkrywają algorytm FASTP;
1986 – utworzenie bazy SWISS-PROT;
1986 – ogłoszono The Human Genome Initiative;
1988 – Lipman i Pearson – algorytm FASTA;

Rozwój bioinformatyki (6)

1988 – powstaje National Center for Biotechnology Information (NCBI);

1990 – powstaje program BLAST;

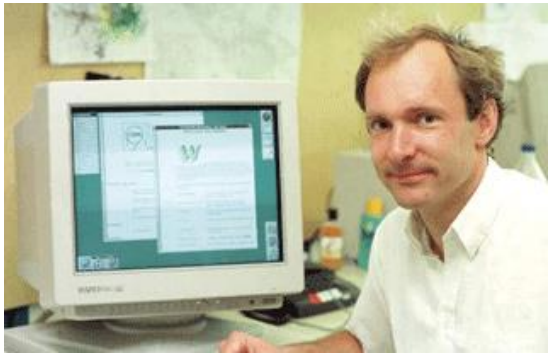
1990 – oficjalnie startuje Human Genome Project;

1991 – instytut badawczy CERN w Genewie zapowiada powstanie protokołów, które utworzą sieć World Wide Web (Berners-Lee);

1991 - opisano utworzenie i użycie sekwencji EST;

1992 - The Institute for Genomic Research (TIGR) utworzony przez Ventra w Rockville;

1994 – EMBL European Bioinformatics Institute, Hinxton, UK;



Rozwój bioinformatyki (7)

1995 – pierwszy genom bakteryjny (*Haemophilus influenzea*) zostaje zsekwencjonowany (1.8 Mb);

1996 – Affymetrix produkuje pierwszą komercyjną mikromacierz DNA;

1996 – zsekwencjonowanie genomu drożdży (pierwszy kompletny genom eukariotyczny);

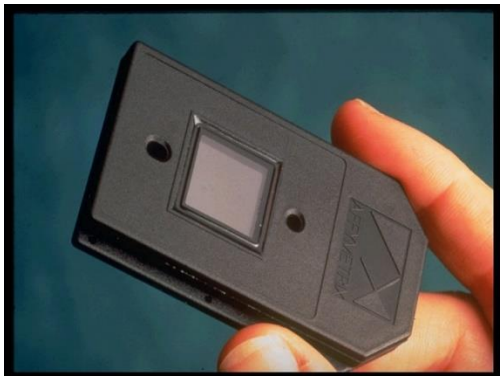
1997 – opublikowano algorytm PSI-BLAST;

1997 – genom E.coli zsekwencjonowany (4,6 Mb);

1998 – genom C. elegans zsekwencjonowany (pierwszy kompletny genom organizmu wielokomórkowego, 97 Mb);

1998 - Venter zakłada Celera w Rockville;

1998 - The Swiss Institute of Bioinformatics powstaje w Genewie;



CELERA

SIB Swiss Institute of Bioinformatics

Rozwój bioinformatyki (8)

1999 – pierwszy kompletny chromosom ludzki (HGP);

2000 – genom *Drosophila melanogaster* kompletny;

2000 – chromosom 21 kompletny;

2001 – opublikowanie sekwencji genomu ludzkiego (3,000 Mb);

2003 – genom ludzki kompletny;

2007 – Human Metabolome Project

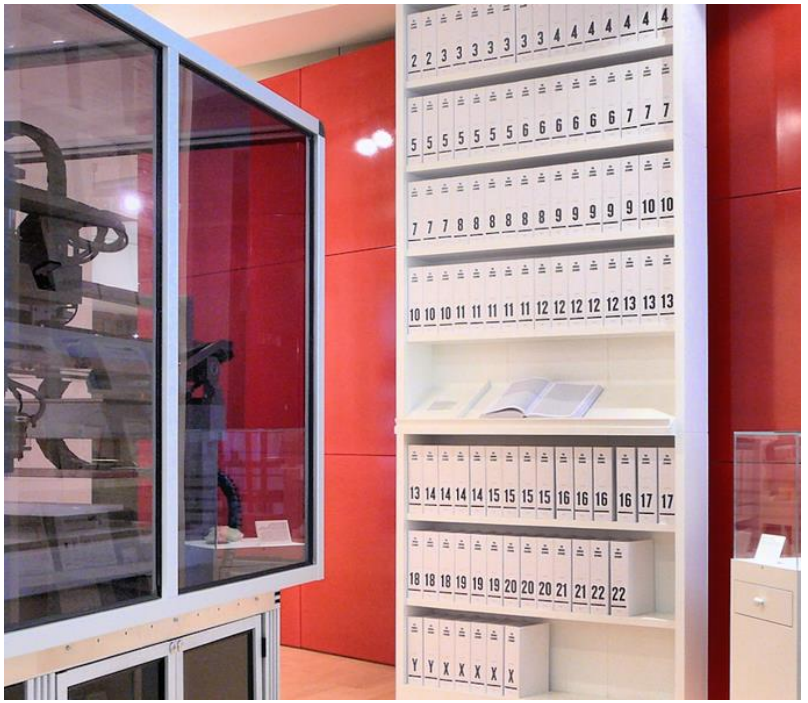
2008 – startuje European Genotype Archive

2010 – mapa ekspresji genów ludzkich

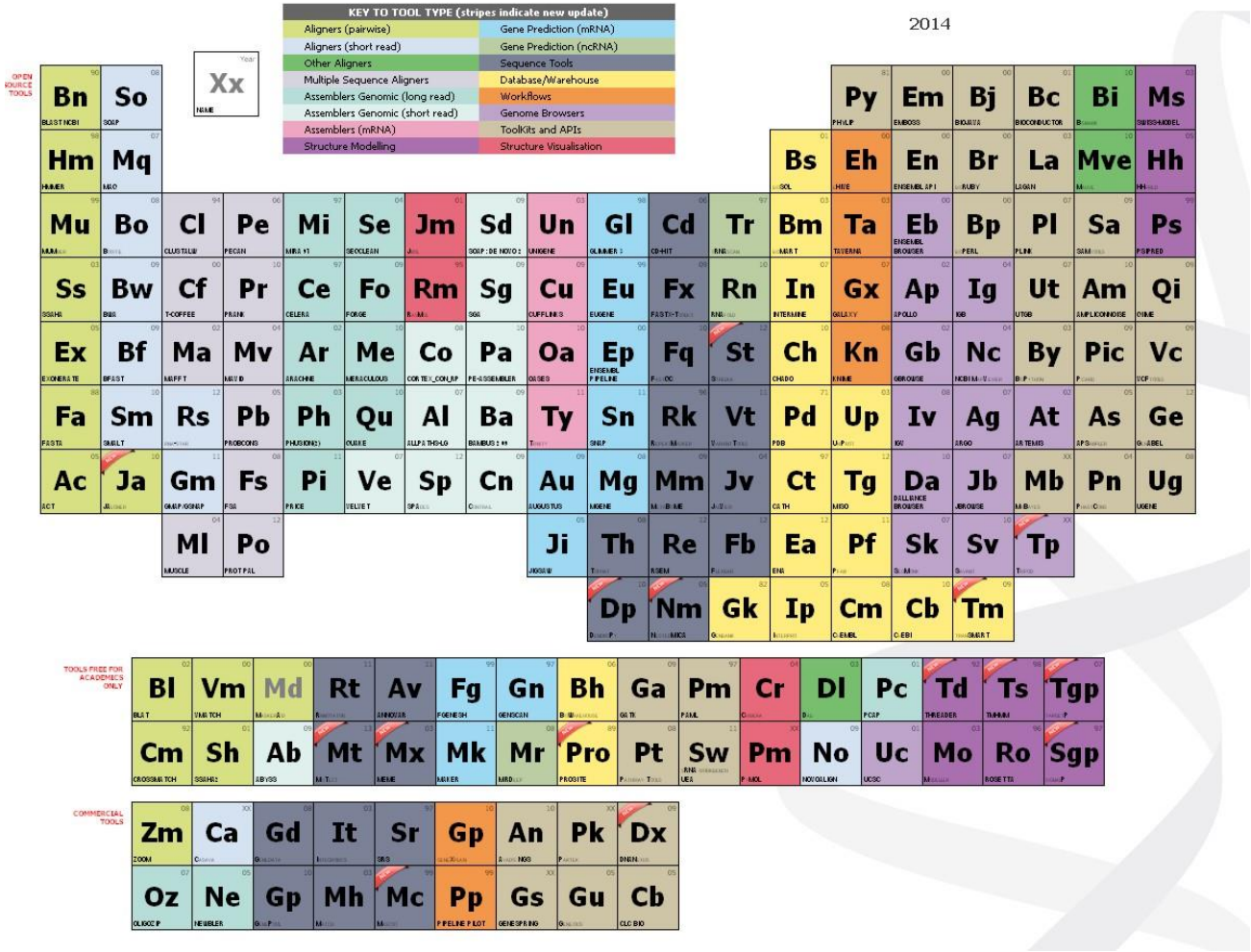
2012 – 1000 Genomes Project

2014 – startuje Elixir

…



Układ okresowy bioinformatyki (rozwój oprogramowania)



Układ przedstawia narzędzia bioinformatyczne pogrupowane wg klucza tematycznego. Warto zajrzeć na stronę źródłową, gdzie jest symulacja dynamiki rozwoju tych narzędzi oraz szereg dodatkowych informacji.

http://elements.eaglegenomics.com

Ewolucja bionformatyki

- Data storage
- Data science
- Meta datasets
- Web serwices
- Standarization

Bioinformatyka w praktyce

Główne obszary zastosowań

Medycyna
np. medycyna personalizowana

Farmaceutyka
np. projektowanie leków

Kryminalistyka
np. identyfikacja sprawców

Sądownictwo
np. ustalanie ojcostwa

Rolnictwo
np. tworzenie nowych odmian

Archeologia
np. badania paleontologiczne

Zagadnienia powiązane

Zarządzanie dużą ilością danych *(Big Data)*

Eksploracja danych *(Data Mining)*

Uczenie maszynowe *(Machine Learning)*

Teoria grafów *(Graph Theory)*

Problemy optymalizacyjne

Algorytmika

Programowanie

Bioinformatyka II UJ: organizacja przedmiotu

Harmonogram wykładów

1. Wprowadzenie do bioinformatyki
2. Zadania bioinformatyki
3. Bioinformatyczne bazy danych
4. Globalne dopasowanie par sekwencji
5. Lokalne dopasowanie par sekwencji, istotność statystyczna
6. Przeszukiwanie baz sekwencyjnych (BLAST)
7. Dopasowania wielosekwencyjne
8. Analizy filogenetyczne
9. Sekwencjonowanie DNA, składanie genów i genomów
10. Transkryptomika; eksperymenty mikromacierzowe
11. Aminokwasy i białka
12. Przewidywanie struktur drugorzędowych
13. Przewidywanie struktur przestrzennych
14. RNA
15. Przewidywanie interakcji, dokowanie, modelowanie sieci

Zasady zaliczenia

50+ pkt	laboratoria
50 pkt	projekt
+	wykłady

Laboratoria:

- na każdym spotkaniu można otrzymać max 4 pkt
- specyfikacja w materiałach do poszczególnych laboratoriów

Projekt:

- temat wybieramy z listy udostępnionej przez wykładowcę
- implementacja: python 3 (algorytm, testy, dokumentacja)
- publiczne repozytorium na githubie
- obowiązkowe konsultacje w trakcie realizacji
- obrona w sesji na prawach egzaminu

Szczegółowe zasady na repozytorium przedmiotu:
https://github.com/dadoskawina/Bioinformatics_lecture_2019

Polskie Towarzystwo Bioinformatyczne
http://ptbi.org.pl

Konferencje:

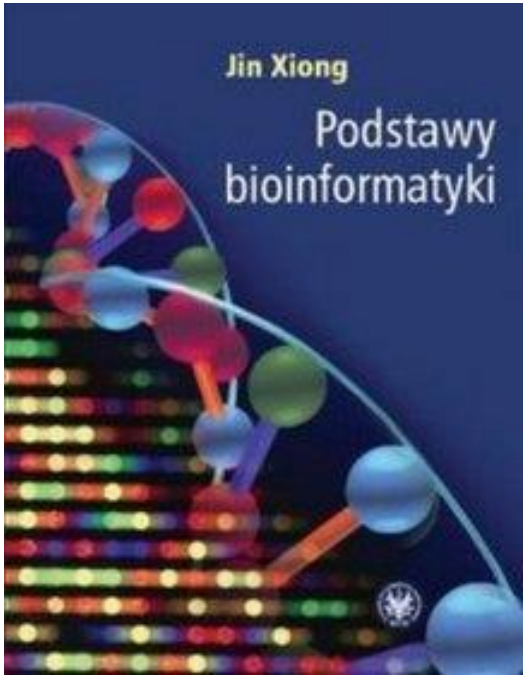
- BIT (Bioinformatics in Torun), czerwiec
- Sympozjum PTBI, wrzesień

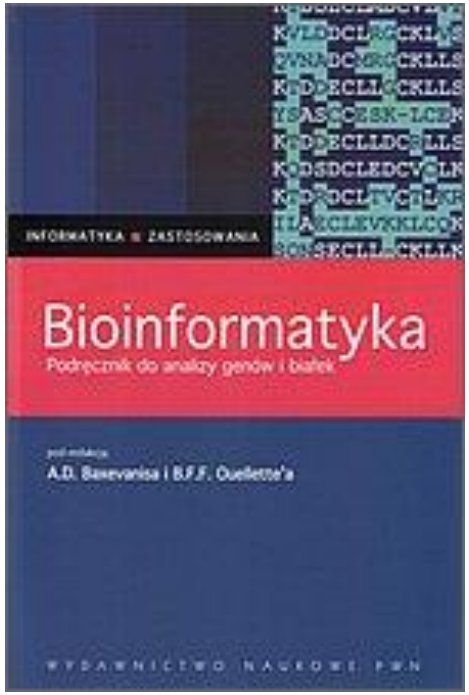
Konkurs prac magisterskich.

Literatura uzupełniająca w języku polskim

W języku polskim wydane zostały zaledwie trzy książki. Wszystkie dosyć dawno i wszystkie raczej dla biologów niż informatyków:

1. Jin Xiong, Podstawy bioinformatyki (2011)
2. Paul G. Higgs, Teresa K. Attwood, Bioinformatyka i ewolucja molekularna (2008)
3. A. D. Baxevanis, B. F. F. Ouellette, Bioinformatyka: podręcznik do analizy genów i białek (2005)





Poziomy rozważań i centralny dogmat

Poziomy organizacji (szczegółowość reprezentacji)

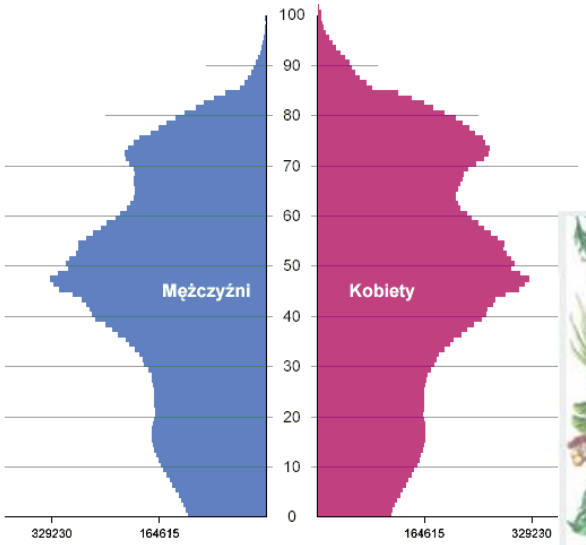
Gatunek

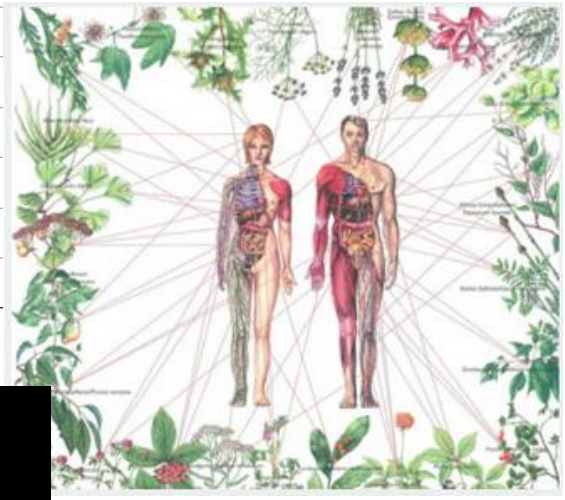
Populacja

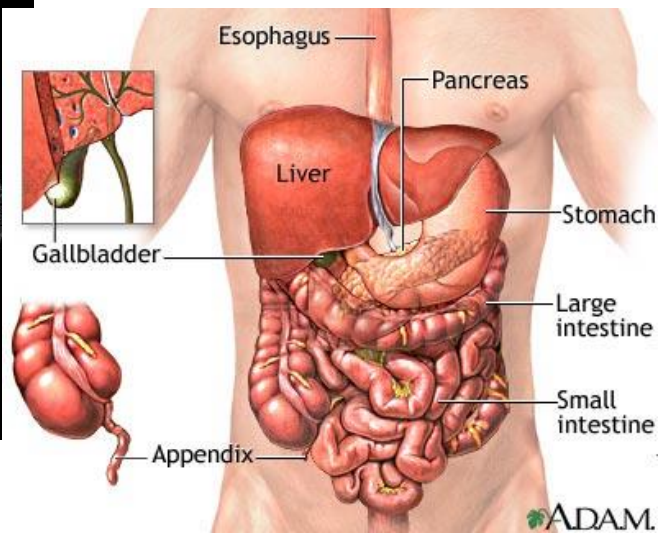
Organizm

Tkanka

Narząd

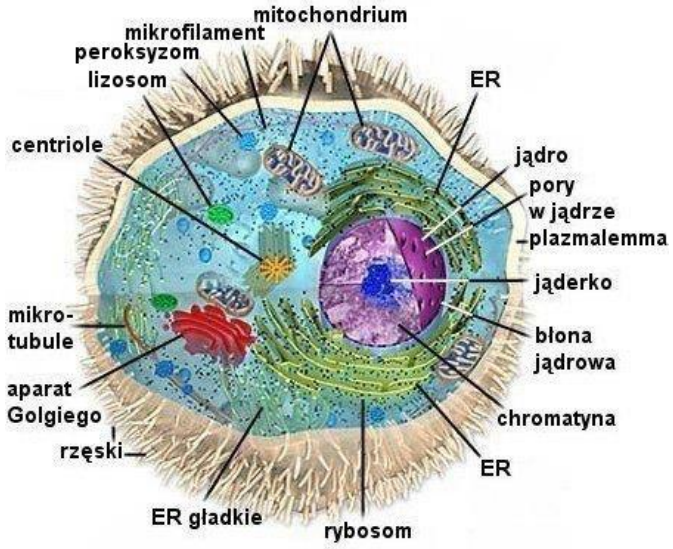




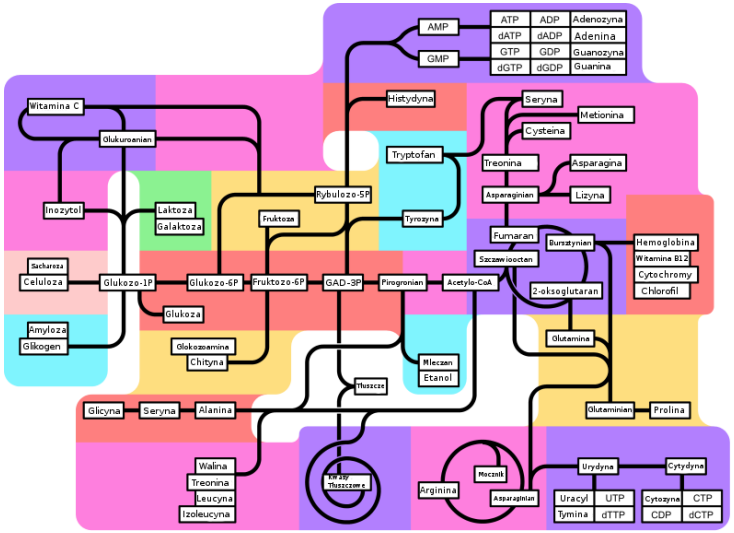


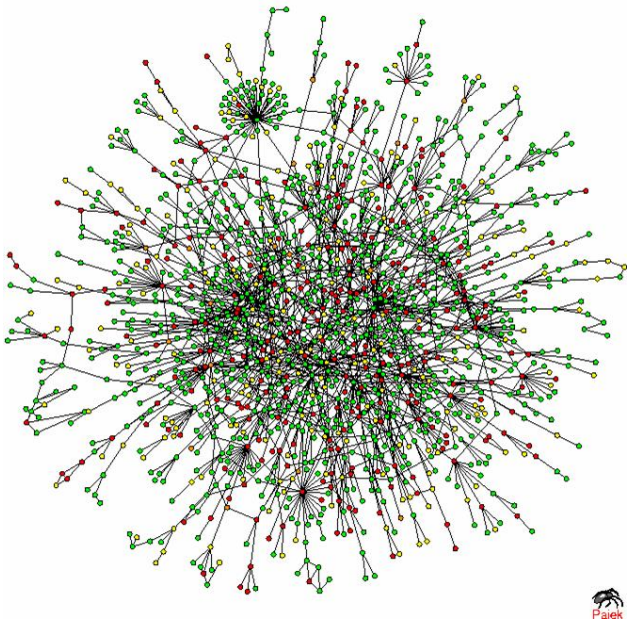
Poziomy organizacji (2)

Komórka



Szlak metaboliczny

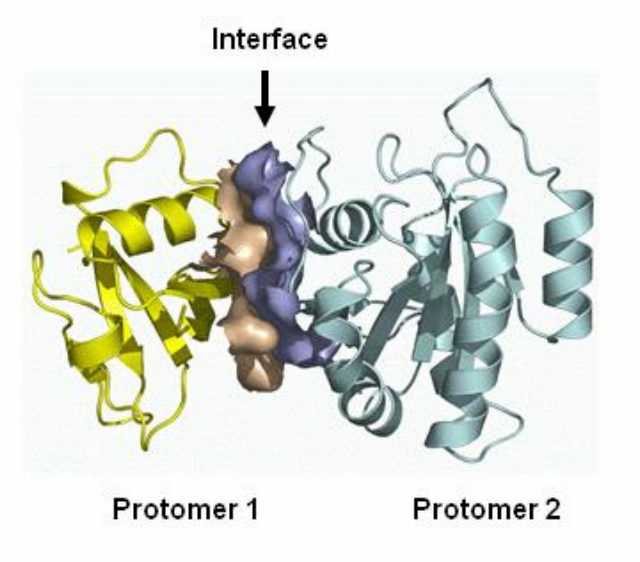


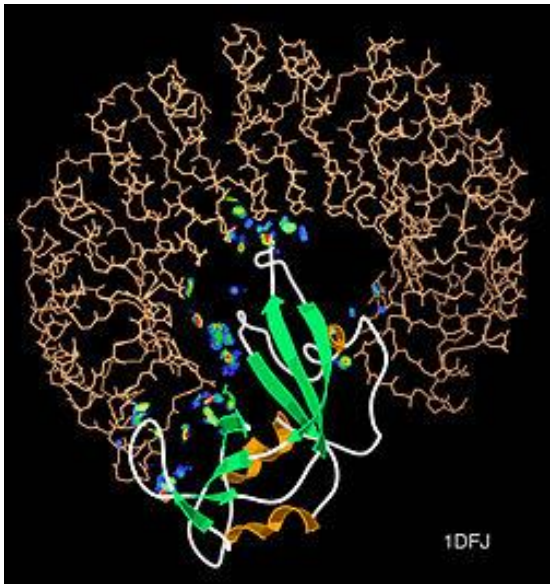


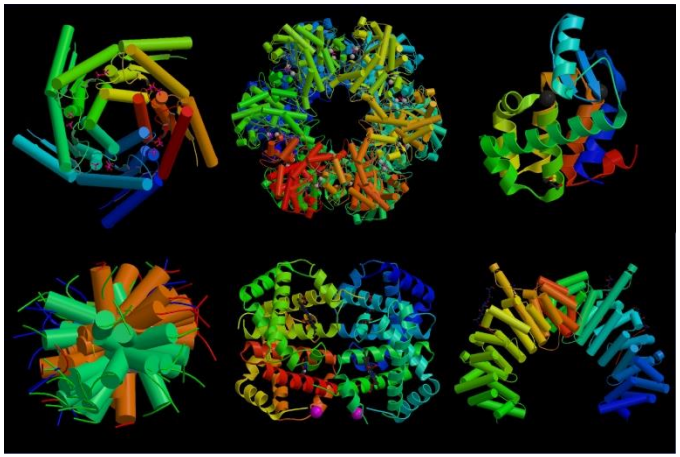
Poziomy organizacji (3)

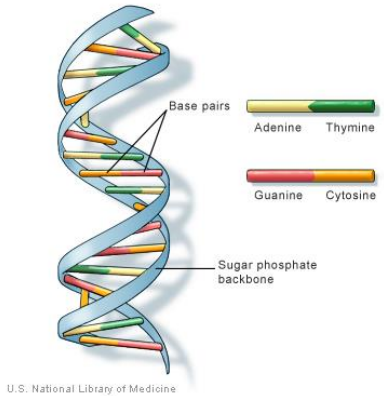
Interakcja

Cząsteczka





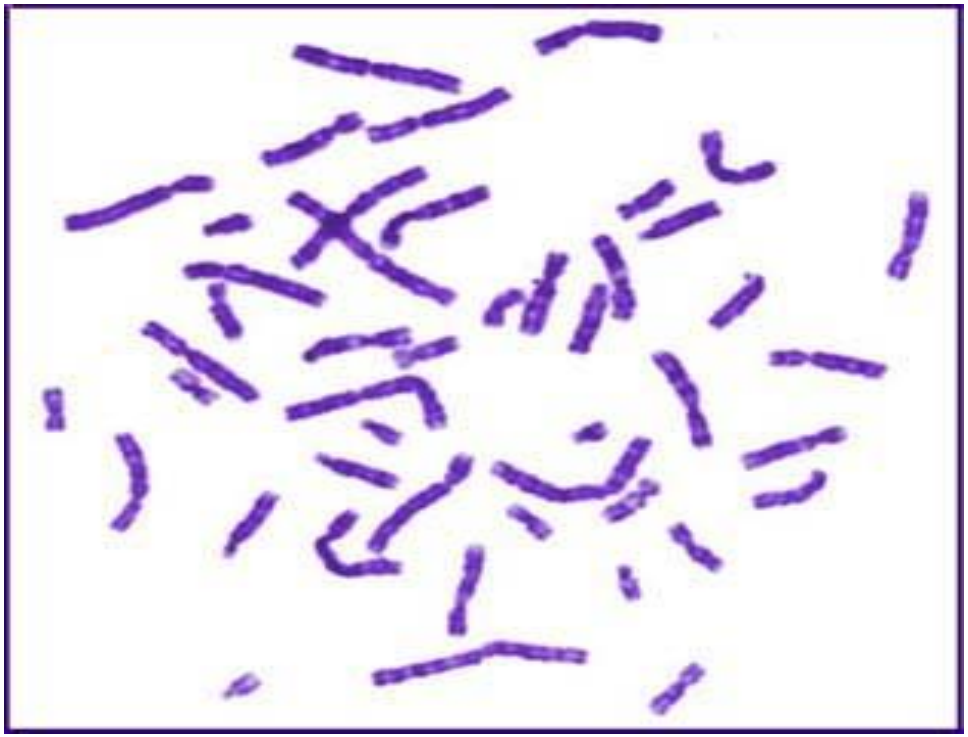


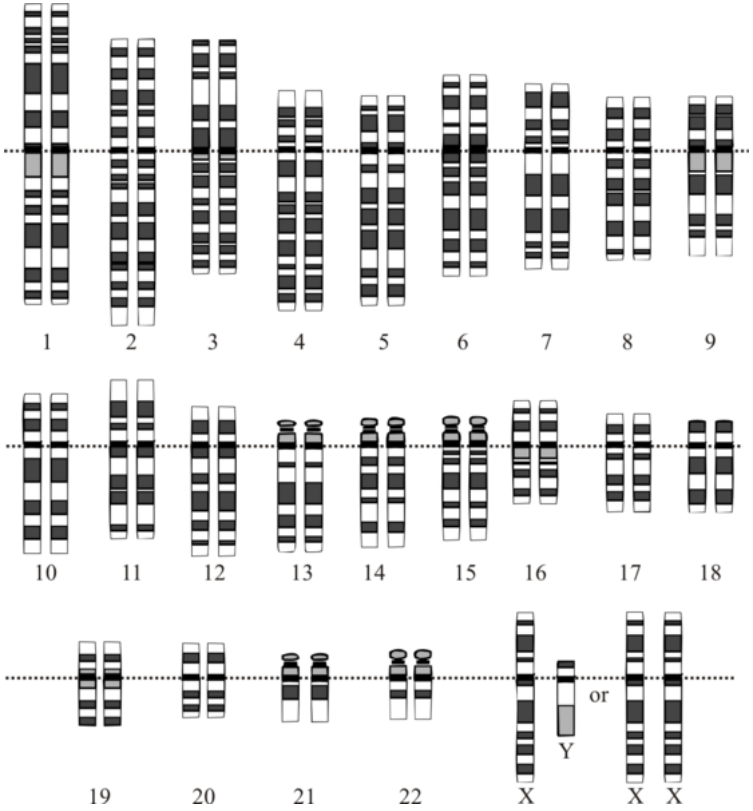


Genom

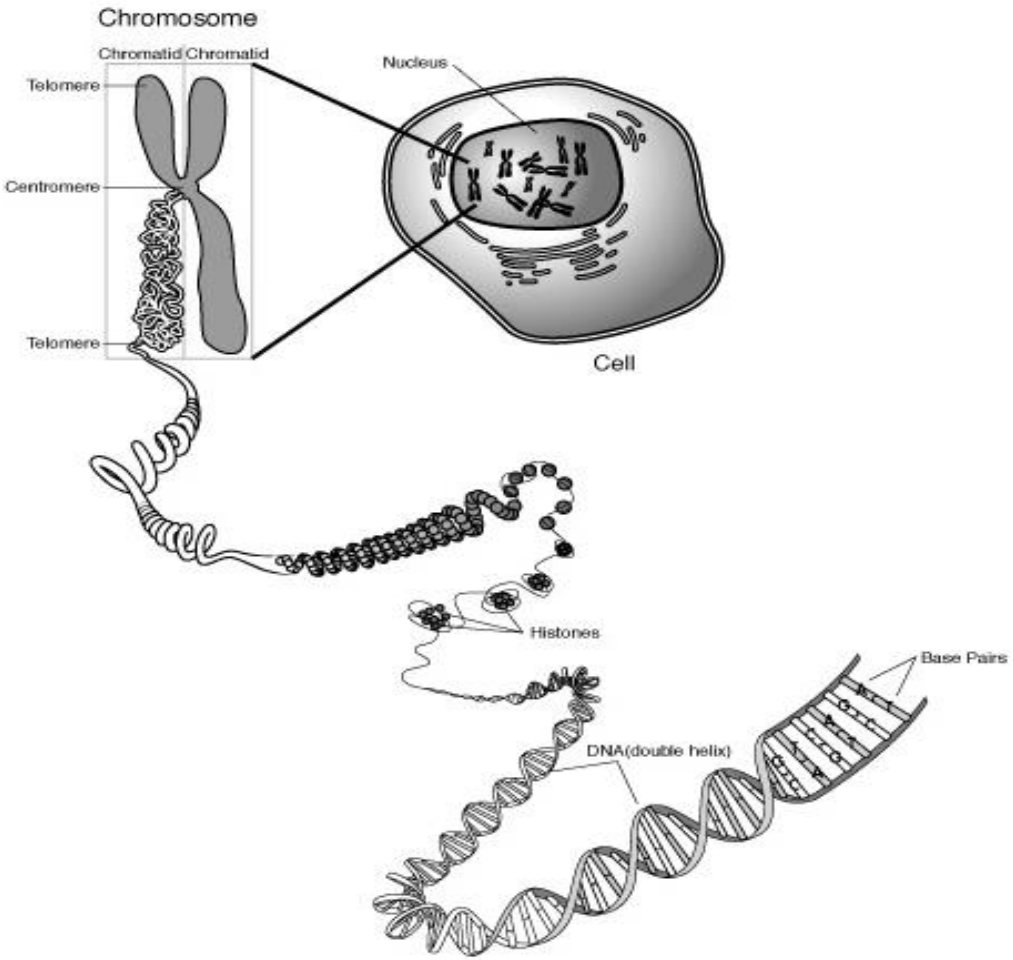
Materiał genetyczny organizmu, zbudowany z DNA. Praktycznie każda komórka posiada pełną kopię swojego genomu.

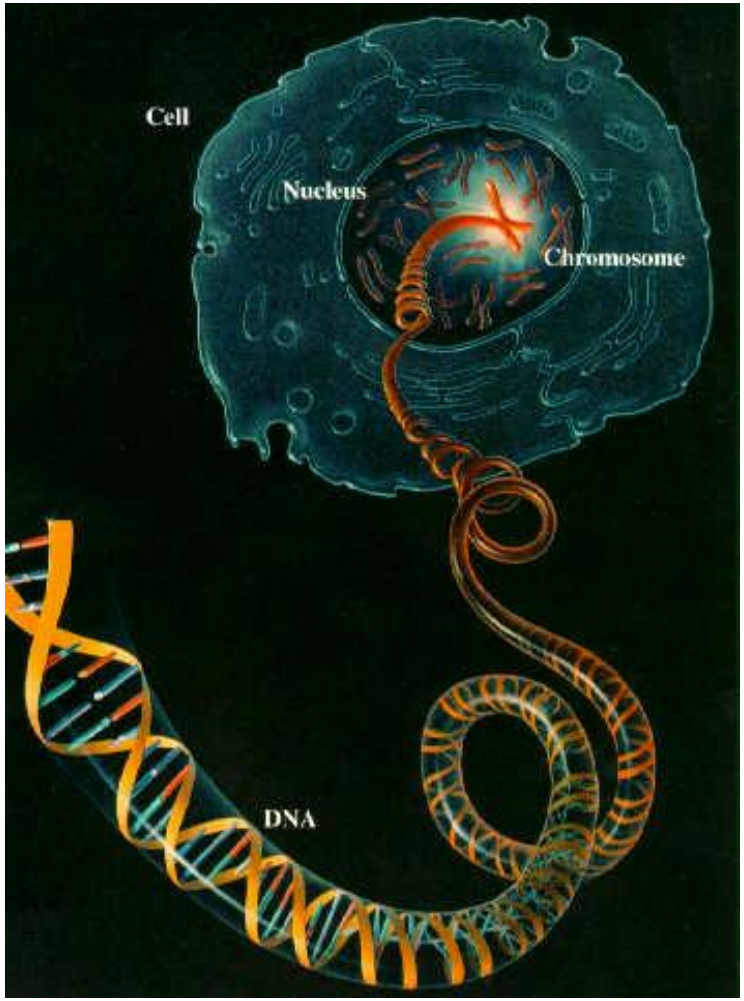
U organizmów wyższych, genom znajduje się w jądrze komórkowym, upakowany w zestawie chromosomów (liczba chromosomów jest stała dla każdego gatunku; u człowieka są to 23 pary).





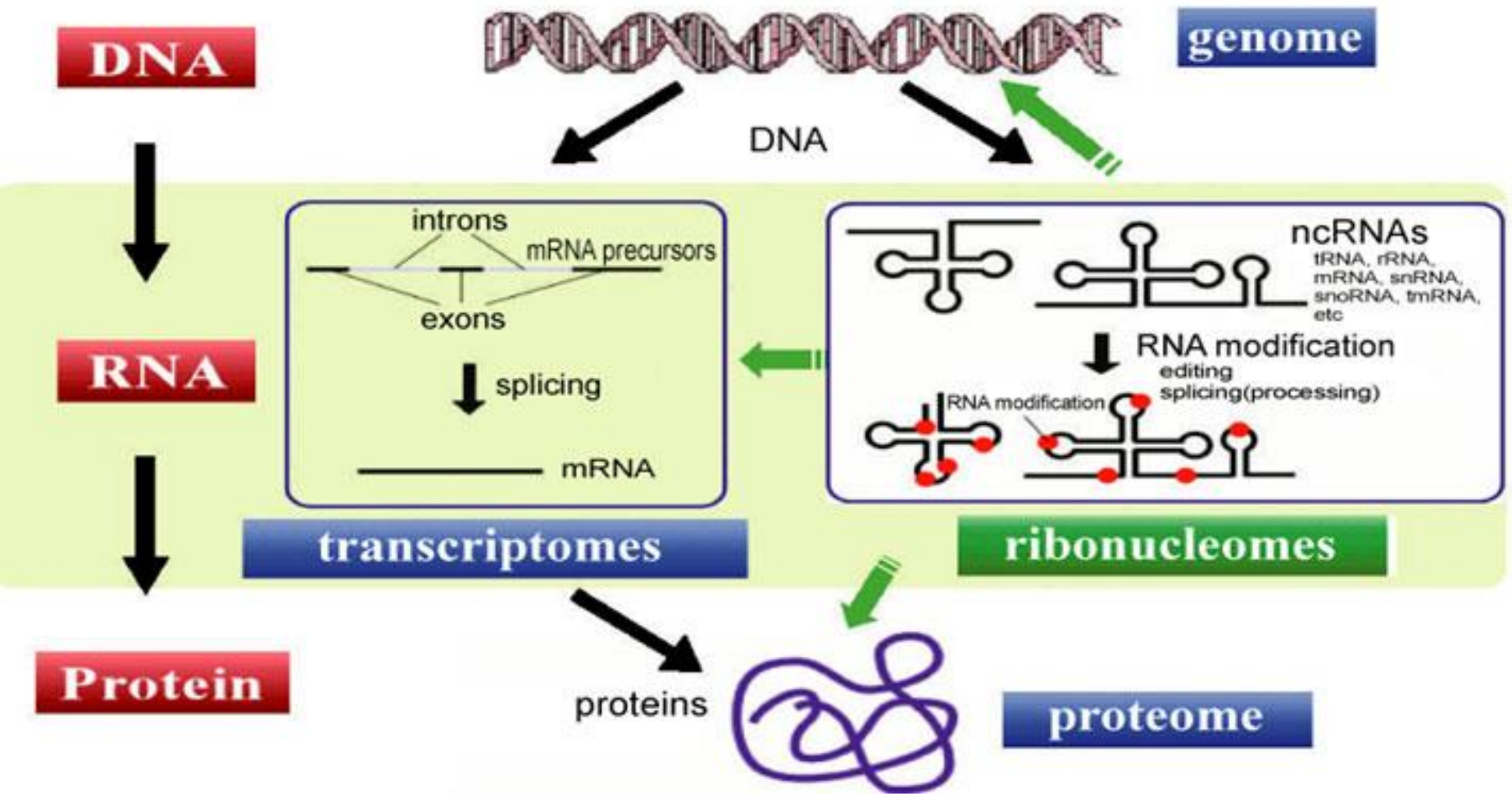
Hierarchiczna organizacja chromosomów





Upakowanie i lokalizacja w odrębnej przestrzeni komórkowej (jądro) zapewnia ochronę przechowywanej w DNA informacji.

Obszary zainteresowań na poziomie molekularnym

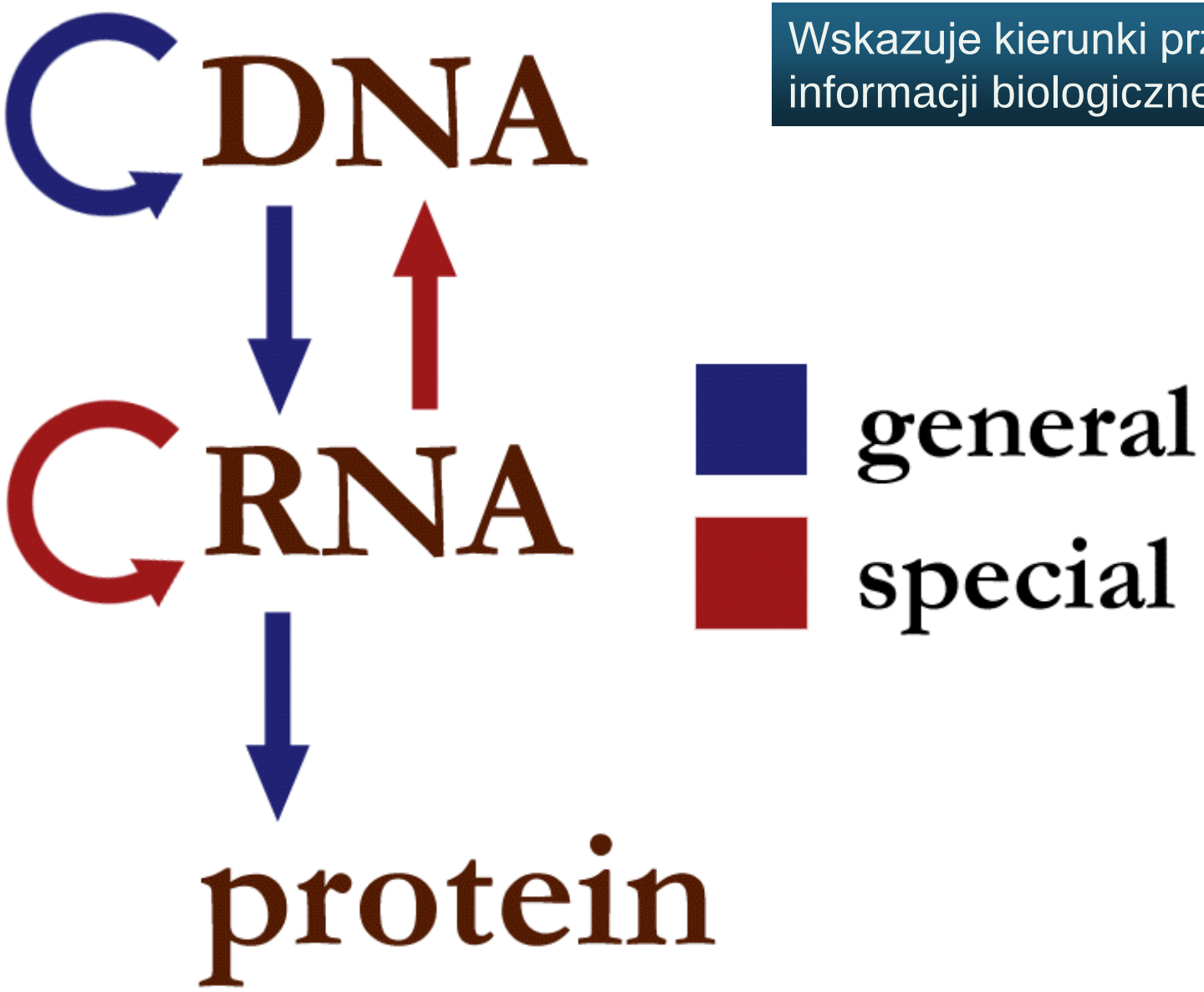


Źródło: http://www.whatisepigenetics.com/wp-content/uploads/2013/07/ncrna.jpg

Omy i omiki

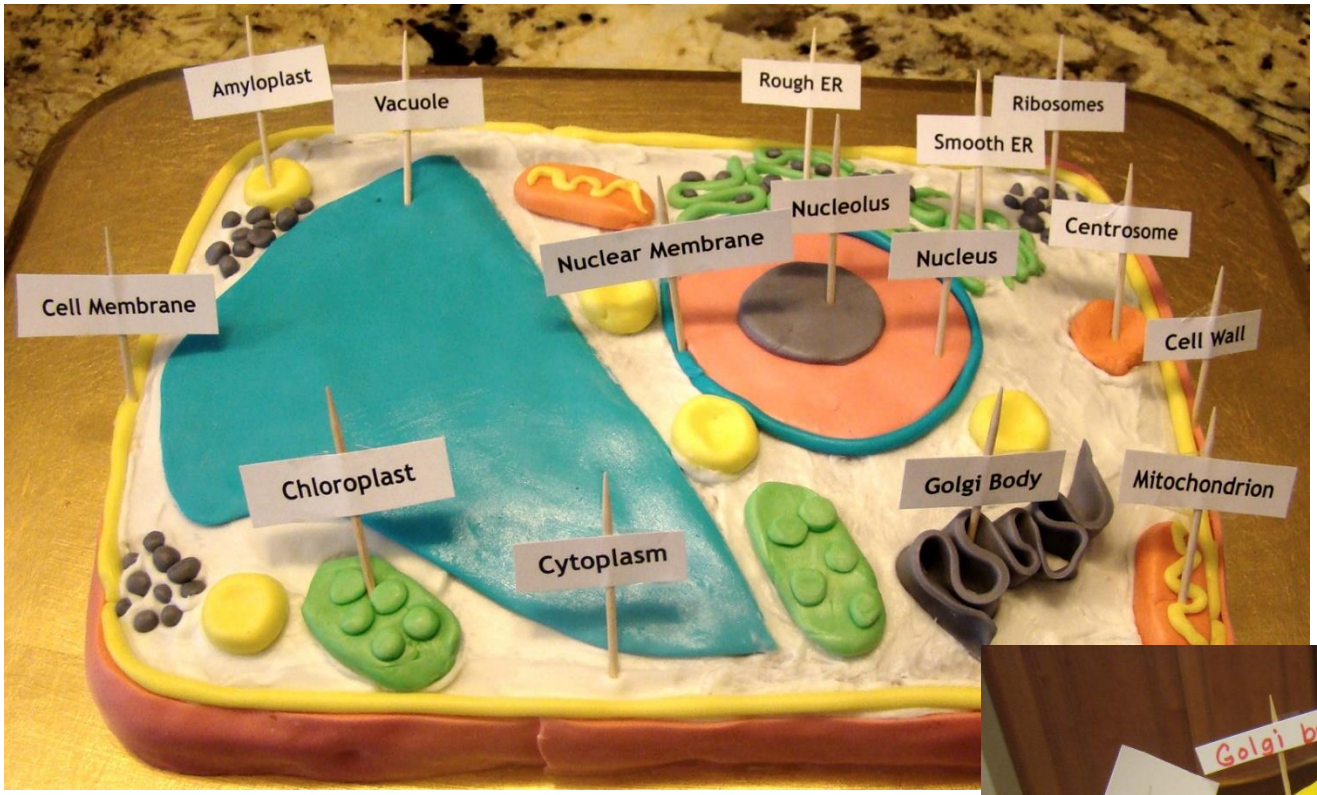
poziom badań	przedmiot badań	dziedzina badań	tematy badań
genom	wszystkie sekwencjie DNA zawarte w organizmie, geny, sekwencje regulatorowe	genomika	poszukiwanie sekwencji kodujących, rozpoznawanie eksonów i intronów, organizacja genomów, porównanie sekwencji
transkryptom	wszystkie sekwencie RNA zawarte w organizmie	transkryptomika	analiza ekspresji genów
proteom	wszystkie białka zawarte w organizmie	proteomika	porównanie sekwencji, identyfikacja zachowanych regionów, przewidywannie struktury, oddziaływania
metabolom	wszystkie procesy metaboliczne zachodzące w organizmie, metabolity	metabolomika	określanie sieci i szlaków metabolicznych, symulacje

Centralny dogmat biologii molekularnej



Wskazuje kierunki przepływu informacji biologicznej.

Przestrzeń



Komórka roślinna (powyżej)
I zwierzęca (po prawej)

Źródło: http://www.sciencecartoonsplus.com