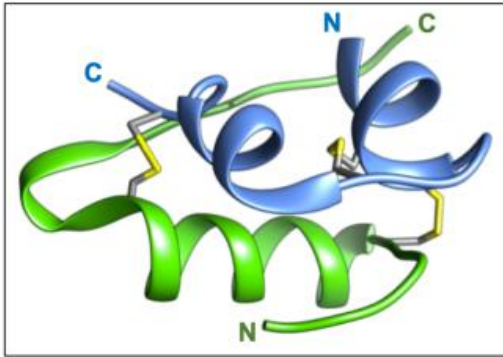
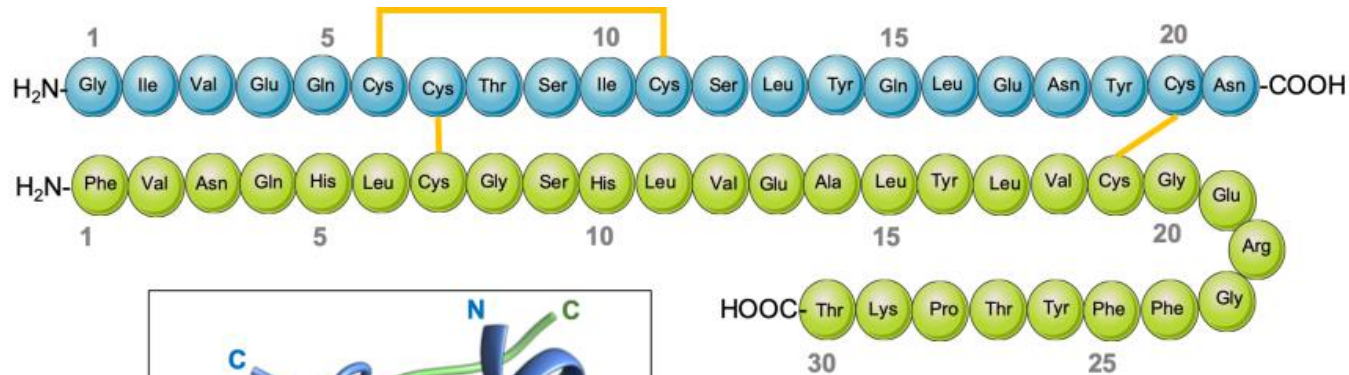


Bazy danych stosowane w analizie wyników badań molekularnych

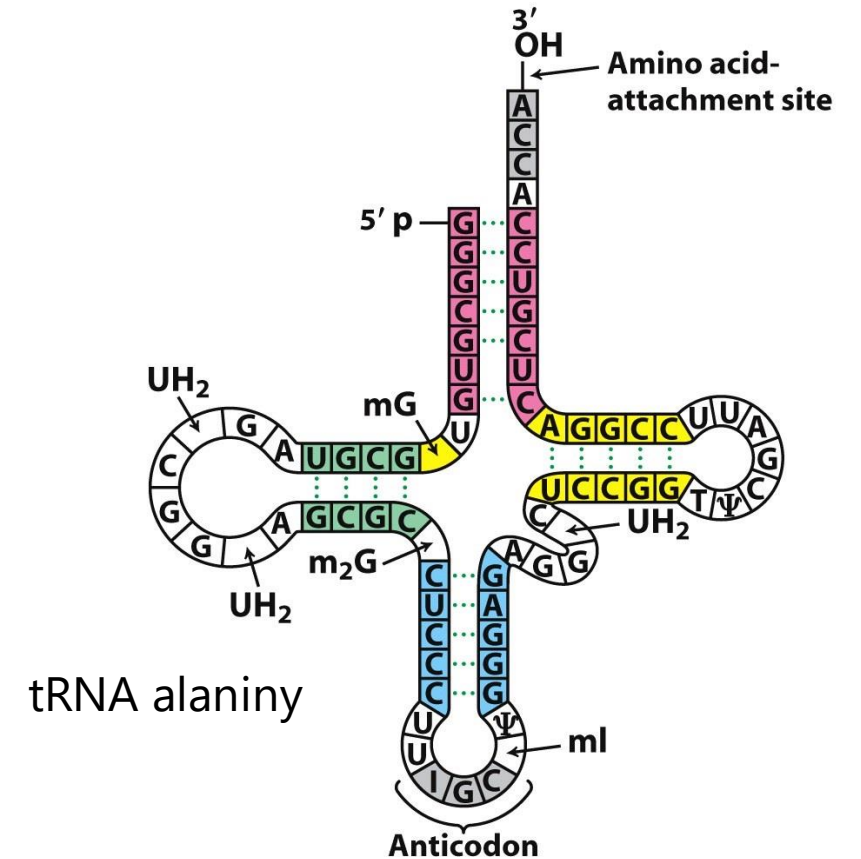
Mateusz Dawidziuk
Zakład Genetyki Medycznej
Instytut Matki i Dziecka w Warszawie

Sekwencje cząsteczek biologicznych

- Pierwsza sekwencja białkowa – insulina
- Pierwsza sekwencja nukleotydowa – tRNA alaniny

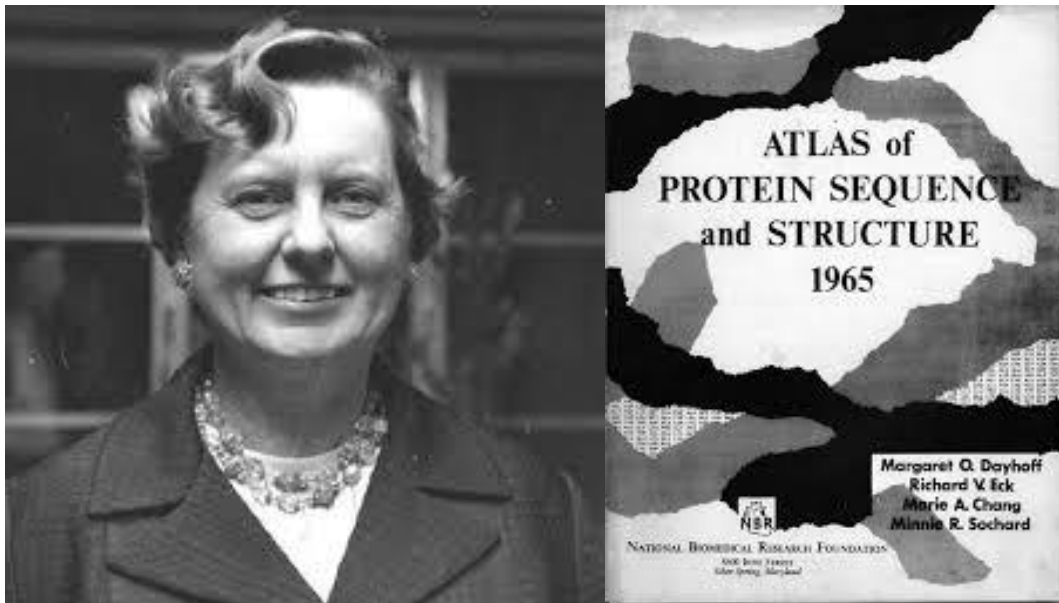


Insulina

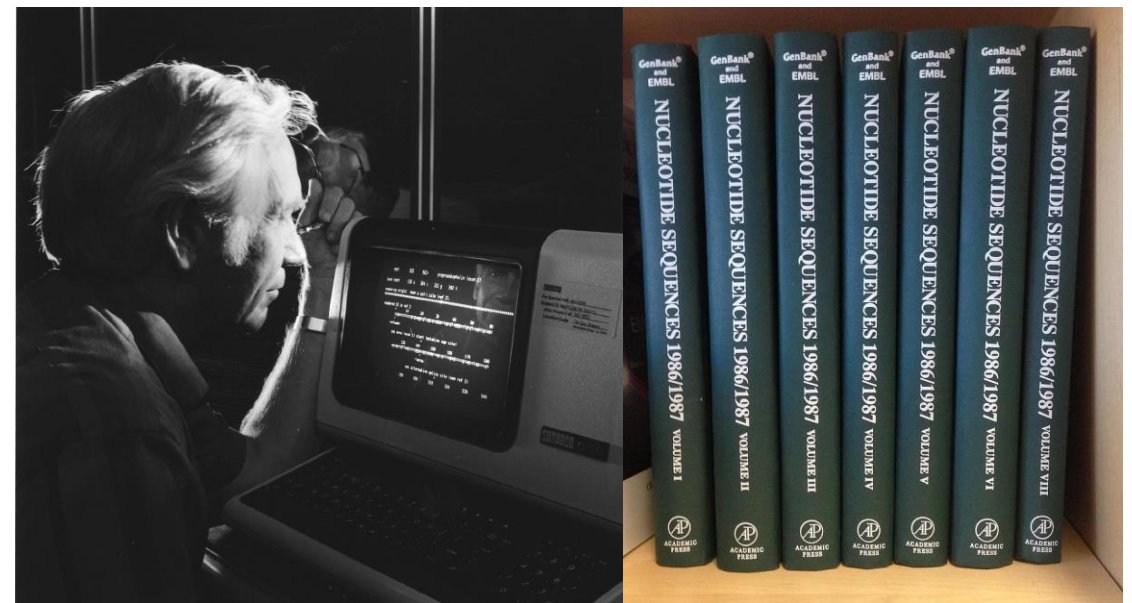


Pierwsze biologiczne bazy danych

- Atlas of Protein Sequence – 65 sekwencji białkowych, podstawa do stworzenia bazy Protein Information Resource w roku 1984
- GenBank – 606 sekwencji nukleotydowych, rok 1982



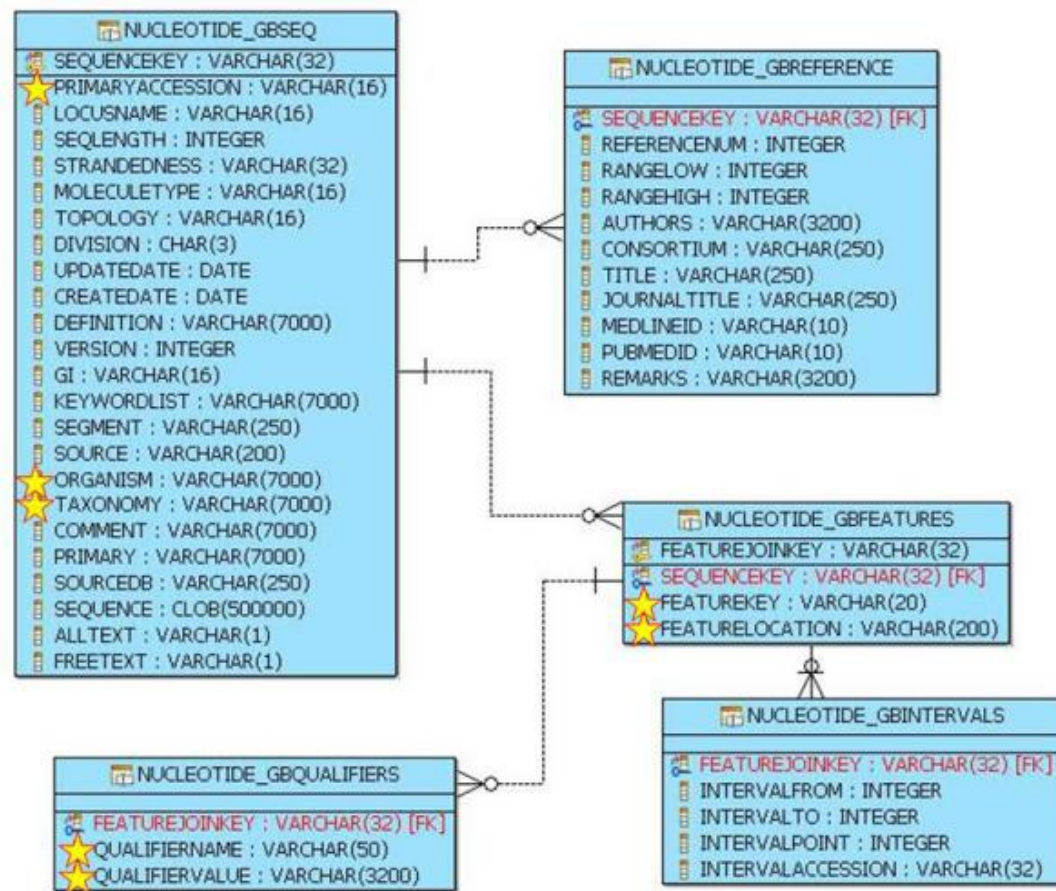
Margaret Dayhoff



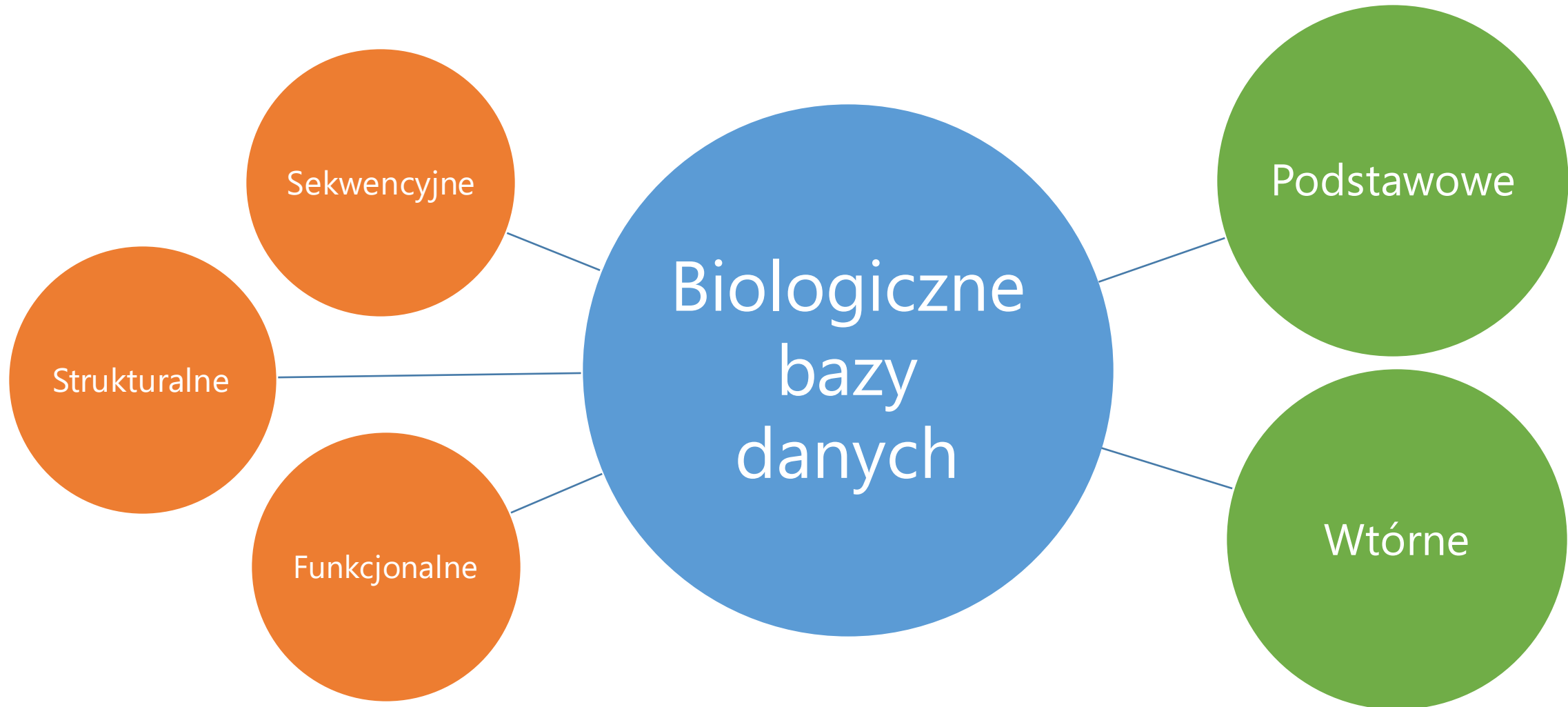
Walter Goad

Czym jest biologiczna baza danych

- Zbiór danych zaprojektowany w sposób umożliwiający pozyskiwanie informacji, ich wprowadzanie czy aktualizację
- Większość baz biologicznych to relacyjne bazy danych



Rodzaje biologicznych baz danych



Bazy podstawowe

- Dane uzyskane w sposób eksperymentalny
- Dane wprowadzane są do systemu przez badaczy
- Dane raz wprowadzone nie ulegają zmianie, mogą być aktualizowane ale informacje o wcześniejszej wersji nie są nadpisywane

- Przykłady:

- GenBank – sekwencje nukleotydowe



- UniProt – sekwencje aminokwasowe



- Protein Data Bank – struktury 3D białek

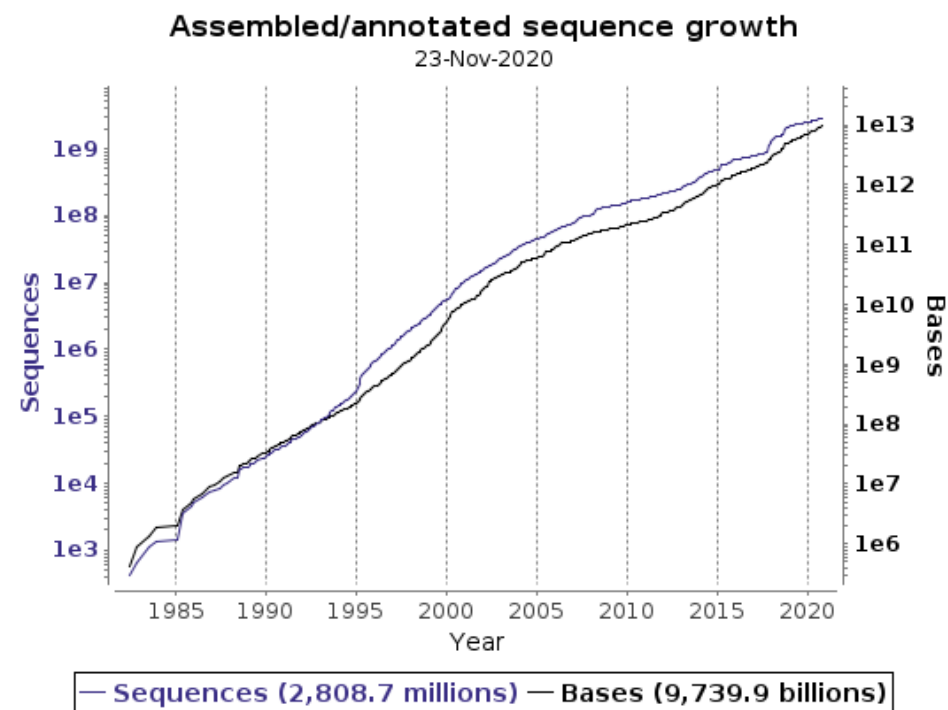
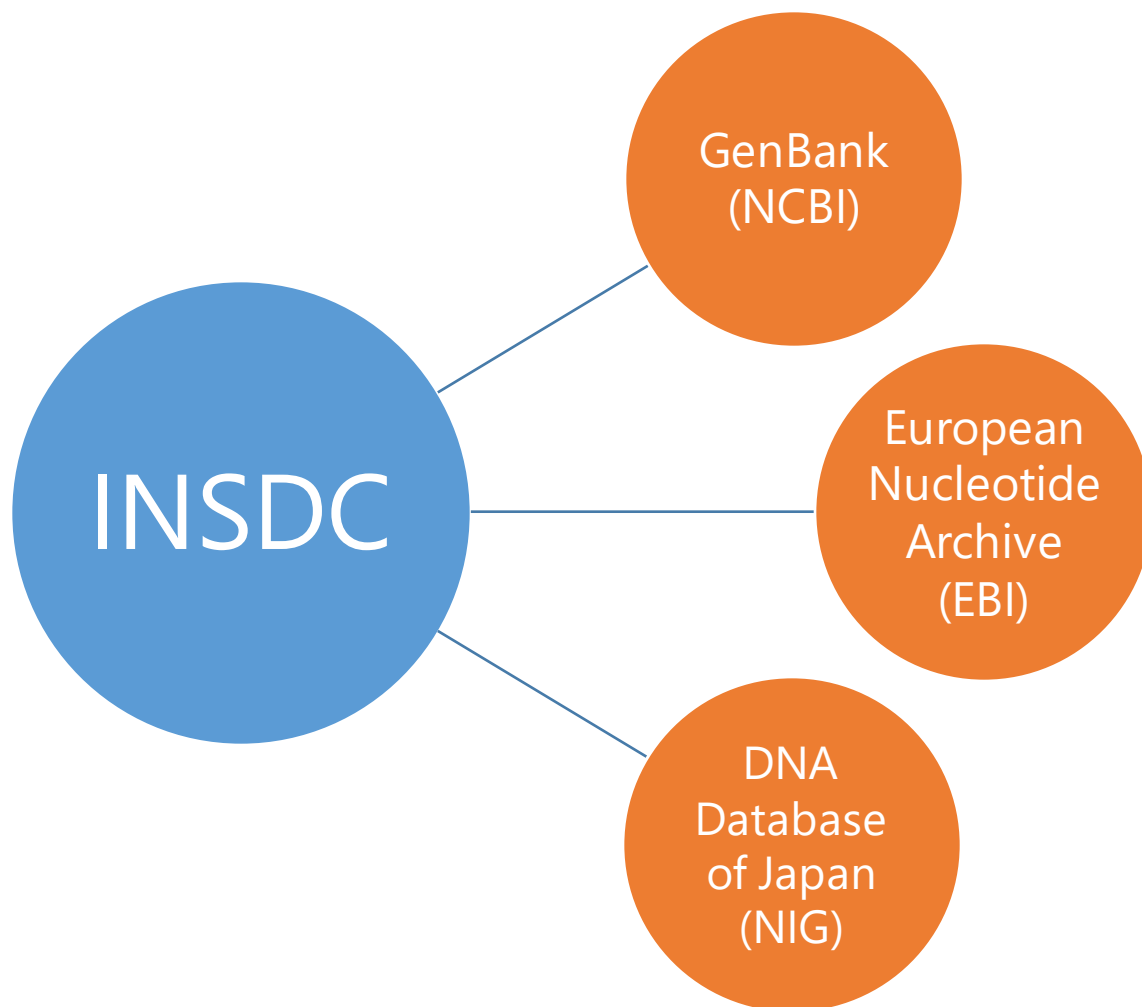


Bazy wtórne

- Informacje uzyskane na podstawie analizy danych pochodzących z baz podstawowych lub innych baz wtórnych
- Duży nacisk kładzie się na kurację danych
 - Algorytmy komputerowe
 - Manualna analiza i interpretacja uzyskanych danych
- Przykłady:
 - Ensembl - sekwencje, warianty, funkcje, regulacja
 - InterPro - sekwencje, motywy, domeny i rodziny białek

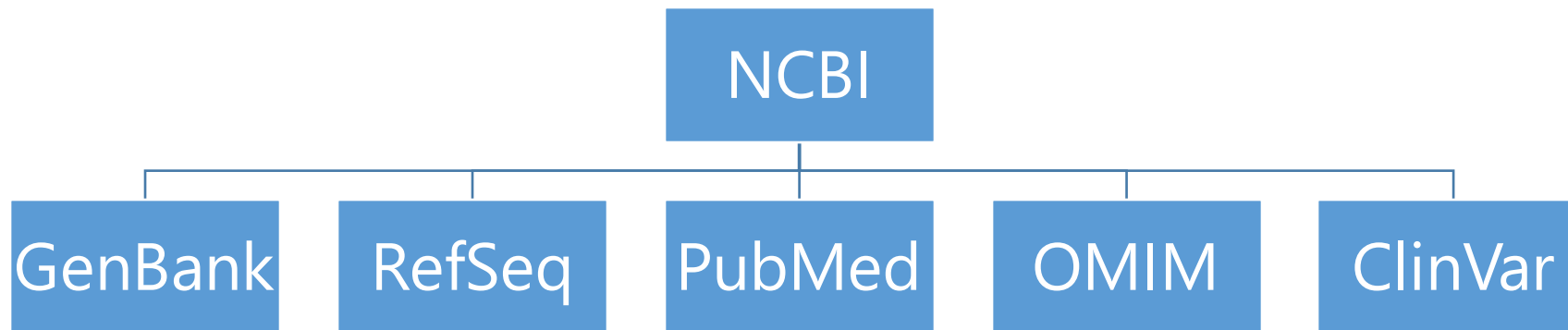


Bazy sekwencji nukleotydowych



National Center for Biotechnology Information

- Główne zadania:
 - Przechowywanie informacji
 - Tworzenie narzędzi do analizy danych
 - Rozpowszechnianie wiedzy



www.ncbi.nlm.nih.gov

Identyfikatory sekwencji w NCBI

- Numer GI (GenInfo Identifier)
 - Ciąg liczb kolejno nadawanych każdej sekwencji w bazie
 - Gdy sekwencja ulegnie zmianie, nadawany jest nowy numer
- Numer wersji sekwencji (Accession version)
 - Grupuje numery GI konkretnej sekwencji w uporządkowane serie
 - Składa się z przedrostka, numeru sekwencji oraz przyrostka oznaczającego numer wersji sekwencji

Accession version	GI
NC_000001.8	51511461
NC_000001.7	42406218
NC_000001.6	42405892
NC_000001.5	37623929
NC_000001.4	29824572
NC_000001.3	29824110
NC_000001.2	27777714
NC_000001.1	22539468

Identyfikatory sekwencji w NCBI

- N**G**_numer.wersja gen
 - N**M**_numer.wersja mRNA
 - N**P**_numer.wersja białko
 - N**R**_numer.wersja niekodujące RNA
 - N**C**_numer.wersja kontig, sekwencja genomowa
-
- N - sekwencje pozyskane eksperymentalnie, podlegają kuracji
 - X - sekwencje uzyskane najczęściej w sposób automatyczny, nie podlegają kuracji

Wyszukiwanie w bazach NCBI

(MCPH1[Preferred Symbol]) AND Human[Organism]

[Edit](#) [Clear](#)

Builder

	Preferred Symbol ▼	MCPH1	–	Show index list
AND ▼	Organism ▼	Human	–	Show index list
AND ▼	All Fields ▼		– +	Show index list

[Search](#) or [Add to history](#)

www.ncbi.nlm.nih.gov/books/NBK49540/

PubMed

- Wyszukiwarka artykułów naukowych
- Oparta na bazie MEDLINE zawierającej dane bibliograficzne
- Poszczególne rekordy indeksowane są za pomocą słownika MeSH

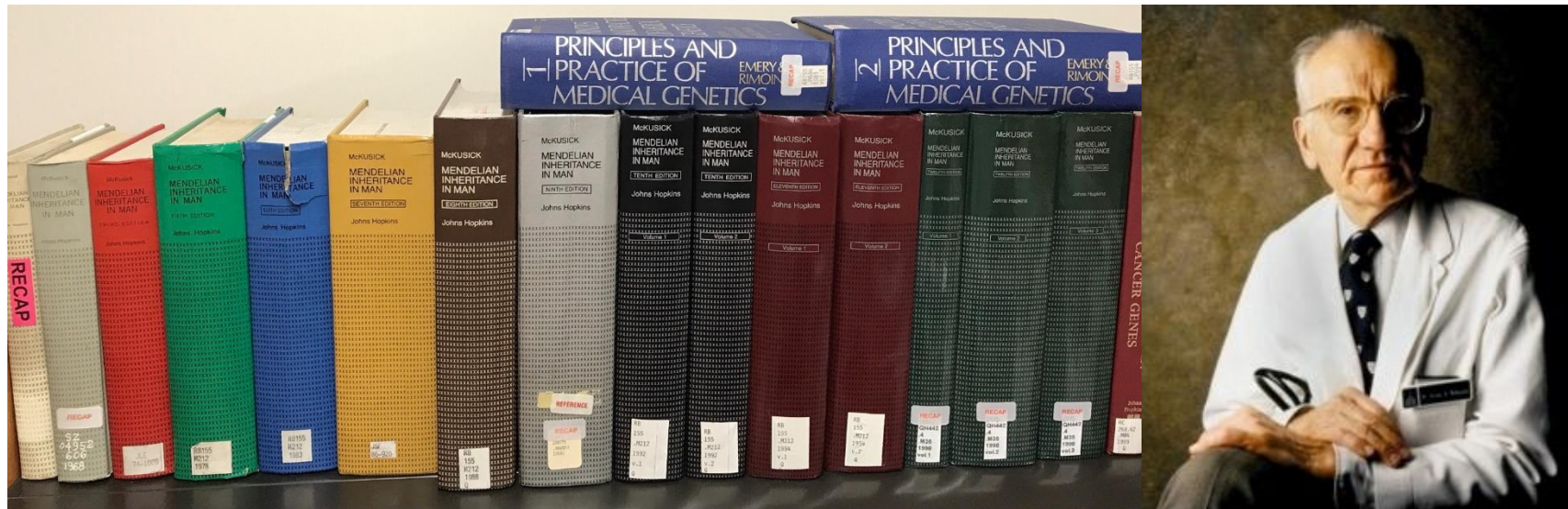
www.pubmed.ncbi.nlm.nih.gov

<https://meshb.nlm.nih.gov/search>

OMIM

Online Mendelian Inheritance in Man

- Kompendium chorób człowieka o podłożu genetycznym
- Zawiera opisy chorób genetycznych o dziedziczeniu mendlowskim
- Informację o ponad 16000 genów



www.omim.org

OMIM kategorie zespołów genetycznych

- 1 – Choroba została umiejscowiona na mapie genetycznej na podstawie powiązania z genem, ale konkretny defekt nie został poznany
- 2 – Choroba została opisana na podstawie analizy sprzężeń, nie jest znana konkretna mutacja powodująca występowanie choroby
- 3 – Podstawa molekularna choroby jest znana, warianty patogenne powodujące chorobę zostały znalezione w genie
- 4 – Choroba spowodowana jest delecją lub duplikacją większego regionu genomu, może obejmować wiele genów

dbSNP

- Baza wariantów genetycznych
- Podstawienia pojedynczych nukleotydów, małe warianty INDEL, powtórzenia mikrosatelitarne
- Informacje o pozycji genomowej, częstości w populacji, konsekwencjach zmiany na poziomie molekularnym oraz znaczeniu klinicznym

<https://www.ncbi.nlm.nih.gov/snp>

rs748011724

Current Build 154
Released April 21, 2020

Organism	<i>Homo sapiens</i>	Clinical Significance	Reported in ClinVar
Position	chr8:6445347 (GRCh38.p12) ?	Gene : Consequence	MCPH1 : Stop Gained
Alleles	T>C / T>G	Publications	0 citations
Variation Type	SNV Single Nucleotide Variation	Genomic View	See rs on genome
Frequency	G=0.000032 (8/249570, GnomAD_exome) G=0.000048 (6/125568, TOPMED) G=0.000017 (2/120628, ExAC) (+ 3 more)		

Variant Details

Clinical Significance

Frequency

HGVS

Submissions

History

Publications

Flanks

ALFA Allele Frequency (New)
The ALFA project provide aggregate allele frequency from dbGaP. More information is available on the project [page](#) including descriptions, data access, and terms of use.
Release Version: 20200227123210

Search:

Population	Group	Sample Size	Ref Allele	Alt Allele
Total	Global	11176	T=0.99991	C=0.00000, G=0.00009
European	Sub	8134	T=0.9999	C=0.0000, G=0.0001
African	Sub	676	T=1.000	C=0.000, G=0.000
African Others	Sub	14	T=1.00	C=0.00, G=0.00
African American	Sub	662	T=1.000	C=0.000, G=0.000

ClinVar

- Baza wariantów genetycznych
- Powiązania pomiędzy konkretnymi wariantami a zespołami genetycznymi
- Informacje o patogenności zmiany oraz wiarygodności źródła informacji

<https://www.ncbi.nlm.nih.gov/clinvar>

NM_024596.5(MCPH1):c.1625T>G (p.Leu542Ter)[Cite this record](#)

Interpretation: Pathogenic/Likely pathogenic

Review status: ★★☆☆ criteria provided, multiple submitters, no conflicts

Submissions: 2 (Most recent: Nov 14, 2018)

Last evaluated: Oct 31, 2018

Accession: VCV000586140.1

Variation ID: 586140

Description: single nucleotide variant

Variant details

Conditions

Gene(s)

Aggregate interpretations per condition

Interpreted condition	Interpretation	Number of submissions	Review status	Last evaluated	Variation/condition record
not provided	Likely pathogenic	1	criteria provided, single submitter	Jun 13, 2018	RCV000712273.1
Primary autosomal recessive microcephaly 1	Pathogenic	1	criteria provided, single submitter	Oct 31, 2018	RCV000763602.1

dbNSFP

database of Non-Synonymous SNPs and their Functional Predictions

- Baza danych służąca funkcjonalnej analizie predykcyjnej wszystkich niesynonimicznych wariantów kodujących pojedynczego nukleotydu
- Kompilacja 37 algorytmów predykcyjnych takich jak PolyPhen, SIFT, MutationTaster, CADD czy SpliceAI
- Pierwotnie była dostępna wraz z narzędziami do adnotacji wariantów, takich jak VEP czy Annovar, obecnie możliwe jest także przeszukiwanie jej zasobów on-line

<http://database.liulab.science/dbNSFP>

European Bioinformatics Institute

- Zadania zbliżone do NCBI
- Główne bazy danych:
 - EMBL-Bank - baza danych genetycznych
 - Ensembl - przeglądarka genomowa
 - UniProt - baza sekwencji białkowych

<https://www.ebi.ac.uk/services>

Identyfikatory sekwencji w Ensembl

Dla człowieka:

- ENS**G**numer.wersja Gen
- ENS**T**numer.wersja Transkrypt
- ENS**P**numer.wersja Białko
- ENS**E**numer Ekson



Dla innych gatunków w nazwie umieszczony jest skrót:

- ENS**DARG**numer.wersja Gen danio pręgowany
- ENS**MUST**numer.wersja Transkrypt mysz domowa
- ENS**RNOE**numer Ekson szczur

Ensembl BioMart

- Narzędzie służące do pozyskiwania informacji z różnych baz danych
- Bardzo rozbudowane możliwości przeszukiwania
- Możliwość pozyskiwania i przekształcania danych, bez potrzeby znajomości narzędzi bioinformatycznych
- Dostęp przez stronę internetową



Ensembl VEP

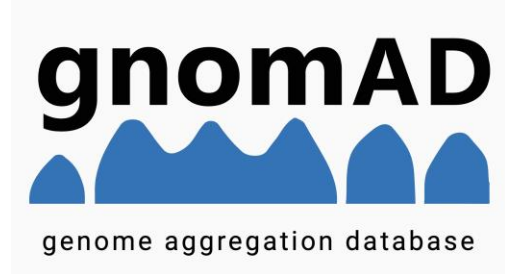
- Narzędzie służące do adnotacji wariantów z plików VCF (pliki z wariantami znalezionymi w sekwencjonowaniu NGS)
- Możliwość opisanie wariantów korzystając z informacji z różnych baz danych
- Rozbudowane możliwości, duży wybór baz danych
- Dostępny w formie narzędzia bioinformatycznego oraz przez stronę internetową



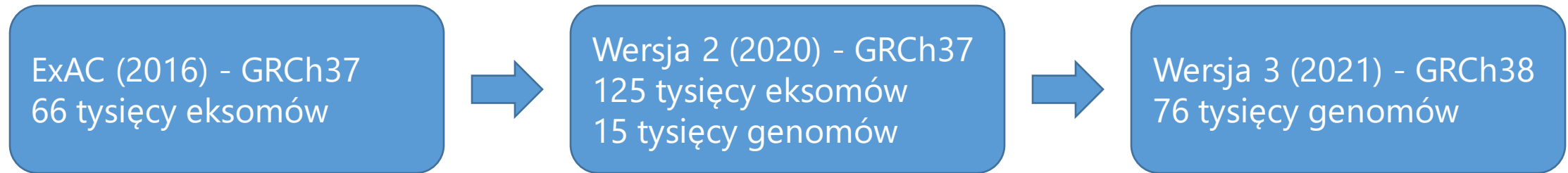
<https://www.ensembl.org/info/docs/tools/vep/index.html>

gnomAD

Genome Aggregation Database



- Baza agregacyjna
- Dane z różnych wielkoskalowych projektów sekwencjonowania eksomowego i genomowego człowieka
- Informacje o częstości występowania danych wariantów w populacji, jakości sekwencjonowania czy też przeglądarkę genomową.

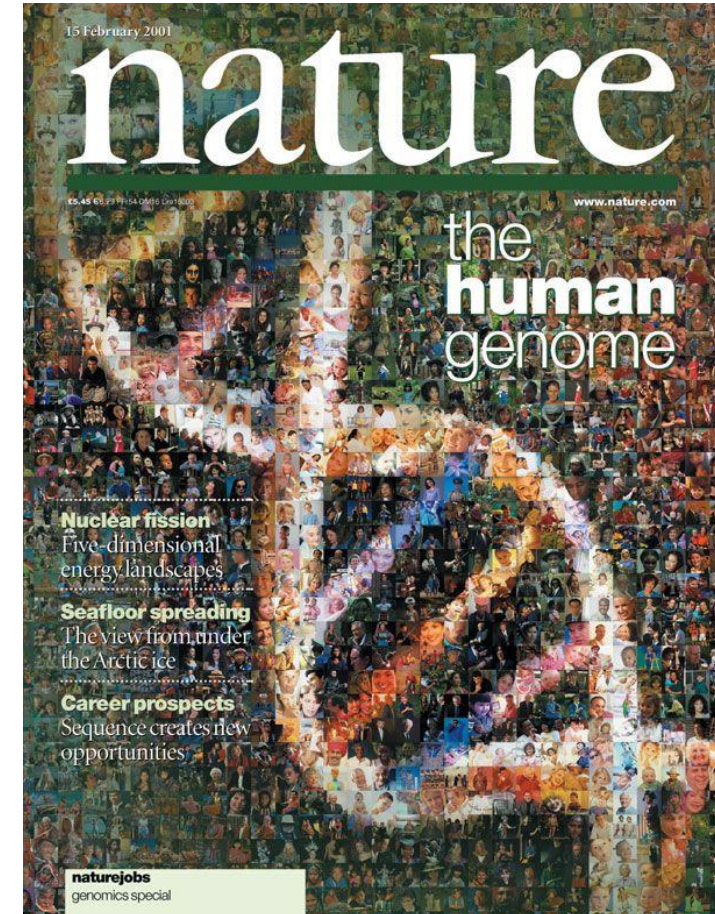


<https://gnomad.broadinstitute.org>

UCSC

- Baza sekwencji genomowych
- Przydatne narzędzia do analizy danych
 - Przeglądarka genomowa
 - Blat
 - Table browser
 - PCR *in-silico*
 - LiftOver

<https://genome-euro.ucsc.edu>



HGMD

Human Gene Mutation Database

- Informację o mutacjach związanych z chorobami genetycznymi
- Dane te z artykułów naukowych
- Główna baza informacji o konsekwencjach klinicznych mutacji
- Płatny dostęp
- Wersja darmowa - dane publikowane z 3 letnim opóźnieniem, ograniczona funkcjonalność

<https://portal.biobase-international.com>

<http://www.hgmd.cf.ac.uk>



Mastermind

- Baza łącząca warianty z publikacjami naukowymi
- Częściowo komercyjna



MASTERMIND
By: GENOMENON

<https://mastermind.genomenon.com>

COSMIC

Catalogue Of Somatic Mutations In Cancer

- Największa baza zbierająca informacje o mutacjach somatycznych
- Dwa główne źródła danych:
 - Manualnie weryfikowane dane pochodzące z publikacji, celowanych badań panelowych NGS, informacji o historii choroby u pacjenta
 - Analizy wielkoskalowych baz danych z różnego rodzaju projektów badawczych dotyczących nowotworów
- Informacje o wariantach, powodowanych nowotworach, częstości, pochodzeniu tkankowym, fuzjach genowych

<https://cancer.sanger.ac.uk/cosmic>



VarSome

- Baza agregacyjna
- Informacje o wariantach
- Klasyfikacja ACMG
- Informacje z algorytmów predykcyjnych
- Powiązania wariantów/genów z publikacjami
- Częściowo komercyjna

<https://varsome.com/>



The human genetics search engine
Supported by the global community of geneticists



UniProt

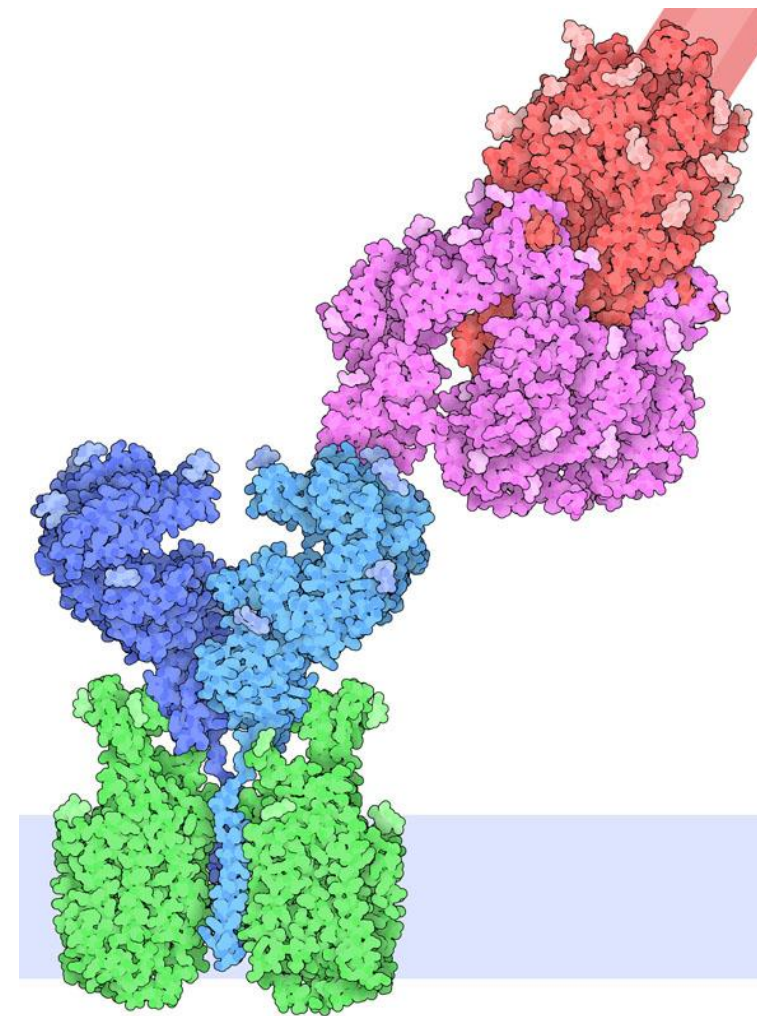
- Dwie części:
 - Swiss-Prot - dane kurowane
 - TrEMBL - dane uzyskane automatycznie
- Informacje o sekwencjach aminokwasowych, motywach, domenach czy rodzinach białek
- Narzędzia do analizy - BLAST, Clustal Omega

<https://www.uniprot.org>

Protein Databank

- Zawiera struktury 3D białek
- Uzyskane metodą krystalografii, spektroskopii NMR czy mikroskopii krioelektronowej
- Pliki z koordynatami dla konkretnych cząsteczek
 - Lista atomów w cząsteczce
 - Lokalizacja 3D w przestrzeni
- Narzędzia do wizualizacji białek oraz narzędzia do ich analizy

<https://www.rcsb.org>



Dziękuję za uwagę

dr n. med. Mateusz Dawidzук
Zakład Genetyki Medycznej
kontakt: mateusz.dawidziuk@imid.med.pl