Bioinformatika

C6215 Pokročilá biochemie a její metody

Podzim 2020

Michaela Wimmerová

Osnova

- **Úvod do bioinformatiky**
  - Definice, molekulárně biologická data, databáze
  - Rozdělení databází, bioinformatická centra
- **Manipulace se sekvencemi**
  - Sekvence biomakromolekul, aminokyseliny, báze, alignment
  - Význam alignmentu, přiložení páru sekvencí a vícenásobné přiložení
- **Predikce struktury proteinů**
  - Predikce 2-D struktury proteinů, predikce 3-D struktury proteinů
  - Threading, Homology modeling, *Ab initio*
- **Predikce genů**
  - Predikce genů u prokaryot a eukaryot, predikční nástroje a postupy

Osnova

- **Úvod do bioinformatiky**
  - Definice, molekulárně biologická data, databáze
  - Rozdělení databází, bioinformatická centra
- **Manipulace se sekvencemi**
  - Sekvence biomakromolekul, aminokyseliny, báze, alignment
  - Význam alignmentu, přiložení páru sekvencí a vícenásobné přiložení
- **Predikce struktury proteinů**
  - Predikce 2-D struktury proteinů, predikce 3-D struktury proteinů
  - Threading, Homology modeling, *Ab initio*
- **Predikce genů**
  - Predikce genů u prokaryot a eukaryot, predikční nástroje a postupy

Bioinformatika – definice

- Existuje mnoho různých definic – nejednotnost odráží dynamický rozvoj oboru.
- **Bioinformatika** – vědní disciplína, která využívá výpočetní techniku (počítače) pro shromažďování, vyhledávání, manipulaci a distribuci informací o biologických makromolekulách (DNA, RNA, proteiny). *Luscombe et al.*
- **Bioinformatika** – nová disciplína na rozhraní počítačových věd, informačních technologií, matematiky a biologie a zahrnuje studium a praktické uchovávání, vyhledávání, zobrazování, manipulaci a modelování biologických dat. *R. Pantůček*
- **Bioinformatika** (zaměření na sekvence) vs. výpočetní biologie (všechny oblasti biologie zahrnující výpočty).
- **Bioinformatika**: vývoj výpočetních nástrojů a databází + jejich aplikace

Bioinformatika – aplikace

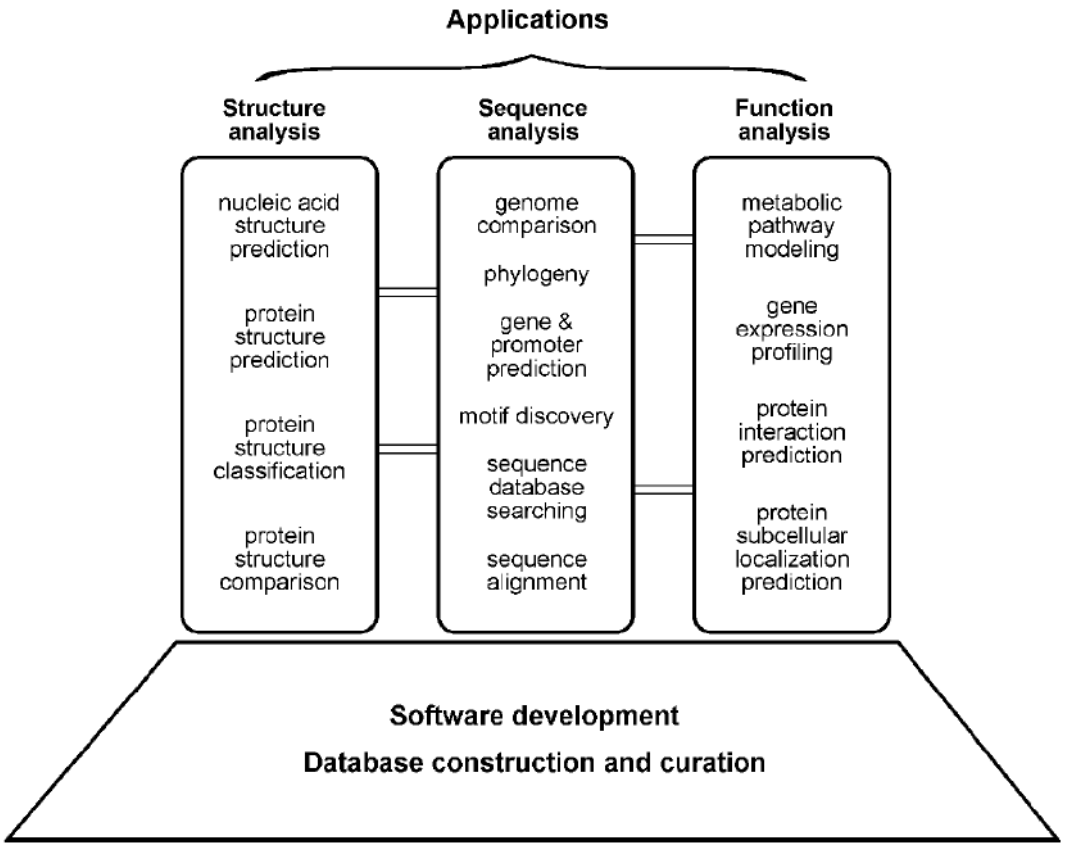


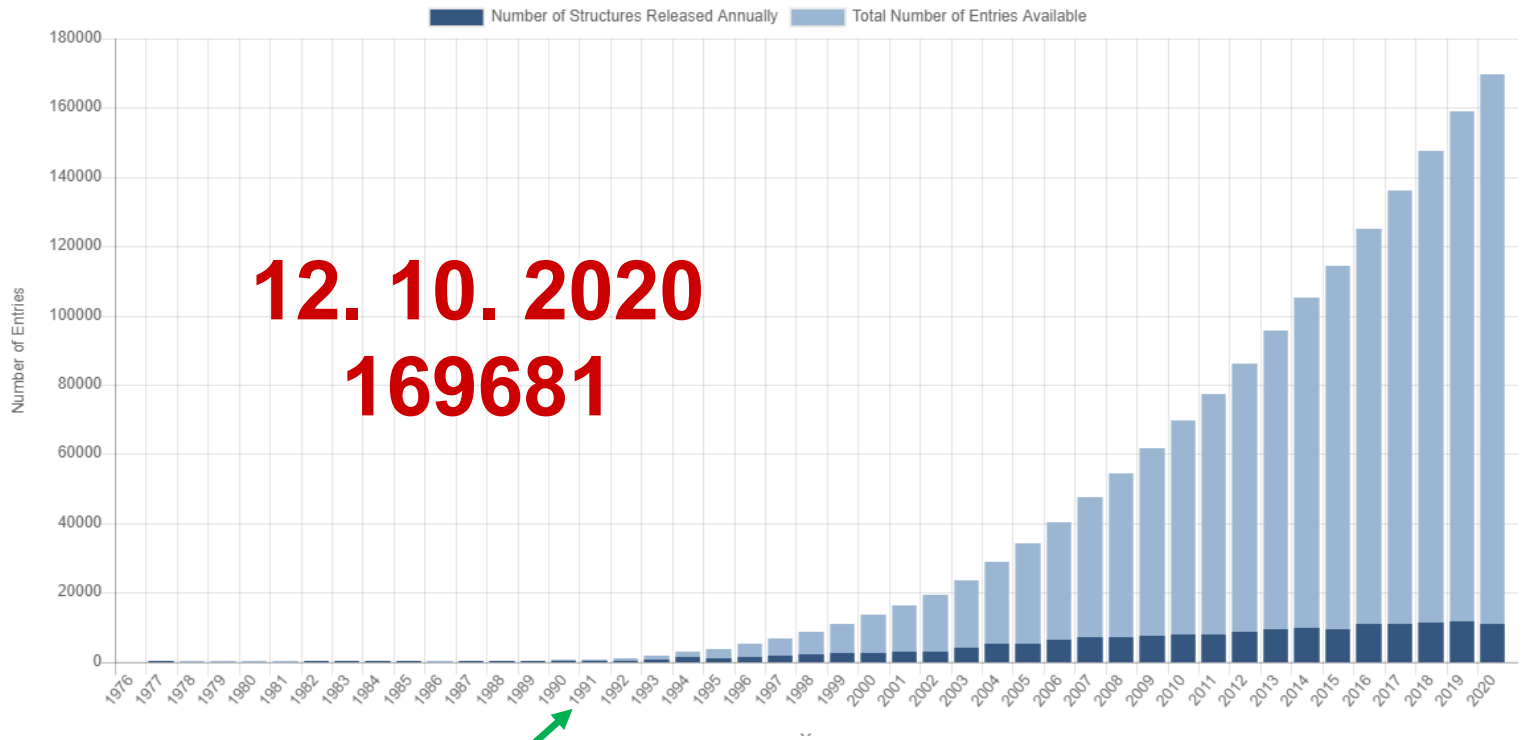
Figure 1.1: Overview of various subfields of bioinformatics. Biocomputing tool development is at the foundation of all bioinformatics analysis. The applications of the tools fall into three areas: sequence analysis, structure analysis, and function analysis. There are intrinsic connections between different areas of analyses represented by bars between the boxes.

ESSENTIAL BIOINFORMATICS,
Jin Xiong, 2006

Molekulárně biologická data, databáze

- **Molekulárně biologická data**: sekvence a struktury proteinů a nukleových kyselin, genomy, struktury (introny, exony) a funkce genů, metabolické a signální dráhy, organely...
- Rozvoj výkonných technologií (**automatické sekvencování, MALDI-TOF, NMR spektroskopie, proteinová krystalografie)** koncem minulého století vedl k **obrovskému** nárůstu množství biologických dat.
- **Nutnost organizovaného ukládání, skladování a manipulace s velkým množstvím dat vedla ke vzniku bioinformatiky.**

Molekulárně biologická data, databáze



První výskyt termínu bioinformatika

https://www.rcsb.org/stats/growth/growth-released-structures

Rozdělení databází

- **Primární databáze:** anotované sekvence nukleových kyselin nebo proteinů
- **Sekundární databáze:** informace odvozené z primárních databází ve formě charakteristických vzorů sekvencí, tj. funkčních nebo strukturních motivů získaných srovnáním primárních dat (sekvencí).
- **Strukturní databáze:** struktury proteinů (nukleových kyselin) a jejich anotace.
- **Genomové databáze:** genomy organismů.
- Databáze **specializované** vs. **univerzální**

Rozdělení databází

Primární

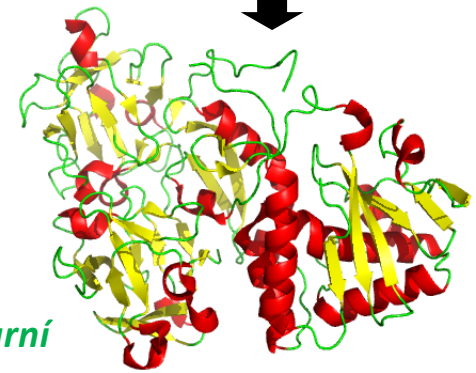
```
EDRPIKFSTEGATSQSYKQFIEALRERLRGGLIHDIPVLPDPTTLQERNRYIT
VELSNSDTESIEVGIDVTNAYVVAYRAGTQSYFLRDAPSSASDYLFTGTDQHS
LPFYGTYGDLERWAHQSRQQIPLGLQALTHGISFFRSGGNDNEEKARTLIVII
QMVAEAARFRYISNRVRVSIQTGTAFQPDAAMISLENNWDNLSRGVQESVQDT
FPNQVTLTNIRNEPVIVDSLSHPTVAVLALMLFVCNPPNIVEKSKICSSRYEP
TVRIGGRDGMCVDVYDNGYHNGNRIIMWKCKDRLEENQLWTLKSDKTIRSNGK
```

Ribosome-inactivating protein, subdomain 1

Ribosome-inactivating protein, subdomain 2

Ricin B-like lectins

Sekundární



Strukturní

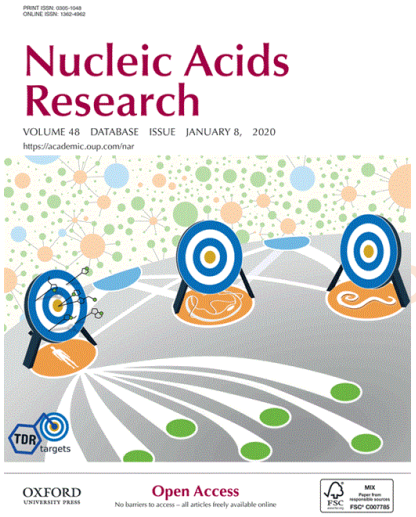
Specializované

Univerzální

Rozdělení databází

Nucleic Acids Research

http://www.oxfordjournals.org/nar/database/a/



2020: 1637 databází

- Nucleotide Sequence Databases
- International Nucleotide Sequence Database Collaboration
- Coding and non-coding DNA
- Gene structure, introns and exons, splice sites
- Transcriptional regulator sites and transcription factors
- RNA sequence databases
- Protein sequence databases
- Structure Databases
- Genomics Databases (non-vertebrate)
- Metabolic and Signaling Pathways
- Human and other Vertebrate Genomes
- Human Genes and Diseases
- Microarray Data and other Gene Expression Databases
- Proteomics Resources
- Other Molecular Biology Databases
- Organelle databases
- Plant databases
- Immunological databases

The 27th annual Nucleic Acids Research database issue and molecular biology database collection

Daniel J. Rigden[1,*] and Xosé M. Fernández[2]

[1]Institute of Integrative Biology, University of Liverpool, Crown Street, Liverpool L69 7ZB, UK and [2]Institut Curie, 25 rue d'Ulm, 75005 Paris, France

ABSTRACT

The 2020 Nucleic Acids Research Database Issue contains 148 papers spanning molecular biology. They include 59 papers reporting on new databases and 79 covering recent changes to resources previously published in the issue. A further ten papers are updates on databases most recently published elsewhere. This issue contains three breakthrough articles: AntiBodies Chemically Defined (ABCD) curates antibody sequences and their cognate antigens; SCOP returns with a new schema and breaks away from a purely hierarchical structure; while the new Alliance of Genome Resources brings together a number of Model Organism databases to pool knowledge and tools. Major returning nucleic acid databases include miRDB and miRTarBase. Databases for protein sequence analysis include CDD, DisProt and ELM, alongside no fewer than four newcomers covering proteins involved in liquid–liquid phase separation. In metabolism and signaling, Pathway Commons, Reactome and Metabolights all contribute papers. PATRIC and MicroScope update in microbial genomes while human and model organism genomics resources include Ensembl, Ensembl genomes and UCSC Genome Browser. Immune-related proteins are covered by updates from IPD-IMGT/HLA and AFND, as well as newcomers VDJbase and OGRDB. Drug design is catered for by updates from the IUPHAR/BPS Guide to Pharmacology and the Therapeutic Target Database. The entire Database Issue is freely available online on the Nucleic Acids Research website (https://academic.oup.com/nar). The NAR online Molecular Biology Database Collection has been revised, updating 305 entries, adding 65 new resources and eliminating 125 discontinued URLs; so bringing the current total to 1637 databases. It is available at http://www.oxfordjournals.org/nar/database/c/.

NEW AND UPDATED DATABASES

The year 2020 sees the Nucleic Acids Research Database Issue reach its 27th annual issue. As usual, the 148 papers included span the full range of biological research. This year there are papers on 59 new databases (Table 1) while 79 resources provide Update papers covering recent developments. A further 10 papers cover updates of databases most recently published elsewhere (Table 2). The issue begins with reports from the major database providers at the U.S. National Center for Biotechnology Information (NCBI), the European Bioinformatics Institute (EBI) and the National Genomics Data Center (NGDC) in China, a new venture encompassing the previously published Beijing Institute of Genomics Data Center. Further papers are grouped in the now-familiar fashion: (i) nucleic acid sequence and structure, transcriptional regulation; (ii) protein sequence and structure; (iii) metabolic and signaling pathways, enzymes and networks; (iv) genomics of viruses, bacteria, protozoa and fungi; (v) genomics of human and model organisms plus comparative genomics; (vi) human genomic variation, diseases and drugs; (vii) plants and (viii) other topics, such as proteomics databases. As ever, the discipline-spanning nature of many modern resources means that readers are encouraged to browse the whole issue. The Nucleic Acids Research online Molecular Biology Database Collection, classifies databases more finely using 15 categories and 41 subcategories, and can be found at http://www.oxfordjournals.org/nar/database/c/.

Among the major global centers, the NCBI (1) reports updates across many databases and interfaces. For example, gene searches can now cleverly retrieve orthologs from (subsets of) vertebrates. The EBI paper (2) includes striking figures that illustrate the deep inter-connectedness of its hosted databases, as well as their myriad links to external resources. It also describes a significant new arrival, the BioImage Archive. The paper from the National Genomics Data Center (3) includes descriptions of their rapidly expanding suite of databases, some featured in detail elsewhere in this Issue. They report that their database for raw sequence reads, the Genome Sequence Archive, now occupies more than a petabyte.

https://academic.oup.com/nar/issue/48/D1

EBI/NCBI/DDBJ

Instituce zabývající se shromažďováním, správou a poskytováním dat a informací a vývojem analytických nástrojů.

EBI

Evropský institut pro bioinformatiku

European Bioinformatics Institute

NCBI

Národní centrum pro biotechnologické informace

NCBI
National Center for Biotechnology Information

National Center for Biotechnology Information

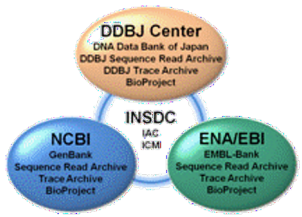
DDBJ Center

DDBJ
DNA Data Bank of Japan

The DNA Data Bank of Japan Center

http://www.ebi.ac.uk/ — http://www.ncbi.nlm.nih.gov/ — http://www.ddbj.nig.ac.jp/

ENA ⟷ GenBank ⟷ DDBJ



EBI

The European Bioinformatics Institute in 2020: building a global infrastructure of interconnected data resources for the life sciences

Charles E. Cook*, Oana Stroe, Guy Cochrane, Ewan Birney and Rolf Apweiler

European Molecular Biology Laboratory, European Bioinformatics Institute (EMBL-EBI), Wellcome Genome Campus, Hinxton, Cambridge CB10 1SD, UK

Received September 21, 2019; Revised October 18, 2019; Editorial Decision October 21, 2019; Accepted November 06, 2019

ABSTRACT

Data resources at the European Bioinformatics Institute (EMBL-EBI, https://www.ebi.ac.uk/) archive, organize and provide added-value analysis of research data produced around the world. This year's update for EMBL-EBI focuses on data exchanges among resources, both within the institute and with a wider global infrastructure. Within EMBL-EBI, data resources exchange data through a rich network of data flows mediated by automated systems. This network ensures that users are served with as much information as possible from any search and any starting point within EMBL-EBI's websites. EMBL-EBI data resources also exchange data with hundreds of other data resources worldwide and collectively are a key component of a global infrastructure of interconnected life sciences data resources. We also describe the BioImage Archive, a deposition database for raw images derived from primary research that will supply data for future knowledgebases that will add value through curation of primary image data. We also report a new release of the PRIDE database with an improved technical infrastructure, a new API, a new webpage, and improved data exchange with UniProt and Expression Atlas. Training is a core mission of EMBL-EBI and in 2018 our training team served more users, both in-person and through web-based programmes, than ever before.

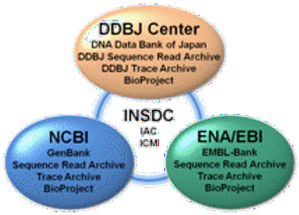
INTRODUCTION: ARCHIVAL RESOURCES AND KNOWLEDGEBASES

EMBL-EBI data resources cover the entire range of molecular biology and include nucleotide sequence data, protein sequences and families, chemical biology, structural biology, systems, pathways, ontologies and the scientific literature. EMBL-EBI's data resources c
and make freely available to the pub
data.

Our resources (www.ebi.ac.uk/ser
or deposition databases that store
data submitted by researchers, as w
that integrate and add value to ex
many having both functions (1,2). A
sources, are open access and freely
worldwide at any time, and EMBL-
the concept of FAIR data (findable,
ble, and resuable) (3). In the case of t
Phenome Archive (www.ebi.ac.uk/eg
man data consented for research, re
access from a data access committee

Deposition databases are repositor
mental data on behalf of the entire sc
serve as a key component of the sci
data types. These open and searchab
researchers with direct access to the
access to and re-use of experiment
nal results and, by combining multip
analytical insights. Deposition datab
ence data for the research communi
of search tools also allow researche
their own unpublished data with op

Storing experimental data in arc
the first step in extracting knowle
search. Added-value databases, or
on archival resources by providing
notation, reanalysis, and integratic
mental data. Knowledgebases may
functionality such as searching, an
sualization, and linking to related
archival resources and knowledgeba
and integrated analysis of archival
provide an opportunity to reuse
discoveries. The BioImage Archive
bioimage-archive/), introduced belo



NCBI

Database resources of the National Center for Biotechnology Information

Eric W. Sayers*, Jeff Beck, J. Rodney Brister, Evan E. Bolton, Kathi Canese, Donald C. Comeau, Kathryn Funk, Anne Ketter, Sunghwan Kim, Avi Kimchi, Paul A. Kitts, Anatoliy Kuznetsov, Stacy Lathrop, Zhiyong Lu, Kelly McGarvey, Thomas L. Madden, Terence D. Murphy, Nuala O'Leary, Lon Phan, Valerie A. Schneider, Françoise Thibaud-Nissen, Bart W. Trawick, Kim D. Pruitt and James Ostell

National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Building 38A, 8600 Rockville Pike, Bethesda, MD 20894, USA

Received September 16, 2019; Editorial Decision October 01, 2019; Accepted October 09, 2019

ABSTRACT

The National Center for Biotechnology Information (NCBI) provides a large suite of online resources for biological information and data, including the GenBank® nucleic acid sequence database and the PubMed database of citations and abstracts published in life science journals. The Entrez system provides search and retrieval operations for most of these data from 35 distinct databases. The E-utilities serve as the programming interface for the Entrez system. Custom implementations of the BLAST program provide sequence-based searching of many specialized datasets. New resources released in the past year include a new PubMed interface, a sequence database search and a gene orthologs page. Additional resources that were updated in the past year include PMC, Bookshelf, My Bibliography, Assembly, RefSeq, viral genomes, the prokaryotic genome annotation pipeline, Genome Workbench, dbSNP, BLAST, Primer-BLAST, IgBLAST and PubChem. All of these resources can be accessed through the NCBI home page at www.ncbi.nlm.nih.gov.

INTRODUCTION

NCBI overview

The National Center for Biotechnology Information (NCBI), a center within the National Library of Medicine at the National Institutes of Health, was created in 1988 to develop information systems for molecular biology. Since that time the amount and variety of data that NCBI maintains has expanded enormously and can be generally grouped into six categories: Literature, Health, Genomes, Genes, Proteins and Chemicals (Table 1). NCBI provides facilities for submitting and downloading data, analysis and visualization software, educational events and materials about NCBI products, and software and services to support an expanding developer community. These services, along with all other data resources, are available through the NCBI home page at www.ncbi.nlm.nih.gov/. In most cases, the data underlying these resources and executables for the software described are available for download at ftp.ncbi.nlm.nih.gov.

This article provides a brief overview of the NCBI Entrez system of databases, followed by a summary of resources that were either introduced or significantly updated in the past year. More complete discussions of NCBI resources can be found on the home pages of individual databases, on the NCBI Learn page (www.ncbi.nlm.nih.gov/learn/) or in the NCBI Handbook (www.ncbi.nlm.nih.gov/books/NBK143764/).

The Entrez system

Entrez (1) is an integrated database retrieval system that provides access to a diverse set of 35 databases that together contain 2.7 billion records (Table 1 and Figure 1). Links to the web portal for each of these databases are provided on the Entrez global search page (www.ncbi.nlm.nih.gov/search/). Entrez supports text searching using simple Boolean queries, downloading of data in various formats and linking records between databases based on asserted relationships. The LinkOut service expands the range of links to include external resources, such as organism-specific genome databases. The records retrieved in Entrez can be displayed in many formats and downloaded singly

DDBJ

DDBJ Database updates and computational infrastructure enhancement

Osamu Ogasawara*, Yuichi Kodama*, Jun Mashima, Takehide Kosuge and Takatomo Fujisawa

The Bioinformation and DDBJ Center, National Institute of Genetics, Mishima, Shizuoka, 411-8540, Japan

Received September 26, 2019; Revised October 10, 2019; Editorial Decision October 11, 2019; Accepted October 21, 2019

ABSTRACT

The Bioinformation and DDBJ Center (https://www.ddbj.nig.ac.jp) in the National Institute of Genetics (NIG) maintains a primary nucleotide sequence database as a member of the International Nucleotide Sequence Database Collaboration (INSDC) in partnership with the US National Center for Biotechnology Information and the European Bioinformatics Institute. The NIG operates the NIG supercomputer as a computational basis for the construction of DDBJ databases and as a large-scale computational resource for Japanese biologists and medical researchers. In order to accommodate the rapidly growing amount of deoxyribonucleic acid (DNA) nucleotide sequence data, NIG replaced its supercomputer system, which is designed for big data analysis of genome data, in early 2019. The new system is equipped with 30 PB of DNA data archiving storage; large-scale parallel distributed file systems (13.8 PB in total) and 1.1 PFLOPS computation nodes and graphics processing units (GPUs). Moreover, as a starting point of developing multi-cloud infrastructure of bioinformatics, we have also installed an automatic file transfer system that allows users to prevent data lock-in and to achieve cost/performance balance by exploiting the most suitable environment from among the supercomputer and public clouds for different workloads.

INTRODUCTION

The DNA Data Bank of Japan (DDBJ) (https://www.ddbj.nig.ac.jp) (1) is a public database of nucleotide sequences established at the National Institute of Genetics (NIG) (https://www.nig.ac.jp/nig). Since 1987, the DDBJ Center has been collecting annotated nucleotide sequences as its traditional database service. This endeavour has been conducted in collaboration with GenBank (2) at the US National Center for Biotechnology Information (NCBI) and in partnership with the European Nucleotide Archive (ENA) (3) at the European Bioinformatics Institute (EBI). The collaborative framework is called the International Nucleotide Sequence Database Collaboration (INSDC) (4), and the product database from this framework is called the International Nucleotide Sequence Database (INSD).

Within the INSDC framework, the DDBJ Center also services the DDBJ Sequence Read Archive (DRA) for raw sequencing data and alignment information from high-throughput sequencing platforms (5), the BioProject for sequencing project metadata, and BioSample for sample information (1,6). This comprehensive resource of nucleotide sequences and associated biological information complies with the INSDC policy that guarantees free and unrestricted access to data archives (7). In addition to these INSDC databases, the DDBJ Center has accepted functional genomics experiments in the Genomic Expression Archive (GEA) which is counterpart of the Gene Expression Omnibus at NCBI (8) and the ArrayExpress at EBI (9). For human individual genotype and phenotype data requiring authorized access, the DDBJ Center has provided the controlled-access database Japanese Genotype-phenotype Archive (JGA) in collaboration with the National Bioscience Database Center (NBDC) in the Japan Science and Technology Agency (JST) since 2013 (10).

The supercomputer system operated by the NIG as a computational infrastructure for developing the DDBJ databases is also provided for use as large-scale computational resources to Japanese researchers in the fields of medicine and biology (11). In early 2019, the NIG supercomputer system was replaced in order to accommodate the recent rapid growth of the genome data archives.

In the present article, we report on updates to the above-mentioned services at the DDBJ Center, and on the new supercomputer system. All of the resources described here are available from https://www.ddbj.nig.ac.jp, and most of the archival data can be downloaded at ftp://ftp.ddbj.nig.ac.jp.

EBI/NCBI/DDBJ

Instituce zabývající se shromažďováním, správou a poskytováním dat a informací a vývojem analytických nástrojů.

EBI

Evropský institut pro bioinformatiku

European Bioinformatics Institute

NCBI

Národní centrum pro biotechnologické informace

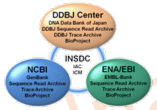
National Center for Biotechnology Information

DDBJ Center

The DNA Data Bank of Japan Center

http://www.ebi.ac.uk/ http://www.ncbi.nlm.nih.gov/ http://www.ddbj.nig.ac.jp/

ENA ⟷ GenBank ⟷ DDBJ



LSD 3.0

Leaf Senescence DataBase

View details

MethBank

Methylation Bank

View details

PigGIS

Pig Genomic Informatics System

View details

ChickVD

Chicken Variation Database

View details

国家生物信息中心
China National Center for Bioinformation

National Genomics Data Center

The National Genomics Data Center (NGDC), part of the China National Center for Bioinformation (CNCB), advances life & health sciences by providing open access to a suite of resources, with the aim to translate big data into big discoveries and support worldwide activities in both academia and industry.

NGDC

https://bigd.big.ac.cn/

Database Resources of the National Genomics Data Center in 2020

National Genomics Data Center Members and Partners*,†

Received September 15, 2019; Revised September 30, 2019; Editorial Decision October 01, 2019; Accepted October 02, 2019

ABSTRACT

The National Genomics Data Center (NGDC) provides a suite of database resources to support worldwide research activities in both academia and industry. With the rapid advancements in higher-throughput and lower-cost sequencing technologies and accordingly the huge volume of multi-omics data generated at exponential scales and rates, NGDC is continually expanding, updating and enriching its core database resources through big data integration and value-added curation. In the past year, efforts for update have been mainly devoted to BioProject, BioSample, GSA, GWH, GVM, NONCODE, LncBook, EWAS Atlas and IC4R. Newly released resources include three human genome databases (*PGG*.SNV, *PGG*.Han and CGVD), eLMSG, EWAS Data Hub, GWAS Atlas, iSheep and PADS Arsenal. In addition, four web services, namely, eGPS Cloud, BIG Search, BIG Submission and BIG SSO, have been significantly improved and enhanced. All of these resources along with their services are publicly accessible at https://bigd.big.ac.cn.

INTRODUCTION

The National Genomics Data Center (NGDC), officially approved by the Ministry of Science & Technology and the Ministry of Finance of the People's Republic of China in June 2019, is a national-level center dedicated to advancing life and health sciences by archiving, managing and processing a wide range of genomics related data. NGDC is established based on the BIG Data Center (1–3) at Beijing Institute of Genomics (BIG) of Chinese Academy of Sciences (CAS), jointly in close collaboration with two CAS institutions, namely, Institute of Biophysics (IBP) and Shanghai Institute of Nutrition and Health (SINH). Considering the rapid advancements in higher-throughput and lower-cost sequencing technologies, huge amounts of multi-omics data are generated at ever-growing rates and scales. Therefore, the primary mission of NGDC is to build archive platforms and information systems, develop advanced algorithms and tools to translate big data into big discovery, and provide open access to a suite of database resources in support of research activities of global users from both academia and industry.

During the past year, NGDC has expanded, updated and enriched the amount and type of data through big data integration and value-added curation, particularly by close collaboration with IBP and SINH, with significant improvements and advances over the previous release. In terms of data attribute and curation intensity, database resources in NGDC can be generally divided into three categories: Data—raw sequence data and metadata, Information—value-added standardized information, and Knowledge—curated knowledge and knowledge graphs. Here, we provide a brief summary of new developments and recent updates, and describe the core resources and services of NGDC (Figure 1). All resources, along with their services, are publicly accessible through the home page of NGDC at https://bigd.big.ac.cn.

NEW DEVELOPMENTS

Human genome resources

PGG.SNV (http://www.pggsnv.org) (4) is a human genome database, which gives much higher weight to previously under-investigated indigenous populations in Asia, as these genomes harbor an enormous number of variants that have not been observed in the extensively studied populations of European ancestry. In the current version, *PGG*.SNV archives 265 million single nucleotide variants (SNVs) across 220 147 present-day human genomes and 1018 ancient genomes and estimates their frequencies in 977 diverse populations, including 1009 newly sequenced genomes rep-

Strukturní databáze

- **PDB – Protein Data Bank.** Databáze obsahuje experimentálně získané struktury proteinů, nukleových kyselin a komplexů informačních biomakromolekul.

Molecular Type	X-ray	NMR	EM	Multiple methods	Neutron	Other	Total
Protein (only)	132997	11479	4105	160	67	32	148840
Other	8027	92	479	6	0	4	8608
Protein/NA	7063	265	1442	3	0	0	8773
Nucleic acid (only)	2095	1309	47	6	2	1	3460
Total	150182	13145	6073	175	69	37	169681

- **NDB – Nucleic Acid Database**

A Portal for Three-dimensional Structural Information about Nucleic Acids
As of 7-Oct-2020 number of released structures: 11005

Osnova

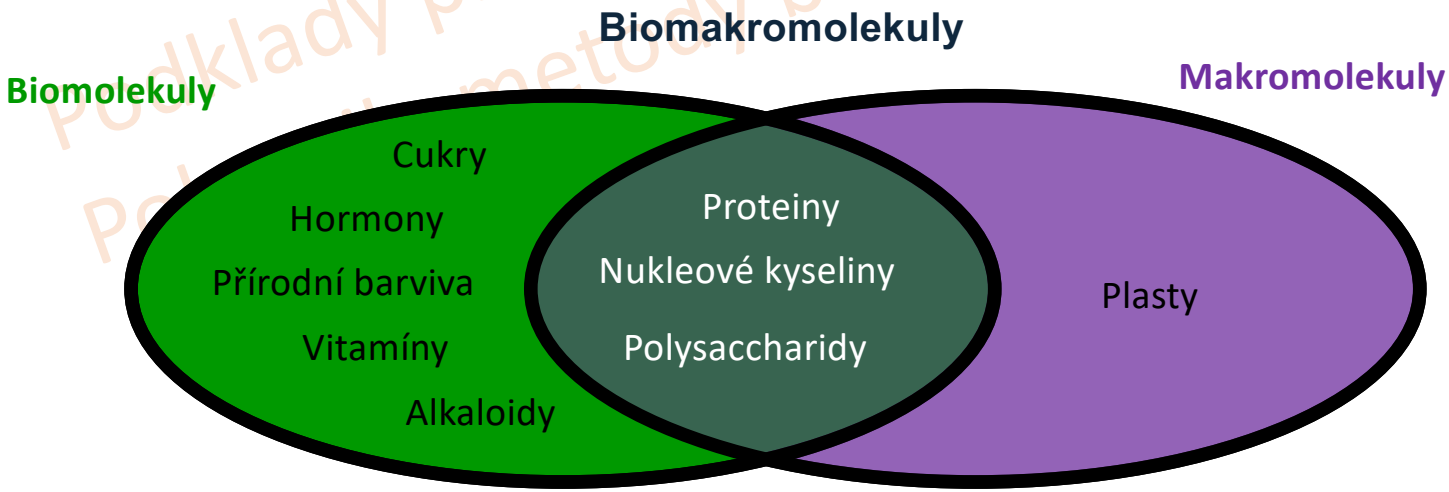
- **Úvod do bioinformatiky**
  - Definice, molekulárně biologická data, databáze
  - Rozdělení databází, bioinformatická centra
- **Manipulace se sekvencemi**
  - Sekvence biomakromolekul, aminokyseliny, báze, alignment
  - Význam alignmentu, přiložení páru sekvencí a vícenásobné přiložení
- **Predikce struktury proteinů**
  - Predikce 2-D struktury proteinů, predikce 3-D struktury proteinů
  - Threading, Homology modeling, *Ab initio*
- **Predikce genů**
  - Predikce genů u prokaryot a eukaryot, predikční nástroje a postupy

Biomakromolekuly

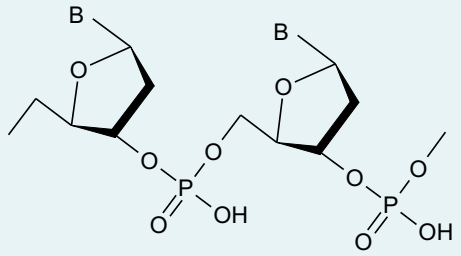
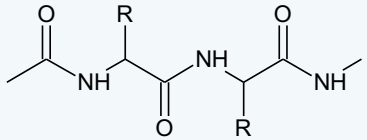
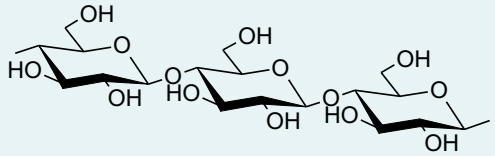
Biomolekuly jsou přirozenou součástí živých organismů.

Velké molekuly. Typické malé molekuly jsou tvořeny několika atomy až několika sty atomů. Makromolekuly tvoří tisíce až miliony atomů.

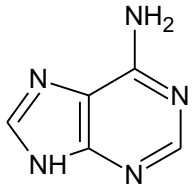
Základní stavební jednotky hmoty. Jsou tvořeny atomy, které navzájem spojují kovalentní vazby.



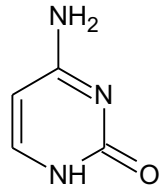
Sekvence biomakromolekul

Makromolekula	Stavební jednotky	Typ vazby	Schéma
Nukleová kyselina	Nukleotidy	Esterová	
Protein	Aminokyseliny	Peptidová	
Polysacharid	Monosacharidy	Glykosidická	

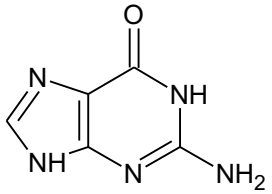
Nukleové báze



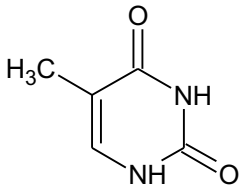
Adenine



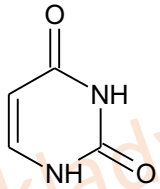
Cytosine



Guanine



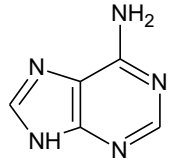
Thymine



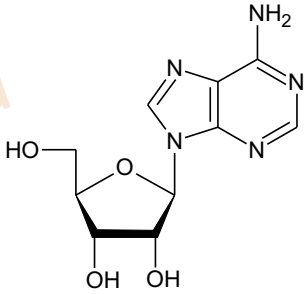
Uracil

adenin	cytosin	guanin	thymin	uracil
A	C	G	T	U

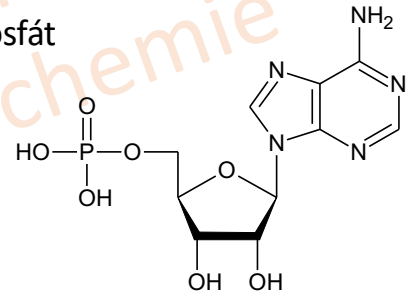
Nukleová báze
Adenin



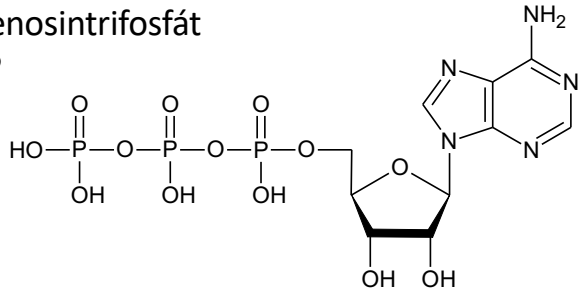
Nukleosid
Adenosin



Nukleotid
Adenosinmonofosfát
AMP



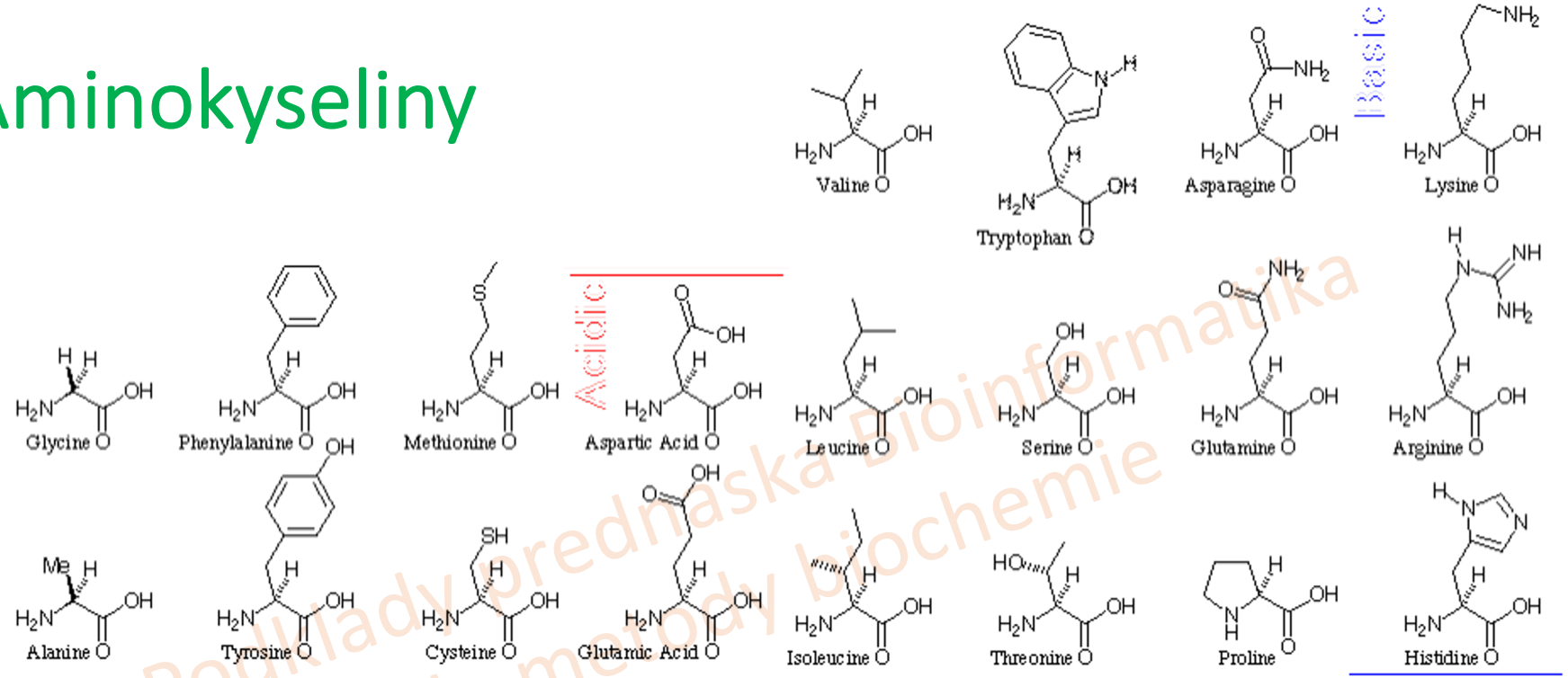
Nukleotid
Adenosintrifosfát
ATP



Při základních manipulacích se sekvencemi v bioinformatice uvažujeme vždy pouze Watson-Crickovo párování bazí

(neplatí pro 3D predikce a struktury)

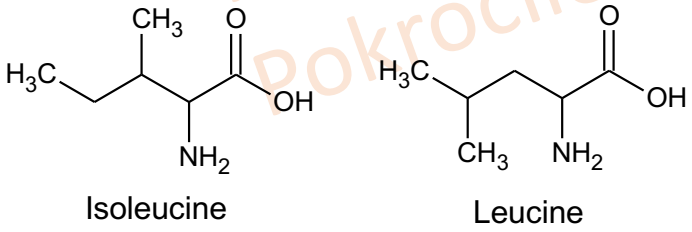
Aminokyseliny

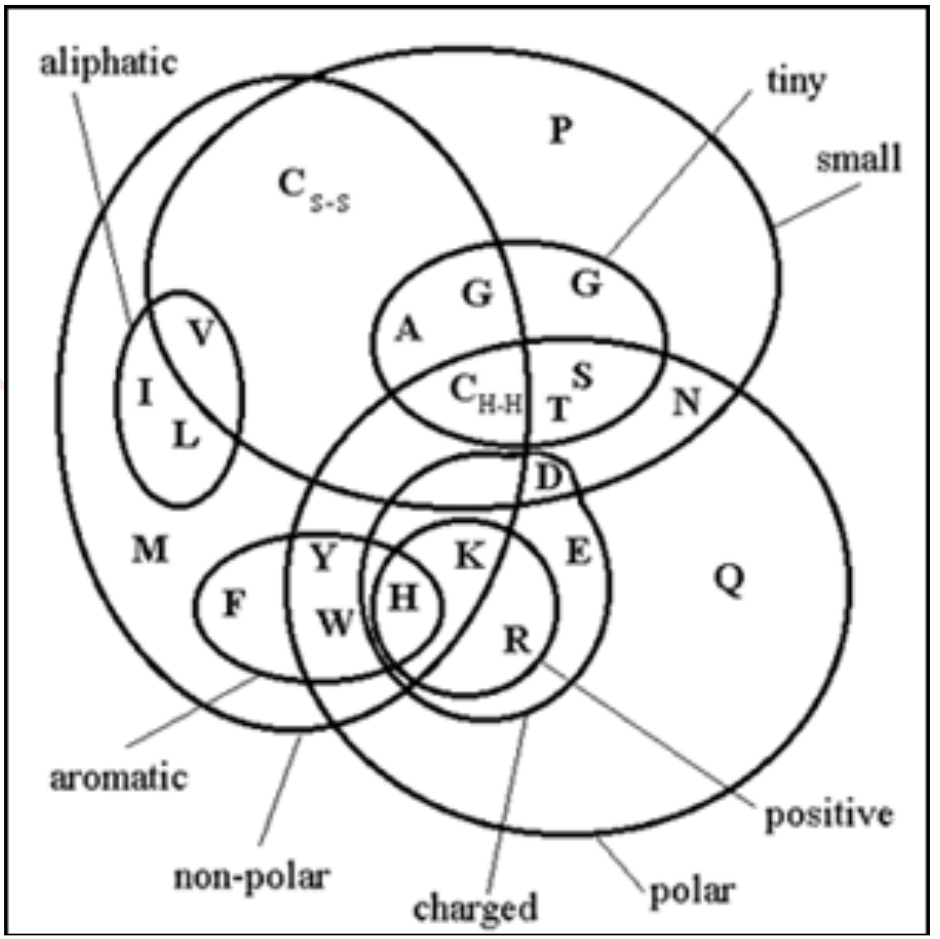


glycin	alanin	valin	leucin	izoleucin	asparagová kys.	asparagin	glutamová kys.	glutamin	arginin	lysin	histidin	fenylalanin	serin	threonin	tyrozin	tryptofan	methionin	cystein	prolin	selenocystein	pyrolysin
Gly	Ala	Val	Leu	Ile	Asp	Asn	Glu	Gln	Arg	Lys	His	Phe	Ser	Thr	Tyr	Trp	Met	Cys	Pro	Sec	Pyr
G	**A**	**V**	**L**	**I**	**D**	**N**	**E**	**Q**	**R**	**K**	**H**	**F**	**S**	**T**	**Y**	**W**	**M**	**C**	**P**	**U**	**O**

Aminokyseliny

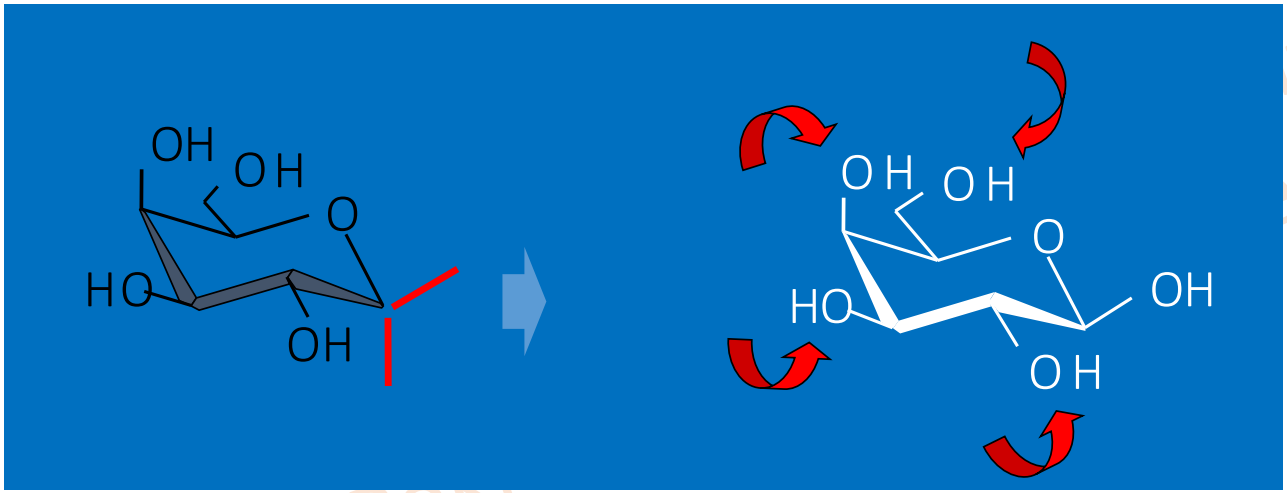
Aminokyseliny s podobnými vlastnostmi mohou plnit v proteinu stejné funkce – bývají vzájemně zastupitelné





Polysacharidy

Komplikované sekvence – alignment se neprovádí



Fucα1-2
|
Galα1-3Galβ1-4/GlcNAcβ–
|
Fucα1-3

Galα1-3(Fucα1-2)Galβ1-4(Fucα1-3)
GlcNAcβ–

	Hexasaccharides	Hexapeptides
Nb Isomers	> 10^{12} (for 6 hexopyranoses)	64 10^{6} (20 amino acids)

Polysacharidy

Komplikované sekvence – alignment se neprovádí

Polymer	**Protein**	**Nukleová kyselina**	**Polysacharid**
Počet druhů základních stavebních jednotek	20 (22)	4 (DNA) 4 (RNA)	”desítky”
Počet typů vzájemných vazeb	1	1	2 x 4 (pro hexosu)

Práce se sekvencemi

- Vyskytuje se shodná/podobná sekvence (protein/DNA) v databázi?
- Jak podobné jsou podobné sekvence?
- Jsou podobné, shodné, odlišné?
- **Alignment** – srovnání (**přiložení**) dvou či více sekvencí (aminokyselinových, nukleotidových) na základě jejich vzájemné podobnosti.

ATGTCTACTCCTGGAGCACAGCAAGTCCTCTTCCGCACCGGAATTGCCGCGGTCAACTCAACCAACCATCTCCGTGTTTACTTCCAGGATGTCTATG
GCAGTATTCGCGAGAGTCTCTACGAGGGCAGCTGGGCTAACGGCACCGAAAAGAACGTTATCGGCAATGCTAAGCTTGGCAGCCCTGTGGCCGC
GACTTCTAAGGAGCTGAAGCATATCCGTGTCTACACCCTCACTGAAGGAAACACCCTACAGGAGTTCGCCTACGACTCCGGAACCGGATGGTACAA
CGGCGGGCTGGGCGGTGCAAAGTTCCAAGTCGCACCCTACTCTCGCATTGCTGCCGTGTTCCTAGCCGGAACAGATGCATTGCAGTTGCGAATCTA
TGCACAGAAGCCAGATAACACAATCCAGGAGTATATGTGGAACGGCGATGGCTGGAAGGAGGGCACCAACCTGGGAGGTGCTCTCCCCGGCACT
GGAATCGGAGCCACCTCCTTCCGCTATACCGACTACAATGGCCCAAGCATCCGGATCTGGTTCCAAACTGACCTCAAACTCGTCCAAAGAGCCTAC
GACCCGCACAAAGGCTGGTACCCGGACCTCGTCACCATCTTTGACAGGGCACCGCCACGTACGGCCATTGCAGCCACCAGCTTTGGAGCCGGCAA
CAGTTCCATCTACATGCGTATCTACTTTGTCAATTCGGACAACACTATCTGGCAGGTCTGCTGGGACCACGGCAAGGGCTATCACGACAAGGGAAC
CATCACCCCAGTCATTCAGGGCTCGGAGGTCGCCATTATCAGCTGGGGCAGTTTCGCCAATAACGGGCCGGATCTGCGTCTGTACTTTCAGAATGG
AACATACATTAGTGCTGTGAGCGAGTGGGTTTGGAATCGGGCACATGGGTCGCAGTTGGGCAGAAGTGCTCTTCCTCCTGCTTGA

ATGGCTGATTCTCAAACGTCATCCAACCGCGCCGGCGAATTCTCGATTCCGCCGAATACCGATTTCCGCGCGATTTTCTTCGCGAATGCCGCCGAGC
AACAGCACATCAAATTGTTCATCGGCGACAGCCAGGAACCCGCCGCGTATCACAAGCTGACGACGCGCGACGGCCCGCGCGAAGCCACGCTGAAT
TCCGGCAACGGCAAGATCCGTTTCGAGGTGTCGGTGAACGGCAAGCCGTCGGCGACCGACGCGCGTCTCGCGCCGATCAACGGCAAGAAGTCGG
ACGGCTCGCCGTTCACGGTCAACTTCGGGATCGTCGTGTCGGAAGACGGCCACGACAGCGACTACAACGACGGCATCGTCGTGCTCCAGTGGCCG
ATCGGCTGA

ATGCTGGTGATTGTGGATGCCGTTACCCTGCTGAGCGCCTATCCGGAAGCCAGCCGTGATCCGGCCGCCCCGACCGTGATTGATGGTCGCCACCTG
TATGTTGTTAGCCCGGGCGATGCCGCGCAGCTGGGCCATAACGATAGCCGTCTGTTTACCGGTCTGAGCCCGGGTGATCAGCTGCATCTGCGCGA
AACCGCGCTGGCGCTGCGCGCGGAAGTGAGCGTGCTGTTTATTCGCTTTGCCCTGAAAGATGCCGGCATTGTTGCCCCGATCGAACTGGAAGTGC
GTGATGCCGCCACCGCCGTTCCGGATGCGGATGATCTGCTGCATCCGAGCTGTCGTCCGCTGAAAGATCATTATTGGCGCAGCGATGTGCTGGCG
GCGGGCGCGACCACCTGTACCGCCGATTTTGCGGTGTGCGATCGTGATGGCACCGTGAGCGGTTATTTTCGTTGGGAAACCAGCATTGAAATTGC
GGGCAGCCAGCCGGATACCAAACAGCCGGGCTTTAAACCGAGCAGCGATCGCAATGGCAACTTTAGCCTGCCGCCGAATACCGCCTTTAAAGCGA
TCTTCTATGCGAACGCGGCGGATCGTCAGGATCTGAAACTGTTTATTGATGATGCGCCGGAACCGGCCGCCACCTTTGTGGGTAACAGCGAAGAT
GGTGTGCGTCTGTTTACCCTGAATAGCAAAGGTGGTAAAATTCGTATTGAAGCGAGCGCGAACGGCCGTCAGAGCGCGACCGATGCCCGTCTGGC
GCCGCTGAGCGCGGGCGATACCGTGTGGCTGGGCTGGCTGGGCGCGGAAGATGGTGCCGATGCGGATTATAATGATGGCATTGTTATTCTGCAG
TGGCCGATTACCTAA

ATGTCGAGCGTTCAAACCGCTGCCACTTCGTGGGGAACCGTACCGTCGATCCGTGTGTACACGGCCAATAATGGCAAGATCACCGAGCGATGCTG
GGACGGGAAGGGGTGGTACACCGGTGCCTTCAACGAGCCCGGCGATAACGTCTCCGTAACCAGCTGGCTGGTCGGCAGCGCGATCCATATCCGC
GTCTATGCAAGCACCGGCACCACGACCACGGAGTGGTGCTGGGACGGCAACGGCTGGACCAAGGGCGCCTACACCGCCACGAACTGA

ATGCCGCTGCTGAGCGCCAGTATCGTGAGCGCGCCGGTGGTGACCAGCGAAACCTATGTGGATATTCCGGGCCTGTATCTGGATGTTGCGAAAGC
CGGTATCCGTGATGGCAAACTGCAGGTTATCCTGAATGTGCCGACCCCGTATGCGACGGGCAATAACTTTCCGGGTATTTATTTTGCGATCGCCAC
CAACCAGGGCGTGGTGGCGGATGGTTGCTTTACGTATAGTAGCAAAGTGCCGGAAAGTACGGGCCGTATGCCGTTTACCCTGGTTGCGACCATTG
ATGTGGGTAGCGGTGTTACCTTCGTGAAAGGTCAGTGGAAATCTGTTCGCGGCTCTGCGATGCATATTGATAGCTATGCAAGCCTGAGTGCGATTT
GGGGCACCGCGGCACCGAGTTCTCAGGGTTCTGGTAACCAGGGTGCGGAAACGGGTGGCACCGGTGCCGGTAATATTGGTGGCGGCGGTGAAC
GTGATGGCACCTTTAATCTGCCGCCGCATATTAAATTCGGTGTTACCGCGCTGACCCACGCGGCGAACGATCAGACCATTGATATTTATATTGATGA
TGATCCGAAACCGGCAGCCACCTTTAAAGGCGCGGGCGCGCAGGATCAGAACCTGGGTACCAAAGTGCTGGATTCTGGCAATGGCCGTGTTCGC
GTTATCGTTATGGCGAACGGCCGTCCGAGCCGCCTGGGTTCTCGTCAGGTGGATATTTTTAAAAAATCTTATTTCGGTATTATTGGCTCTGAAGATG
GTGCGGATGATGATTATAACGATGGCATCGTGTTTCTGAACTGGCCGCTGGGCTAA

ATGCCGCTCCTGAGCGCCAGTATCGTGAGCGCGCCGGTGGTGACCAGCCAAACCTATGTGGATATTCCGGGCCTGTATCTGGATGTTGCGAAAGC
CGGTATCCGTGATGGCAAACTGCAGGTTATCCTGAATGTGCCGACCCCGTATGCGACGGGCAATAACTTTCCGGGTATTTATTTTGCGATCGCCAC
CAACCAGGGCGTGGTGGCGGATGGTTGCTTTACGTATAGTAGCAAAGTGCCGGAAAGTACGGGCCGTATGCCGTTTACCCTGGTTGCGACCATTG
ATGTGGGTAGCGGTGTTACCTTCGTGAAAGGTCAGTGGAAATCTGTTCGCGGCTCTGCGATGCATATTGATAGCTATGCAAGCCTGAGTGCGATTT
GGGGCACCGCGGCACCGAGTTCTCAGGGTTCTGGTAACCAGGGTGCGGAAACGGGTGGCACCGGTGCCGGTAATATTGGTGGCGGCGGTAAGCT
TGCGGCCGCACTCGAGATCAAACGGGCTAGCCAGCCAGAACTCGCCCCGGAAGACCCCGAGGATGTCGAGCACCACCACCACCACCACTGA

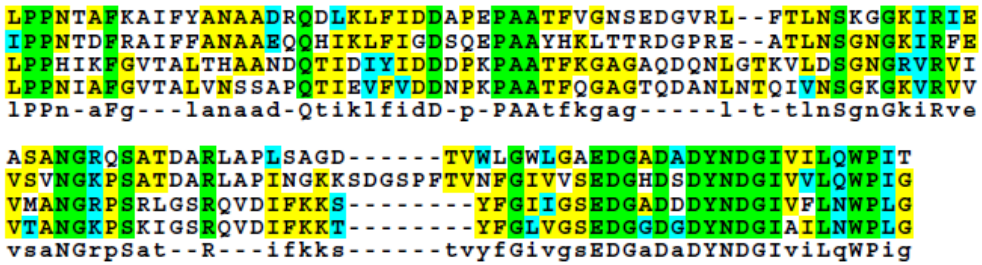
Práce se sekvencemi

- Vyskytuje se shodná/podobná sekvence (protein/DNA) v databázi?
- Jak podobné jsou podobné sekvence?
- Jsou podobné, shodné, odlišné?
- **Alignment** – srovnání (**přiložení**) dvou či více sekvencí (aminokyselinových, nukleotidových) na základě jejich vzájemné podobnosti.

```
MSTPGAQQVLFRTGIAAVNLTNHLRVYFQDVYGSIRESLYEGSWANGTEKNVIGNAKLGSPVAATSKEL
KHIRVYTLTEGNTLQEFAYDSGTGWYNGGLGGAKFQVAPYSRIAAVFLAGTDALQLRIYAQKPDNTIQE
YMWNGDGWKEGTNLGGALPGTGIGATSFRYTDYNGPSIRIWFQTDDLKLVQRAYDPHKGWYPDLVTIFD
RAPPRTAIAATSFGAGNSSIYMRIYFVNSDNTIWQVCWDHGKGYHDKGTITPVIQGSEVAIISWGSFAN
NGPDLRLYFQNGTYISAVSEWVWNRAHGSQLGRSALPPA

MADSQTSSNRAGEFSIPPNTDFRAIFFANAAEQQHIKLFIGDSQEPAAYHKLTTRDGPREATLNSGNGK
IRFEVSVNGKPSATDARLAPINGKKSDGSPFTVNFGIVVSEDGHDSDYNDGIVVLQWPIG

MLVIVDAVTLLSAYPEASRDPAAPTVIDGRHLYVVSPGDAAQLGHNDSRLFTGLSPGDQLHLRETALAL
RAEVSVLFIRFALKDAGIVAPIELEVRDAATAVPDADDLLHPSCRPLKDHYWRSDVLAAGATTCTADFA
VCDRDGTVSGYFRWETSIEIAGSQPDTKQPGFKPSSDRNGNFSLPPNTAFKAIFYANAADRQDLKLFID
DAPEPAATFVGNSEDGVRLFTLNSKGGKIRIEASANGRQSATDARLAPLSAGDTVWLGWLGAEDGADAD
YNDGIVILQWPIT

MSSVQTAATSWGTVPSIRVYTANNGKITERCWDGKGWYTGAFNEPGDNVSVTSWLVGSAIHIRVYASTG
TTTTEWCWDGNGWTKGAYTATN

MPLLSASIVSAPVVTSETYVDIPGLYLDVAKAGIRDGKLQVILNVPTPYATGNNFPGIYFAIATNQGVV
ADGCFTYSSKVPESTGRMPFTLVATIDVGSGVTFVKGQWKSVRGSAMHIDSYASLSAIWGTAAPSSQGS
GNQGAETGGTGAGNIGGGGERDGTFNLPPHIKFGVTALTHAANDQTIDIYIDDDPKPAATFKGAGAQDQ
NLGTKVLDSGNGRVRVIVMANGRPSRLGSRQVDIFKKSYFGIIGSEDGADDDYNDGIVFLNWPLG

MPLLSASIVSAPVVTSQTYVDIPGLYLDVAKAGIRDGKLQVILNVPTPYATGNNFPGIYFAIATNQGVV
ADGCFTYSSKVPESTGRMPFTLVATIDVGSGVTFVKGQWKSVRGSAMHIDSYASLSAIWGTAAPSSQGS
GNQGAETGGTGAGNIGGGGKLAAALEIKRASQPELAPEDPEDVEHHHHHH
```

Význam alignmentu

- Identifikace sekvence v databázi
- Hledání podobných sekvencí v databázi
- Detekce mutací
- Hledání konzervovaných částí sekvence
- Odhalování příbuzenských vztahů
- Předpověď funkce makromolekuly
- Předpověď vyšších struktur



Typy alignmentu

Pairwise alignment (párové přiložení) – dvě sekvence

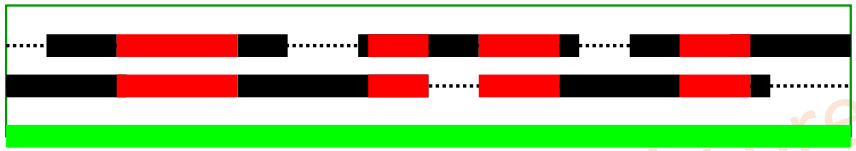
```
WLAKALKYLMETAQASSISTELARHHPRAVDAKRKSEMKRKTAMWLAKALKYLMETAQASSISTELARHHPRAVDAKRKSEMKRKTAM
WLAKALKYLMETAQASSISTELARHHPRAVDAKRKSEMKRKTAMWLAKALKYLMETAQASSISTELARHHPRAVDAKRKSEMKRKTAM
```

Multiple sequence alignment (vícenásobné přiložení) – více sekvencí

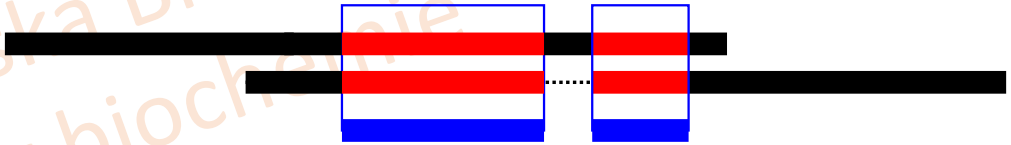
```
WLAKALKYLMETAQASSISTELARHHPRAVDAKRKSEMKRKTAMWLAKALKYLMETAQASSISTELARHHPRAVDAKRKSEMKRKTAM
 WLAKALKYLMETAQASSISTELARHHPRAVDAKRKSEMKRKTAMWLAKALKYLMETAQASSISTELARHHPRAVDAKRKSEMKRKTAM
  WLAKALKYLMETAQASSISTELARHHPRAVDAKRKSEMKRKTAMWLAKALKYLMETAQASSISTELARHHPRAVDAKRKSEMKRKTAM
   WLAKALKYLMETAQASSISTELARHHPRAVDAKRKSEMKRKTAMWLAKALKYLMETAQASSISTELARHHPRAVDAKRKSEMKRKTAM
    WLAKALKYLMETAQASSISTELARHHPRAVDAKRKSEMKRKTAMWLAKALKYLMETAQASSISTELARHHPRAVDAKRKSEMKRKTAM
     WLAKALKYLMETAQASSISTELARHHPRAVDAKRKSEMKRKTAMWLAKALKYLMETAQASSISTELARHHPRAVDAKRKSEMKRKTAM
```

Pairwise alignment

- Srovnání dvou sekvencí.
- Sekvence mohou být přiloženy v celé své délce (**global alignment**) nebo jen v určitém regionu (**local alignment**).



Vychází z předpokladu, že obě srovnávané sekvence jsou víceméně shodné v celé své délce. Alignment k sobě přikládá celé sekvence (od počátku do konce) a to včetně částí, které si příliš neodpovídají.



Hledá úseky dvou sekvencí, které si podle zvolených kritérií dobře odpovídají. Nesnaží se zahrnout celé sekvence, pokud si jejich některé části neodpovídají.

Algoritmy

- Témeř výhradně se užívají **heuristické algoritmy** – nalezení výsledku v dostatečně krátkém čase.
- Vývoj algoritmů je prováděn v návaznosti na srovnávání výsledků s tzv. zlatým standardem – alignment na základě známých 3D struktur.

FASTA formát

>název(␣popis dle vlastní volby)↵
SEKVENCESEKVENCESEKVENCESEKVENCESEKVENCESEKVENCE↵

POVINNÉ VOLITELNÉ

http://emboss.sourceforge.net/docs/themes/SequenceFormats.html

```
>AFL
MSTPGAQQVLFRTGIAAVNLTNHLRVYFQDVYGSIRESLYEGSWANGTEKNVIGNAKLGSPVAATSKELKHIRVYTLTEGNTLQ
EFAYDSGTGWYNGGLGGAKFQVAPYSRIAAVFLAGTDALQLRIYAQKPDNTIQEYMWNGDGWKEGTNLGGALPGTGIGATSFRY
TDYNGPSIRIWFQTDDLKLVQRAYDPHKGWYPDLVTIFDRAPPRTAIAATSFGAGNSSIYMRIYFVNSDNTIWQVCWDHGKGYH
DKGTITPVIQGSEVAIISWGSFANNGPDLRLYFQNGTYISAVSEWVWNRAHGSQLGRSALPPA

>BC2LA
MADSQTSSNRAGEFSIPPNTDFRAIFFANAAEQQHIKLFIGDSQEPAAYHKLTTRDGPREATLNSGNGKIRFEVSVNGKPSATD
ARLAPINGKKSDGSPFTVNFGIVVSEDGHDSDYNDGIVVLQWPIG

> BC2LD
MLVIVDAVTLLSAYPEASRDPAAPTVIDGRHLYVVSPGDAAQLGHNDSRLFTGLSPGDQLHLRETALALRAEVSVLFIRFALKD
AGIVAPIELEVRDAATAVPDADDLLHPSCRPLKDHYWRSDVLAAGATTCTADFAVCDRDGTVSGYFRWETSIEIAGSQPDTKQP
GFKPSSDRNGNFSLPPNTAFKAIFYANAADRQDLKLFIDDAPEPAATFVGNSEDGVRLFTLNSKGGKIRIEASANGRQSATDAR
LAPLSAGDTVWLGWLGAEDGADADYNDGIVILQWPIT

>RSL
MSSVQTAATSWGTVPSIRVYTANNGKITERCWDGKGWYTGAFNEPGDNVSVTSWLVGSAIHIRVYASTGTTTTEWCWDGNGWTK
GAYTATN

>gi|444369855|ref|ZP_21169562.1| fucose-binding lectin II [Burkholderia cenocepacia
K56-2Valvano]
MPLLSASIVSAPVVTSETYVDIPGLYLDVAKAGIRDGKLQVILNVPTPYATGNNFPGIYFAIATNQGVVADGCFTYSSKVPEST
GRMPFTLVATIDVGSGVTFVKGQWKSVRGSAMHIDSYASLSAIWGTAAPSSQGSGNQGAETGGTGAGNIGGGGERDGTFNLPPH
IKFGVTALTHAANDQTIDIYIDDDPKPAATFKGAGAQDQNLGTKVLDSGNGRVRVIVMANGRPSRLGSRQVDIFKKSYFGIIGS
EDGADDDYNDGIVFLNWPLG

>gi|283806765|pdb|2WQ4|A Chain A
MPLLSASIVSAPVVTSQTYVDIPGLYLDVAKAGIRDGKLQVILNVPTPYATGNNFPGIYFAIATNQGVVADGCFTYSSKVPEST
GRMPFTLVATIDVGSGVTFVKGQWKSVRGSAMHIDSYASLSAIWGTAAPSSQGSGNQGAETGGTGAGNIGGGGKLAAALEIKRA
SQPELAPEDPEDVEHHHHHH
```

Jak poznat dobré přiložení?

```
MAM--UZDOST--STAROSTISHAMIZ--NOSTIRATOLESTI
|||    ||||  |||||         |  ||          |||
MAMRA--DOSTZESTARO--------ZITNO---------STI
```

18 shod

```
1 MAMUZDOST--STAROSTISHAMIZNOSTIRATOLESTI  37
  ||| .||||  ||||       . |    |  |. |||
1 MAMRADOSTZESTAR-------O-Z----I--TNO-STI  24
```

17 shod, 3 podobnosti

```
1 MAMUZDOST--STAROSTISHAMIZNOSTIRATOLESTI   37
  ||| .||||  |||||.:.:.                |||
1 MAMRADOSTZESTAROZITNO--------------STI   24
```

15 shod, 6 podobností

Scoring matrix (skórovací matice)

- Dvě sekvence považujeme za **příbuzné**, vycházejí-li ze společného předka; pak dobu potřebnou k jejich evoluci můžeme odvodit z množství rozdílů mezi nimi
- **Záměna** aa je častější než inserce/delece. Pravděpodobnost změny jedné aminokyseliny na jinou je **přímo úměrná podobnosti** obou aminokyselin.
- **Matice** vzniká přiřazením hodnoty (pravděpodobnosti) jednotlivým dvojicím aminokyselin v závislosti na jejich vzájemné „zastupitelnosti“ – pravděpodobnosti substituce

Substituční skórovací matice

víceméně dva typy:

1. založené na záměnnosti genetického kódu nebo vlastností aminokyselin
2. odvozené z **empirických** studií aminokyselinových substitucí (přesnější)

Nejvíce používané jsou empirické matice PAM a BLOSUM

PAM – Point Accepted Mutation

Constructed by Margaret Dayhoff in 1978.

Zahrnuje pravděpodobnost záměny jedné aminokyseliny v druhou během evoluce

Předpokládá, že každá další mutace nezávisí na předchozí.

Odvozena z globálního alignmentu rodin proteinů
(Podobnost sekvencí v rodině > 85%, vypočtena na základě 1572 změn v aminokyselinovém složení v 71 proteinových rodinách)

- vysoká spolehlivost alignmentu
- vysoká pravděpodobnost, že záměna aminokyseliny je dána jedinou mutací

Vypočtena pravděpodobnost s jakou jedna AA se změní na jakoukoliv jinou

PAM1 reflektuje průměrnou záměnu 1% všech aminokyselinových pozic

PAM250 (20% identita) je odvozena od PAM1
její 250-tinásobnou multiplikací (250 mutací na 100 aminokyselin)

Vyšší číslo PAM matrice znamená větší evoluční vzdálenost

k vysvětlení

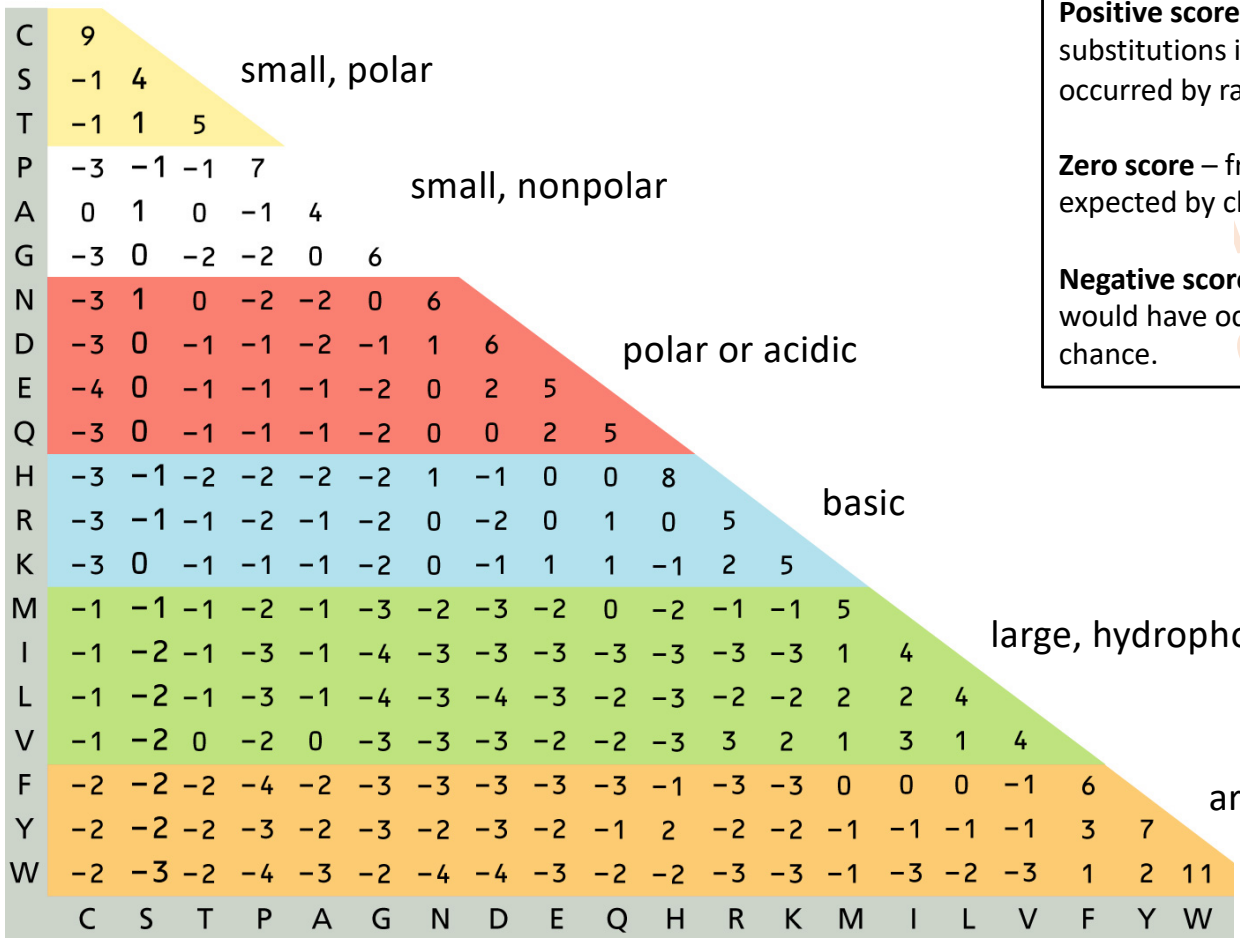
PAM 1 matice

	A	R	N	D	C
A	9867	2	9	10	3
R	1	9913	1	0	1
N	4	1	9822	36	0
D	6	0	42	9859	0
C	1	1	0	0	9973

All entries $\times 10^4$

k vysvětlení

PAM250 matice



	C	S	T	P	A	G	N	D	E	Q	H	R	K	M	I	L	V	F	Y	W
C	9																			
S	−1	4																		
T	−1	1	5																	
P	−3	−1	−1	7																
A	0	1	0	−1	4															
G	−3	0	−2	−2	0	6														
N	−3	1	0	−2	−2	0	6													
D	−3	0	−1	−1	−2	−1	1	6												
E	−4	0	−1	−1	−1	−2	0	2	5											
Q	−3	0	−1	−1	−1	−2	0	0	2	5										
H	−3	−1	−2	−2	−2	−2	1	−1	0	0	8									
R	−3	−1	−1	−2	−1	−2	0	−2	0	1	0	5								
K	−3	0	−1	−1	−1	−2	0	−1	1	1	−1	2	5							
M	−1	−1	−1	−2	−1	−3	−2	−3	−2	0	−2	−1	−1	5						
I	−1	−2	−1	−3	−1	−4	−3	−3	−3	−3	−3	−3	−3	1	4					
L	−1	−2	−1	−3	−1	−4	−3	−4	−3	−2	−3	−2	−2	2	2	4				
V	−1	−2	0	−2	0	−3	−3	−3	−2	−2	−3	3	2	1	3	1	4			
F	−2	−2	−2	−4	−2	−3	−3	−3	−3	−3	−1	−3	−3	0	0	0	−1	6		
Y	−2	−2	−2	−3	−2	−3	−2	−3	−2	−1	2	−2	−2	−1	−1	−1	−1	3	7	
W	−2	−3	−2	−4	−3	−2	−4	−4	−3	−2	−2	−3	−3	−1	−3	−2	−3	1	2	11

Positive score – frequency of substitutions is greater than would have occurred by random chance.

Zero score – frequency is equal to that expected by chance.

Negative score – frequency is less than would have occurred by random chance.

k vysvětlení

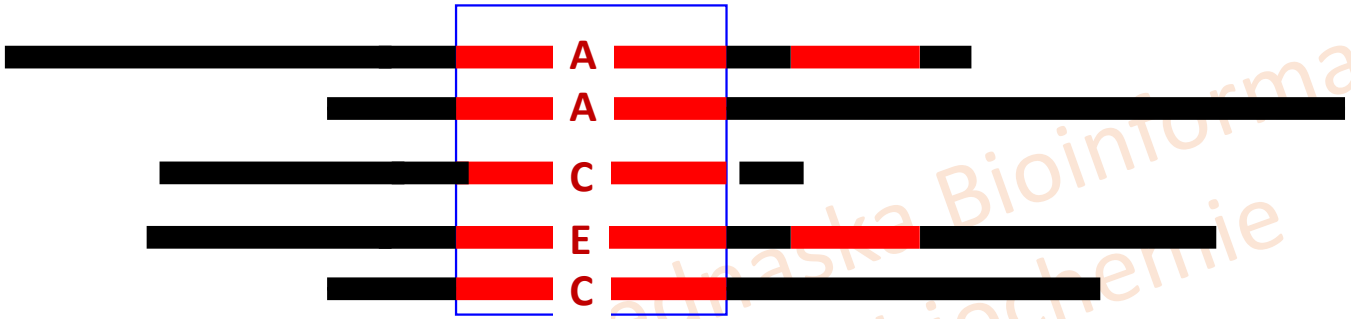
BLOSUM (Blocks Amino Acid Substitution)

- 1992, Henikoff and Henikoff
- database BLOCKS – používá koncept „bloků“ k identifikaci proteinových rodin
- **sekvenční motiv**
  - konzervovaný aminokyselinový úsek (conserved stretch of amino acids) spojený se specifickou funkcí proteinu
- **sekvenční blok**
  - spárované motivy ze stejné proteinové rodiny bez mezer
- BLOSUM matice byly vytvořeny na základě substitučních vzorů více než > 2 000 bloků (< 60 residuí) z 500 skupin proteinů

- nebere v potaz evoluci

k vysvětlení

- BLOSUM62 – znamená, že ke konstrukci matice byly použity proteiny s průměrnou identitou 62%.




```
A - C = 4
A - E = 2
C - E = 2
A - A = 1
C - C = 1
```

- výskyt každého AMK páru v každém sloupci každého bloku je sečten
- čísla získána ze všech bloků slouží pro výpočet BLOSUM matricí

k vysvětlení

Odlišné substituční matice jsou pro odlišné účely

Matrix	Best use	Similarity (%)*
Pam40	Short highly similar alignments	70-90
PAM160	Detecting members of a protein family	50-60
PAM250	Longer alingments of more divergent sequences	~30
BLOSUM90	Short highly similar alignments	70-90
BLOSUM80	Detecting members of a protein family	50-60
BLOSUM62	Most effective in finding all potential similarities	30-40
BLOSUM30	Longer alingments of more divergent sequences	<30

Číslování BLOSUM jde v obráceném pořadí oproti PAM (čím menší číslo, tím odlišnější sekvence byly použity)

- BLOSUM matice pracují obvykle lépe než PAM pro lokální vyhledávání podobností (Henikoff & Henikoff, 1993)
- Pro porovnání blízce příbuzných proteinů by se měla používat nižší číslo PAM a vyšší BLOSUM, pro vzdálenější vyšší číslo PAM a nižší BLOSUM
- Pro prohledávání databází je nejběžnější BLOSUM62

k vysvětlení

Mezery (gaps)

- Mezery umožňují alignment sekvencí, kdy v jedné z nich došlo k deleci. Zvyšují však také možnost alignmentu náhodných sekvencí. Jejich přítomnost je proto vždy „**penalizována**“, často více než substituce.

- Čím nižší je penalizace mezer, tím lepší (dokonalejší) bude alignment, ovšem **z biologického hlediska může jít o nesmysl**.

- Jednotlivé programy obvykle penalizují **přítomnost mezery** (gap open) a také zvyšují penalizaci s **délkou mezery** (gap ext).

Krátká mezera:

```
ATCTTCAGTGTTTCCCCTGTTTTGCCC-ATTTAGTTCGCTC
|||||||||||||||||||||||||| ||||||||||||
ATCTTCAGTGTTTCCCCTGTTTTGCCCGATTTAGTTCGCTC
```

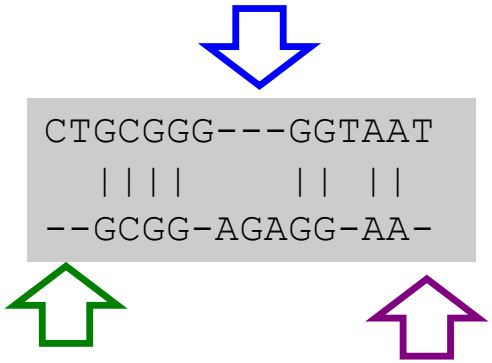
Dlouhá mezera:

```
ATCTTCAGTGTTTCCCCTGTTTTGCCC--------------------ATTTAGTTCGCTC
|||||||||||||||||||||||||                      ||||||||||||
ATCTTCAGTGTTTCCCCTGTTTTGCCCGCCCCCCCCCCCCCCCCCCCATTTAGTTCGCTC
```

Příčiny vzniku mezer:

- **Bodová mutace** (velmi častá příčina)
- Nepřesný crossover při meióze (inzerce nebo delece řetězce bází)
- DNA slippage během replikace (vzniká repetice – opakující se sekvence v řetězci)
- Inzerce retroviru
- Translokace DNA mezi chromozomy

Mezery nacházíme na **začátku** řetězce, **uprostřed** nebo na jeho **konci**.



Mezery (gaps)

- Mezery umožňují alignment sekvencí, kdy v jedné z nich došlo k deleci. Zvyšují však také možnost alignmentu náhodných sekvencí. Jejich přítomnost je proto vždy „**penalizována**“, často více než substituce.
- Čím nižší je penalizace mezer, tím lepší (dokonalejší) bude alignment, ovšem **z biologického hlediska může jít o nesmysl**.
- Jednotlivé programy obvykle penalizují **přítomnost mezery** (gap open) a také zvyšují penalizaci s **délkou mezery** (gap ext).

Krátká mezera:

```
ATCTTCAGTGTTTCCCCTGTTTTGCCC-ATTTAGTTCGCTC
|||||||||||||||||||||||||| ||||||||||||
ATCTTCAGTGTTTCCCCTGTTTTGCCCGATTTAGTTCGCTC
```

Dlouhá mezera:

```
ATCTTCAGTGTTTCCCCTGTTTTGCCC--------------------ATTTAGTTCGCTC
|||||||||||||||||||||||||                      ||||||||||||
ATCTTCAGTGTTTCCCCTGTTTTGCCCGCCCCCCCCCCCCCCCCCCCATTTAGTTCGCTC
```

Vysoká penalizace mezer:

Hledání sekvencí velmi striktně zaměřených na podobnost s hledanou sekvencí - najde oblasti velmi příbuzných sekvencí

Nízká penalizace mezer:

Hledání podobností mezi sekvencemi vzdáleně příbuzných.

Skóre

Každé dvojici sekvencí je ve výsledku přiřazeno číslo – **skóre**, které **určuje míru** jejich **podobnosti**

1. identita (identity)
2. podobnost (similarity)
3. mezery (gaps)

Čím vyšší je skóre, tím vyšší je podobnost.
Podle použité matice může být skóre i záporné.

AAEECCDDEEF
AADDKKKEFGG

Ve chvíli, kdy zafixujeme pozici dvou sekvencí, pak můžeme snadno vypočítat skóre pro dané přiložení (příklad BLOSUM 62):

```
A A E E C C D D E E F
A A D D K K K E F G G
4+4+2+2-3-3-1+2-3-2-3        = -1
```

```
A A E E C C D D - - E E F
A A - - - - D D K K K E F G G
4+4         +6+6    +1+5+6        = 32
```

```
A A E E C C D D - - E E F
A A - - - - D D K K K E F G G
   -10-1-1-1   -10-1            = -24
```

Celkové skóre 32 – 24 = 8

```
A A E E C C D D E E F
A A - - - - D D K K K E F G G
4+4-10-1-1-1+6+6+1+1-3           = 6
```

Skóre

```
1 MAMUZDOST--STAROSTISHAMIZNOSTIRATOLESTI  37
  ||| .|||| |||| .|  | |. |||
1 MAMRADOSTZESTAR-------O-Z----I--TNO-STI  24
```

Gap_penalty: 1

Extend_penalty: 2

Score: 55

```
1 MAMUZDOST--STAROSTISHAMIZNOSTIRATOLESTI   37
  ||| .|||| ||||||...:.               |||
1 MAMRADOSTZESTAROZITNO---------------STI   24
```

Gap_penalty: 12

Extend_penalty: 2

Score: 4

Alignment DNA

U nukleových kyseliny nemá smysl posuzovat podobnost:

Frekvence mutací všech bází je obdobná, takže nejjednodušší hodnocení je: shoda (1), neshoda (0)

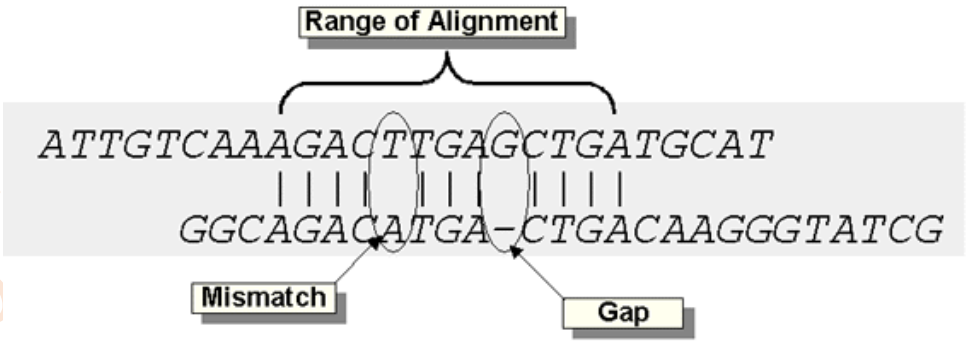
tím se nerozliší výborný alignment krátkých a mizerný dlouhých sekvencí:
proto **penalizace záměn,** např.:

match score +5

mismatch score -4

gap penalty: opening -10, extending -2

Každé dvojici sekvencí je ve výsledku přiřazeno číslo – **skóre**, které **určuje míru** jejich **podobnosti**



$S = \sum(\text{identities, mismatches}) - \sum(\text{gap penalties})$

$\text{Score} = \text{Max}(S)$

Čím vyšší je skóre, tím vyšší je podobnost.
Podle použité matice může být skóre i záporné.

Přesto:

Jak statisticky významné je skóre?

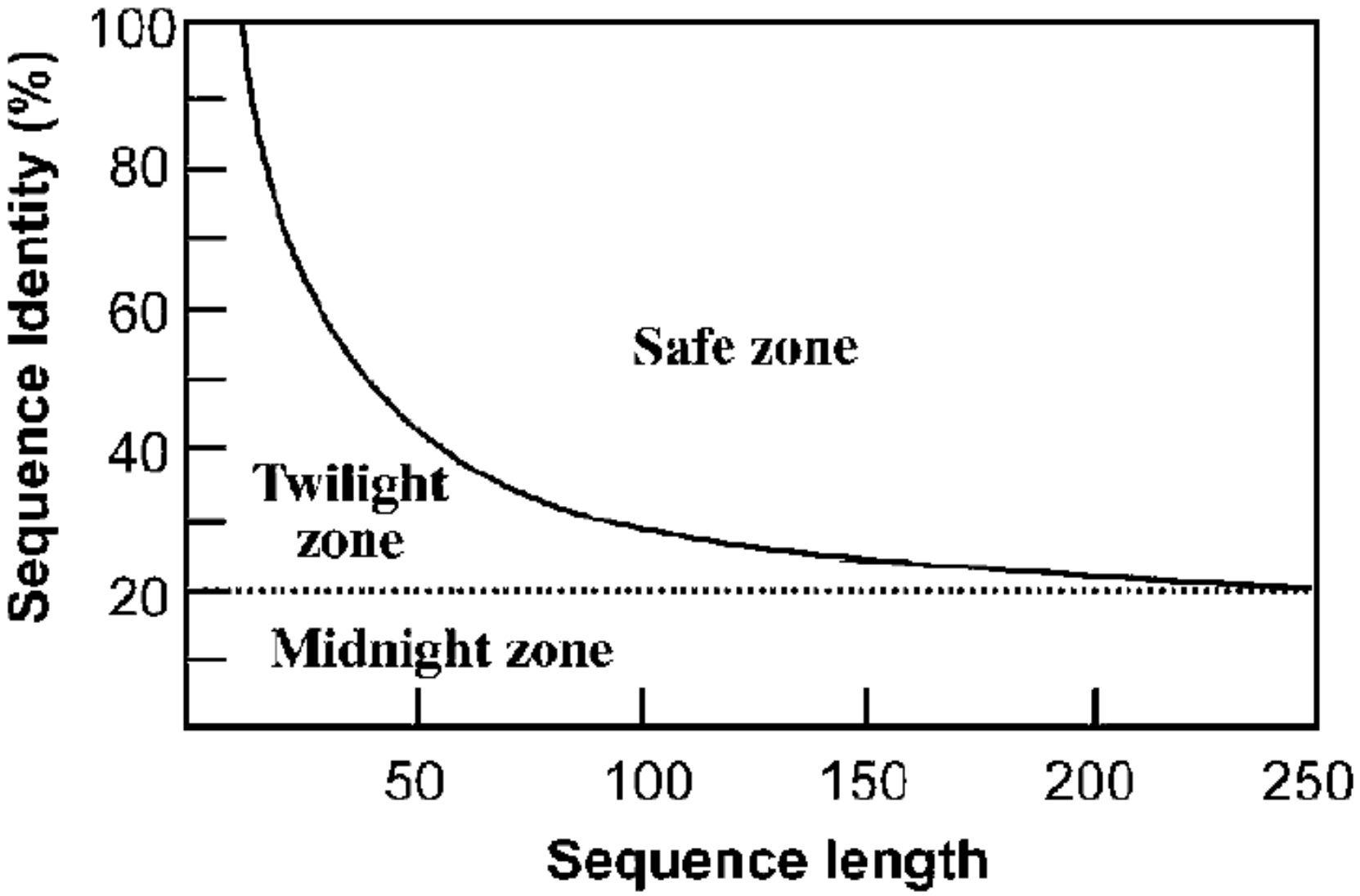
Pokud je podobnost dostatečně významná lze usuzovat na společné evoluční vztahy . Ale co je DOSTATEČNĚ?

závisí na typu sekvence a její délce

- Pravděpodobnost, že dvě rezidua v nepříbuzných sekvencích jsou identická je:

  25% v NA, 5% v proteinech

- Vliv délky sekvence
  - čím kratší sekvence, tím větší je šance, že alignment je dán náhodnou shodou. Čím delší, tím je méně pravděpodobné, že je stejná úroveň podobnosti výsledkem náhody.
  - kratší sekvence vyžadují vyšší cut-off pro zjištění příbuznosti než u delších sekvencí



Essential bioinformatics, Xiong

Multiple sequence alignment - MSA

(mnohonásobné sekvenční přiložení)

Multiple alignment slouží k:

- Nalezení „diagnostického vzoru“ (diagnostic patterns) na jehož základě jsou **charakterizovány proteinové rodiny**
- Odhalení či dokázání **homologie** mezi novou sekvencí a sekvencemi v databázích
- Určení vzájemné příbuznosti sekvencí v rámci skupiny – tvorba **fylogenetických stromů**
- **Predikci** sekundární a terciární **struktury** nových proteinů
- Navržení primerů (oligonukleotidů) pro PCR

Multiple sequence alignment - MSA

(mnohonásobné přiložení)

- Dynamické programování (dynamic programming) – rozšíření pairwise alignmentu - náročné na paměť a čas, nevhodné pro více než 3-4 sekvence (n=rozměrný prostor)

- **Progresivní alignment** (progressive sequence alignment) – nejčastěji používaný k vytvoření alignmentu; využívá fylogenetické informace – hierarchický, nejdříve identifikuje nejpodobnější sekvence a následně inkorporuje ostatní

- Iterativní alignment (iterative sequence alignment) – odstraňuje problémy progresivního alignmentu, který je závislý na prvotním přiložení nejpodobnějších sekvencí pomocí opakování alignmentu pro podskupiny sekvencí následující po globálním alignmentu

- Hledání motivů – nalezení částí konzervovaných sekvenčních motivů pomocí globálního přiložení a následně „hodnocení“ těchto úseků nezávisle na celé sekvenci

Výstup

```
CLUSTAL 2.0.10 multiple sequence alignment


PAIIL           ------------------------------------------------------------
RSIIL           ------------------------------------------------------------
CVIIL           ------------------------------------------------------------
BCLB            ---LVEKLPQYDVFVDIATIPYSFDVGSWQNKVKTDAAGEVVACTVTWAGAPGVLPGAAA
BCLC            AIATNQGVVADGCFTYSSKVPESTGRMPFTLVATIDVGSGVTFVKGQWKSVRGSAMHIDS
BCLA            ------------------------------------------------------------
BCLD            LRETALALRAEVSVLFIRFALKDAGIVAPIELEVRDAATAVPDADDLLHPSCRPLKDHYW


PAIIL           ----------------------------------------------------------ATQGVFT
RSIIL           ----------------------------------------------------------AQQGVFT
CVIIL           ----------------------------------------------------------AQQGVFT
BCLB            KFGVGAVVN----------------YFSKATPQPVQPAPVP--------TGGGERDGIFT
BCLC            YASLSAIWG----------------TAAPSSQGSGNQGAETGGTGAGNIGGGGERDGTFN
BCLA            ------------------------------------ADSQT--------SSNRAGEFS
BCLD            RSDVLAAGATTCTADFAVCDRDGTVSGYFRWETSIEIAGSQPDTKQPGFKPSSDRNGNFS
                                                                  * *.


PAIIL           LPANTRFGVTAFANSSGTQTVNVLVNNETA--ATFSGQSTNNAVIGTQVLNSGSSGKVQV
RSIIL           LPANTSFGVTAFANAANTQTIQVLVDNVVK--ATFTGSGTSDKLLGSQVLNSGS-GAIKI
CVIIL           LPARINFGVTVLVNSAATQHVEIFVDNEPR--AAFSGVGTGDNNLGTKVINSGS-GNVRV
BCLB            LPPNIAFGVTALVNSSAPQTIEVFVDDNPKPAATFQGAGTQDANLNTQIVNSGK-GKVRV
BCLC            LPPHIKFGVTALTHAANDQTIDIYIDDDPKPAATFKGAGAQDQNLGTKVLDSGN-GRVRV
BCLA            IPPNTDFRAIFFANAAEQQHIKLFIGDSQEPAAYHKLTTRDGPRE--ATLNSGN-GKIRF
BCLD            LPPNTAFKAIFYANAADRQDLKLFIDDAPEPAATFVGNSEDGVRL--FTLNSKG-GKIRI
                :*..  * .   .::: * :.: :.:    * .     .      ::*   * ::.
```

MSA – programové balíky

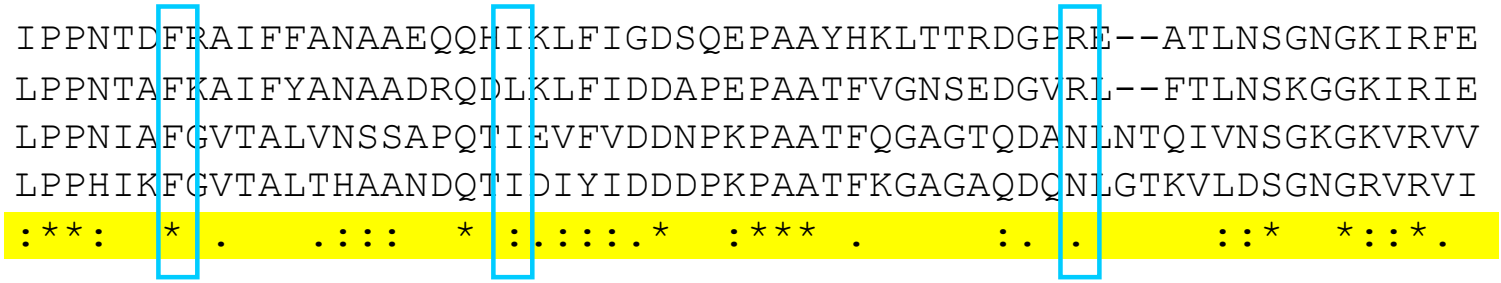
Za posledních 25 let vzniklo přes 50 MSA programových balíků
(Wallace, I. M., O'Sullivan, O., Higgins, D. G. and Notredame, C. (2006). M-Coffee: combining multiple sequence alignment methods with T-Coffee. Nucleic Acids Res. **34**, 1692-1699.)

- Clustal W (Thompson et al., 1994)
- Clustal X (Thompson et al., 1997)
- Dialign2 (Morgenstern, 1999)
- T-Coffee (Notredame et al., 2000)
- MAFFT (Katoh et al., 2002)
- MUSCLE (Edgar, 2004)
- Kalign (Lassmann, 2005)

* - identické residuum ve všech sekvencích

: - silně konzervovaný sloupec

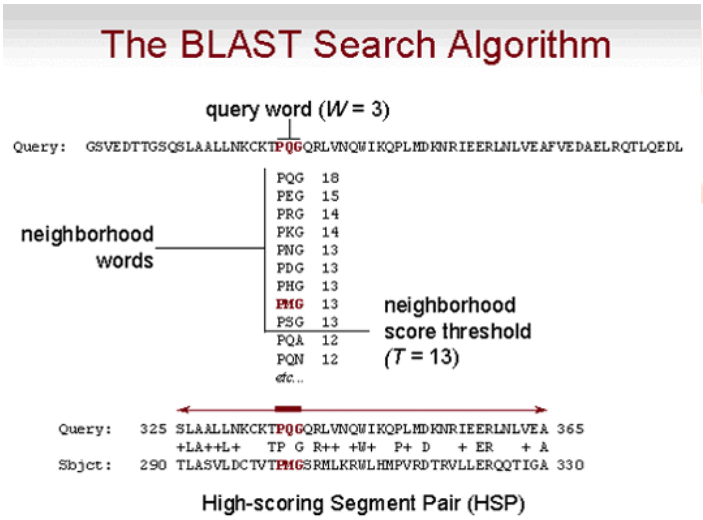
. - slabě konzervovaný sloupec



BLAST algoritmus

BLAST (Basic Local Alignment Search Tool)

Heuristický algoritmus jehož základem je **hledání slov** (několikapísmenných sekvencí), s dostatečnou podobností (poskytují dostatečně vysoké skóre v substituční matici).



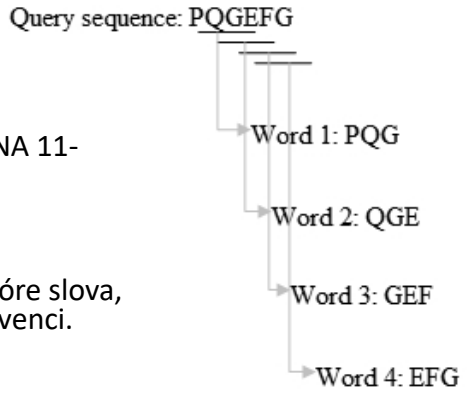
- **Tvorba k-písmenných slov ze vstupní sekvence**
  pro proteiny typicky 3-písmenných (v případě DNA 11-písmenných)

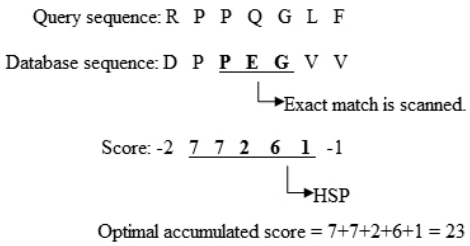
- **Porovnání slov na základě substituční matice**
  algoritmus BLAST hledá na základě vloženého skóre slova, která jsou podobná každému slovu v zadané sekvenci. Vyhovující slova jsou následně uspořádána.

- **Prohledání databázových sekvencí**
  Je hledána shoda s nalezenými vysoce podobnými slovy.

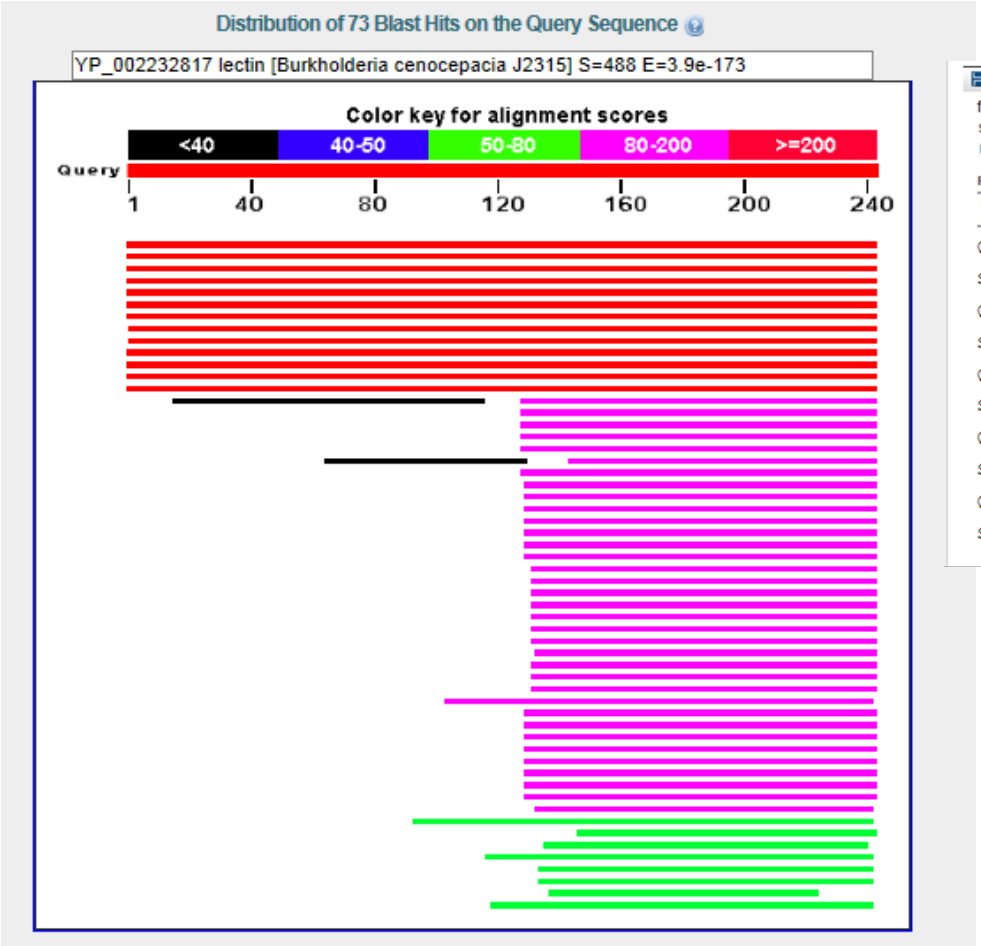
- **Rozšíření slov na segmenty**
  Přesné shody slov s databázovými sekvencemi jsou rozšiřovány oběma směry. To pokračuje dokud skóre pro tuto dvojici sekvencí je dostatečně vysoké.

Novější verze BLASTu (BLAST2) má mj. níže nastavenu hladinu pro hledání podobných slov, což rozšiřuje možnost nalezení vzdálenějších homologů.





Výstup z BLASTu



Download GenPept Graphics

fucose-binding lectin II [Burkholderia multivorans ATCC BAA-247]

Sequence ID: ref|ZP_15916739.1| Length: 274 Number of Matches: 1

See 1 more title(s)

Range 1: 31 to 274 GenPept Graphics

Score	Expect	Method
443 bits(1140)	4e-155	Compositional matrix adjus

```
Query  2    QPFTHDDLYALLQLAGNDATAVC
            QPFTHDDLYALLQLAGNDA AVC
Sbjct  31   QPFTHDDLYALLQLAGNDAKAVC

Query  62   SFDVGSWQNKVKTDAAGEVVAC1
            SFDVGSWQNKVKTDAAG+VVAC1
Sbjct  91   SFDVGSWQNKVKTDAAGQVVAC1

Query  120  PAPVPTGGGERDGIFTLPPNIAI
             P     GGGERDG+F LPPNIAI
Sbjct  151  PDTATAGGGERDGVFNLPPNIAI

Query  180  LNTQIVNSGKGKVRVVVTANGKI
             LNTQIVNSG GKVRVVVT NGKI
Sbjct  211  LNTQIVNSGNGKVRVVVTVNGKI

Query  240  WPLG  243
             WPLG
Sbjct  271  WPLG  274
```

Download GenPept Graphics

sugar-binding lectin protein [Ralstonia solanacearum PSI07]

Sequence ID: ref|YP_003750856.1| Length: 114 Number of Matches: 1

See 3 more title(s)

Range 1: 3 to 114 GenPept Graphics Next Match Previous Match

Score	Expect	Method	Identities	Positives	Gaps
124 bits(312)	2e-32	Compositional matrix adjust.	62/114(54%)	80/114(70%)	2/114(1%)

```
Query  130  RDGIFTLPPNIAFGVTALVNSSAPQTIEVFVDDNPKPAATFQGAGTQDANLNTQIVNSGK  189
            + G+FTLP N   FGVTA  N++   QTI+V VD+  K  ATF G+GT D  L +Q++NSG+
Sbjct  3    QQGVFTLPANTNFGVTAFANAANTQTIKVLVDNVVK--ATFSGSGTSDKLLGSQVLNSGR  60

Query  190  GKVRVVVTANGKPSKIGSRQVDIFKKTYFGLVGSEDGGDGDYNDGIAILNWPLG  243
            G V++ V+ NGKPS + S Q  +  K  F +VGSED  D DYNDGIA+LNWPLG
Sbjct  61   GAVQIQVSVNGKPSDLVSNQTILANKLNFAMVGSEDSSDNDYNDGIAVLNWPLG  114
```

Download GenPept Graphics

fucose-binding lectin PA-IIL [Pseudomonas aeruginosa ATCC 25324]

Sequence ID: ref|ZP_15618368.1| Length: 115 Number of Matches: 1

See 1 more title(s)

Range 1: 5 to 115 GenPept Graphics Next Match Previous Match

Score	Expect	Method	Identities	Positives	Gaps
117 bits(294)	7e-30	Compositional matrix adjust.	61/113(54%)	77/113(68%)	3/113(2%)

```
Query  132  GIFTLPPNIAFGVTALVNSSAPQTIEVFVDDNPKPAATFQGAGTQDANLNTQIVNSGK-G  190
            G+FTLP N  FGVTA  NSS  QT+ V V  N + AATF G  T +A + TQ++NSG  G
Sbjct  5    GVFTLPANTQFGVTAFANSSGTQTVNVLV--NNETAATFSGQSTNNAVIGTQVLNSGSSG  62

Query  191  KVRVVVTANGKPSKIGSRQVDIFKKTYFGLVGSEDGGDGDYNDGIAILNWPLG  243
            KV+V V+ NG+PS + S QV +  +  F LVGSEDG D DYND + ++NWPLG
Sbjct  63   KVQVQVSVNGRPSDLVSAQVILTNELNFALVGSEDGTDNDYNDAVVVINWPLG  115
```

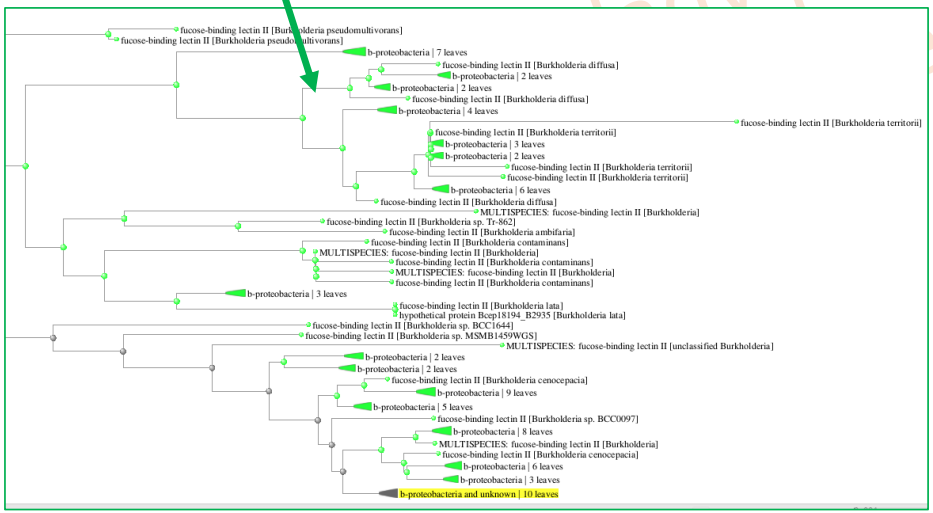
Search Parameters	
Program	blastp
Word size	6
Expect value	0.05
Hitlist size	100
Gapcosts	11,1
Matrix	BLOSUM62
Filter string	F
Genetic Code	1
Window Size	40
Threshold	21
Composition-based stats	2

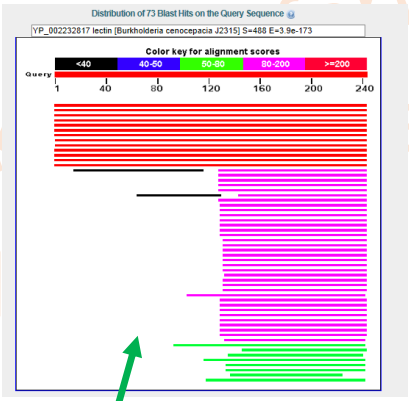
Database	
Posted date	Oct 11, 2020 12:42 AM
Number of letters	116,124,398,076
Number of sequences	322,194,847
Entrez query	None

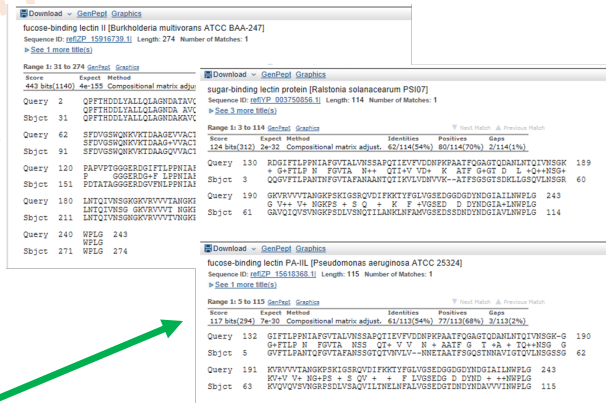
‹ Edit Search Save Search Search Summary

Job Title	**Protein Sequence**
RID	SB9C6PUH014 *Search expires on 10-14 21:03 pm* Download All
Program	BLASTP Citation
Database	nr See details
Query ID	lcl|Query_36959
Description	None
Molecule type	amino acid
Query Length	244
Other reports	Distance tree of results Multiple alignment MSA viewer

Title: All non-redundant GenBank CDS translations+PDB+SwissProt+PIR+PRF excluding environmental samples from WGS projects
Molecule Type: Protein
Update date: 2020/10/12
Number of sequences: 321587928







Descriptions | Graphic Summary | Alignments | Taxonomy

Sequences producing significant alignments Download Manage Columns Show 100

select all *100 sequences selected* GenPept Graphics Distance tree of results Multiple alignment

Description	Max Score	Total Score	Query Cover	E value	Per. Ident	Accession
fucose-binding lectin II [Burkholderia cenocepacia]	490	490	100%	6e-175	100.00%	WP_006490839.1
fucose-binding lectin II [Burkholderia cenocepacia]	488	488	100%	3e-174	99.59%	WP_077181202.1
fucose-binding lectin II [Burkholderia cenocepacia]	488	488	100%	4e-174	99.59%	WP_060264297.1
fucose-binding lectin II [Burkholderia cenocepacia]	488	488	100%	5e-174	99.59%	WP_069353106.1
photopexin A [Burkholderia cenocepacia]	488	488	100%	7e-174	99.59%	WP_006494487.1
MULTISPECIES: fucose-binding lectin II [Burkholderia]	488	488	100%	8e-174	99.18%	WP_011548596.1

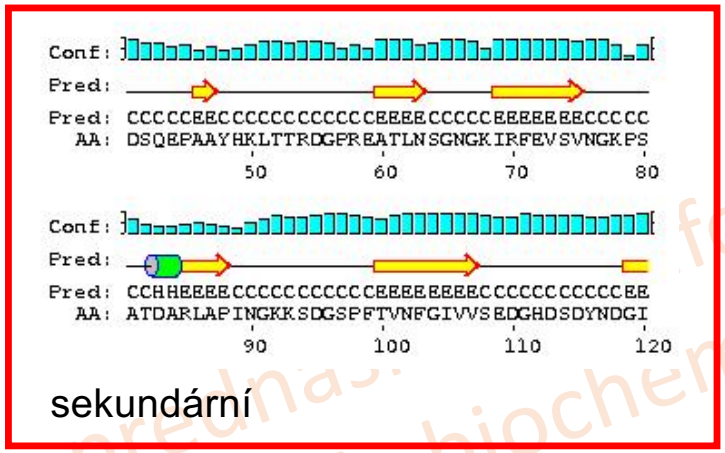
Osnova

- **Úvod do bioinformatiky**
  - Definice, molekulárně biologická data, databáze
  - Rozdělení databází, bioinformatická centra
- **Manipulace se sekvencemi**
  - Sekvence biomakromolekul, aminokyseliny, báze, alignment
  - Význam alignmentu, přiložení páru sekvencí a vícenásobné přiložení
- **Predikce struktury proteinů**
  - Predikce 2-D struktury proteinů, predikce 3-D struktury proteinů
  - Threading, Homology modeling, *Ab initio*
- **Predikce genů**
  - Predikce genů u prokaryot a eukaryot, predikční nástroje a postupy

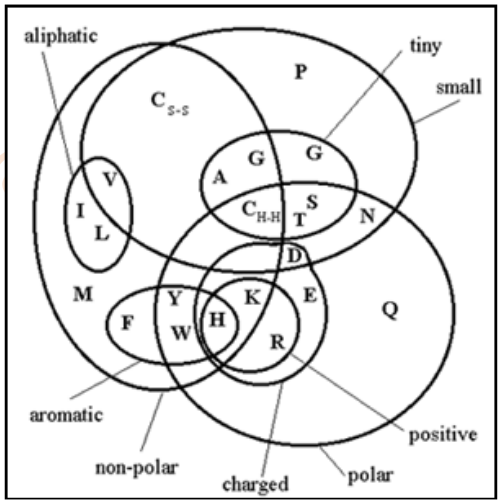
Predikce struktury proteinů

ADSQTSSNRAGEFSIPPNTDFRAIF
FANAAEQQHIKLFIGDSQEPAAYHK
LTTRDGPREATLNSGNGKIRFEVSV
NGKPSATDARLAPINGKKSDGSPF
TVNFGIVVSEDGHDSDYNDGIVVL
QWPIG

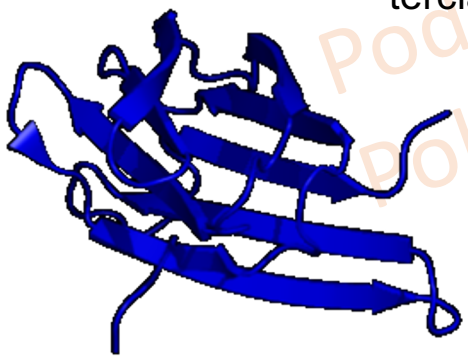
primární
(sekvence)



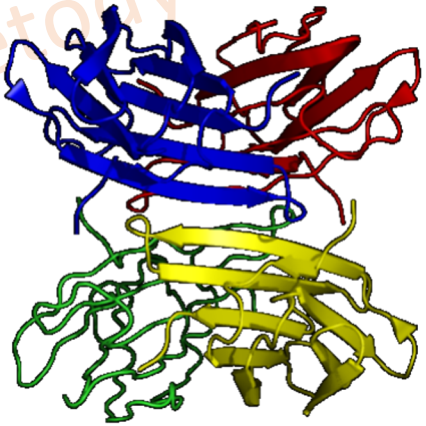
sekundární



terciární



kvartérní



Aminokyseliny s podobnými vlastnostmi mohou plnit v proteinu stejné funkce – bývají vzájemně zastupitelné

Predikce 2-D struktury proteinů

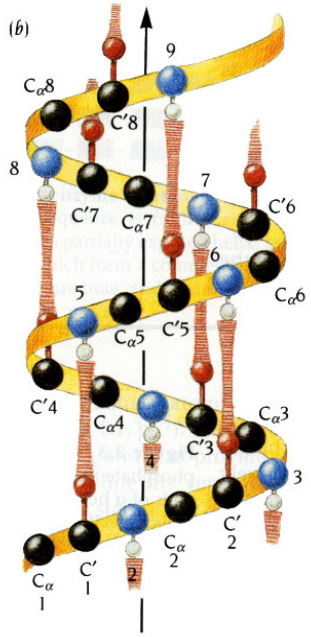
- **Stabilní** konformace **polypeptidového** řetězce.
- Důležité pro udržení proteinové 3-D struktury.
- Cca 50 % aa residuí je součástí **α-helixů** nebo **β-skládaných listů.**
- Predikce sekundárních struktur znamená **předpověď** zda residuum spadá mezi H (helix), E (list) nebo C (smyčka).
- Důležité pro klasifikaci proteinů.
- Separace domén a funkčních motivů.
- **Sekundární struktury** jsou mnohem konzervovanější než aminokyselinová sekvence.
- Předpověď sekundárních struktur předchází obvykle jako **mezikrok** při předpovědi terciární struktury při threadingových metodách.

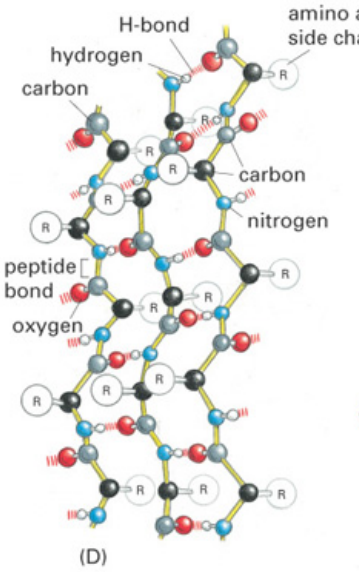
Predikce 2-D struktury proteinů

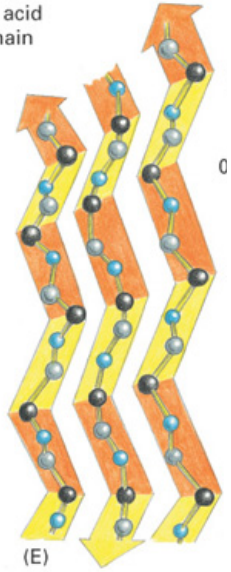
- Rozlišujeme tři základní typy
  - **H** – helix
  - **E** – β-list
  - **C/(-)** – smyčka/náhodné klubko (coil) – někdy jsou rozlišovány tyto dvě varianty

- S dobrou přesností lze určit helix (jejich tvorba je určena interakcemi „krátkého“ dosahu), u β -listu (interakce „dlouhého“ dosahu) úspěšnost určení 2D struktury klesá.

- Některé programy přidávají i číslo vyjadřující pravděpodobnost pro daný AK zbytek (např. H 60% - znamená, že s 60% pravděpodobností se jedná o helix).



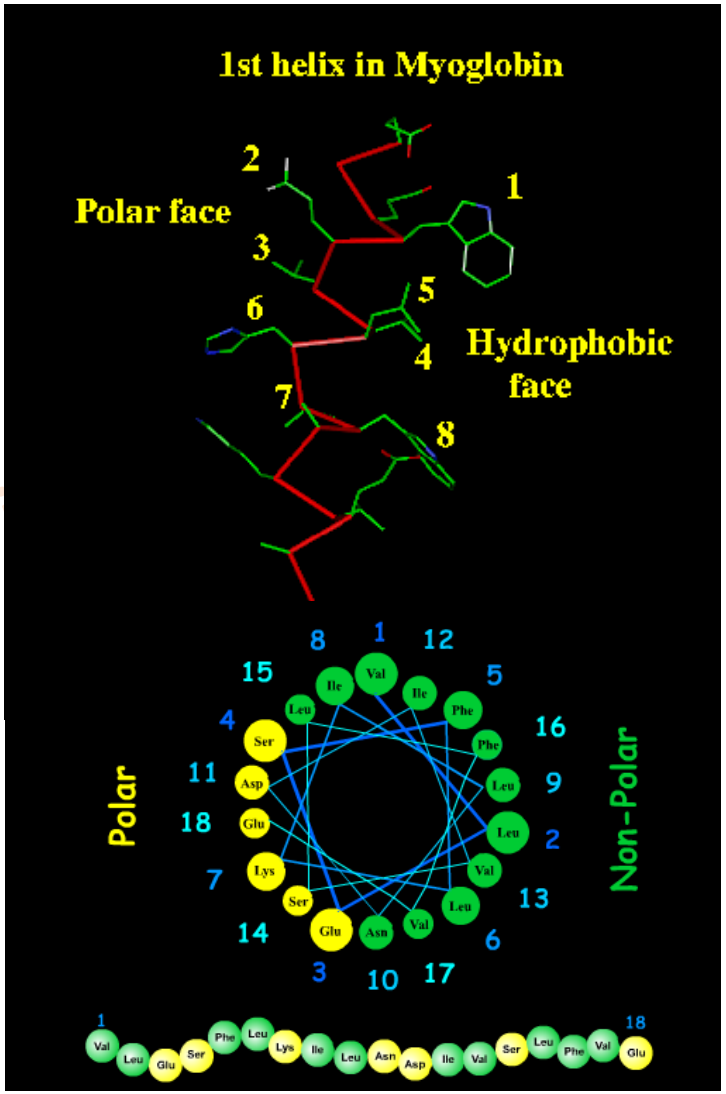




Typické znaky α -helix

Často je helix částečně exponovaný – tj. jedna strana je otočena dovnitř proteinu (hydrofobní), druhá ven (hydrofilní)

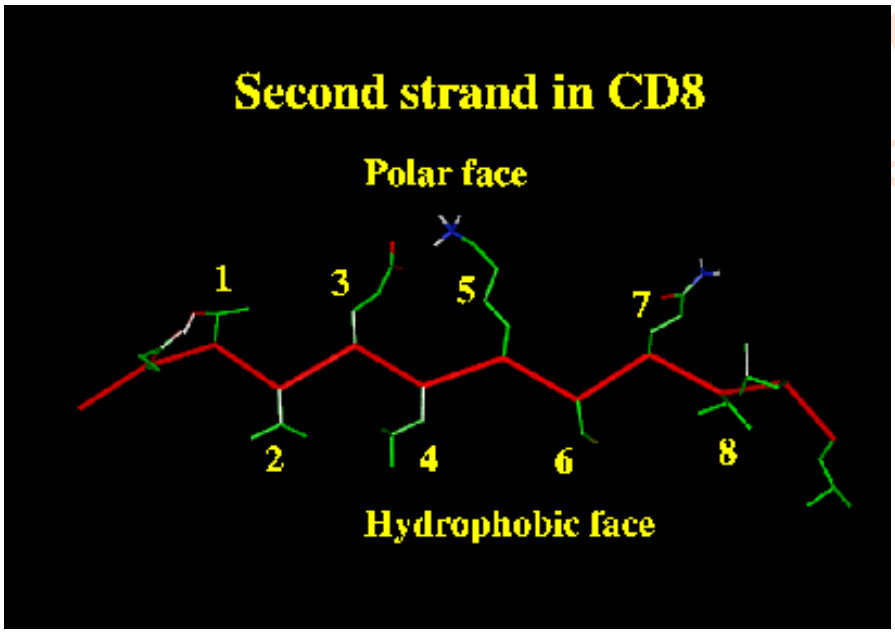
Potom pro 3.6 helix (α-helix) platí, že i, i+3, i+4 & i+7 -té reziduum míří na tutéž stranu. Jsou-li všechna hydrofobní či naopak hydrofilní = zřejmě α -helix

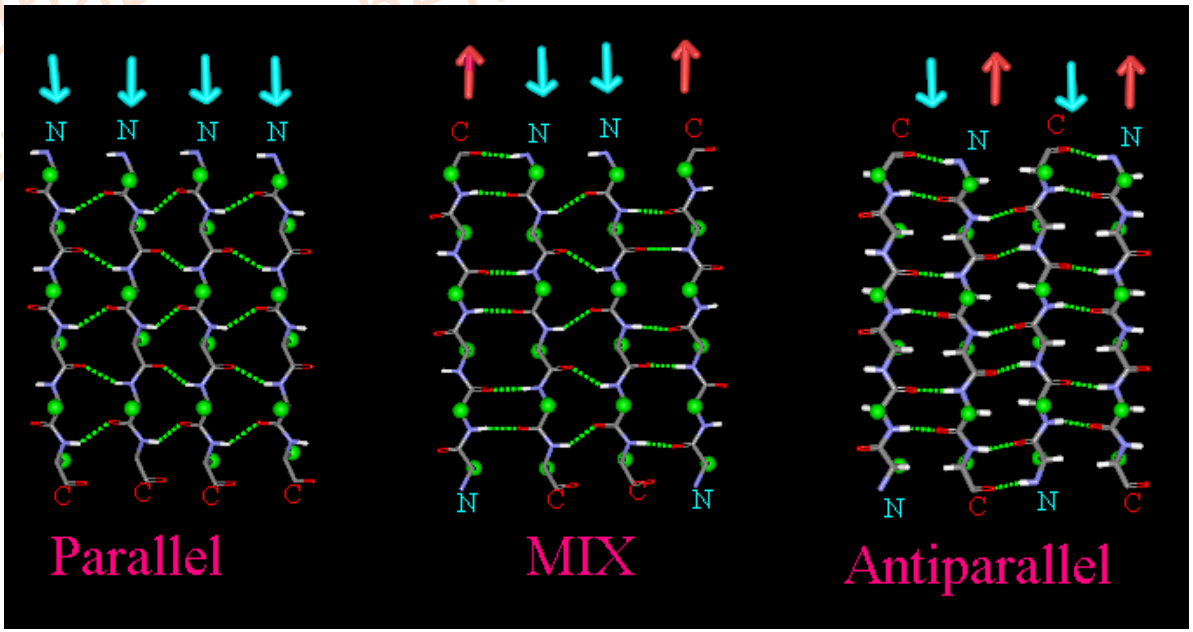


Typické znaky β –list (musí být stabilizován jinou částí polypeptidového řetězce!)

U β -listu se střídají boční řetězce po 180°

pro částečně zanořený β -list platí, že každé liché reziduum je polární, každé sudé nepolární, u plně zanořeného jsou všechna nepolární… tj. residua směřující na stejnou stranu by měla mít stejný charakter





Predikce 2-D struktury proteinů
Predikční algoritmy

1. **generace: *ab-initio***, vychází z fyzikálně-chemických vlastností a ze statistiky pro jednotlivá rezidua (Chou-Fasman, GOR (Garnier, Osguthorpe, Robson))
2. **generace: *plus incorporation of more local residue interactions***, zahrnovala i vliv nejbližších AK na zkoumané reziduum – předpověď max. 60% správnost, u β -listu do 40%
3. **generace: *homology-based models***, zahrnuje navíc multiple sequence alignment a využívá skutečnosti, že 2D struktura se zachovává déle než sekvenční podobnost – až 80% spolehlivost (závisí na metodě)

Plus využití skrytých Markovových modelů a neuronových sítí

1. Generace – *ab initio*

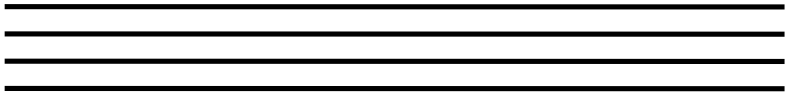
Relative Amino acid Propensity Values for Secondary Structure Elements Used in the Chou-Fasman Methods

Amino Acid	(α-Helix)	P (β-Strand)	P (Turn)
Alanine	1.42	0.83	0.66
Arginine	0.98	0.93	0.95
Asparagine	0.67	0.89	1.56
Aspartic acid	1.01	0.54	1.46
Cysteine	0.70	1.19	1.19
Glutamic acid	1.51	0.37	0.74
Glutamine	1.11	1.11	0.98
Glycine	0.57	0.75	1.56
Histidine	1.00	0.87	0.95
Isoleucine	1.08	1.60	0.47
Leucine	1.21	1.30	0.59
Lysine	1.14	0.74	1.01
Methionine	1.45	1.05	0.60
Phenylalanine	1.13	1.38	0.60
Proline	0.57	0.55	1.52
Serine	0.77	0.75	1.43
Threonine	0.83	1.19	0.96
Tryptophan	0.83	1.19	0.96
Tyrosine	0.69	1.47	1.14
Valine	1.06	1.70	0.50

$$\frac{\frac{R_i(SS)}{R_t\,(SS)}}{\frac{\Sigma\, R_i}{\Sigma\, R_t}}$$

3. Generace - Homology-based methods

MSA



Predikce sekundárních struktur pro každou sekvenci

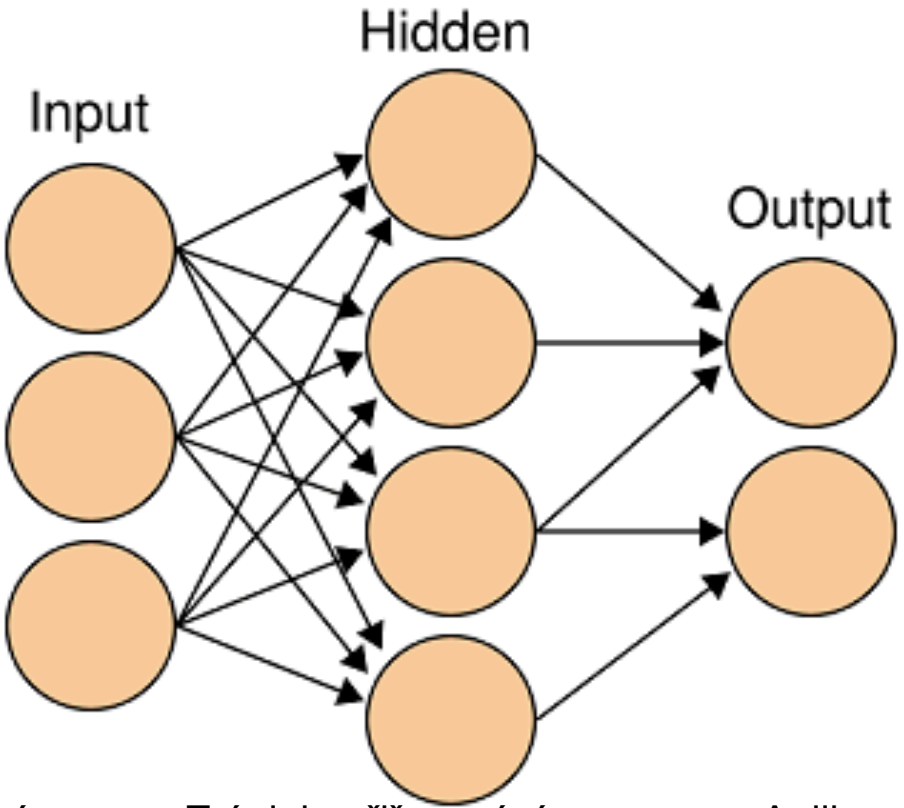
fitování předpovězené sekundární struktury do AA přiložení

HHHCHCCEEEECCHH
HHHHHCCEEEECCHH
ECCHHCEEEECCCEE
HHHHHCCCCEEECCH
HHHHCCCEEEECHHC

Konečná předpověď
Založená na konsensuální sekvenci

HHHHHCCEEEECCHH

3. Generace – neuronové sítě



Sekvence se známou sek. strukturou

Trénink, přiřazování Váh jednotlivým funkcím

Aplikace nalezených algoritmů na neznámou sekvenci

Predikce 2-D struktury proteinů
Programové balíky

- AGADIR - An algorithm to predict the helical content of peptides
- APSSP - Advanced Protein Secondary Structure Prediction Server
- GOR - Garnier et al, 1996
- HNN - Hierarchical Neural Network method (Guermeur, 1997)
- HTMSRAP - Helical TransMembrane Segment Rotational Angle Prediction
- Jpred - A consensus method for protein secondary structure prediction at University of Dundee
- JUFO - Protein secondary structure prediction from sequence (neural network)
- nnPredict - University of California at San Francisco (UCSF)
- Porter - University College Dublin
- PredictProtein - PHDsec, PHDacc, PHDhtm, PHDtopology, PHDthreader, MaxHom, EvalSec from Columbia University
- Prof - Cascaded Multiple Classifiers for Secondary Structure Prediction
- PSA - BioMolecular Engineering Research Center (BMERC) / Boston
- PSIpred - Various protein structure prediction methods at Brunel University
- SOPMA - Geourjon and Deléage, 1995
- SSpro - Secondary structure prediction using bidirectional recurrent neural networks at University of California
- DLP-SVM - Domain linker prediction using SVM at Tokyo University of Agriculture and Technology

Který program je však nejlepší???

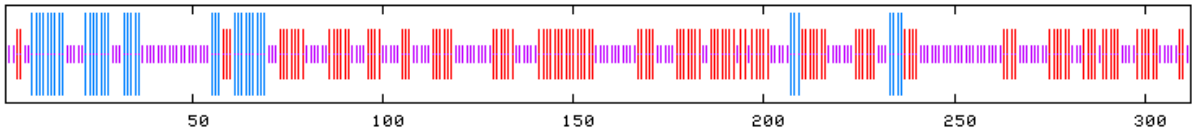
Soubor Úpravy Zobrazit Historie Záložky Nástroje Nápověda

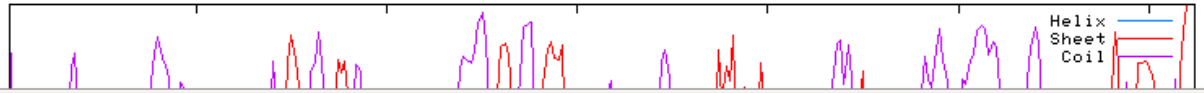
http://npsa-pbil.ibcp.fr/cgi-bin/secpred_gor4.pl

Nejnavštěvovanější Jak začít Přehled zpráv http://www.ncbi.nlm.... http://www.glycoscie... CHMI Radar Departme...

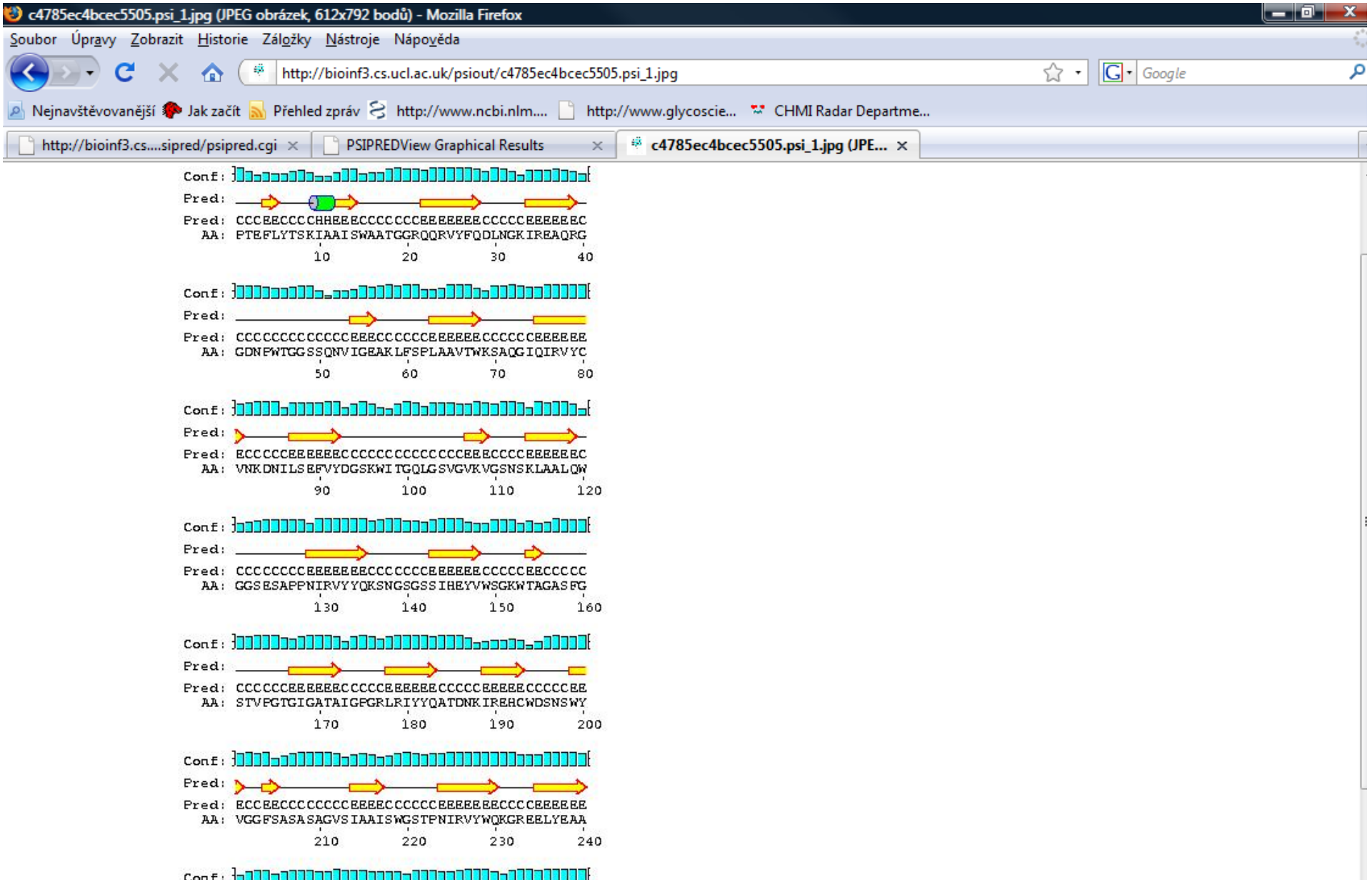
NPS@ : GOR4 SECONDARY STRU... | PSIPREDView Graphical Results | c4785ec4bcec5505.psi_1.jpg (JPEG o...

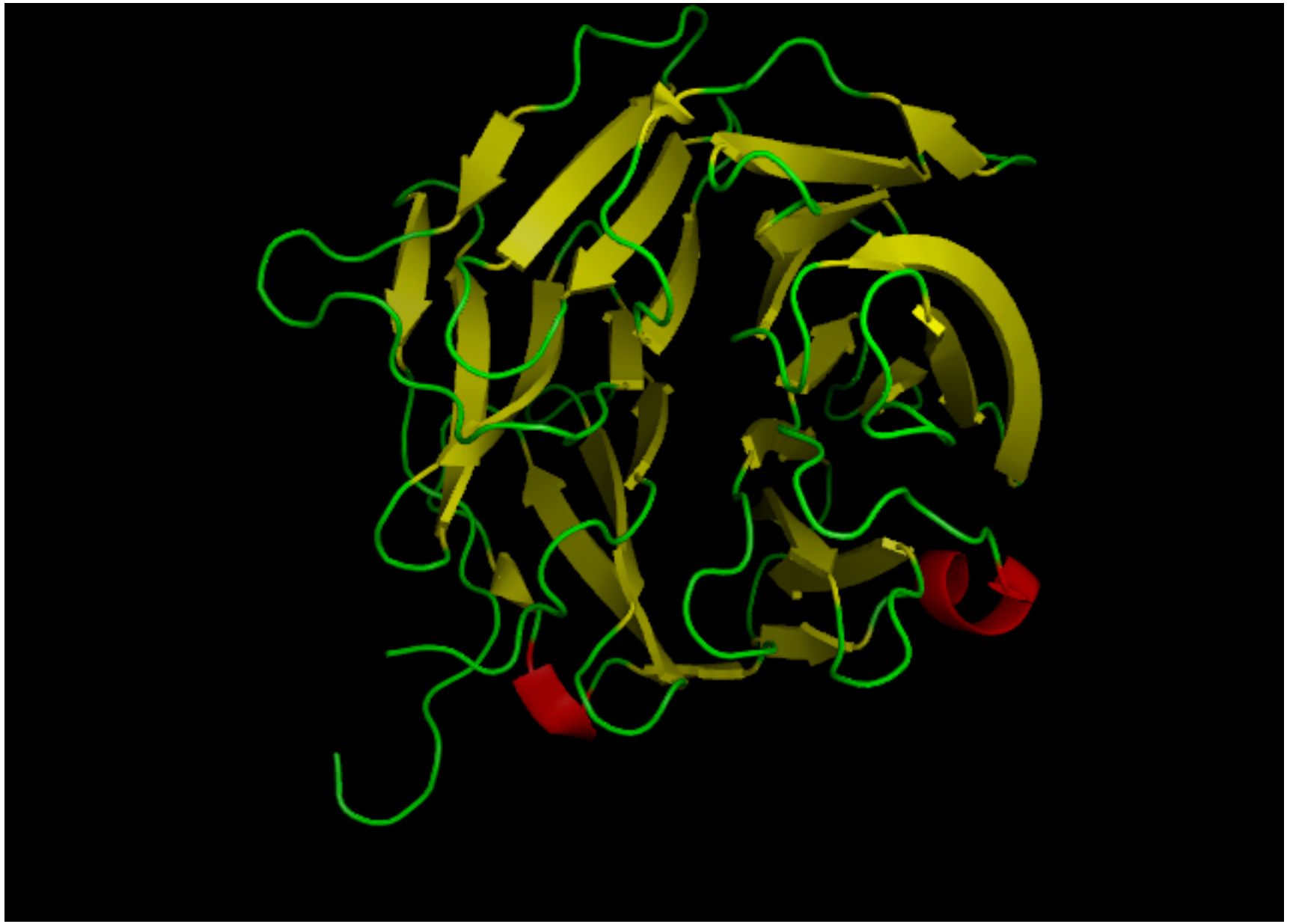
```
        10        20        30        40        50        60        70
         |         |         |         |         |         |         |
PTEFLYTSKIAAISWAATGGRQQRVYFQDLNGKIREAQRGGDNPWTGGSSQNVIGEAKLFSPLAAVTWKS
ccccceeechhhhhhhhhccccchhhhhhhccchhhhhccccccccccccccccccchhheeehhhhhhhhh
AQGIQIRVYCVNKDNILSEFVYDGSKWITGQLGSVGVKVGSNSKLAALQWGGSESAPPNIRVYYQKSNGS
ccceeeeeeecccccceeeeeecccceeeeccccceeeccccceeeeeecccccccccceeeeeeccccc
GSSIHEYVWSGKWTAGASFGSTVPGTGIGATAIGPGRLRIYYQATDNKIREHCWDSNSWYVGGFSASASA
ceeeeeeeeeeeeeeccccccccccceeeeecccccceeeeeeecceeeeeeeceeceeeeecccccchhh
GVSIAAISWGSTPNIRVYWQKGREELYEAAYGGSWNTPGQIKDASRPTPSLPDTFIAANSSGNIDISVFF
eeeeeeecccccccceeeeeeccchhhheeeecccccccccccccccccccccceeeecccccccceeeee
QASGVSLQQWQWISGKGWSIGAVVPTGTPAGW
eccceeeeceeeeecccceeeeeeccccceec

Sequence length :   312

GOR4 :
   Alpha helix      (Hh) :    40 is  12.82%
   3₁₀ helix        (Gg) :     0 is   0.00%
   Pi helix         (Ii) :     0 is   0.00%
   Beta bridge      (Bb) :     0 is   0.00%
   Extended strand  (Ee) :   122 is  39.10%
   Beta turn        (Tt) :     0 is   0.00%
   Bend region      (Ss) :     0 is   0.00%
   Random coil      (Cc) :   150 is  48.08%
   Ambigous states  (?)  :     0 is   0.00%
   Other states          :     0 is   0.00%
```





Hotovo


c4785ec4bcec5505.psi_1.jpg (JPEG obrázek, 612x792 bodů) - Mozilla Firefox
Soubor Úpravy Zobrazit Historie Záložky Nástroje Nápověda
http://bioinf3.cs.ucl.ac.uk/psiout/c4785ec4bcec5505.psi_1.jpg
Google
Nejnavštěvovanější Jak začít Přehled zpráv http://www.ncbi.nlm.... http://www.glycoscie... CHMI Radar Departme...
http://bioinf3.cs....sipred/psipred.cgi
PSIPREDView Graphical Results
c4785ec4bcec5505.psi_1.jpg (JPE...
Conf:
Pred:
Pred: CCCEECCCCHHEEECCCCCCCCEEEEEEECCCCCEEEEEEC
AA: PTEFLYTSKIAAISWAATGGRQQRVYFQDLNGKIREAQRG
10 20 30 40
Conf:
Pred:
Pred: CCCCCCCCCCCCCEEECCCCCCCEEEEEECCCCCCEEEEEE
AA: GDNPWTGGSSQNVIGEAKLFSPLAAVTWKSAQGIQIRVYC
50 60 70 80
Conf:
Pred:
Pred: ECCCCCEEEEEECCCCCCCCCCCCCEEECCCCCEEEEEEC
AA: VNKDNILSEFVYDGSKWITGQLGSVGVKVGSNSKLAALQW
90 100 110 120
Conf:
Pred:
Pred: CCCCCCCCEEEEEEECCCCCCCCEEEEEECCCCCEECCCCC
AA: GGSESAPPNIRVYYQKSNGSGSSIHEYVWSGKWTAGASFG
130 140 150 160
Conf:
Pred:
Pred: CCCCCCEEEEEECCCCCEEEEEECCCCCEEEEECCCCCEE
AA: STVPGTGIGATAIGPGRLRIYYQATDNKIREHCWDSNSWY
170 180 190 200
Conf:
Pred:
Pred: ECCEECCCCCCCCEEEECCCCCCEEEEEECCCCEEEEEE
AA: VGGFSASASAGVSIAAISWGSTPNIRVYWQKGREELYEAA
210 220 230 240
Conf:



GOR4 result for : UNK_78160

Abstract GOR secondary structure prediction method version IV, J. Garnier, J.-F. Gibrat, B. Robson, Methods in Enzymology,R.F. Doolittle Ed., vol 266, 540-553, (1996)

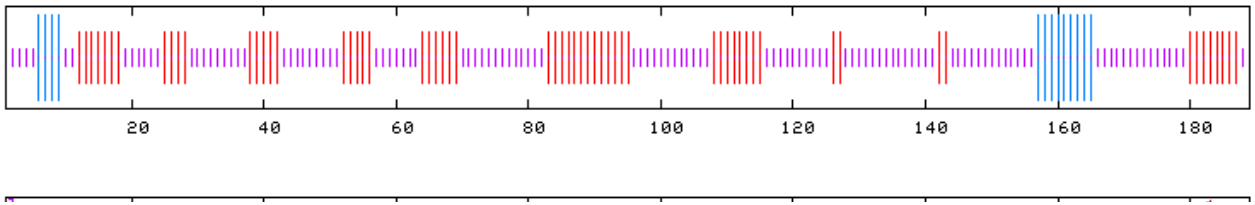
View GOR4 in: [AnTheProt (PC) , Download...] [HELP]

```
         10        20        30        40        50        60        70
          |         |         |         |         |         |         |
SDVDIEAQDAGQTLVQVISIPSGETWVAIQLPSQYRYFDFVFENVSPTSSGSVLVAQMAPQSGGVYGSNY
ccccchhhhcceeeeeeccccceeeeccccccccceeeeeccccccccceeeeeccccccceeeeee
SGSGWGNDLGGGGFYGYSEAKWMCLWPANRSGPSSKTGLYGTCKLMNLNQSSAVPSVTSNLFAPTAYKNE
ccccccccccccceeeeeeeeeeeeecccccccccccceeeeeeeecccccccccceecccccccccccc
PGYANVGGCCQKIRGLASSIQFAFALAGGNVPQNTDTFNGGTIKVYGWN
ccecccccccccccchhhhhhhhhcccccccccccccceeeeeeeec
```

Sequence length : 189

GOR4 :

Alpha helix	(Hh) :	13 is	6.88%
3_{10} helix	(Gg) :	0 is	0.00%
Pi helix	(Ii) :	0 is	0.00%
Beta bridge	(Bb) :	0 is	0.00%
Extended strand	(Ee) :	60 is	31.75%
Beta turn	(Tt) :	0 is	0.00%
Bend region	(Ss) :	0 is	0.00%
Random coil	(Cc) :	116 is	61.38%
Ambigous states	(?) :	0 is	0.00%
Other states	:	0 is	0.00%



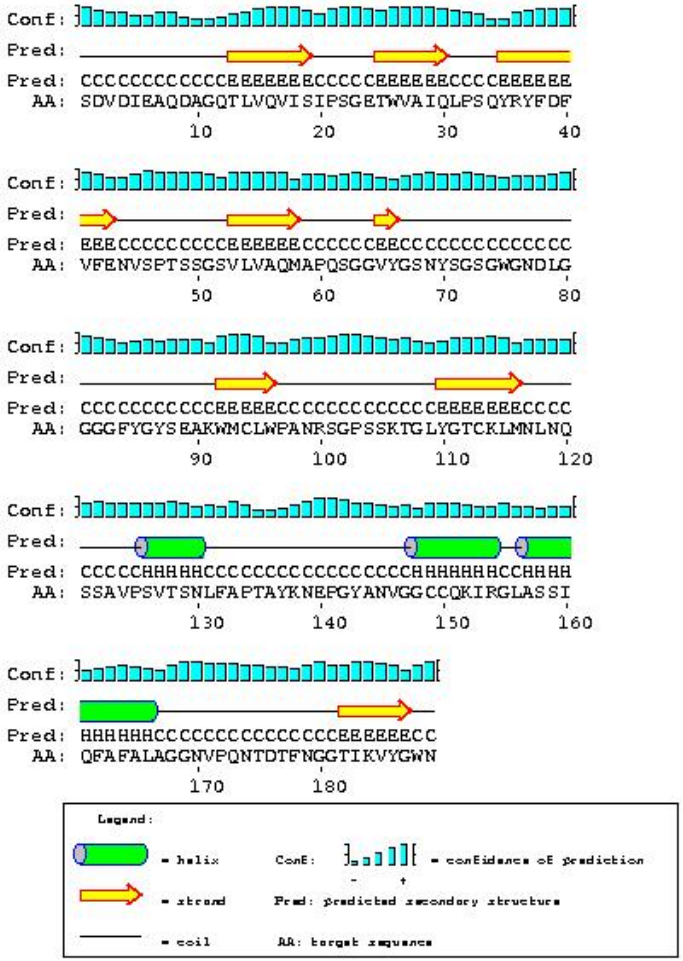
cc37c2fa6403027b.psi_1.jpg (JPEG obrázek, 612x792 bodů) - Mozilla Firefox

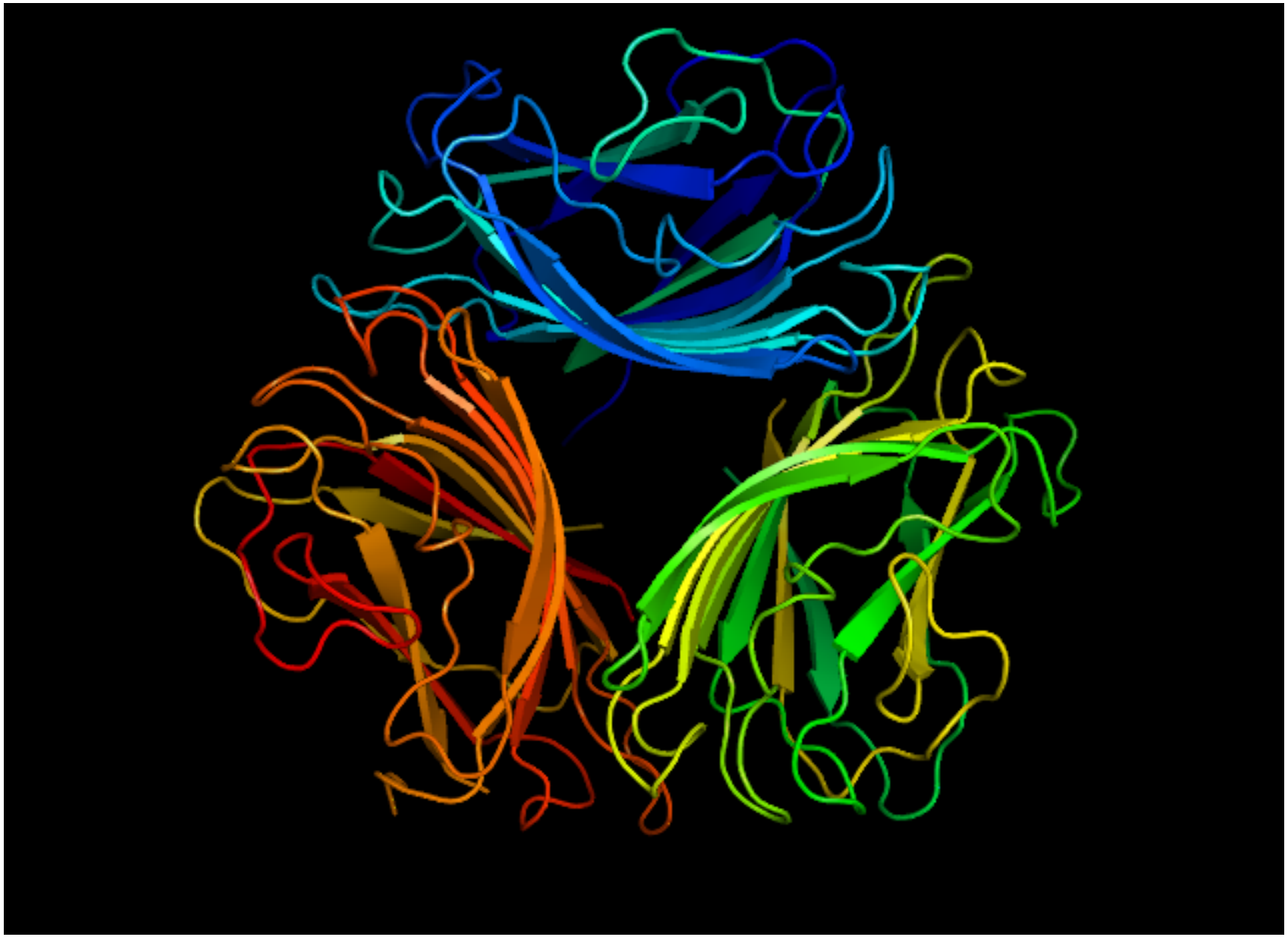
Soubor Úpravy Zobrazit Historie Záložky Nástroje Nápověda

http://bioinf3.cs.ucl.ac.uk/psiout/cc37c2fa6403027b.psi_1.jpg

Nejnavštěvovanější Jak začít Přehled zpráv http://www.ncbi.nlm.... http://www.glycoscie... CHMI Radar Departme...

http://www.igb.uci...r/sql_predict.cgi | cc37c2fa6403027b.psi_1.jpg (JPE...

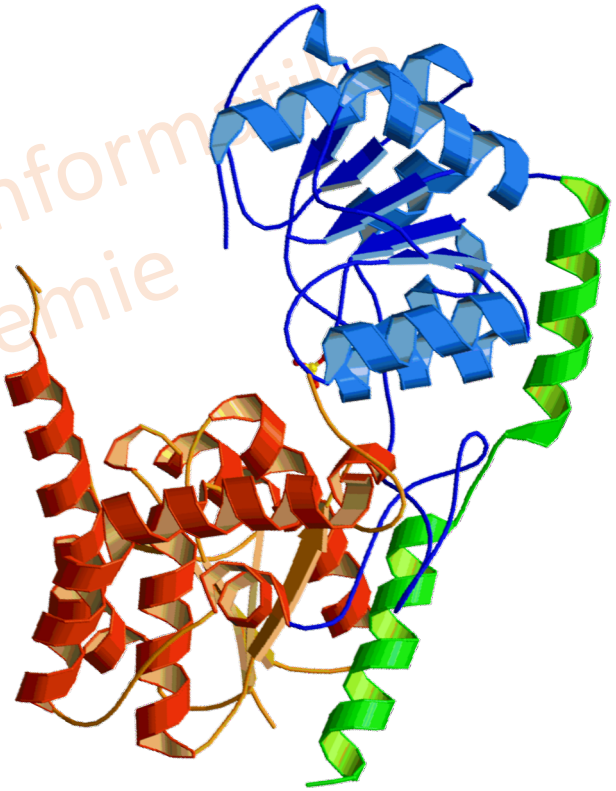




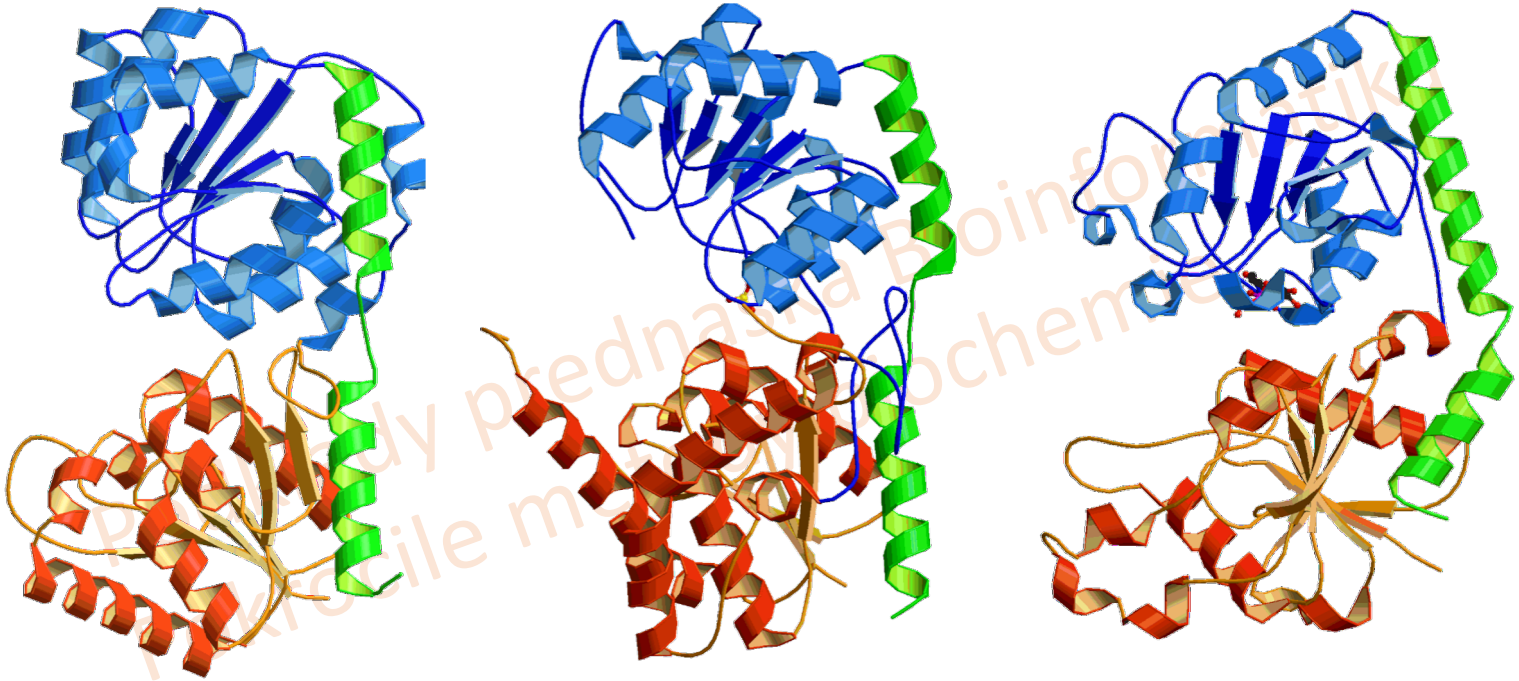
Predikce 3-D struktury/foldu proteinů

- Klasifikace proteinů
- Předpověď funkce
- Vytvoření modelu pro další studium

- **Threading - „navlékání“**
- **Homology modeling**
- ***Ab initio* metody**



Nadrodina s BGT foldem
(mezi rodinami nejsou významné sekvenční podobnosti)



MurG (β-GlcNAcT)
GT28
E. coli
Ha *et al*., 2000

GtfB (β-GlcT)
GT1
A. orientalis
Mulichak *et al*, 2001

BGT (β-GlcT)
n.c.
Phage T4
Vrielink *et al*, 1994

Predikce 3-D struktury/foldu proteinů - Threading

• **„Navlékání“** = rozpoznání a přiřazení proteinového foldu aminokyselinové sekvenci.

• sekvence je porovnávána s databází existujících **foldů (3D profilů)** a na jejich základě jsou konstruovány 3D- modely.

• 3D profil - každému reziduu v 3D struktuře je přiřazena environmentální proměnná (obsah polárních atomů v postranním řetězci, skrytá plocha, sekundární elementy, apod.) vycházející z předpokladu, že okolí rezidua je více konzervováno než aminokyselina samotná.

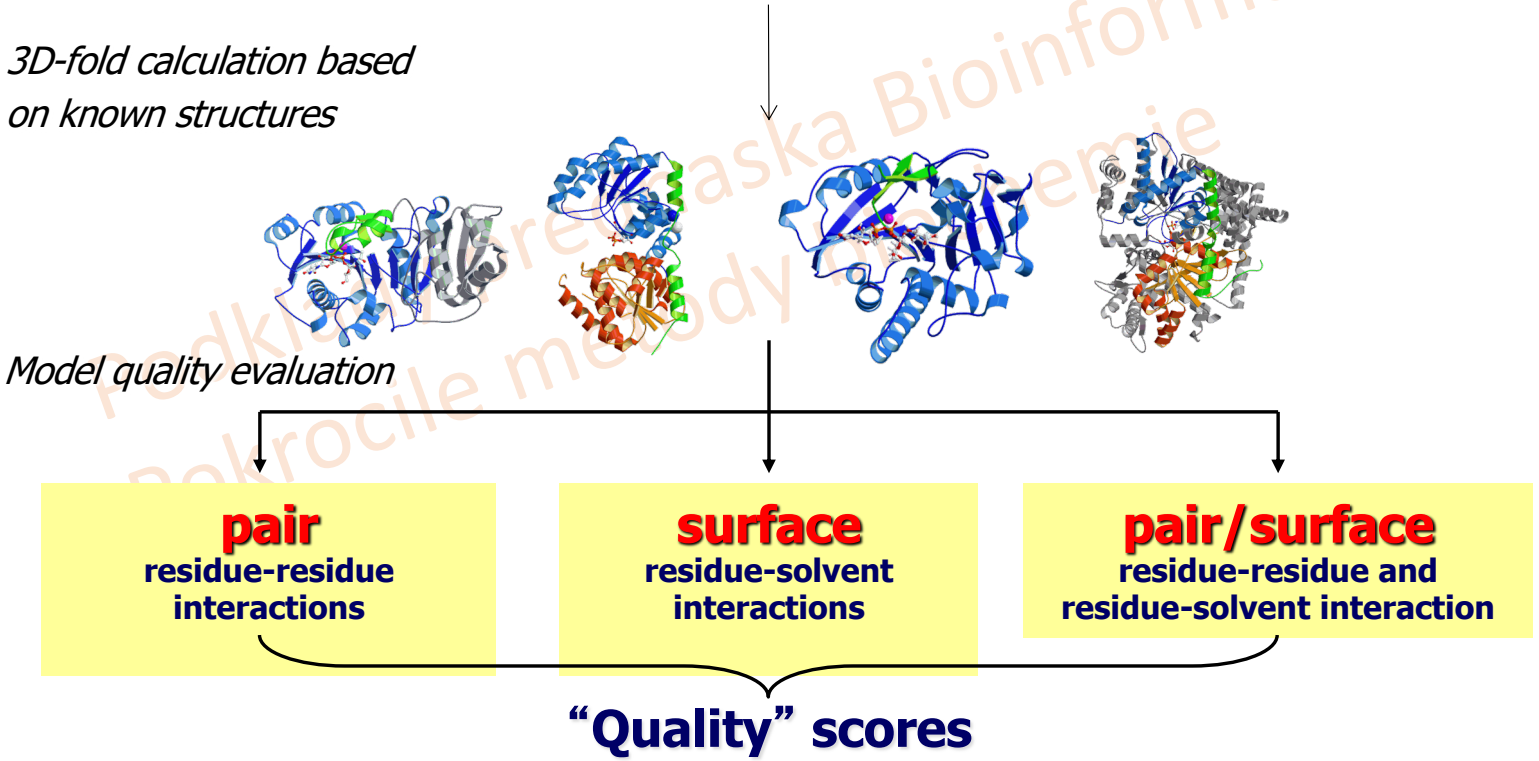
• Reziduum může být také popsáno pomocí svých interakcí.

• Výsledná kvalita modelu shoda je popsána pomocí Z-skóre nebo energie.

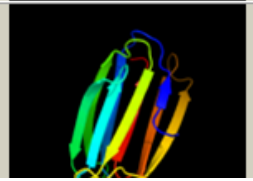
Často využíváme k hledání funkce neznámého proteinu a k odhadu 3D struktury

Predikce 3-D struktury/foldu proteinů - Threading

SDVDIEAGQTLVQVVNISNGETWVAIQLPAQYRSFDLVFENVSPSTSGSVLVAQMAPQSGGVYGSNYS
GSGWGNDLGGGGFYGYSEAKWMCLWPANRSGPNSKTGIYGTCKLMNLNQSNAVPSVTSNLFAPTAY
KNEPGYANVGGCCQKIRGLASSIQFAFALHGGNVPQNTDTFSGGTIKVYGWN



Phyre = http://www.sbg.bio.ic.ac.uk/phyre2

#	Template	Alignment Coverage	3D Model	Confidence	% i.d.	Template Information
1	c2vnvC_	Alignment		100.0	60	**PDB header:**sugar-binding protein **Chain:** C: **PDB Molecule:**bcla; **PDBTitle:** crystal structure of bcla lectin from burkholderia2 cenocepacia in complex with alpha-methyl-mannoside at 1.73 angstrom resolution
2	c2xr4A_	Alignment		100.0	43	**PDB header:**sugar binding protein **Chain:** A: **PDB Molecule:**lectin; **PDBTitle:** c-terminal domain of bc2l-c lectin from burkholderia cenocepacia
3	d2chha1	Alignment		100.0	37	**Fold:**Calcium-mediated lectin **Superfamily:**Calcium-mediated lectin **Family:**Calcium-mediated lectin

Predikce 3-D struktury/foldu proteinů - Homology modeling

- Přiložení cílové sekvence se sekvencí **homologního** proteinu se **známou 3D strukturou**
- Extrakce uhlíkové páteře ze struktury templátu a umístění postranních řetězců
- Modelování otoček a smyček
- Minimalizace energie
- Validace modelované struktury

MODELLER

Mostly used program in academic environment for serious homology modeling

SWISS-MODEL

An automated knowledge-based protein modelling server

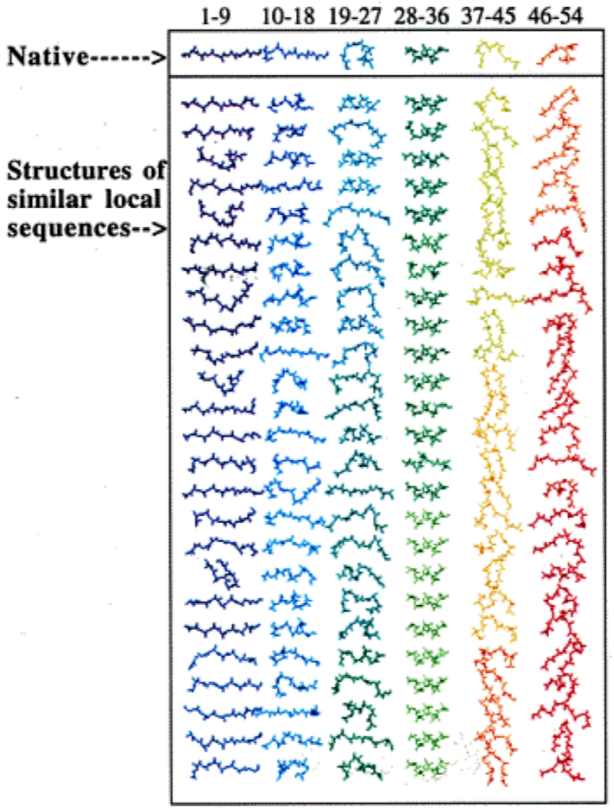
Obvykle se snažíme předpovědět skutečnou strukturu proteinu k další práci (predikce vazebných míst, dokování ligandů,…)

Predikce 3-D struktury/foldu proteinů - *Ab initio*

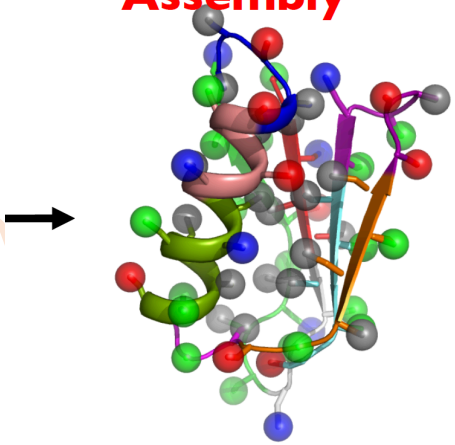
- **Přímý výpočet** nativní konformace (struktury) proteinu pouze ze sekvence.
- Nativní konformace je taková, která má ze všech možných nejnižší energii.
- Navzdory rozvoji výpočetní techniky, prohloubení znalostí o proteinech a vývoji metodiky se stále jedná o nevyřešený problém.
- **Budoucnost???**

De novo modelling with Rossetta

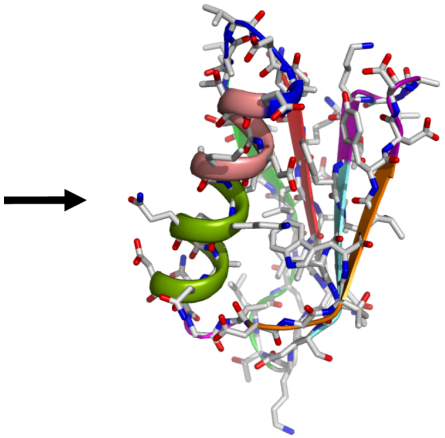
- fragments are selected from known structures
- the window-fragment matches are calculated using
  - PSI-BLAST to build a profile model of the sequence
  - the predicted secondary structure of the sequence



Stage I. Fragment Assembly



Stage II. All-atom refinement

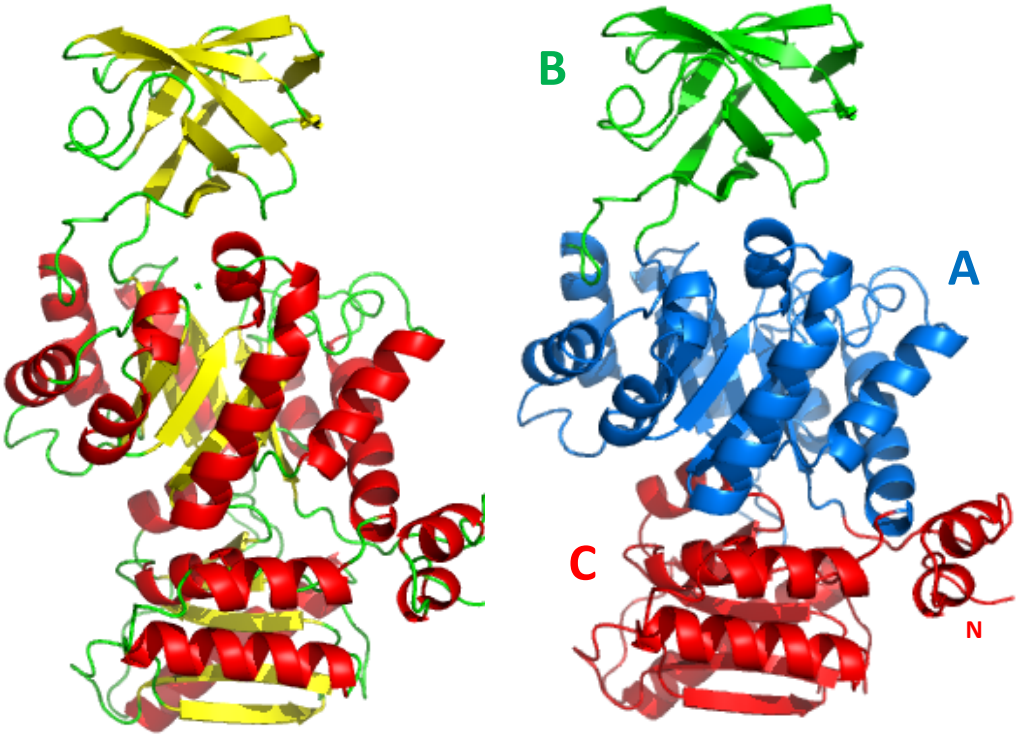


Proteinové domény – klasifikace proteinů

- **Proteinové domény** jsou konzervované funkční a/nebo strukturní části proteinu. Většinou jsou nezávislé, tj. schopné správného sbalení a zachování funkce i po oddělení od zbytku proteinu.

- Určitá konkrétní doména se může vyskytovat v různých proteinech.

- **Doména vs. podjednotka**

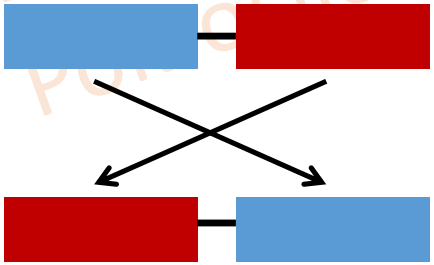


Pyruvátkinasa, tři domény + jedna krátká (A,B,C,N)

Proč detegovat domény?

PLLSASIVSAPVVTSETYVDIPGLYLDVAKAGIRDGKLQVILNVPTPYATGN
NFPGIYFAIATNQGVVADGCFTYSSKVPESTGRMPFTLVATIDVGSGVTFV
KGQWKSVRGSAMHIDSYASLSAIWGTAAPSSQGSGNQGAETGGTGAGN
IGGGGERDGTFNLPPHIKFGVTALTHAANDQTIDIYIDDDPKPAATFKGAGA
QDQNLGTKVLDSGNGRVRVIVMANGRPSRLGSRQVDIFKKSYFGIIGSED
GADDDYNDGIVFLNWPLG

ERDGTFNLPPHIKFGVTALTHAANDQTIDIYIDDDPKPAATFKGAGAQDQN
LGTKVLDSGNGRVRVIVMANGRPSRLGSRQVDIFKKSYFGIIGSEDGADD
DYNDGIVFLNWPLGPLLSASIVSAPVVTSQTYVDIPGLYLDVAKAGIRDGK
LQVILNVPTPYATGNNFPGIYFAIATNQGVVADGCFTYSSKVPESTGRMPF
TLVATIDVGSGVTFVKGQWKSVRGSAMHIDSYASLSAIWGTAAPSSQGSG
NQGAETGGTGAGNIGGGGKLAAALEIKRASQPELAPEDPEDVEHHHHHH



```
#
#=======================================

EMBOSS_001         1 --------------------------------------------------      0

EMBOSS_001         1 ERDGTFNLPPHIKFGVTALTHAANDQTIDIYIDDDPKPAATFKGAGAQDQ     50

EMBOSS_001         1 --------------------------------------------------      0

EMBOSS_001        51 NLGTKVLDSGNGRVRVIVMANGRPSRLGSRQVDIFKKSYFGIIGSEDGAD    100

EMBOSS_001         1 ---------------PLLSASIVSAPVVTSETYVDIPGLYLDVAKAGIRD     35
                                    |||||||||||||||:||||||||||||||||||
EMBOSS_001       101 DDYNDGIVFLNWPLGPLLSASIVSAPVVTSQTYVDIPGLYLDVAKAGIRD    150

EMBOSS_001        36 GKLQVILNVPTPYATGNNFPGIYFAIATNQGVVADGCFTYSSKVPESTGR     85
                     ||||||||||||||||||||||||||||||||||||||||||||||||||||
EMBOSS_001       151 GKLQVILNVPTPYATGNNFPGIYFAIATNQGVVADGCFTYSSKVPESTGR    200

EMBOSS_001        86 MPFTLVATIDVGSGVTFVKGQWKSVRGSAMHIDSYASLSAIWGTAAPSSQ    135
                     ||||||||||||||||||||||||||||||||||||||||||||||||||||
EMBOSS_001       201 MPFTLVATIDVGSGVTFVKGQWKSVRGSAMHIDSYASLSAIWGTAAPSSQ    250

EMBOSS_001       136 GSGNQGAETGGTGAGNIGGGGERDGTFNLPPHIKFGVTALTHAANDQTID    185
                     ||||||||||||||||||||
EMBOSS_001       251 GSGNQGAETGGTGAGNIGGGG----------------------------    271

EMBOSS_001       186 IYIDDDPKPAATFKGAGAQDQNLGTKVLDSGNGRVRVIVMANGRPSRLGS    235
                           |.||:          |..|                 |:|
EMBOSS_001       272 -------KLAAA----------LEIK----------------RAS----    283

EMBOSS_001       236 RQVDIFKKSYFGIIGSEDGADDDYNDGIVFLNWPLG    271
                      |.:         ::.||..|.:.:.       ...
EMBOSS_001       284 -QPE---------LAPEDPEDVEHHH-------HHH    302
```

Klasifikace proteinů - databáze

CATH-Gene3D database describes protein families and domain architectures in complete genomes. Protein families are formed using a Markov clustering algorithm, followed by multi-linkage clustering according to sequence identity. Mapping of predicted structure and sequence domains is undertaken using hidden Markov models libraries representing CATH and Pfam domains. CATH-Gene3D is based at University College, London, UK.

CDD is a protein annotation resource that consists of a collection of annotated multiple sequence alignment models for ancient domains and full-length proteins. These are available as position-specific score matrices (PSSMs) for fast identification of conserved domains in protein sequences via RPS-BLAST. CDD content includes NCBI-curated domain models, which use 3D-structure information to explicitly define domain boundaries and provide insights into sequence/structure/function relationships, as well as domain models imported from a number of external source databases.

MobiDB offers a centralized resource for annotations of intrinsic protein disorder. The database features three levels of annotation: manually curated, indirect and predicted. The different sources present a clear tradeoff between quality and coverage. By combining them all into a consensus annotation, MobiDB aims at giving the best possible picture of the "disorder landscape" of a given protein of interest.

HAMAP stands for High-quality Automated and Manual Annotation of Proteins. HAMAP profiles are manually created by expert curators. They identify proteins that are part of well-conserved proteins families or subfamilies. HAMAP is based at the SIB Swiss Institute of Bioinformatics, Geneva, Switzerland.

PANTHER is a large collection of protein families that have been subdivided into functionally related subfamilies, using human expertise. These subfamilies model the divergence of specific functions within protein families, allowing more accurate association with function, as well as inference of amino acids important for functional specificity. Hidden Markov models (HMMs) are built for each family and subfamily for classifying additional protein sequences. PANTHER is based at at University of Southern California, CA, US.

Pfam is a large collection of multiple sequence alignments and hidden Markov models covering many common protein domains. Pfam is based at EMBL-EBI, Hinxton, UK.

PIRSF protein classification system is a network with multiple levels of sequence diversity from superfamilies to subfamilies that reflects the evolutionary relationship of full-length proteins and domains. PIRSF is based at the Protein Information Resource, Georgetown University Medical Centre, Washington DC, US.

PRINTS is a compendium of protein fingerprints. A fingerprint is a group of conserved motifs used to characterise a protein family or domain. PRINTS is based at the University of Manchester, UK.

ProDom protein domain database consists of an automatic compilation of homologous domains. Current versions of ProDom are built using a novel procedure based on recursive PSI-BLAST searches. ProDom is based at PRABI Villeurbanne, France.

PROSITE is a database of protein families and domains. It consists of biologically significant sites, patterns and profiles that help to reliably identify to which known protein family a new sequence belongs. PROSITE is base at the Swiss Institute of Bioinformatics (SIB), Geneva, Switzerland.

SFLD (Structure-Function Linkage Database) is a hierarchical classification of enzymes that relates specific sequence-structure features to specific chemical capabilities.

SMART (a Simple Modular Architecture Research Tool) allows the identification and annotation of genetically mobile domains and the analysis of domain architectures. SMART is based at at EMBL, Heidelberg, Germany.

SUPERFAMILY is a library of profile hidden Markov models that represent all proteins of known structure. The library is based on the SCOP classification of proteins: each model corresponds to a SCOP domain and aims to represent the entire SCOP superfamily that the domain belongs to. SUPERFAMILY is based at the University of Bristol, UK.

TIGRFAMs is a collection of protein families, featuring curated multiple sequence alignments, hidden Markov models (HMMs) and annotation, which provides a tool for identifying functionally related proteins based on sequence homology. TIGRFAMs is based at the J. Craig Venter Institute, Rockville, MD, US.

Databáze jsou sdruženy do integrovaného nástroje InterPro

What is InterPro?

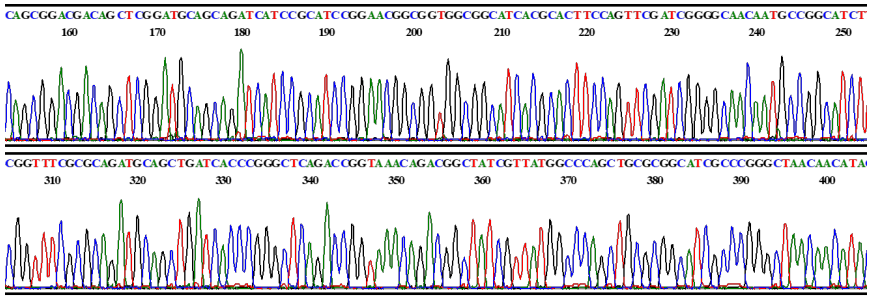
InterPro is a resource that provides functional analysis of protein sequences by classifying them into families and predicting the presence of domains and important sites. To classify proteins in this way, InterPro uses predictive models, known as signatures, provided by several different databases (referred to as member databases) that make up the InterPro consortium.

https://www.ebi.ac.uk/interpro/

Osnova

- **Úvod do bioinformatiky**
  - Definice, molekulárně biologická data, databáze
  - Rozdělení databází, bioinformatická centra
- **Manipulace se sekvencemi**
  - Sekvence biomakromolekul, aminokyseliny, báze, alignment
  - Význam alignmentu, přiložení páru sekvencí a vícenásobné přiložení
- **Predikce struktury proteinů**
  - Predikce 2-D struktury proteinů, predikce 3-D struktury proteinů
  - Threading, Homology modeling, *Ab initio*
- **Predikce genů**
  - Predikce genů u prokaryot a eukaryot, predikční nástroje a postupy

Predikce genů



Sekvenace

GATAGCGTAATGATCGGCTGGCTGCCGCATTTCATGCTGGTTTCCCAACGAAAATAACCGCTCACGGTGCCATCACGATCGCACACCGCAAAATCGGCGG

TACAGGTGGTCGCGCCCGCCGCCAGCACATCGCTGCGCCAATAATGATCTTTCAGCGGACGACAGCTCGGATGCAGCAGATCATCCGCATCCGGAACGGC

GGTGGCGGCATCACGCACTTCCAGTTCGATCGGGGCAACAATGCCGGCATCTTTCAGGGCAAAGCGAATAAACAGCACGCTCACTTCCGCGCGCAGCGCC

AGCGCGGTTTCGCGCAGATGCAGCTGATCACCCGGGCTCAGACCGGTAAACAGACGGCTATCGTTATGGCCCAGCTGCGCGGCATCGCCCGGGCTAACAA

CATACAGGTGGCGACCATCAATCACGGTCGGGGCGGCCGGATCACGGCTGGCTTCCGGATAGGCGCTCAGCAGGGTAACGGCATCCACAATCACCAGCAT

Surové sekvence DNA

Identifikace a anotace genů a proteinů

Table 1
Software commonly used for bacterial genome annotation and comparison

DNA level annotation		
GeneMark	http://exon.gatech.edu/genemark/	Protein gene prediction
Glimmer	http://www.genomics.jhu.edu/Glimmer/	Protein gene prediction
SHOW	http://genome.jouy.inra.fr/ssb/SHOW/	Protein gene prediction
tRNAscan-SE	http://lowelab.ucsc.edu/tRNAscan-SE/	tRNA gene prediction
RNAmmer	http://www.cbs.dtu.dk/services/RNAmmer/	rRNA gene prediction
RepSeek	http://www.abi.snv.jussieu.fr/%98public/RepSeek/	Search for approximate repeats in complete DNA sequences
IslandPath	http://www.pathogenomics.sfu.ca/islandpath/	Identification of genomic islands
Protein level annotation		
BLAST	http://www.ebi.ac.uk/blast/	Compare a novel sequence with those contained in nucleotide and protein databases
InterProScan	http://www.ebi.ac.uk/InterProScan/	Search for domains/motifs in the InterPro database
COGNITOR	http://www.ncbi.nlm.nih.gov/COG/old/xognitor.html	Compare a query sequence to the COG (Cluster of Orthologous Groups of proteins) database
PRIAM	http://bioinfo.genopole-toulouse.prd.fr/priam/	Detection of enzymatic function in a fully sequenced genome, based on all sequences available in the ENZYME database
GOAnno	http://bips.u-strasbg.fr/GOAnno/	BLAST search on the Gene Ontology database
PSORTb	http://www.psort.org/psortb/	Prediction of bacterial protein subcellular localization
TMHMM	http://www.cbs.dtu.dk/services/TMHMM/	Prediction of transmembrane helices in protein sequences
SignalP	http://www.cbs.dtu.dk/services/SignalP/	Prediction of signal peptide cleavage sites in protein sequences
Comparative genomic tools		
Mauve	http://gel.ahabs.wisc.edu/mauve/	Multiple genome alignments in the presence of large-scale evolutionary events
MOSAIC	http://mig.jouy.inra.fr/mig/mig_eng/presentation/project/mosaic	Define the set of backbones and loops in closely related bacterial genomes
ACT	http://www.sanger.ac.uk/Software/ACT/	Comparative genome analysis and visualization tools for multiple genome alignments
CGAT	http://mbgd.genome.ad.jp/CGAT/	
MaGe	http://www.genoscope.cns.fr/agc/mage/	Computation of gene order conservation (syntenies) between available bacterial genomes
Pathologic	http://biocyc.org/	Metabolic network reconstruction and comparative pathway analysis
PUMA2	http://compbio.mcs.anl.gov/puma2/	Metabolic pathway reconstruction
The SEED	http://theseed.uchicago.edu/FIG/	Comparative analysis and annotation tools using the subsystem approach
STRING	http://string.embl.de/	Search Tool for the Retrieval of Interacting Proteins
PyPhy	http://www.cbs.dtu.dk/staff/thomas/pyphy/	Reconstruction of phylogenetic relationships of complete microbial genomes
HoSeqI	http://pbil.univ-lyon1.fr/software/HoSeqI/	Automatically assign sequences to homologous gene families from the HOGENOM database

Predikce genů
(Predikce *kódující* části genu)

- **Prokaryotické geny** – nepřerušované úseky DNA mezi **startovním kodonem** (ATG, GTG, TTG, CTG) a **stop kodonem** (TAA, TGA, TAG).

- **Eukaryotické geny** – Přerušovány introny. Průměrná délka exonu je 50 kodonů, některé jsou mnohem kratší. Některé introny extrémně dlouhé, geny zabírají mbp v genomové DNA. **Predikce je mnohem složitější a vzniká velké množství chyb!**

Predikce prokaryotických genů

GTATGCTGGTGATTGTGGATGCCGTTACCCTGCTGAGCGCCTATCCGGAAGCCAGCCGTGATCCGGCCGCCCCGA
CCGTGATTGATGGTCGCCACCTGTATGTTGTTAGCCCGGGCGATGCCGCGCAGCTGGGCCATAACGATAGCCGTC
TGTTTACCGGTCTGAGCCCGGGTGATCAGCTGCATCTGCGCGAAACCGCGCTGGCGCTGCGCGCGGAAGTGAGCG
TGCTGTTTATTCGCTTTGCCCTGAAAGATGCCGGCATTGTTGCCCCGATCGAACTGGAAGTGCGTGATGCCGCCA
CCGCCGTTCCGGATGCGGATGATCTGCTGCATCCGAGCTGTCGTCCGCTGAAAGATCATTATTGGCGCAGCGATG
TGCTGGCGGCGGGCGCGACCACCTGTACCGCCGATTTTGCGGTGTGCGATCGTGATGGCACCGTGAGCGGTTATT
TTCGTTGGGAAACCAGCATTGAAATTGCGGGCAGCCAGCCGGATACCAAACAGCCGGGCTTTAAACCGAGCAGCG
ATCGCAATGGCAACTTTAGCCTGCCGCCGAATACCGCCTTTAAAGCGATCTTCTATGCGAACGCGGCGGATCGTC
AGGATCTGAAACTGTTTATTGATGATGCGCCGGAACCGGCCGCCACCTTTGTGGGTAACAGCGAAGATGGTGTGC
GTCTGTTTACCCTGAATAGCAAAGGTGGTAAAATTCGTATTGAAGCGAGCGCGAACGGCCGTCAGAGCGCGACCG
ATGCCCGTCTGGCGCCGCTGAGCGCGGGCGATACCGTGTGGCTGGGCTGGCTGGGCGCGGAAGATGGTGCCGATG
CGGATTATAATGATGGCATTGTTATTCTGCAGTGGCCGATTACCTAATGGG

The table shows the 64 codons and the amino acid for each. The direction of the mRNA is 5' to 3'.

1st base	2nd base U	2nd base C	2nd base A	2nd base G
U	UUU (Phe/F) Phenylalanine	UCU (Ser/S) Serine	UAU (Tyr/Y) Tyrosine	UGU (Cys/C) Cysteine
	UUC (Phe/F) Phenylalanine	UCC (Ser/S) Serine	UAC (Tyr/Y) Tyrosine	UGC (Cys/C) Cysteine
	UUA (Leu/L) Leucine	UCA (Ser/S) Serine	UAA Ochre (*Stop*)	UGA Opal (*Stop*)
	UUG (Leu/L) Leucine	UCG (Ser/S) Serine	UAG Amber (*Stop*)	UGG (Trp/W) Tryptophan
C	CUU (Leu/L) Leucine	CCU (Pro/P) Proline	CAU (His/H) Histidine	CGU (Arg/R) Arginine
	CUC (Leu/L) Leucine	CCC (Pro/P) Proline	CAC (His/H) Histidine	CGC (Arg/R) Arginine
	CUA (Leu/L) Leucine	CCA (Pro/P) Proline	CAA (Gln/Q) Glutamine	CGA (Arg/R) Arginine
	CUG (Leu/L) Leucine	CCG (Pro/P) Proline	CAG (Gln/Q) Glutamine	CGG (Arg/R) Arginine
A	AUU (Ile/I) Isoleucine	ACU (Thr/T) Threonine	AAU (Asn/N) Asparagine	AGU (Ser/S) Serine
	AUC (Ile/I) Isoleucine	ACC (Thr/T) Threonine	AAC (Asn/N) Asparagine	AGC (Ser/S) Serine
	AUA (Ile/I) Isoleucine	ACA (Thr/T) Threonine	AAA (Lys/K) Lysine	AGA (Arg/R) Arginine
	AUG (Met/M) Methionine, *Start* [A]	ACG (Thr/T) Threonine	AAG (Lys/K) Lysine	AGG (Arg/R) Arginine
G	GUU (Val/V) Valine	GCU (Ala/A) Alanine	GAU (Asp/D) Aspartic acid	GGU (Gly/G) Glycine
	GUC (Val/V) Valine	GCC (Ala/A) Alanine	GAC (Asp/D) Aspartic acid	GGC (Gly/G) Glycine
	GUA (Val/V) Valine	GCA (Ala/A) Alanine	GAA (Glu/E) Glutamic acid	GGA (Gly/G) Glycine
	GUG (Val/V) Valine	GCG (Ala/A) Alanine	GAG (Glu/E) Glutamic acid	GGG (Gly/G) Glycine

Překlad DNA sekvence

Open reading frames are highlighted in red. Please select one of the following frames - in the next page, you will be able to select your initiator and retrieve your amino acid sequence:

5'3' Frame 1
V C W Stop L W Met P L P C Stop A P I R K P A V I R P P R P Stop L Met V A T C Met L L A R A Met P R S W A I T I A V C L P V Stop A R V I S C I C A K P R W R C A R K
Stop A C C L F A L P Stop K Met P A L L P R S N W K C V Met P P P P F R Met R Met I C C I R A V V R Stop K I I I G A A Met C W R R A R P P V P P I L R C A I V Met A P
Stop A V I F V G K P A L K L R A A S R I P N S R A L N R A A I A Met A T L A C R R I P P L K R S S Met R T R R I V R I Stop N C L L Met Met R R N R P P P L W V T A K Met
V C V C L P Stop I A K V V K F V L K R A R T A V R A R P Met P V W R R Stop A R A I P C G W A G W A R K Met V P Met R I I Met Met A L L F C S G R L P N G

5'3' Frame 2
Y A G D C G C R Y P A E R L S G S Q P Stop S G R P D R D Stop W S P P V C C Stop P G R C R A A G P Stop R Stop P S V Y R S E P G Stop S A A S A R N R A G A A R
G S E R A V Y S L C P E R C R H C C P D R T G S A Stop C R H R R S G C G Stop S A A S E L S S A E R S L L A Q R C A G G G R D H L Y R R F C G V R S Stop W H R E
R L F S L G N Q H Stop N C G Q P A G Y Q T A G L Stop T E Q R S Q W Q L Stop P A A E Y R L Stop S D L L C E R G G S S G S E T V Y Stop Stop C A G T G R H L C G
Stop Q R R W C A S V Y P E Stop Q R W Stop N S Y Stop S E R E R P S E R D R C P S G A A E R G R Y R V A G L A G R G R W C R C G L Stop Stop W H C Y S A V A D
Y L Met

5'3' Frame 3
Met L V I V D A V T L L S A Y P E A S R D P A A P T V I D G R H L Y V V S P G D A A Q L G H N D S R L F T G L S P G D Q L H L R E T A L A L R A E V S V L F I R F A L K D
A G I V A P I E L E V R D A A T A V P D A D D L L H P S C R P L K D H Y W R S D V L A A G A T T C T A D F A V C D R D G T V S G Y F R W E T S I E I A G S Q P D T K Q P
G F K P S S D R N G N F S L P P N T A F K A I F Y A N A A D R Q D L K L F I D D A P E P A A T F V G N S E D G V R L F T L N S K G G K I R I E A S A N G R Q S A T D A R L
A P L S A G D T V W L G W L G A E D G A D A D Y N D G I V I L Q W P I T Stop W

3'5' Frame 1
P I R Stop S A T A E Stop Q C H H Y N P H R H H L P R P A S P A T R Y R P R S A A P D G H R S R S D G R S R S L Q Y E F Y H L C Y S G Stop T D A H H L R C Y P Q R W
R P V P A H H Q Stop T V S D P D D P P R S H R R S L Stop R R Y S A A G Stop S C H C D R C S V Stop S P A V W Y P A G C P Q F Q C W F P N E N N R S R C H H D R
T P Q N R R Y R W S R P P P A H R C A N N D L S A D D S S D A A D H P H P E R R W R H H A L P V R S G Q Q C R H L S G Q S E Stop T A R S L P R A A P A R F R A D
A A D H P G S D R Stop T D G Y R Y G P A A R H R P G Stop Q H T G G D H Q S R S G R P D H G W L P D R R S A G Stop R H P Q S P A Y

3'5' Frame 2
P L G N R P L Q N N N A I I I I R I G T I F R A Q P A Q P H G I A R A Q R R Q T G I G R A L T A V R A R F N T N F T T F A I Q G K Q T H T I F A V T H K G G G R F R R I I N K Q F
Q I L T I R R V R I E D R F K G G I R R Q A K V A I A I A A R F K A R L F G I R L A A R N F N A G F P T K I T A H G A I T I A H R K I G G T G G R A R R Q H I A A P I Met I F Q R T
T A R Met Q Q I I R I R N G G G G I T H F Q F D R G N N A G I F Q G K A N K Q H A H F R A Q R Q R G F A Q Met Q L I T R A Q T G K Q T A I V Met A Q L R G I A R A N N I Q V
A T I N H G R G G R I T A G F R I G A Q Q G N G I H N H Q H

3'5' Frame 3
H Stop V I G H C R I T Met P S L Stop S A S A P S S A P S Q P S H T V S P A L S G A R R A S V A L Stop R P F A L A S I R I L P P L L F R V N R R T P S S L L P T K V A A
G S G A S S I N S F R S Stop R S A A F A Stop K I A L K A V F G G R L K L P L R S L L G L K P G C L V S G W L P A I S Met L V S Q R K Stop P L T V P S R S H T A K S A
V Q V V A P A A S T S L R Q Stop Stop S F S G R Q L G C S R S S A S G T A V A A S R T S S S I G A T Met P A S F R A K R I N S T L T S A R S A S A V S R R C S Stop S
P G L R P V N R R L S L W P S C A A S P G L T T Y R W R P S I T V G A A G S R L A S G Stop A L S R V T A S T I T S I

Identifikace ORF (otevřených čtecích rámců)

ExPASy
http://web.expasy.org/translate/
ORF Finder (NCBI)
https://www.ncbi.nlm.nih.gov/orffinder/

Predikce prokaryotických genů

- Opravdu kóduje ORF protein?
- **ORF kóduje protein, který je podobný již dříve popsanému proteinu (prohledávání DATABÁZÍ pomocí ALIGNMENTU).**
- **ORF má typický obsah GC nebo frekvenci kodonů. Srovnání s charakteristickými vlastnostmi známých genů ze stejného organismu.**
- **Před ORF se nachází typické RBS (ribosome-binding site) nebo promotor.** Analýza signálních sekvencí pro transkripci a/nebo translaci.

Predikce eukaryotických genů

- **Rozpoznání exonů/intronů**

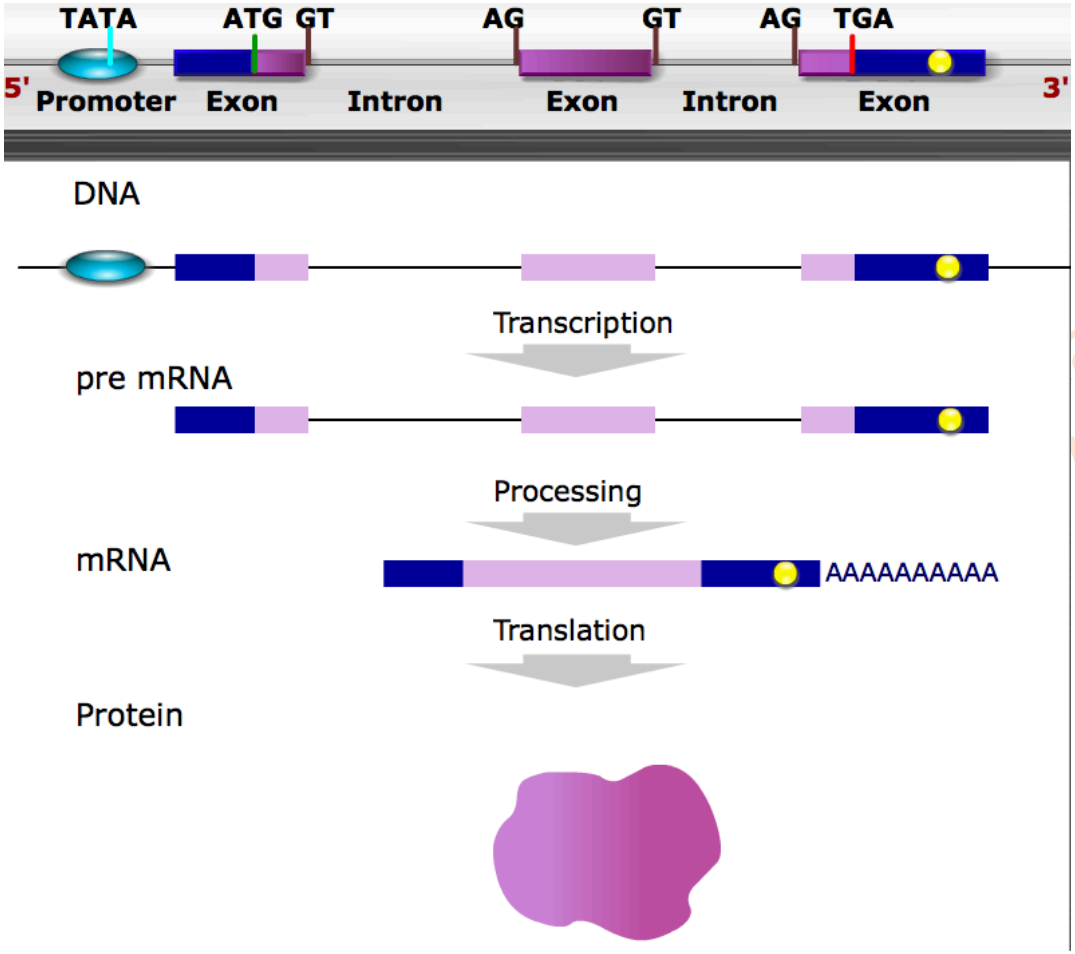
  Identifikace míst sestřihu: GT na 5´konci, AG na 3´konci.

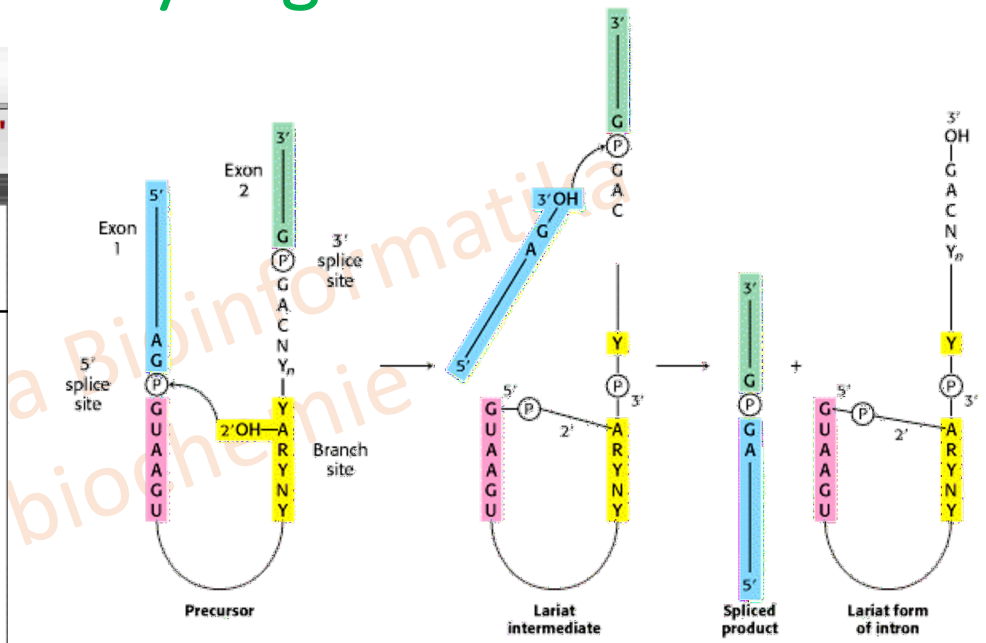
- **Chyby při rozpoznávání exonů/intronů**

  Velké množství chyb. Dlouhé introny – určeny jako intergenové úseky. Krátké intergenové useky – určeny jako introny.

Glyceraldehyd-3-fosfát-dehydrogenasa
Homo sapiens

Predikce eukaryotických genů

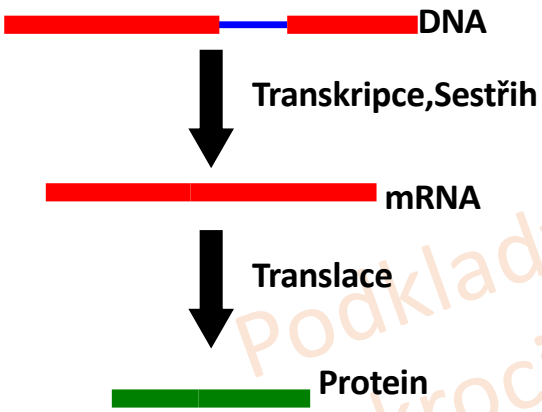




Splicing Mechanism Used for mRNA Precursors. The upstream (5′) exon is shown in blue, the downstream (3′) exon in green, and **the branch site in yellow. R stands for a purine nucleotide, Y for a pyrimidine nucleotide**, and N for any nucleotide. The 5′ splice site is attacked by the 2′-OH group of the branch-site adenosine residue. The 3′ splice site is attacked by the newly formed 3′-OH group of the upstream exon. The exons are joined, and the intron is released in the form of a lariat. [After P. A. Sharp. *Cell* 2(1985):3980.]

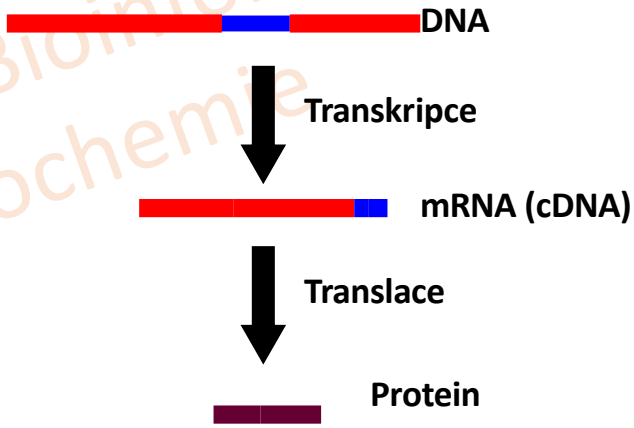
Predikce eukaryotických genů – příklad z praxe

Hypotetický gen/protein, predikovaný při anotaci genomu *Aspergillus fumigatus* Af293




```
MADPEVEADG ELDLEKRASA QTCKIVNVDT
YVNCRYDAKL DAGAIFGFPK GEKLTFACWK
HGDCYNGVCS WDQVTYLKTT CYVNGYFTDS
NCSSSMLSRC
```

Identifikace genu/proteinu na úrovni mRNA (příprava cDNA pro klonování)

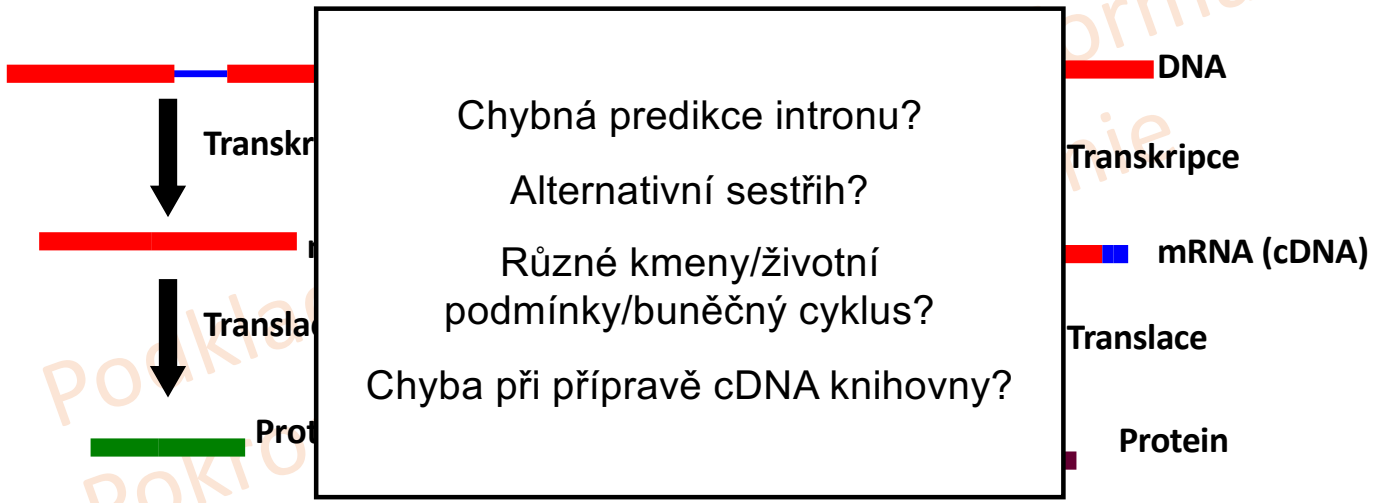



```
MADPEVEADG ELDLEKRASA QTCKIVNVDT
YVNCRYDAKL DAGAIFGFPK GEKLTFACWK HGDCYNGV
```

Predikce eukaryotických genů – příklad z praxe

Hypotetický gen/protein,
predikovaný při anotaci genomu
Aspergillus fumigatus Af293

Identifikace genu/proteinu na úrovni
mRNA (příprava cDNA pro klonování)



MADPEVEADG ELDLEKRASA QTCKIVNVDT
YVNCRYDAKL DAGAIFGFPK GEKLTFACWK
HGDCYNGVCS WDQVTYLKTT CYVNGYFTDS
NCSSSMLSRC

MADPEVEADG ELDLEKRASA QTCKIVNVDT
YVNCRYDAKL DAGAIFGFPK GEKLTFACWK
HGDCYNGV

Predikce genů – algoritmy a nástroje

- **Predikce genů na základě sekvenční homologie – vyhledávání v databázích pomocí algoritmů.**
- **Predikce genů *ab initio* – predikce na základě statistických parametrů DNA sekvence.**
- **Většina běžně používaných metod kombinuje oba dva přístupy.**

Program	Organism	Algorithm*	Website	Homology
GeneID	Vertebrates, plants	DP	http://www1.imim.es/geneid.html	
FGENESH	Human, mouse, Drosophila, rice	HMM	http://www.softberry.com/berry.phtml?topic=fgenesh&group=programs&subgroup=gfind	
GeneParser	Vertebrates	NN	http://beagle.colorado.edu/~eesnyder/GeneParser.html	EST
Genie	Drosophila, human, other	GHMM	http://www.fruitfly.org/seq_tools/genie.html	protein
GenLang	Vertebrates, Drosophila, dicots	Grammar rule	http://www.cbil.upenn.edu/genlang/genlang_home.html	
GENSCAN	Vertebrates, Arabidopsis, maize	GHMM	http://genes.mit.edu/GENSCAN.html	
GlimmerM	Small eukaryotes, Arabidopsis, rice	IMM	http://www.tigr.org/tdb/glimmerm/glmr_form.html	
GRAIL	Human, mouse, Arabidopsis, Drosophila	NN, DP	http://compbio.ornl.gov/Grail-bin/EmptyGrailForm	EST, cDNA
HMMgene	Vertebrates, *C. elegans*	CHMM	http://www.cbs.dtu.dk/services/HMMgene/	
AUGUSTUS	Human, Arabidopsis	IMM,WWAM	http://augustus.gobics.de/	
MZEF	Human, mouse, Arabidopsis, Fission yeast	Quadratic discriminant analysis	http://rulai.cshl.org/tools/genefinder/	

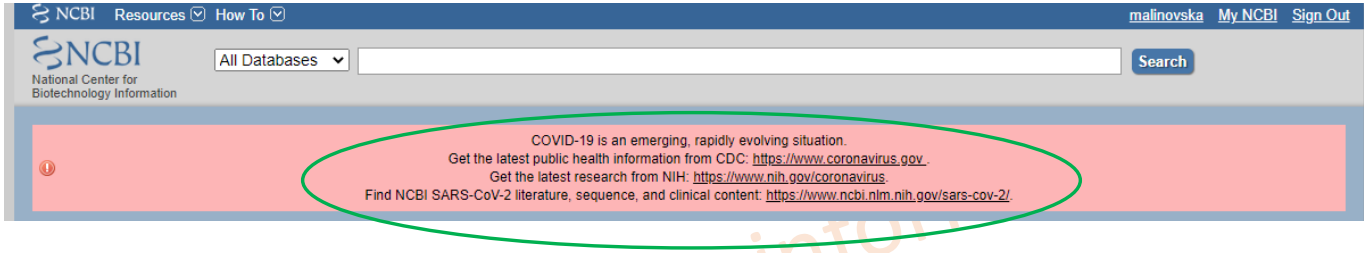
*DP, dynamic programming; NN, neural network; MM, Markov model; HMM, Hidden Markov model; CHMM, class HMM; GHMM, generalized HMM; IMM, interpolated MM.

Bioinformatika a Covid-19

NCBI

Národní centrum pro biotechnologické informace

National Center for Biotechnology Information



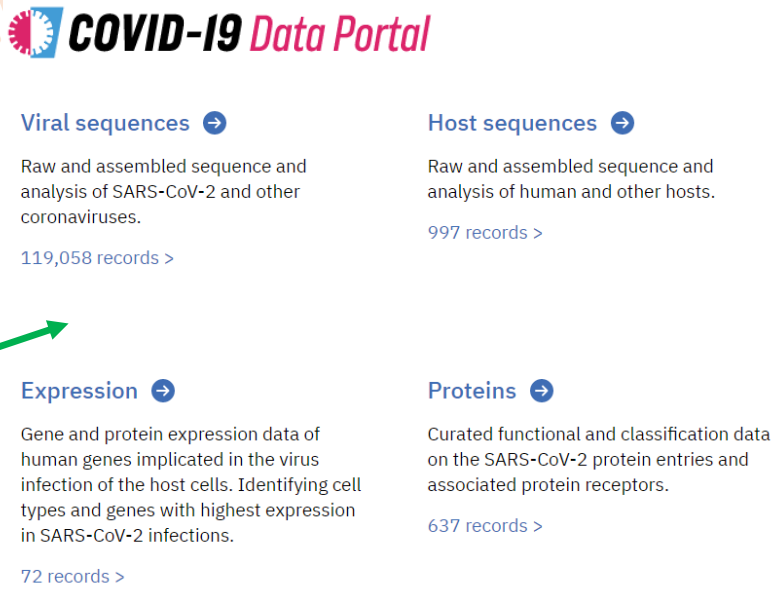
http://www.ncbi.nlm.nih.gov/

National Center for Biotechnology Information

EBI

Evropský institut pro bioinformatiku

http://www.ebi.ac.uk/



Bioinformatika a Covid-19

国家生物信息中心
China National Center for Bioinformation
National Genomics Data Center
The National Genomics Data Center (NGDC), part of the China National Center for Bioinformation (CNCB), advances life & health sciences by providing open access to a suite of resources, with the aim to translate big data into big discoveries and support worldwide activities in both academia and industry.

NGDC https://bigd.big.ac.cn/

Resources

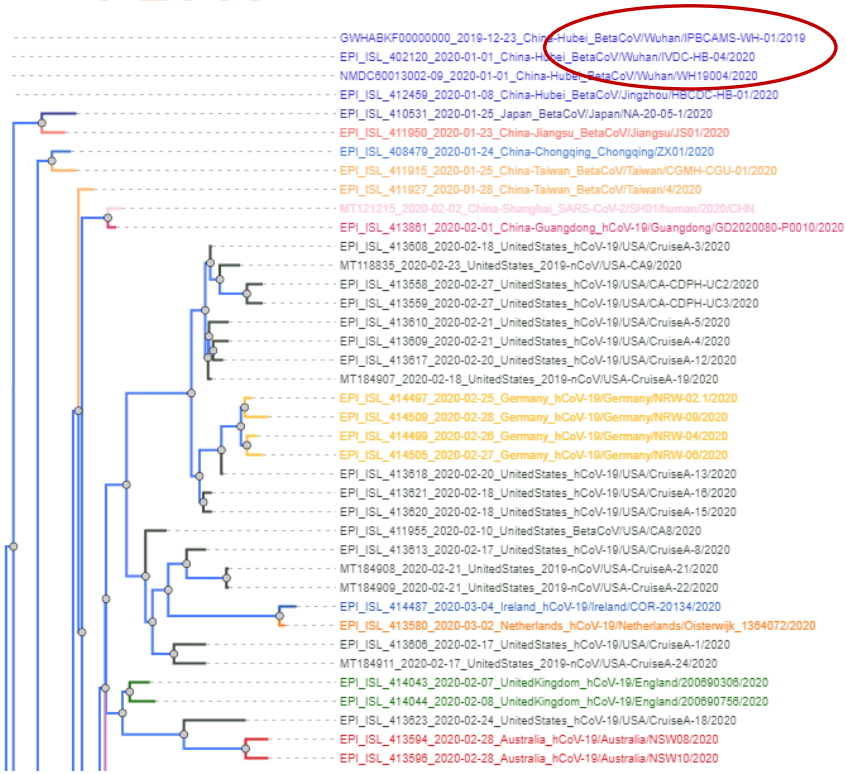
- 2019-nCoV Sequences (144798)
- Coronavirus Sequences (425685)
- Genome Variations (19226)
- AI Diagnosis & Online Tools
- Clinical Records (208)
- Literature (75573)

Phylogenetic Tree

Phylogeny tree of novel coronavirus across the whole world are constructed using BEAST, based on the genome variation information obtained from available high-quality genome sequences.

View Realtime Graph

China National Center for Bioinformation
2019 Novel Coronavirus Resource (2019nCoVR)



Bioinformatika a Covid-19

国家生物信息中心
China National Center for Bioinformation
National Genomics Data Center
The National Genomics Data Center (NGDC), part of the China National Center for Bioinformation (CNCB), advances life & health sciences by providing open access to a suite of resources, with the aim to translate big data into big discoveries and support worldwide activities in both academia and industry.

NGDC https://bigd.big.ac.cn/

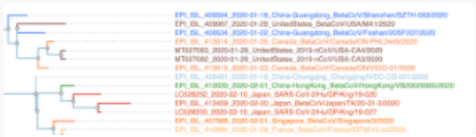
Resources

- 2019-nCoV Sequences (144798)
- Coronavirus Sequences (425685)
- Genome Variations (19226)
- AI Diagnosis & Online Tools
- Clinical Records (208)
- Literature (75573)

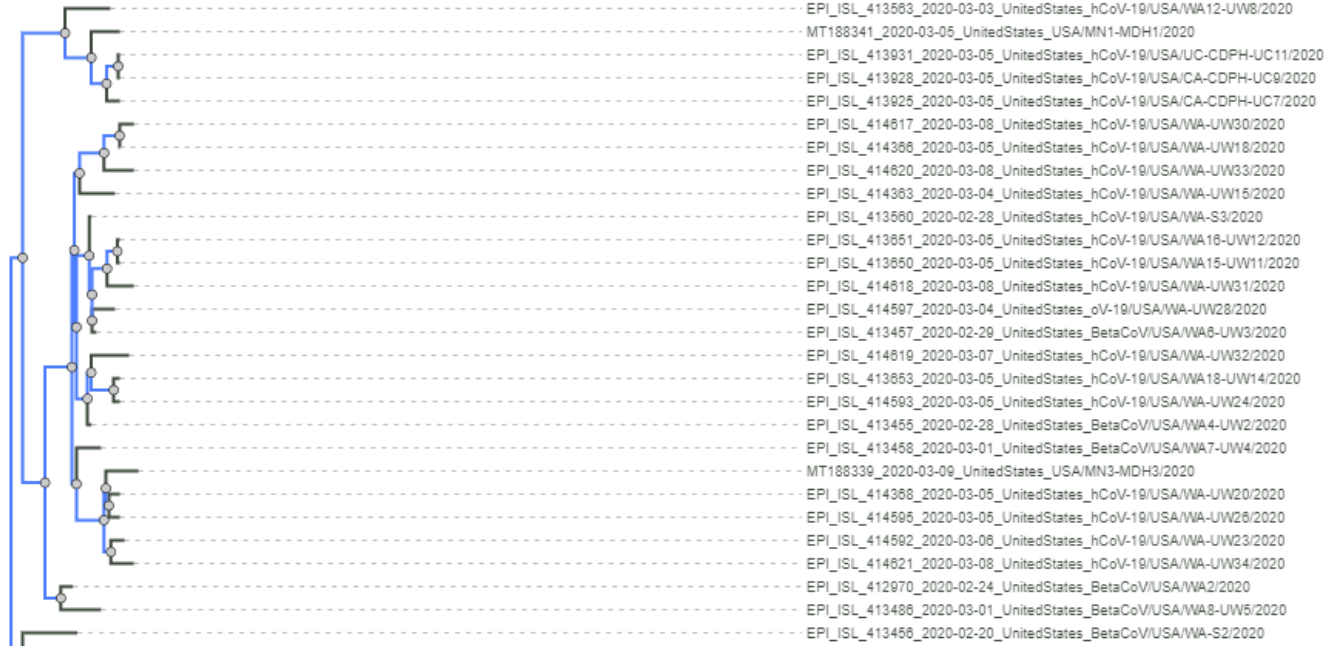
Phylogenetic Tree

Phylogeny tree of novel coronavirus across the whole world are constructed using BEAST, based on the genome variation information obtained from available high-quality genome sequences.

View Realtime Graph

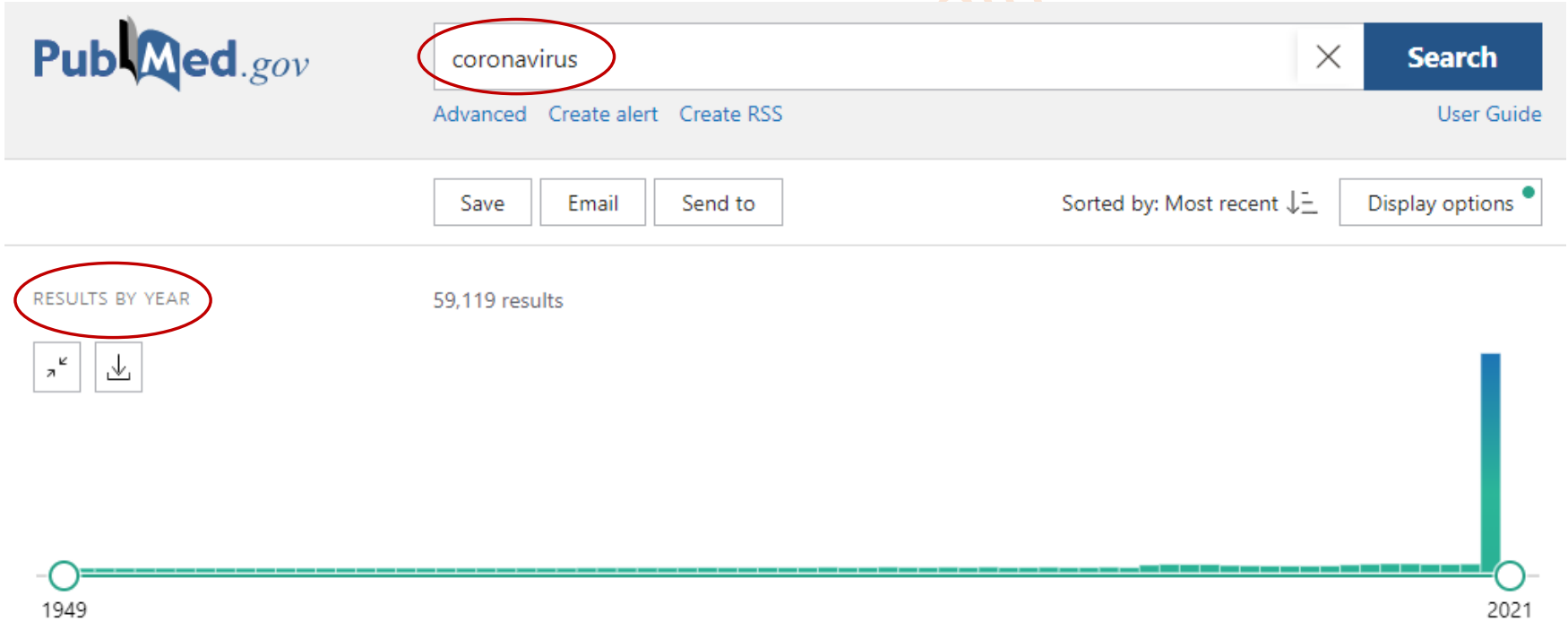


China National Center for Bioinformation
2019 Novel Coronavirus Resource (2019nCoVR)



Pokud Vás zajímají detaily, odkazy na použité články:

viz Učební materiály v ISu, adresář Bioinformatika/Materialy_pro_studenty (není nutno studovat ke zkoušce, pouze detailnější informace, pokud Vás zajíma něco blíže…)



Pokud Vás zajímají detaily, odkazy na použité články:

viz Učební materiály v ISu, adresář Bioinformatika/Materialy_pro_studenty (není nutno studovat ke zkoušce, pouze detailnější informace, pokud Vás zajíma něco blíže…)

Aktuální kurzy (všechny JS):

C2131 Úvod do bioinformatiky (vhled do oboru)
C2132 Úvod do bioinformatiky - seminář
C2135 Bioinformatika v praxi (pokud si chcete ošahat základy bioinformatiky prakticky)
C2138 Pokročilá bioinformatika
C2139 Pokročilá bioinformatika – seminář
C3211 Aplikovaná bioinformatika (pokud Vás zajímá jaké experimentální metody jsou propojeny s bioinformatickými nástroji)