Bioinformatika – rozšířené opakování

Nezbytné databáze, práce se sekvencemi, přiložení, predikce genu

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). *J. Xiong*
- **Bioinformatika** – nová disciplína na rozhraní počítačových věd, informačních technologií, matematiky a biologie; 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

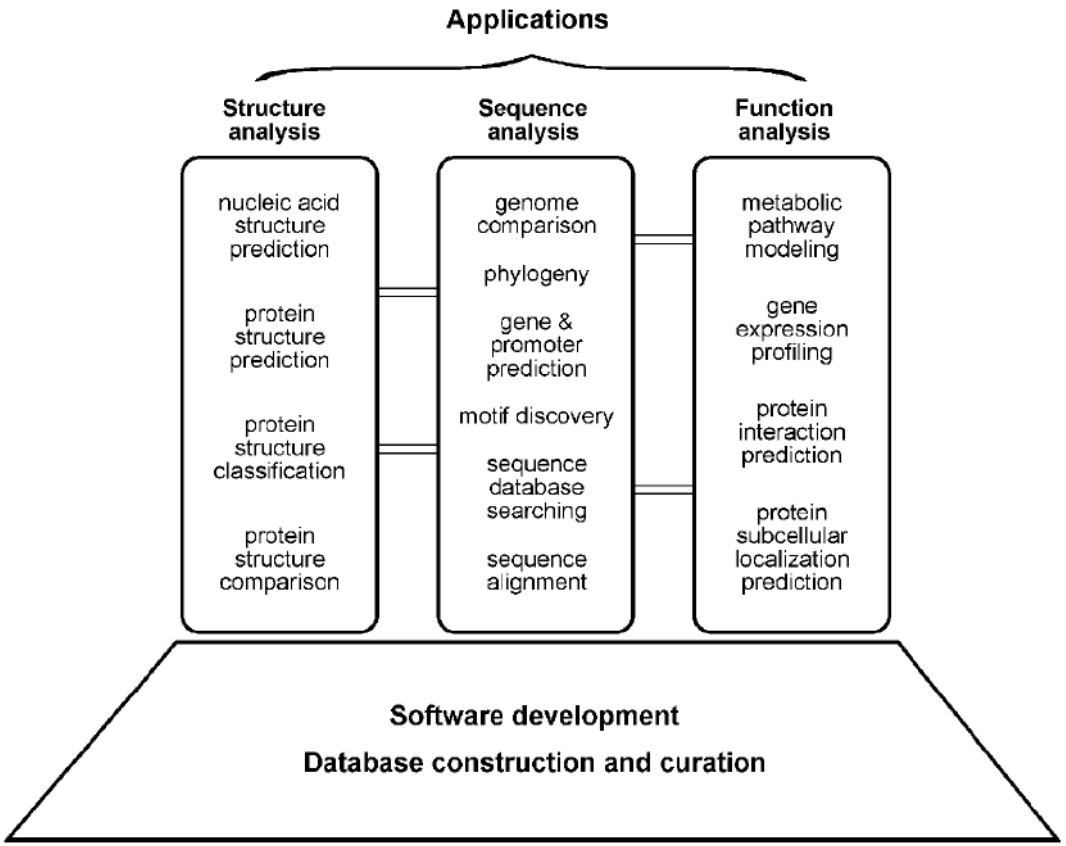


Figure 1.1: Overview of various subfields of bioinformatics. Biocomputing tool development is at the foundation of all bioinformatics analysis. The applications of the tools fall into three areas: sequence analysis, structure analysis, and function analysis. There are intrinsic connections between different 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, proteinová krystalografie, NMR spektroskopie)** koncem minulého století vedl k **obrovskému** nárůstu množství biologických dat.
- **Nutnost organizovaného ukládání, skladování a manipulace s velkým množstvím dat vedla ke vzniku bioinformatiky.**

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

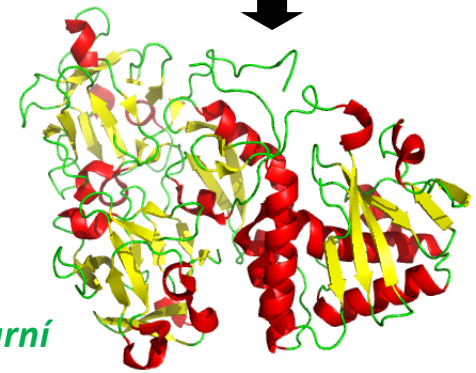
```
EDRPIKFSTEGATSQSYKQFIEALRERLRGGLIHDIPVLPDPTTLQERNRYIT
VELSNSDTESIEVGIDVTNAYVVAYRAGTQSYFLRDAPSSASDYLFTGTDQHS
LPFYGTYGDLERWAHQSRQQIPLGLQALTHGISFFRSGGNDNEEKARTLIVII
QMVAEAARFRYISNRVRVSIQTGTAFQPDAAMISLENNWDNLSRGVQESVQDT
FPNQVTLTNIRNEPVIVDSLSHPTVAVLALMLFVCNPPNIVEKSKICSSRYEP
TVRIGGRDGMCVDVYDNGYHNGNRIIMWKCKDRLEENQLWTLKSDKTIRSNGK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

Ribosome-inactivating protein, subdomain 1

Ribosome-inactivating protein, subdomain 2

Ricin B-like lectins

Sekundární



Strukturní

Specializované

Univerzální

Databáze

Nucleic Acids Research

http://www.oxfordjournals.org/our_journals/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/.

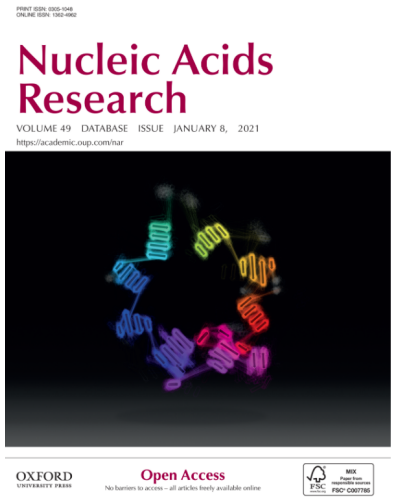
Among the major global centers, the NCBI (1) reports updates across many databases and interfaces. For example, gene searches can now cleverly retrieve orthologs from (subsets of) vertebrates. The EBI paper (2) includes striking figures that illustrate the deep inter-connectedness of its hosted databases, 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

Databáze

Nucleic Acids Research

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



2021: 1641 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 2021 *Nucleic Acids Research* database issue and the online molecular biology database collection

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

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

ABSTRACT

The 2021 *Nucleic Acids Research* database Issue contains 189 papers spanning a wide range of biological fields and investigation. It includes 89 papers reporting on new databases and 90 covering recent changes to resources previously published in the Issue. A further ten are updates on databases most recently published elsewhere. Seven new databases focus on COVID-19 and SARS-CoV-2 and many others offer resources for studying the virus. Major returning nucleic acid databases include NONCODE, Rfam and RNAcentral. Protein family and domain databases include COG, Pfam, SMART and Panther. Protein structures are covered by RCSB PDB and dispersed proteins by PED and MobiDB. In metabolism and signalling, STRING, KEGG and WikiPathways are featured, along with returning KLIFS and new DKK and KinaseMD, all focused on kinases. IMG/M and IMG/VR update in the microbial and viral genome resources section, while human and model organism genomics resources include Flybase, Ensembl and UCSC Genome Browser. Cancer studies are covered by updates from canSAR and PINA, as well as newcomers CNCdatabase and Oncovar for cancer drivers. Plant comparative genomics is catered for by updates from Gramene and GreenPhylDB. 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 substantially updated, revisiting nearly 1000 entries, adding 90 new resources and eliminating 86 obsolete databases, bringing the current total to 1641 databases. It is available at https://www.oxfordjournals.org/nar/database/c/.

NEW AND UPDATED DATABASES

The 28th annual *Nucleic Acids Research* Database Issue contains 189 papers spanning, as usual, a wide range of biology. Unsurprisingly, COVID-19 casts a long shadow over the Issue. Seven new databases specifically address the pandemic and the SARS-CoV-2 virus responsible (Table 1) but new and returning databases in all areas have rushed to support research into the viral pandemic: the reader will find reference to it throughout the Issue, sometimes in quite unexpected places. The Issue contains a further 82 papers (Table 2) on new databases as well as 90 update papers on databases previously published in NAR. To complete the Issue, resources previously published elsewhere update in a further 10 papers (Table 3).

As is customary, the Issue starts 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 (1–3). Thereafter, the usual categorisation applies: (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. Many resources are not easily pigeon-holed so browsing of the whole Issue is strongly encouraged.

The COVID-19 papers span a number of sections clearly indicating the multidisciplinary nature of the huge scientific response to the pandemic. Navigating the deluge of COVID-19 papers is a significant challenge in its own right and one addressed by the NCBI's LitCovid database (4) which features manual curation supported by sophisticated machine-learning assistance. SARS-CoV-2 nucleic acid sequence data and associated curated metadata can be conveniently obtained from the ViruSurf database (5) which also covers other human pathogenic viruses. SARS-Cov-2 comparative genomics is covered by the GESS database (6) where temporal and geographical patterns of SNVs can be analysed. SARS-Cov-2 protein structures – alone and in complex with antibodies, receptors, and small molecules – are collected at the CoV3D database (7) and made available with a variety of bespoke analyses of sequential and conformational diversity. Obviously, drug and vaccine de-

https://academic.oup.com/nar/issue/49/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

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

ENA

NCBI

Národní centrum pro biotechnologické informace

NCBI
National Center for Biotechnology Information

National Center for Biotechnology Information

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

GenBank

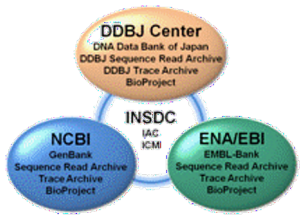
DDBJ Center

DDBJ
DNA Data Bank of Japan

The DNA Data Bank of Japan Center

http://www.ddbj.nig.ac.jp/

DDBJ



Primární databáze nukleových kyselin

- **ENA** – **Evropský institut pro bioinformatiku**

- **GenBank** – **Národní centrum pro biotechnologické informace**

- **DDBJ** – **Národní genetický institut (NIG)**

Formát ENA databáze

```
ID - identification              (begins each entry; 1 per entry)
AC - accession number            (>=1 per entry)
PR - project identifier          (0 or 1 per entry)
DT - date                        (2 per entry)
DE - description                 (>=1 per entry)
KW - keyword                     (>=1 per entry)
OS - organism species            (>=1 per entry)
OC - organism classification     (>=1 per entry)
OG - organelle                   (0 or 1 per entry)
RN - reference number            (>=1 per entry)
RC - reference comment           (>=0 per entry)
RP - reference positions         (>=1 per entry)
RX - reference cross-reference   (>=0 per entry)
RG - reference group             (>=0 per entry)
RA - reference author(s)         (>=0 per entry)
RT - reference title             (>=1 per entry)
RL - reference location          (>=1 per entry)
DR - database cross-reference    (>=0 per entry)
CC - comments or notes           (>=0 per entry)
AH - assembly header             (0 or 1 per entry)
AS - assembly information        (0 or >=1 per entry)
FH - feature table header        (2 per entry)
FT - feature table data          (>=2 per entry)
XX - spacer line                 (many per entry)
SQ - sequence header             (1 per entry)
CO - contig/construct line       (0 or >=1 per entry)
bb - (blanks) sequence data      (>=1 per entry)
// - termination line            (ends each entry; 1 per entry)
```

```
3.4.1  The ID Line
The ID (IDentification) line is always the first line of an entry. The
format of the ID line is:
ID   <1>; SV <2>; <3>; <4>; <5>; <6>; <7> BP.
The tokens represent:
   1. Primary accession number
   2. Sequence version number
   3. Topology: 'circular' or 'linear'
   4. Molecule type (see note 1 below)
   5. Data class (see section 3.1)
   6. Taxonomic division (see section 3.2)
   7. Sequence length (see note 2 below)
```

```
ID   CD789012; SV 4; linear; genomic DNA; HTG; MAM; 500 BP.
```

ftp://ftp.ebi.ac.uk/pub/databases/embl/doc/usrman.txt

Formát ENA databáze

3.1 Data Class

The data class of each entry, representing a methodological approach to the generation of the data or a type of data, is indicated on the first (ID) line of the entry. Each entry belongs to exactly one data class.

Class	Definition
CON	Entry constructed from segment entry sequences; if unannotated, annotation may be drawn from segment entries
PAT	Patent
EST	Expressed Sequence Tag
GSS	Genome Survey Sequence
HTC	High Thoughput CDNA sequencing
HTG	High Thoughput Genome sequencing
MGA	Mass Genome Annotation
WGS	Whole Genome Shotgun
TSA	Transcriptome Shotgun Assembly
STS	Sequence Tagged Site
STD	Standard (all entries not classified as above)

Division	Code
Bacteriophage	PHG
Environmental Sample	ENV
Fungal	FUN
Human	HUM
Invertebrate	INV
Other Mammal	MAM
Other Vertebrate	VRT
Mus musculus	MUS
Plant	PLN
Prokaryote	PRO
Other Rodent	ROD
Synthetic	SYN
Transgenic	TGN
Unclassified	UNC
Viral	VRL

ftp://ftp.ebi.ac.uk/pub/databases/embl/doc/usrman.txt

EMBL (ENA)
„entry“

```
ID   X56734; SV 1; linear; mRNA; STD; PLN; 1859 BP.
XX
AC   X56734; S46826;
XX
DT   12-SEP-1991 (Rel. 29, Created)
DT   25-NOV-2005 (Rel. 85, Last updated, Version 11)
XX
DE   Trifolium repens mRNA for non-cyanogenic beta-glucosidase
XX
KW   beta-glucosidase.
XX
OS   Trifolium repens (white clover)
OC   Eukaryota; Viridiplantae; Streptophyta; Embryophyta; Tracheophyta;
OC   Spermatophyta; Magnoliophyta; eudicotyledons; core eudicotyledons; rosids;
OC   eurosids I; Fabales; Fabaceae; Papilionoideae; Trifolieae; Trifolium.
XX
RN   [5]
RP   1-1859
RX   PUBMED; 1907511.
RA   Oxtoby E., Dunn M.A., Pancoro A., Hughes M.A.;
RT   "Nucleotide and derived amino acid sequence of the cyanogenic
RT   beta-glucosidase (linamarase) from white clover (Trifolium repens L.)";
RL   Plant Mol. Biol. 17(2):209-219(1991).
XX
RN   [6]
RP   1-1859
RA   Hughes M.A.;
RT   ;
RL   Submitted (19-NOV-1990) to the EMBL/GenBank/DDBJ databases.
RL   Hughes M.A., University of Newcastle Upon Tyne, Medical School, Newcastle
RL   Upon Tyne, NE2 4HH, UK
```

Translation =
***proteinová* databáze**

```
FT   source          1..1859
FT                   /organism="Trifolium repens"
FT                   /mol_type="mRNA"
FT                   /clone_lib="lambda gt10"
FT                   /clone="TRE361"
FT                   /tissue_type="leaves"
FT                   /db_xref="taxon:3899"
FT   CDS             14..1495
FT                   /product="beta-glucosidase"
FT                   /EC_number="3.2.1.21"
FT                   /note="non-cyanogenic"
FT                   /db_xref="GOA:P26204"
FT                   /db_xref="HSSP:P26205"
FT                   /db_xref="InterPro:IPR001360"
FT                   /db_xref="UniProtKB/Swiss-Prot:P26204"
FT                   /protein_id="CAA40058.1"
FT                   /translation="MDFIVAIFALFVISSFTITSTNAVEASTLLDIGNLSRSSFPRGFI
FT                   FGAGSSAYQFEGAVNEGGRGPSIWDTFTHKYPEKIRDGSNADITVDQYHRYKEDVGIMK
FT                   DQNMDSYRFSISWPRILPKGKLSGGINHEGIKYYNNLINELLANGIQPFVTLFHWDLPQ
FT                   VLEDEYGGFLNSGVINDFRDYTDLCFKEFGDRVRYWSTLNEPWVFSNSGYALGTNAPGR
FT                   CSASNVAKPGDSGTGPYIVTHNQILAHAEAVHVYKTKYQAYQKGKIGITLVSNWLMPLD
FT                   DNSIPDIKAAERSLDFQFGLFMEQLITGDYSKSMRRIVKNRLPKFSKFESSLVNGSFDF
FT                   IGINYYSSSYISNAPSHGNAKPSYSTNPMTNISFEKHGIPLGPRAASIWIYVYPYMFIQ
FT                   EDFEIFCYILKINITILQFSITENGMNEFNDATLPVEEALLNTYRIDYYYRHLYYIRSA
FT                   IRAGSNVKGFYAWSFLDCNEWFAGFTVRFGLNFVD"
FT   mRNA            1..1859
FT                   /experiment="experimental evidence, no additional details
FT                   recorded"
XX
SQ   Sequence 1859 BP; 609 A; 314 C; 355 G; 581 T; 0 other;
aaacaaacca aatatggatt ttattgtagc catatttgct ctgtttgtta ttagctcatt        60
cacaattact tccacaaatg cagttgaagc ttctactctt cttgacatag gtaacctgag       120
tcggagcagt tttcctcgtg gcttcatctt tggtgctgga tcttcagcat accaatttga       180
aggtgcagta aacgaaggcg gtagaggacc aagtatttgg gataccttca cccataaata       240
tccagaaaaa ataagggatg gaagcaatgc agacatcacg gttgaccaat atcaccgcta       300
caaggaagat gttgggatta tgaaggatca aaatatggat tcgtatagat tctcaatctc       360
ttggccaaga atactcccaa agggaaagtt gagcggaggc ataaatcacg aaggaatcaa       420
```

http://www.insdc.org/documents/feature-table#7.1.1

ENA

```
ID   X64011; SV 1; linear; genomic DNA; STD; PRO; 756 BP.
XX
AC   X64011; S78972;
XX
SV   X64011.1
XX
DT   28-APR-1992 (Rel. 31, Created)
DT   30-JUN-1993 (Rel. 36, Last updated, Version 6)
XX
DE   Listeria ivanovii sod gene for superoxide dismutase
XX
KW   sod gene; superoxide dismutase.
XX
OS   Listeria ivanovii
OC   Bacteria; Firmicutes; Bacillus/Clostridium group;
OC   Bacillus/Staphylococcus group; Listeria.
XX
RN   [1]
RX   MEDLINE; 92140371.
RA   Haas A., Goebel W.;
RT   "Cloning of a superoxide dismutase gene from Listeria ivanovii by
RT   functional complementation in Escherichia coli and characterization of the
RT   gene product.";
RL   Mol. Gen. Genet. 231:313-322(1992).
XX
RN   [2]
RP   1-756
RA   Kreft J.;
RT   ;
RL   Submitted (21-APR-1992) to the EMBL/GenBank/DDBJ databases.
RL   J. Kreft, Institut f. Mikrobiologie, Universitaet Wuerzburg, Biozentrum Am
RL   Hubland, 8700 Wuerzburg, FRG
XX
FH   Key             Location/Qualifiers
FH
FT   source          1..756
FT                   /db_xref="taxon:1638"
FT                   /organism="Listeria ivanovii"
FT                   /strain="ATCC 19119"
FT                   /mol_type="genomic DNA"
FT   regulatory      95..100
FT                   /gene="sod"
FT                   /regulatory_class="ribosome_binding_site"
FT   regulatory      723..746
FT                   /gene="sod"
FT                   /regulatory_class="terminator"
FT   CDS             109..717
FT                   /transl_table=11
FT                   /gene="sod"
FT                   /EC_number="1.15.1.1"
FT                   /db_xref="GOA:P28763"
FT                   /db_xref="HSSP:P00448"
FT                   /db_xref="InterPro:IPR001189"
FT                   /db_xref="UniProtKB/Swiss-Prot:P28763"
FT                   /product="superoxide dismutase"
FT                   /protein_id="CAA45406.1"
FT                   /translation="MTYELPKLPYTYDALEPNFDKETMEIHYTKHHNIYVTKLNEAVSG
FT                   HAELASKPGEELVANLDSVPEEIRGAVRNHGGGHANHTLFWSSLSPNGGGAPTGNLKAA
FT                   IESEFGTFDEFKEKFNAAAAARFGSGWAWLVVNNGKLEIVSTANQDSPLSEGKTPVLGL
FT                   DVWEHAYYLKFQNRRPEYIDTFWNVINWDERNKRFDAAK"
XX
SQ   Sequence 756 BP; 247 A; 136 C; 151 G; 222 T; 0 other;
     cgttatttaa ggtgttacat agttctatgg aaatagggtc tatacctttc gccttacaat   60
     gtaatttctt .........                                               120
//
```

GenBank

```
LOCUS       LISOD                   756 bp    DNA     linear   BCT 30-JUN-1993
DEFINITION  Listeria ivanovii sod gene for superoxide dismutase.
ACCESSION   X64011 S78972
VERSION     X64011.1  GI:44010
KEYWORDS    sod gene; superoxide dismutase.
SOURCE      Listeria ivanovii
  ORGANISM  Listeria ivanovii
            Bacteria; Firmicutes; Bacillales; Listeriaceae; Listeria.
REFERENCE   1  (bases 1 to 756)
  AUTHORS   Haas,A. and Goebel,W.
  TITLE     Cloning of a superoxide dismutase gene from Listeria ivanovii by
            functional complementation in Escherichia coli and characterization
            of the gene product
  JOURNAL   Mol. Gen. Genet. 231 (2), 313-322 (1992)
  MEDLINE   92140371
REFERENCE   2  (bases 1 to 756)
  AUTHORS   Kreft,J.
  TITLE     Direct Submission
  JOURNAL   Submitted (21-APR-1992) J. Kreft, Institut f. Mikrobiologie,
            Universitaet Wuerzburg, Biozentrum Am Hubland, 8700 Wuerzburg, FRG
FEATURES             Location/Qualifiers
     source          1..756
                     /organism="Listeria ivanovii"
                     /strain="ATCC 19119"
                     /db_xref="taxon:1638"
                     /mol_type="genomic DNA"
     regulatory      95..100
                     /gene="sod"
                     /regulatory_class="ribosome_binding_site"
     gene            95..746
                     /gene="sod"
     CDS             109..717
                     /gene="sod"
                     /EC_number="1.15.1.1"
                     /codon_start=1
                     /transl_table=11
                     /product="superoxide dismutase"
                     /db_xref="GI:44011"
                     /db_xref="GOA:P28763"
                     /db_xref="InterPro:IPR001189"
                     /db_xref="UniProtKB/Swiss-Prot:P28763"
                     /protein_id="CAA45406.1"
                     /translation="MTYELPKLPYTYDALEPNFDKETMEIHYTKHHNIYVTKLNEAVS
                     GHAELASKPGEELVANLDSVPEEIRGAVRNHGGGHANHTLFWSSLSPNGGGAPTGNLK
                     AAIESEFGTFDEFKEKFNAAAAARFGSGWAWLVVNNGKLEIVSTANQDSPLSEGKTPV
                     LGLDVWEHAYYLKFQNRRPEYIDTFWNVINWDERNKRFDAAK"
     regulatory      723..746
                     /gene="sod"
                     /regulatory_class="terminator"
ORIGIN
        1 cgttatttaa ggtgttacat agttctatgg aaatagggtc tatacctttc gccttacaat
       61 gtaatttctt .........
//
```

DDBJ

```
LOCUS       LISOD                   756 bp    DNA     linear   BCT 30-JUN-1993
DEFINITION  Listeria ivanovii sod gene for superoxide dismutase.
ACCESSION   X64011 S78972
VERSION     X64011.1  GI:44010
KEYWORDS    sod gene; superoxide dismutase.
SOURCE      Listeria ivanovii
  ORGANISM  Listeria ivanovii
            Bacteria; Firmicutes; Bacillales; Listeriaceae; Listeria.
REFERENCE   1  (bases 1 to 756)
  AUTHORS   Haas,A. and Goebel,W.
  TITLE     Cloning of a superoxide dismutase gene from Listeria ivanovii by
            functional complementation in Escherichia coli and characterization
            of the gene product
  JOURNAL   Mol. Gen. Genet. 231 (2), 313-322 (1992)
  MEDLINE   92140371
REFERENCE   2  (bases 1 to 756)
  AUTHORS   Kreft,J.
  TITLE     Direct Submission
  JOURNAL   Submitted (21-APR-1992) J. Kreft, Institut f. Mikrobiologie,
            Universitaet Wuerzburg, Biozentrum Am Hubland, 8700 Wuerzburg, FRG
FEATURES             Location/Qualifiers
     source          1..756
                     /organism="Listeria ivanovii"
                     /strain="ATCC 19119"
                     /db_xref="taxon:1638"
                     /mol_type="genomic DNA"
     regulatory      95..100
                     /gene="sod"
                     /regulatory_class="ribosome_binding_site"
     gene            95..746
                     /gene="sod"
     CDS             109..717
                     /gene="sod"
                     /EC_number="1.15.1.1"
                     /codon_start=1
                     /transl_table=11
                     /product="superoxide dismutase"
                     /db_xref="GOA:P28763"
                     /db_xref="HSSP:P00448"
                     /db_xref="InterPro:IPR001189"
                     /db_xref="UniProtKB/Swiss-Prot:P28763"
                     /protein_id="CAA45406.1"
                     /translation="MTYELPKLPYTYDALEPNFDKETMEIHYTKHHNIYVTKLNEAVS
                     GHAELASKPGEELVANLDSVPEEIRGAVRNHGGGHANHTLFWSSLSPNGGGAPTGNLK
                     AAIESEFGTFDEFKEKFNAAAAARFGSGWAWLVVNNGKLEIVSTANQDSPLSEGKTPV
                     LGLDVWEHAYYLKFQNRRPEYIDTFWNVINWDERNKRFDAAK"
     regulatory      723..746
                     /gene="sod"
                     /regulatory_class="terminator"
BASE COUNT          247 a          136 c          151 g          222 t
ORIGIN
        1 cgttatttaa ggtgttacat agttctatgg aaatagggtc tatacctttc gccttacaat
       61 gtaatttctt .........
//
```

Sekundární databáze NA

Sekundární databáze: informace odvozené z primárních databází ve formě charakteristických vzorů sekvencí, tj. funkčních nebo strukturních motivů získaných srovnáním primárních dat (sekvencí).

- **Sekundární databáze NA**

TRANSFAC – databáze eukaryotických transkripčních faktorů, jejich vazebných míst a DNA profilů

JASPAR – databáze eukaryotických transkripčních faktorů, jejich vazebných míst a DNA profilů

http://genexplain.com/transfac/

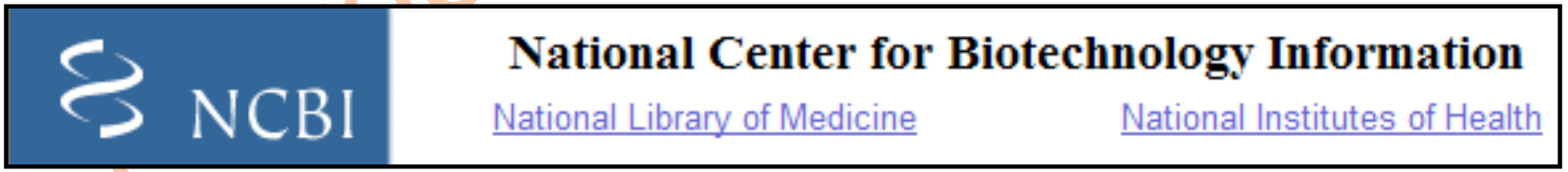
Primární databáze proteinů

- **Univerzální databáze:**

  „Skladiště“ sekvencí – sequence repository

  Manuálně spravovaná – curated database

Příklad: **GenBank *versus* RefSeq**



Primární databáze proteinů

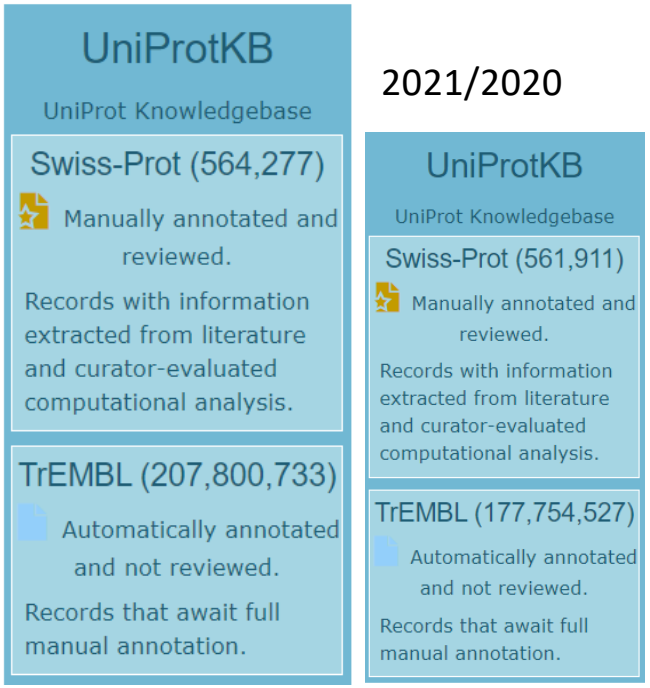
GenBank	RefSeq
Not curated	Curated
Author submits	NCBI creates from existing data
Only author can revise	NCBI revises as new data emerge
Multiple records for same loci common	Single records for each molecule of major organisms
Records can contradict each other	
No limit to species included	Limited to model organisms
Data exchanged among INSDC members	Exclusive NCBI database
Akin to primary literature	Akin to review articles
Proteins identified and linked	Proteins and transcripts identified and linked
Access via NCBI Nucleotide databases	Access via Nucleotide & Protein databases

Swiss-PROT + TrEMBL

- **Swiss-Prot –** „Curated“ databáze založená na Univerzitě v Ženevě v roce 1986. Spravovaná Švýcarským institutem pro bioinformatiku **(SIB - Swiss Institute of Bioinformatics).**

- **Vysoká úroveň anotace**

- **TrEMBL –** Počítačově anotovaná data, odvozená z kódujících úseku sekvencí v DDBJ/EMBL/GenBank, která **ZATÍM** nejsou zařazena v Swiss-Prot.



Swiss-PROT + TrEMBL

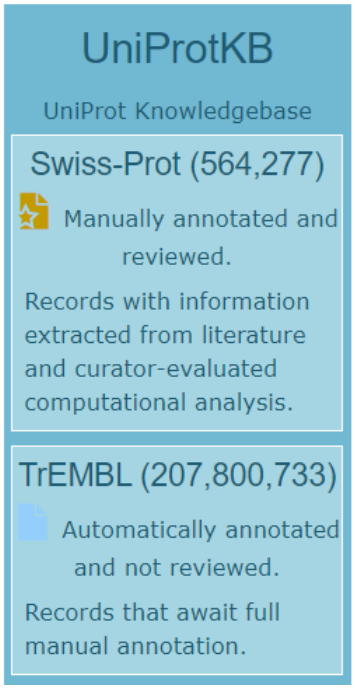
- **Anotace:** Funkce
  Katalytická aktivita
  Podjednotky
  Domény
  Biotechnologické využití
  Sekvenční homologie
  Posttranslační modifikace
  Reference
  atd.

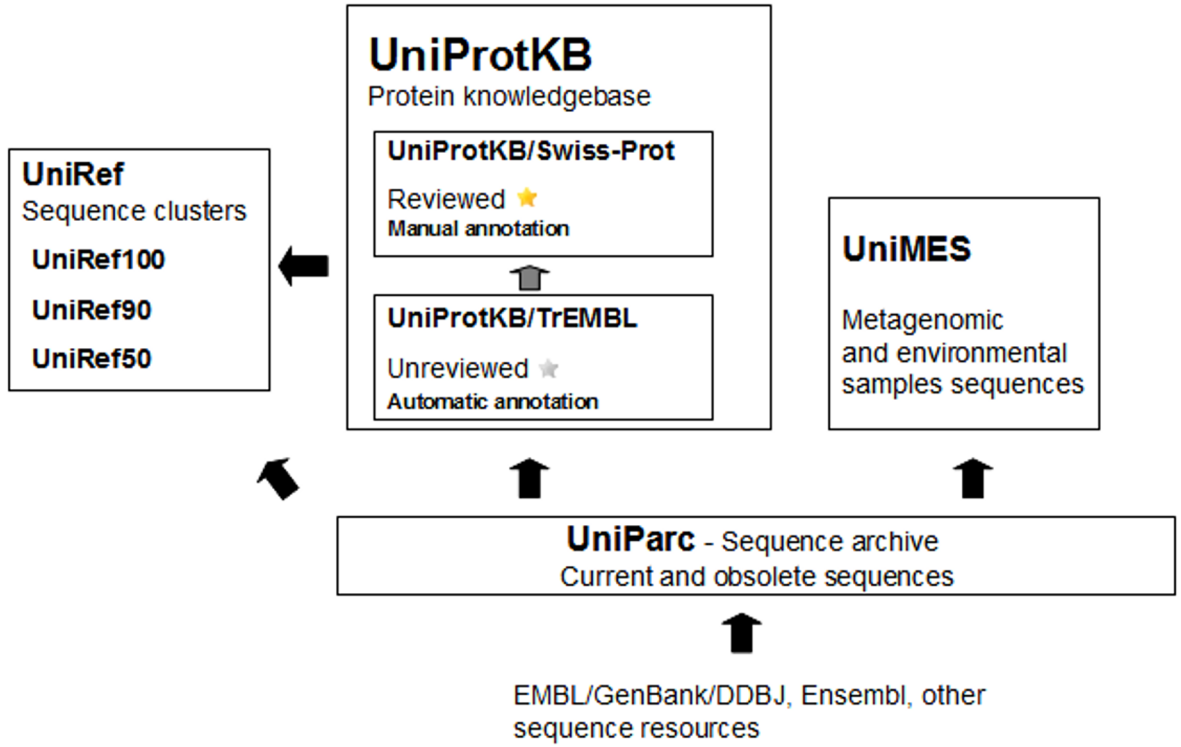
http://www.expasy.org/sprot/

UniProt

2002- spolupráce mezi EBI, SIB a PIR

http:/www.uniprot.org





UniProt

UniProtKB

UniProt Knowledgebase

Swiss-Prot (563,552)

Manually annotated and reviewed.

Records with information extracted from literature and curator-evaluated computational analysis.

- curated by experts
- data from scientific papers
- annotation of sequence features
- collates isoforms in one entry

TrEMBL (195,104,019)

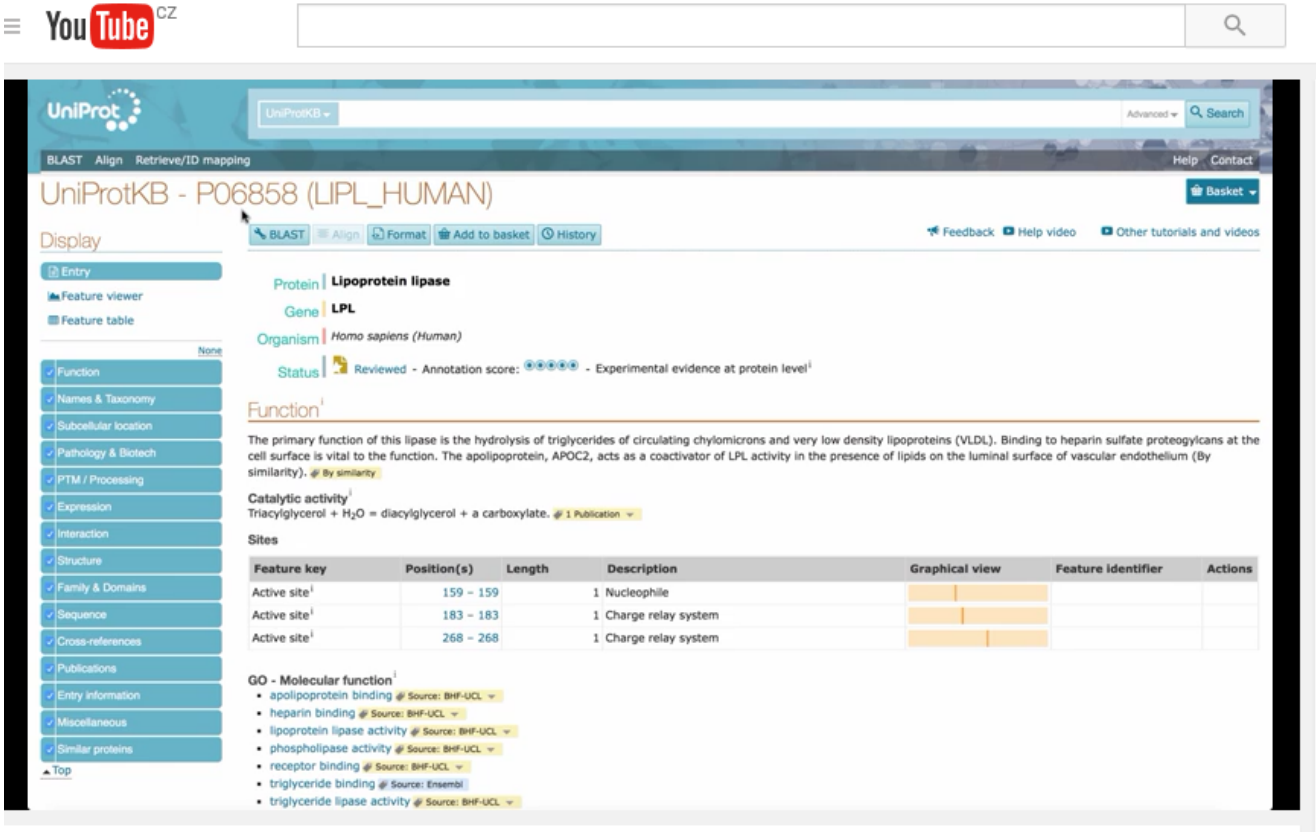
Automatically annotated and not reviewed.

Records that await full manual annotation.

- annotation from rule systems (incl. expert-curated rules)
- mapped experimental sequence features (3D structures)
- isoforms are kept separate

EMBL-EBI PIR SIB

UniProt



https://www.youtube.com/watch?v=x9GNm2DLP-U

Biologické databáze - problémy

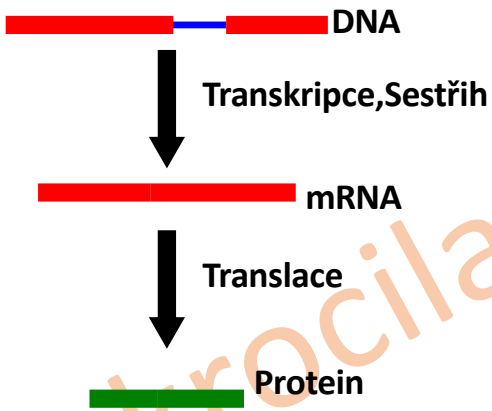
One of the problems associated with biological databases is overreliance on sequence information and related annotations, without understanding the reliability of the information. What is often ignored is the fact that there are many errors in sequence databases. There are also high levels of redundancy in the primary sequence databases. Annotations of genes can also occasionally be false or incomplete. All these types of errors can be passed on to other databases, causing propagation of errors.

ESSENTIAL BIOINFORMATICS,
Jin Xiong, 2006

- Většina chyb v nukleotidových sekvencích pochází již z vlastní **sekvenace** (častější pro sekvence získané cca před rokem 1990). Chyby v nukleotidových sekvencích vedou k chybné translaci do proteinu nebo ji úplně znemožní.
- **Redundance** dat může extrémně zvětšit velikost databáze a vede k problémům při vyhledávání. Lze řešit vytvořením specializovaných databází s vysokou úrovní kontroly.
- **Chybná anotace** – jedna sekvence označena různými názvy, různé (nesouvisející) sekvence mohou mít stejný název. Zdroje chyb: překlepy, nepozornost, čistě hloupost, skutečné neshody mezi odborníky v daném oboru.
- Mnoho informací je pouze **PREDIKOVÁNO** (s využitím bioinformatiky). **Je nutné vyvarovat se slepého spoléhání na informace uvedené v databázi!**
- **Chyby se mohou šířit – nové sekvence s neznámou funkcí jsou často anotovány na základě sekvenční podobnosti s již existujícími záznamy v databázi!** Chybná anotace může ovlivnit celou skupinu podobných sekvencí!

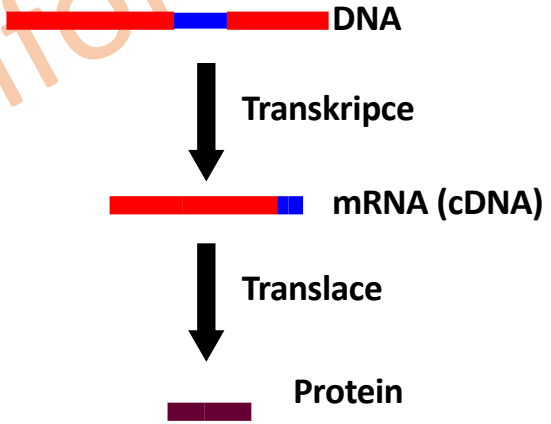
Predikce genů

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)

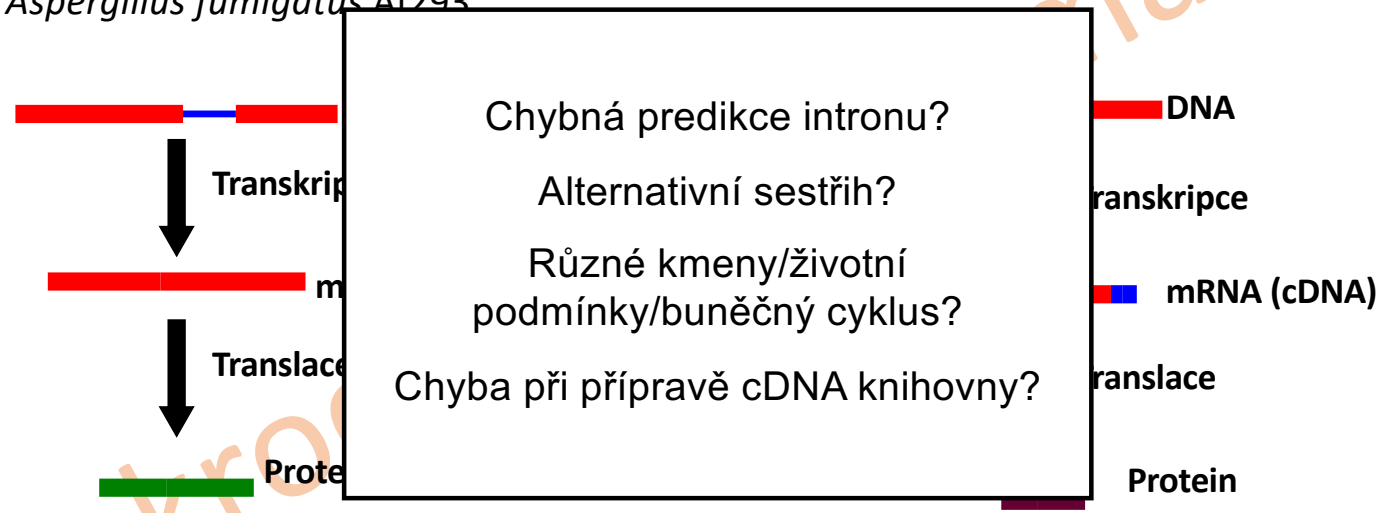


MADPEVEADG ELDLEKRASA QTCKIVNVDT
YVNCRYDAKL DAGAIFGFPK GEKLTFACWK
HGDCYNGV

Predikce genů

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

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



Chybná predikce intronu?

Alternativní sestřih?

Různé kmeny/životní podmínky/buněčný cyklus?

Chyba při přípravě cDNA knihovny?

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

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

Excel vs. genomika, 2004

BMC Bioinformatics

BioMed Central

Correspondence

Open Access

Mistaken Identifiers: Gene name errors can be introduced inadvertently when using Excel in bioinformatics

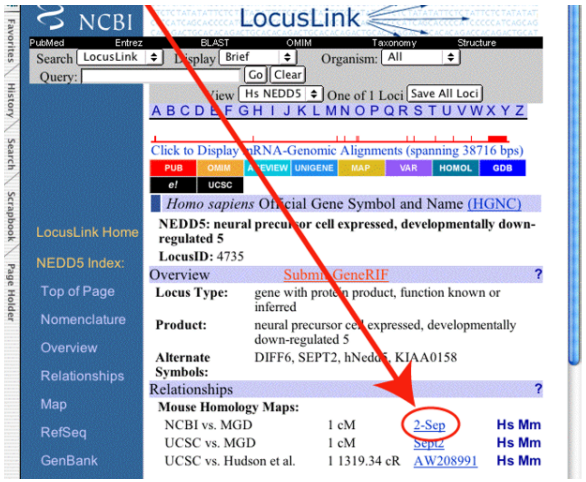
Barry R Zeeberg[†1], Joseph Riss[†2], David W Kane[3], Kimberly J Bussey[1], Edward Uchio[4], W Marston Linehan[4], J Carl Barrett[2] and John N Weinstein[*1]

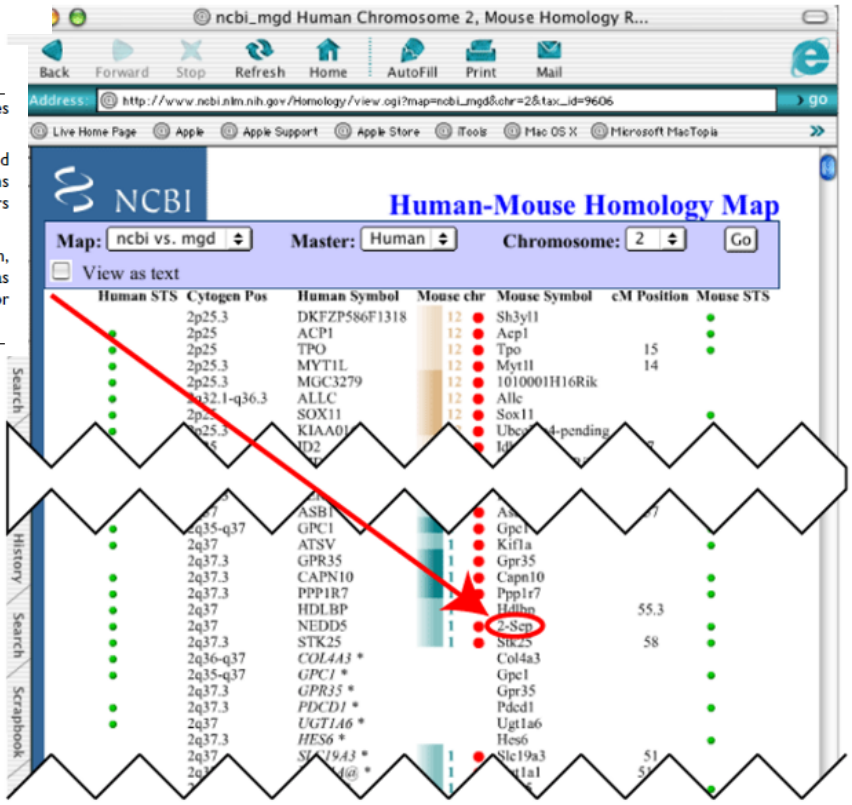
Abstract

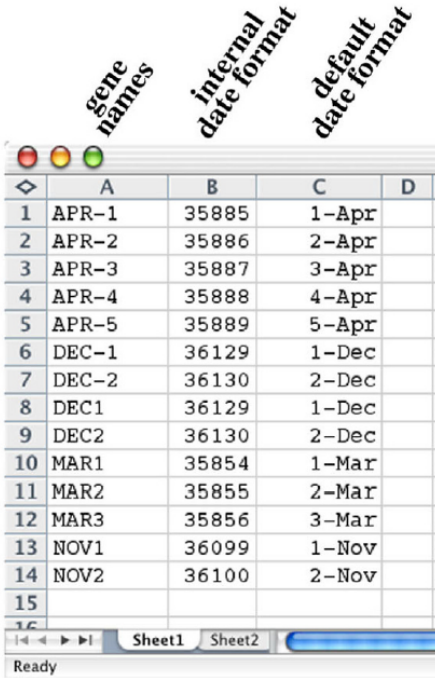
Background: When processing microarray data sets, we recently noticed that some gene names were being changed inadvertently to non-gene names.

Results: A little detective work traced the problem to default date format conversions and floating-point format conversions in the very useful Excel program package. The date conversions affect at least 30 gene names; the floating-point conversions affect at least 2,000 if Riken identifiers are included. These conversions are irreversible; the original gene names cannot be recovered.

Conclusions: Users of Excel for analyses involving gene names should be aware of this problem, which can cause genes, including medically important ones, to be lost from view and which has contaminated even carefully curated public databases. We provide work-arounds and scripts for circumventing the problem.







	A	B	C	D
1	APR-1	35885	1-Apr	
2	APR-2	35886	2-Apr	
3	APR-3	35887	3-Apr	
4	APR-4	35888	4-Apr	
5	APR-5	35889	5-Apr	
6	DEC-1	36129	1-Dec	
7	DEC-2	36130	2-Dec	
8	DEC1	36129	1-Dec	
9	DEC2	36130	2-Dec	
10	MAR1	35854	1-Mar	
11	MAR2	35855	2-Mar	
12	MAR3	35856	3-Mar	
13	NOV1	36099	1-Nov	
14	NOV2	36100	2-Nov	
15				

Excel vs. genomika, 12 let poté...

COMMENT Open Access

Gene name errors are widespread in the scientific literature

Mark Ziemann[1], Yotam Eren[1,2] and Assam El-Osta[1,3*]

Abstract

The spreadsheet software Microsoft Excel, when used with default settings, is known to convert gene names to dates and floating-point numbers. A programmatic scan of leading genomics journals reveals that approximately one-fifth of papers with supplementary Excel gene lists contain erroneous gene name conversions.

Keywords: Microsoft Excel, Gene symbol, Supplementary data

Abbreviations: GEO, Gene Expression Omnibus; JIF, journal impact factor

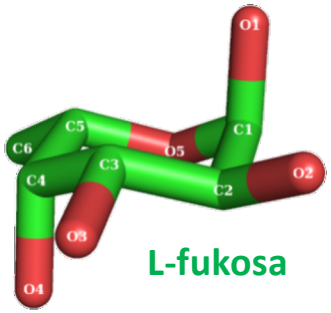
The problem of Excel software (Microsoft Corp., Redmond, WA, USA) inadvertently converting gene symbols to dates and floating-point numbers was originally described in 2004 [1]. For example, gene symbols such as *SEPT2* (Septin 2) and *MARCH1* [Membrane-Associated Ring Finger (C3HC4) 1, E3 Ubiquitin Protein Ligase] are converted by default to '2-Sep' and '1-Mar', respectively. Furthermore, RIKEN identifiers were described to be automatically converted to floating point numbers (i.e. from accession '2310009E13' to '2.31E+13'). Since that report, we have uncovered further instances where gene symbols were converted to dates in supplementary data of recently published papers (e.g. '*SEPT2*' converted to '2006/09/02'). This suggests that gene name errors continue to be a problem in supplementary files accompanying articles. Inadvertent gene symbol conversion is problematic because these supplementary files are an important resource in the genomics community that are frequently reused. Our aim here is to raise awareness of the problem.

Table 1 Results of the systematic screen of supplementary Excel files for gene name conversion errors

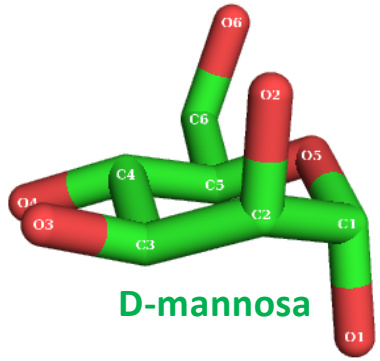
Journal[a]	Number of Excel files screened	Number of gene lists found	Number of papers with gene lists	Number of supplementary files affected	Number of papers affected	Number of gene names converted
PLoS One	7783	2202	994	220	170	4240
BMC Genomics	11464	1650	801	218	158	4932
Genome Res	2607	580	251	114	68	3180
Nucleic Acids Res	2117	540	315	88	67	1661
Genome Biol	2678	664	257	97	63	1878
Genes Dev	932	395	190	75	55	1593
Hum Mol Genet	980	372	168	48	27	1724
Nature	482	150	74	27	23	1375
BMC Bioinformatics	1790	235	152	26	21	534
RNA	569	127	77	20	15	1341
Nat Genet	264	70	37	12	9	178
Bioinformatics	731	112	67	11	6	339
PLoS Comput Biol	177	79	32	6	6	46
PLoS Biol	143	54	29	7	5	206
Mol Biol Evol	995	112	79	7	4	56
Science	172	36	19	7	3	451
Genome Biol Evol	490	32	25	2	2	121
DNA Res	801	57	30	2	2	6
Total	35175	7467	3597	987	704	23861

[a]The 18 journals investigated are ordered by the number of papers affected by gene name conversion errors

Cukr jako cukr?



L-fukosa



D-mannosa

MULTISPECIES: Fucose-binding lectin II [Burkholderia]

NCBI Reference Sequence: WP_014900522.1

Identical Proteins FASTA Graphics

Go to:

```
LOCUS       WP_014900522            129 aa            linear   BCT 15-JAN-2015
DEFINITION  MULTISPECIES: Fucose-binding lectin II [Burkholderia].
ACCESSION   WP_014900522
VERSION     WP_014900522.1
KEYWORDS    RefSeq.
SOURCE      Burkholderia
  ORGANISM  Burkholderia
            Bacteria; Proteobacteria; Betaproteobacteria; Burkholderiales;
            Burkholderiaceae.
COMMENT     REFSEQ: This record represents a single, non-redundant, protein
            sequence which may be annotated on many different RefSeq genomes
            from the same, or different, species.
FEATURES             Location/Qualifiers
     source          1..129
                     /organism="Burkholderia"
                     /db_xref="taxon:32008"
     Protein         1..129
                     /product="Fucose-binding lectin II"
                     /calculated_mol_wt=13768
     Region          14..128
                     /region_name="PA-IIL"
                     /note="Fucose-binding lectin II (PA-IIL); pfam07472"
                     /db_xref="CDD:284811"
ORIGIN
        1 madsqtssnr agefsippnt dfraiffana aeqqhiklfi gdsnepaayh klttrdgpre
       61 atlnsgngkl rfevtvngkt satdarlapi ngkksdgspf tvnfgivvse dghdsdyndg
      121 ivvlqwpig
//
```

UniProtKB - J7JBV3 (J7JBV3_BURCE)

Display

Entry
Publications
Feature viewer
Feature table

None
- Function
- Names & Taxonomy
- Subcellular location
- Pathology & Biotech
- PTM / Processing
- Expression
- Interaction
- Structure
- Family & Domains
- Sequence
- Similar proteins
- Cross-references
- Entry information
- Miscellaneous

Top

BLAST | Align | Format | Add to basket | History

Protein: Submitted name: **Fucose-binding lectin II**
Gene: **GEM_5383**
Organism: *Burkholderia cepacia GG4*
Status: Unreviewed - Annotation score: ●○○○○ - Protein predicted

Function

GO - Molecular function

- carbohydrate binding — Source: UniProtKB-KW

Complete GO annotation on QuickGO ...

Keywords

Ligand: Lectin — Imported

Enzyme and pathway databases

BioCyc: BCEP1009846:G1H9M-5526-MONOMER

Names & Taxonomy

Protein names: ~~Submitted name:~~ Fucose-binding lectin II — Imported
Gene names: ORF Names:GEM_5383
Organism: Burkholderia cepacia GG4
Taxonomic identifier: 1009846 [NCBI]
Taxonomic lineage: Bacteria › Proteobacteria › Betaproteobacteria › Burkholderiales › Burkholderiaceae › Burkholderia
Proteomes: UP000032866 Component: Chromosome 2

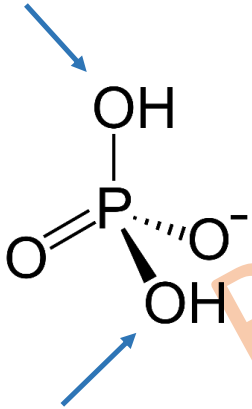
Automatic assertion inferred from database entries
EMBL:AFQ51767.1

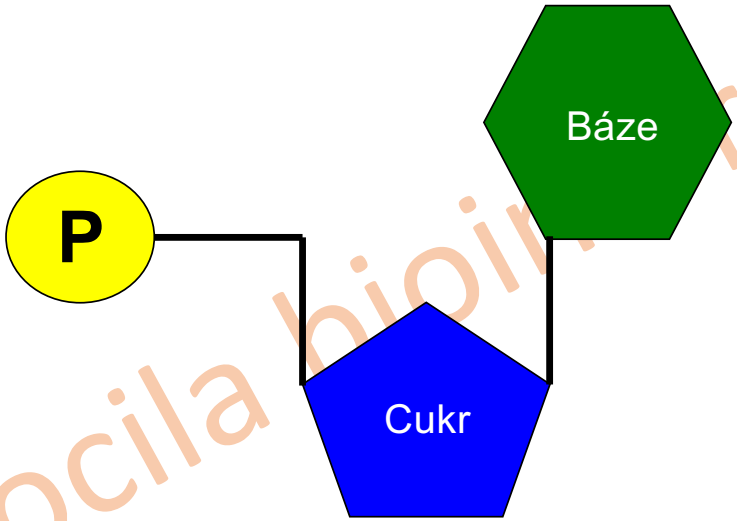
10) Sacharidy a lipidy. Struktura, význam a funkce. Bioinformatický potenciál sacharidů. Glykoproteiny, jejich kódování v genomu. Názvosloví a grafické znázornění. Databáze a nástroje pro glykobioinformatiku a lipidobioinformatiku.

Manipulace se sekvencemi: nukleové kyseliny, proteiny

Složení nukleových kyselin

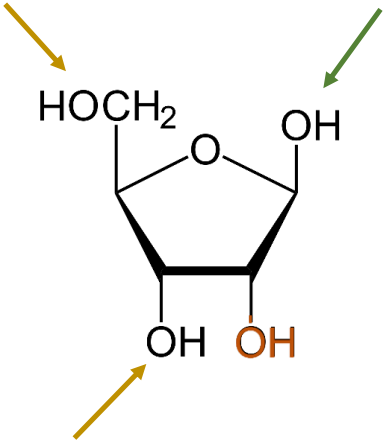
Ester kyseliny fosforečné
= fosfát





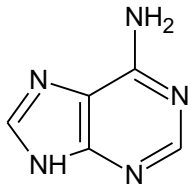
Nukleová báze
(obsahuje amino skupinu = je bazická)

Pentosa: D-ribosa (RNA)
2-deoxy-D-ribosa (DNA)

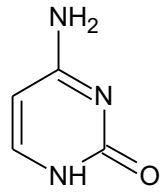


Nukleové báze

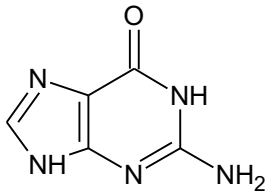
Jako základní součást nukleových kyselin



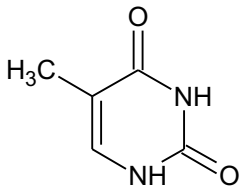
Adenine



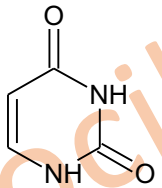
Cytosine



Guanine



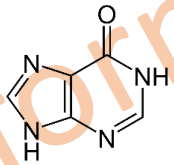
Thymine



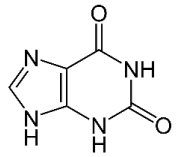
Uracil

adenin	cytosin	guanin	thymin	uracil
A	C	G	T	U

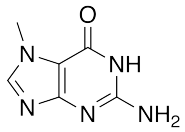
Součást nukleových kyselin (zejm. RNA) po chemické modifikaci



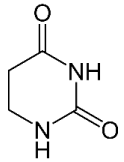
Hypoxanthine



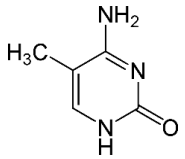
Xanthine



7-Methylguanine

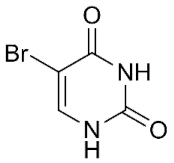


5,6-Dihydrouracil



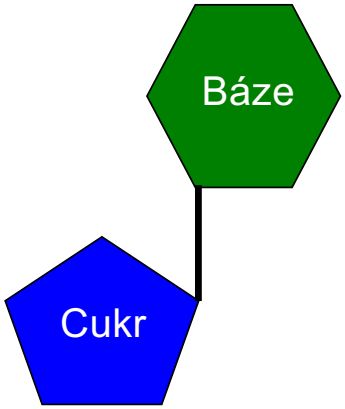
5-Methylcytosine

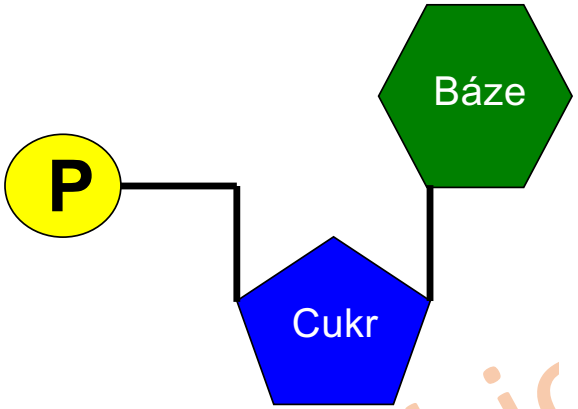
Syntetické



5-Bromouracil

Nukleosid x Nukleotid





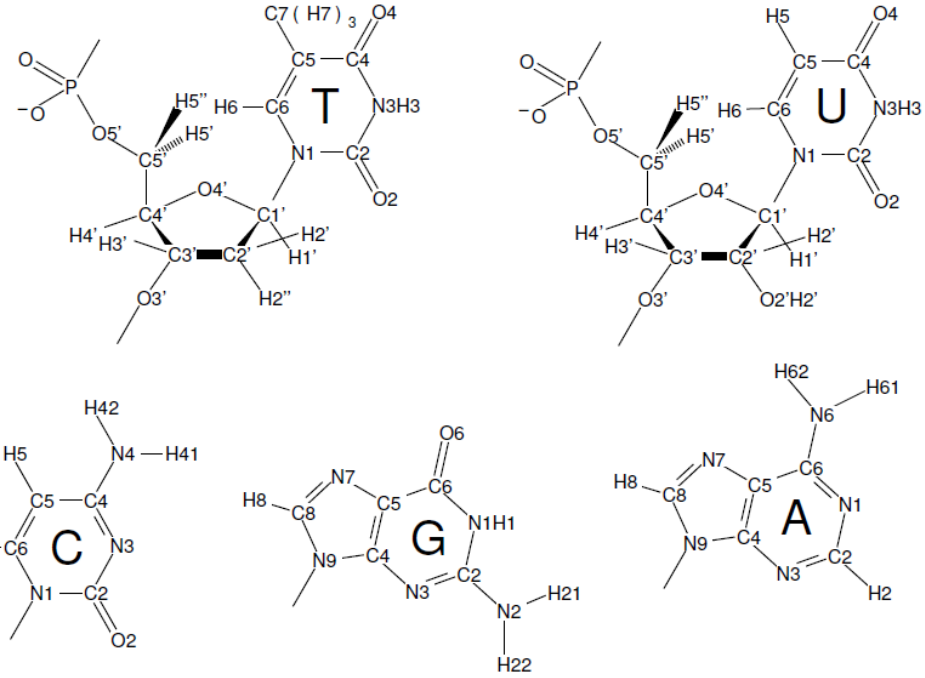
Cukr + báze = nukleosid

dA deoxyadenosin
dG deoxyguanosin
dC deoxycytidin
dT deoxythymidin
U uridin

Cukr + báze + fosfát = nukleotid

dAMP deoxyadenosinmonofosfát
dGMP deoxyguanosinmonofosfát
dCMP deoxycytidinmonofosfát
dTMP deoxythymidinmonofosfát
UMP uridinmonofosfát

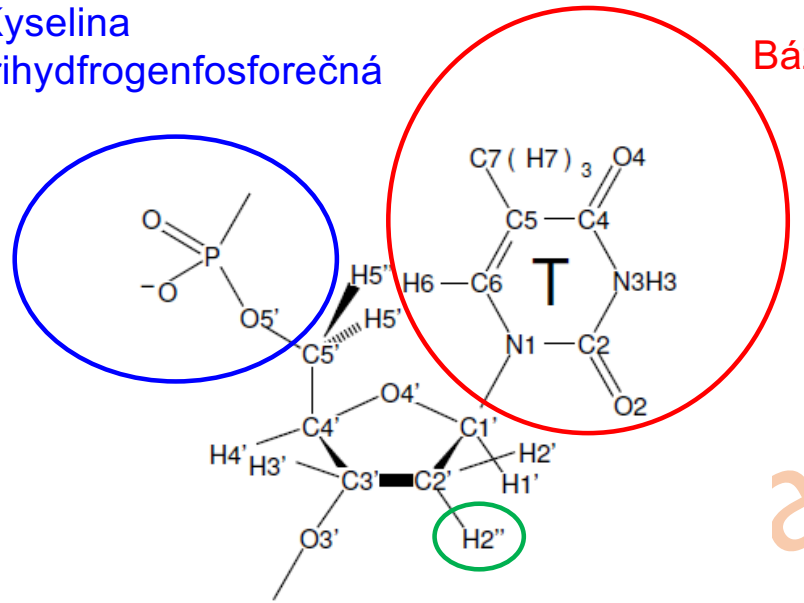
Číslování nukleotidů



Nukleotid – DNA

Kyselina
trihydfrogenfosforečná

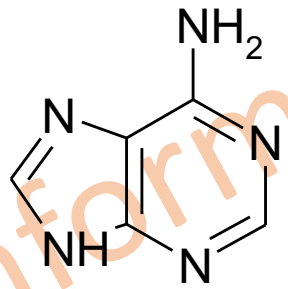
Báze



2-deoxy-β-D-ribosa

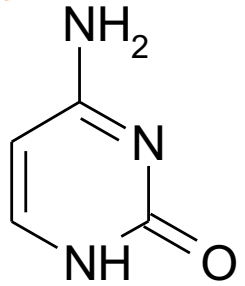
Deoxyribonukleotid

Purinové báze

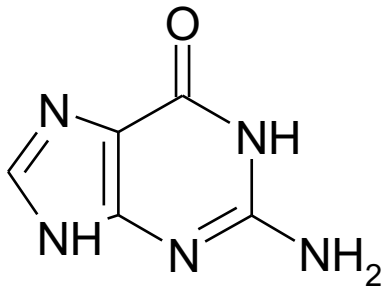


Adenin (ade), A

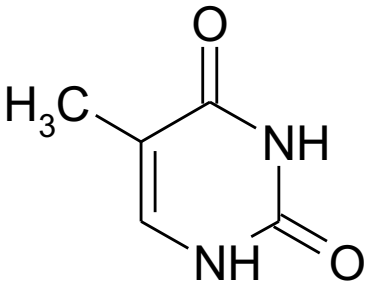
Pyrimidinové báze



Cytosin (cyt), C



Guanin (gua), G



Thymin (thy), T

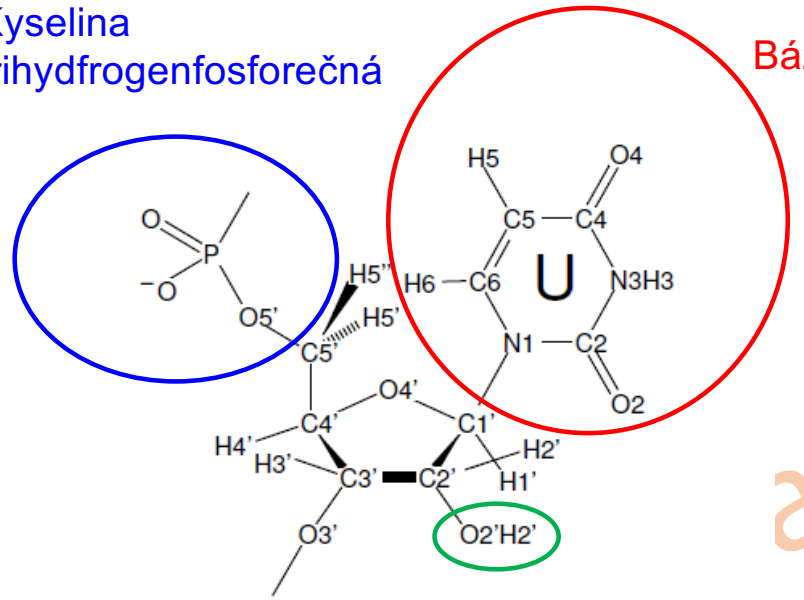
Lukáš Žídek

Skripta předmětu C9530 Strukturní biochemie

Nukleotid – RNA

Kyselina trihydfrogenfosforečná

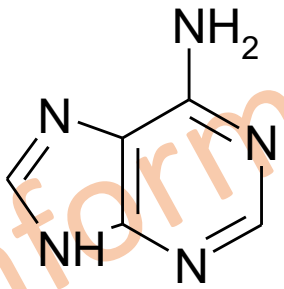
Báze



β-D-ribosa

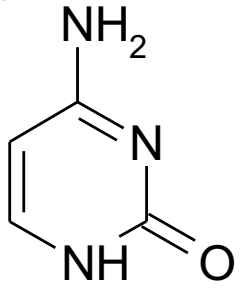
Ribonukleotid

Purinové báze

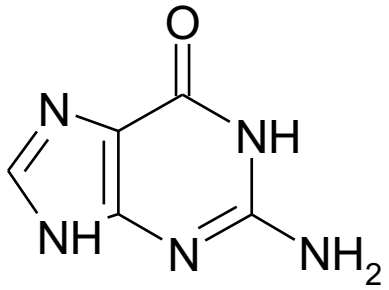


Adenin (ade), A

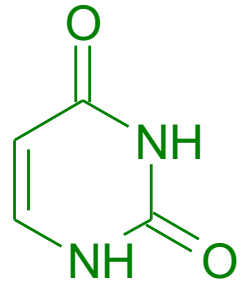
Pyrimidinové báze



Cytosin (cyt), C



Guanin (gua), G

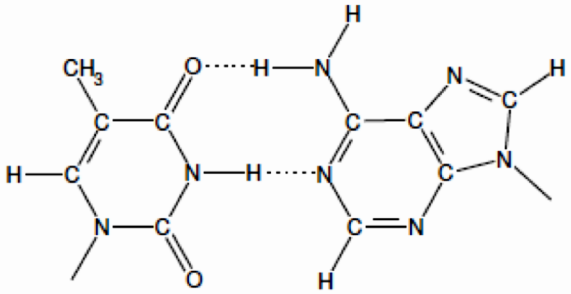


Uracil (ura), U

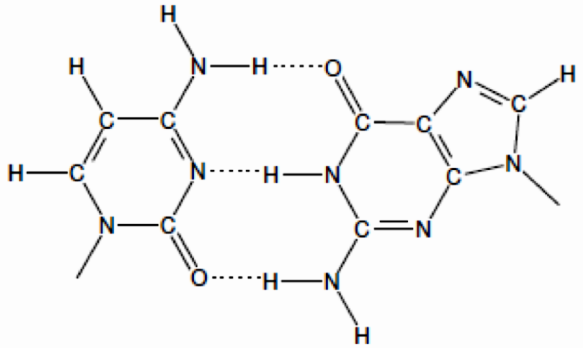
Lukáš Žídek

Skripta předmětu C9530 Strukturní biochemie

Párování bází



T = A

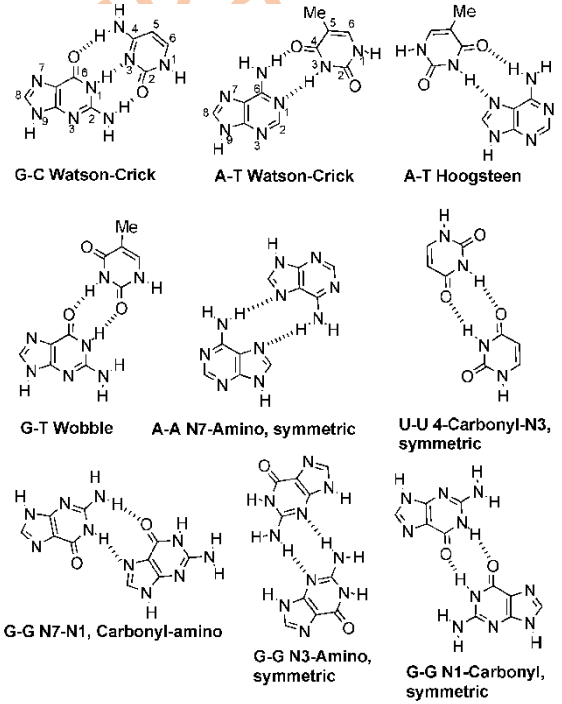


C ≡ G

Watsonovo-Crickovo párování bází (kanonické)

Základní

dsDNA, během transkripce při tvorbě RNA, dsRNA.



Nekanonické párování bází

Funkční RNA, specifické úseky DNA,...

Lukáš Žídek

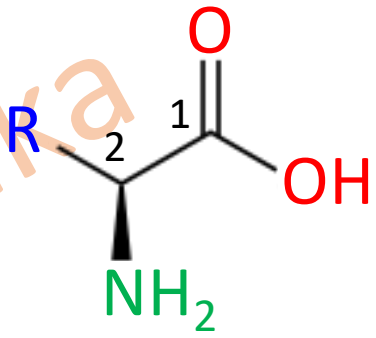
Skripta předmětu C9530 Strukturní biochemie

Nukleové kyseliny

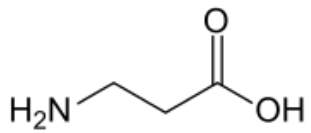
- 4 báze pro DNA
- 4 báze pro RNA (3 totožné)
- 6 zkratek pro dvoubázové kombinace
- 4 zkratky pro tříbázové kombinace
- 1 zkratka pro libovolnou bázi

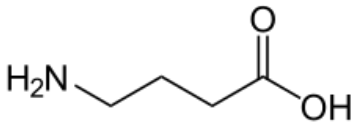
Adenin	Cytosin	Guanin	Thymin	Uracil	A, G	C, T	C, G	G, T	A, C	A, T	A, G, T	A, C, T	A, C, G	C, G, T	A, C, G, T
A	C	G	T	U	R	Y	S	K	M	W	D	H	V	B	N
					purinové	pyrimidinové	„strong“	„keto“	„amino“	„weak“	bez „C“	bez „G“	bez „T/U“	bez „A“	„any“

Aminokyseliny



- Organické látky obsahující:
  - NH_2 skupinu – amino
  - COOH skupinu – karboxylová kyselina
  - Případně další část – tzv. postranní/boční řetězec
- V proteinech se uplatňují tzv. **α-aminokyseliny**, tzn. mající aminoskupinu na druhém uhlíku
- V organismech mají jiné funkce i další aminokyseliny, např.:
  - β-alanin – součást funkčních peptidů, prekurzor vitaminu B5
  - γ-aminomáselná kyselina (GABA) – přenašeč nervových vzruchů

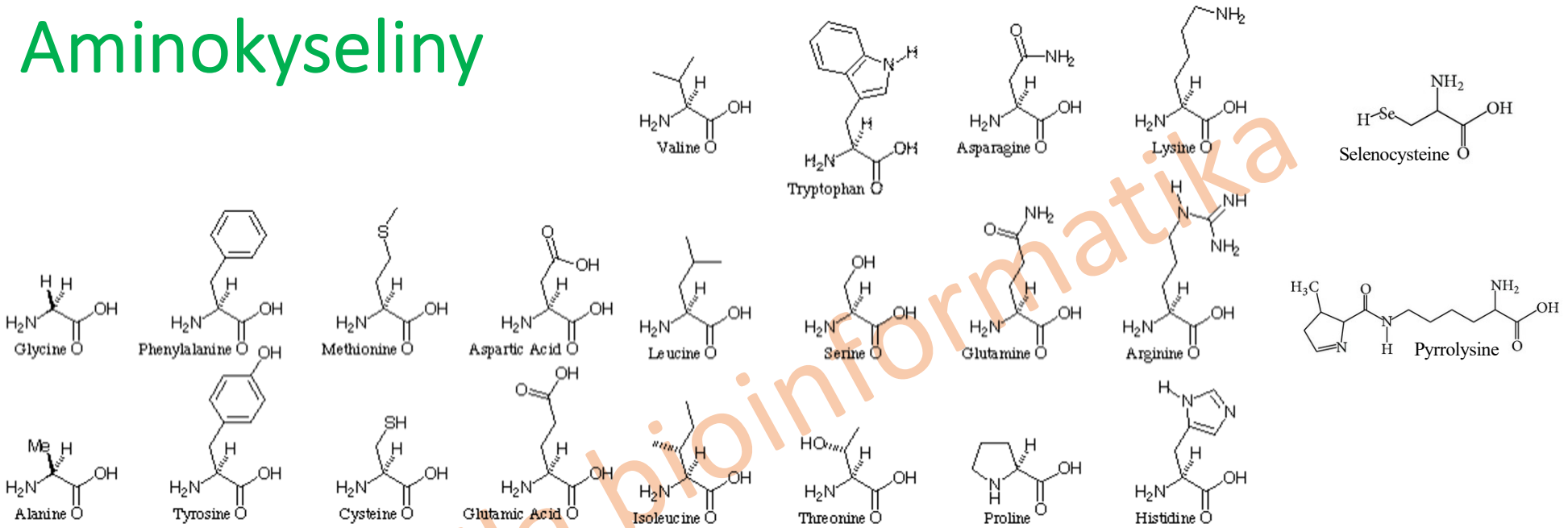




Proteinogenní aminokyseliny

- Stavební jednotky proteinů: α-L-aminokyseliny
- 20 standardních proteinogenních aminokyselin
- Podle charakteru bočního řetězce je můžeme dělit na:
  - Alifatické (Gly, Ala, Val, Leu, Ile)
  - Aromatické (Trp, Tyr, Phe, His)
  - Sirné (Cys, Met)
  - Obsahující OH skupinu (Ser, Thr)
  - Kyselé a z nich odvozené (Glu, Gln, Asp, Asn)
  - Bazické (Lys, Arg, His)

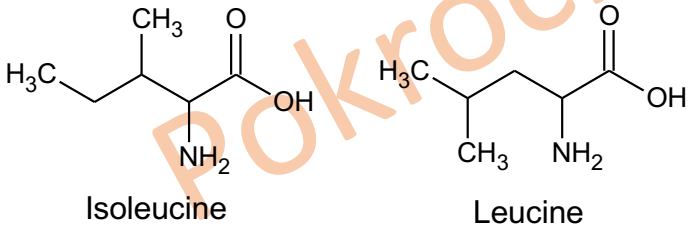
Aminokyseliny

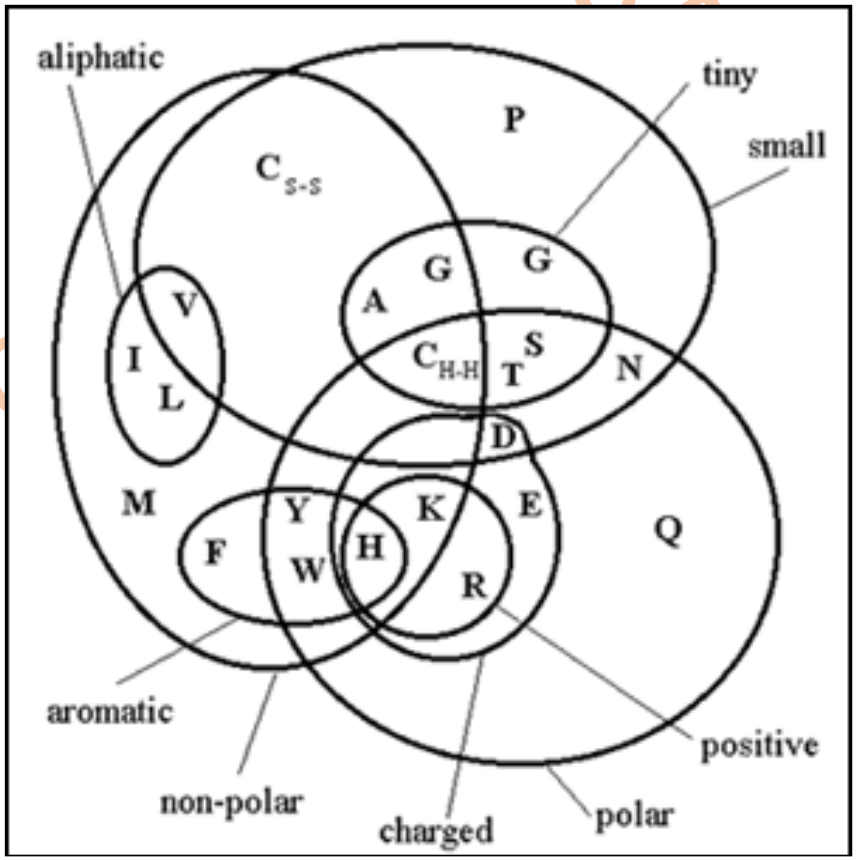


glycin	alanin	valin	leucin	izoleucin	asparagová kys.	asparagin	glutamová kys.	glutamin	arginin	lysin	histidin	fenylalanin	serin	threonin	tyrozin	tryptofan	methionin	cystein	prolin	selenocystein	pyrolysin
Gly	Ala	Val	Leu	Ile	Asp	Asn	Glu	Gln	Arg	Lys	His	Phe	Ser	Thr	Tyr	Trp	Met	Cys	Pro	Sec	Pyr
G	A	V	L	I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

Aminokyseliny

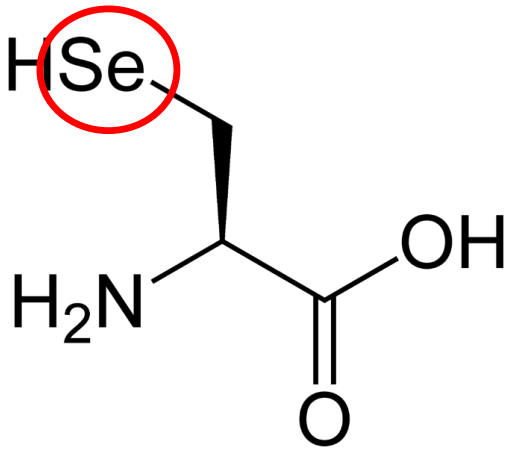
Aminokyseliny s podobnými vlastnostmi mohou plnit v proteinu stejné funkce – bývají vzájemně zastupitelné





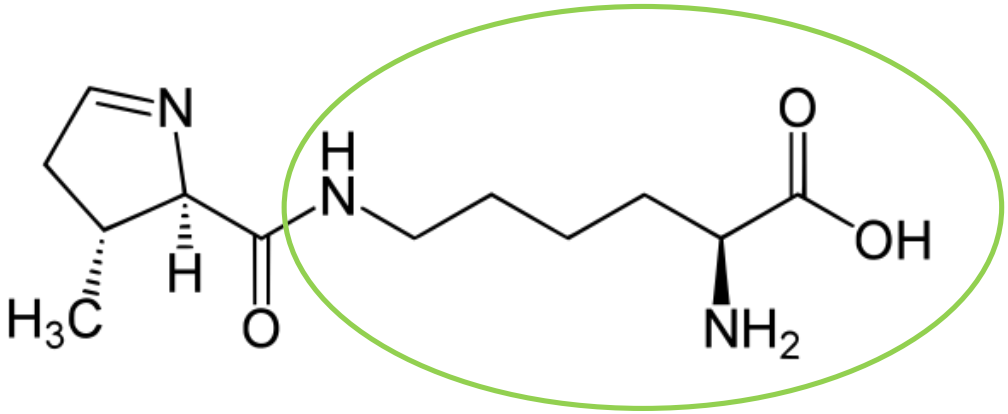
Selenocystein

- „21. aminokyselina“ – (Sec, U)
- Bifunkční kodon UGA
- Vyžaduje přítomnost speciální sekvence
- Využití u různých organismů vč. *E. coli* a člověka
- Výskyt např. v oxidoreduktasách



Pyrrolysin

- „22. aminokyselina“ – (Pyl, O)
- Bifunkční kodon UAG
- Vyžaduje přítomnost speciální sekvence
- Využití zřejmě vzácné – archea, bakterie
- Uplatnění v methyltransferasách



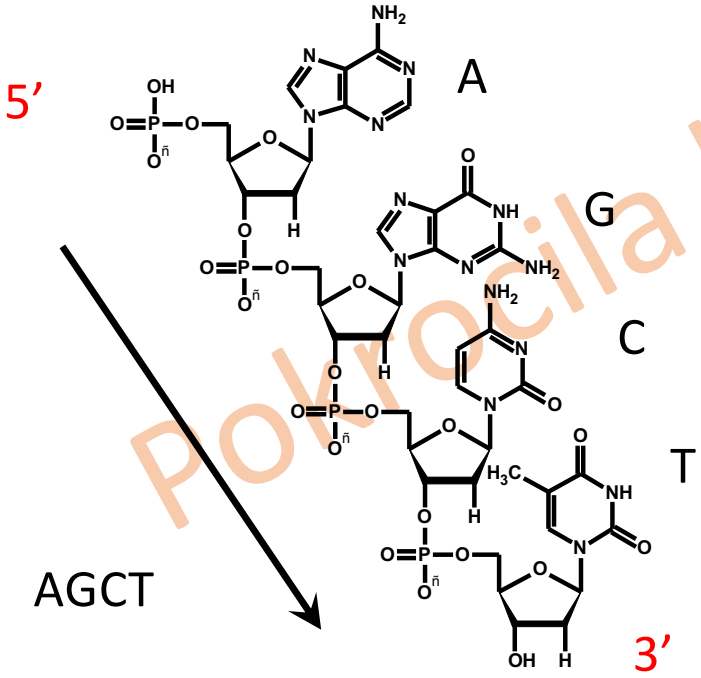
Proteiny

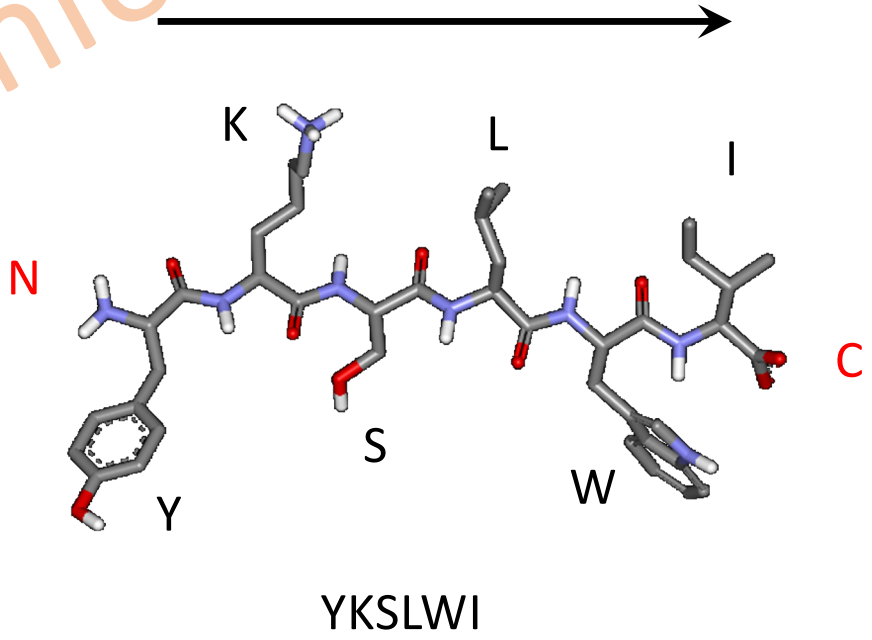
- 20 standardních proteinogenních aminokyselin
- 2 nestandardní proteinogenní aminokyseliny (selenocystein, pyrrolysin)
- 4 zkratky pro nejednoznačnou aminokyselinu

Alanin	**Kyselina asparagová** nebo **Asparagin**	Cystein	Asparagová kyselina	Glutamová kyselina	Fenylalanin	Glycin	Histidin	Isoleucin	**Isoleucin** nebo **Leucin**	Lysin	Leucin	Methionin	Asparagin	Pyrolysin	Prolin	Glutamin	Arginin	Serin	Threonin	Selenocystein	Valin	Tryptofan	**Jakákkoliv aminokyselina**	Tyrosin	**Kyselina glutamová** nebo **Glutamin**
Ala	**Asx**	Cys	Asp	Glu	Phe	Gly	His	Ile		Lys	Leu	Met	Asn	Pyl	Pro	Gln	Arg	Ser	Thr	Sec	Val	Trp		Tyr	**Glx**
A	**B**	C	D	E	F	G	H	I	**J**	K	L	M	N	O	P	Q	R	S	T	U	V	W	**X**	Y	**Z**

Zápis sekvence

- Sekvence jsou vždy orientované
  - DNA/RNA – od 5’ konce ke 3’ konce
  - Proteiny – od N konce k C konci





Formát sekvence

- Sekvence může být zapsána v různých formátech

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

- Nejpoužívanější je tzv. **FASTA formát**

>**NÁZEV**(␣popis dle vlastní volby)↵
SEKVENCESEKVENCESEKVENCESEKVENCESEKVEN ↵
CESEKVENCESEKVENCESEKVENCESEKVENCESEKV ↵
ENCESEKVENCESEKVENCESEKVENCESEKVENCESE ↵
KVENCESEKVENCESEKVENCESEKVENCE↵

POVINNÉ

VOLITELNÉ

Sekvenční přiložení = Alignment

- Přiložení dvou nebo více sekvencí na základě jejich vzájemné podobnosti

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

Párové přiložení (pairwise alignment) – dvě sekvence

WLAKALKYLMETAQASSISTELARHHPRAVDAKRKSEMKRKTAMWLAKALKYLMETAQASSISTELARHHPRAVDAKRKSEMKRKTAM

WLAKALKYLMETAQASSISTELARHHPRAVDAKRKSEMKRKTAMWLAKALKYLMETAQASSISTELARHHPRAVDAKRKSEMKRKTAM

Mnohočetné přiložení (multiple sequence alignment) – více sekvencí

WLAKALKYLMETAQASSISTELARHHPRAVDAKRKSEMKRKTAMWLAKALKYLMETAQASSISTELARHHPRAVDAKRK

WLAKALKYLMETAQASSISTELARHHPRAVDAKRKSEMKRKTAMWLAKALKYLMETAQASSISTELARHHPRAVDAKRK

WLAKALKYLMETAQASSISTELARHHPRAVDAKRKSEMKRKTAMWLAKALKYLMETAQASSISTELARHHPRAVDAKRK

WLAKALKYLMETAQASSISTELARHHPRAVDAKRKSEMKRKTAMWLAKALKYLMETAQASSISTELARHHPRAVDAKRK

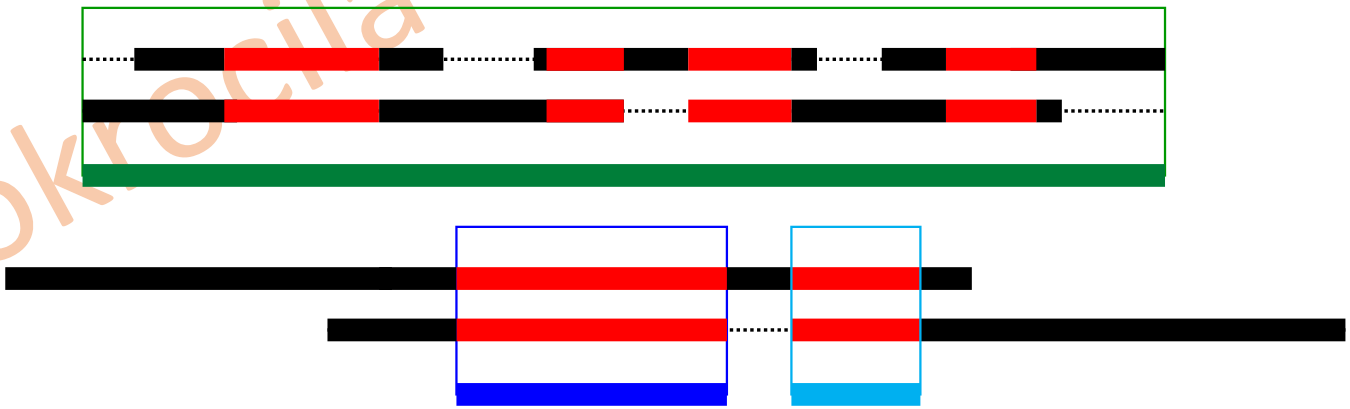
WLAKALKYLMETAQASSISTELARHHPRAVDAKRKSEMKRKTAMWLAKALKYLMETAQASSISTELARHHPRAVDAKRK

...

Párové přiložení (pair-wise alignment)

Srovnání **dvou** sekvencí

- Globální alignment – sekvence jsou přiloženy v celé své délce včetně nepodobných úseků
- Local alignment – sekvence jsou přiloženy pouze v oblasti, kde jsou si podobné



Jak by asi vypadal alignment těchto dvou sekvencí:

```
MAMUZDOSTSTAROSTISHAMIZNOSTIRATOLESTI
MAMRADOSTZESTAROZITNOSTI
```

při absolutnim preferování
A) globálního alignmentu

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

B) lokálního alignmentu

```
MAMUZDOST--STAROSTISHAMIZNOSTIRATOLESTI
MAMRADOSTZESTAROZ------ITNOSTI
```

FASTA algoritmus

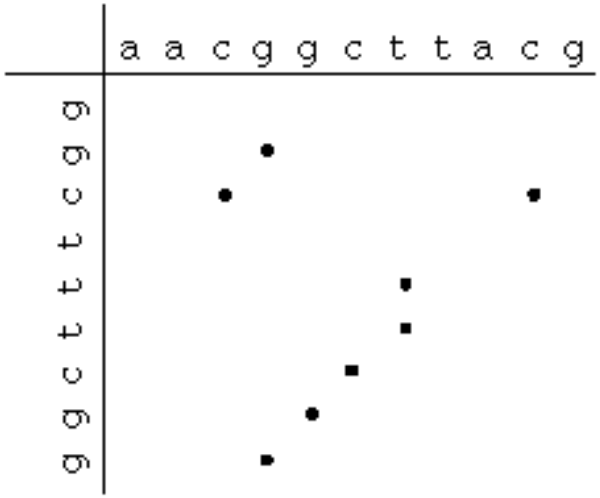
- Lokální přiložení s využitím heuristického přístupu
- Používán od roku 1987

Proces:

Obě porovnávané sekvence tvoří horizontální a vertikální osu grafu.

Následně jsou jednotlivá slova z jedné sekvence porovnávána se slovy sekvence druhé. Odpovídající páry pak vytvoří sadu bodů. Body na úhlopříccе signalizují významnou shodu (či podobnost). Cílem je nalezení nejdelšího shodného úseku (úseku s nejvyšším skóre).

V dalších krocích jsou zahrnuty konzervativní změny pro nejlepší úseky z prvního prohledání. Program pak vyhledává možnost spojení více takových úseků (může mezi nimi být mezera, či jsou na různých diagonálách) a tyto spojené úseky jsou posouzeny z hlediska zadaných kriterií.



Příklad porovnání sekvencí GGCTTTCGG a AACGGCTTACG

Emboss Needle & Water

- vytvořeny 1970

  Needleman S.B. and Wunsch C.D. (1970) A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of Molecular Biology 48: 443-453.*

- využívají dynamické programování
- umožňují vložení mezer

Needle/Stretcher – **globální** pairwise alignment,
Needleman-Wunsch algoritmus

https://www.ebi.ac.uk/Tools/psa/emboss_needle/

https://www.ebi.ac.uk/Tools/psa/emboss_stretcher/

Water – **lokální** pairwise alignment,
Smith-Waterman algoritmus

https://www.ebi.ac.uk/Tools/psa/emboss_water/

Needlman-Wunch algorithm

Shoda +1
Neshoda -1 (po diagonále)

Gap -1 (svisle nebo vodorovně)

		G	C	A	T	G	C	T
	0	-1	-2	-3	-4	-5	-6	-7
G	-1	1						
A	-2							
T	-3							
T	-4							
A	-5							
C	-6							
A	-7							

pokud shoda v diagonále,
nemá smysl řešit mezery

Needlman-Wunch algorithm

Shoda +1
Neshoda -1 (po diagonále)

Gap -1 (svisle nebo vodorovně)

		G	C	A	T	G	C	T
	0	-1	-2	-3	-4	-5	-6	-7
G	-1	1	0					
A	-2	0	0					
T	-3							
T	-4							
A	-5							
C	-6							
A	-7							

Needlman-Wunch algorithm

Shoda +1
Neshoda -1 (po diagonále)

Gap -1 (svisle nebo vodorovně)

		G	C	A	T	G	C	T
	0	-1	-2	-3	-4	-5	-6	-7
G	-1	1	0	-1	-2	-3	-4	-5
A	-2	0	0	1	0	-1	-2	-3
T	-3							
T	-4							
A	-5							
C	-6							
A	-7							

Needlman-Wunch algorithm

Shoda +1
Neshoda -1 (po diagonále)

Gap -1 (svisle nebo vodorovně)

		G	C	A	T	G	C	T
	0	-1	-2	-3	-4	-5	-6	-7
G	-1	1	0	-1	-2	-3	-4	-5
A	-2	0	0	1	0	-1	-2	-3
T	-3	-1	-1	0	2	1	0	-1
T	-4							
A	-5							
C	-6							
A	-7							

Needlman-Wunch algorithm

Shoda +1
Neshoda -1

(po diagonále)

Gap -1 (svisle nebo vodorovně)

		G	C	A	T	G	C	T
	0	-1	-2	-3	-4	-5	-6	-7
G	-1	1	0	-1	-2	-3	-4	-5
A	-2	0	0	1	0	-1	-2	-3
T	-3	-1	-1	0	2	1	0	-1
T	-4	-2	-2	-1	1	1	0	-1
A	-5							
C	-6							
A	-7							

Needlman-Wunch algorithm

Shoda +1
Neshoda -1

(po diagonále)

Gap -1

(svisle nebo vodorovně)

		G	C	A	T	G	C	T
	0	-1	-2	-3	-4	-5	-6	-7
G	-1	1	0	-1	-2	-3	-4	-5
A	-2	0	0	1	0	-1	-2	-3
T	-3	-1	-1	0	2	1	0	-1
T	-4	-2	-2	-1	1	1	0	-1
A	-5	-3	-3	-1	0	0	0	-1
C	-6							
A	-7							

Needlman-Wunch algorithm

Shoda +1
Neshoda -1 (po diagonále)

Gap -1 (svisle nebo vodorovně)

		G	C	A	T	G	C	T
	0	-1	-2	-3	-4	-5	-6	-7
G	-1	1	0	-1	-2	-3	-4	-5
A	-2	0	0	1	0	-1	-2	-3
T	-3	-1	-1	0	2	1	0	-1
T	-4	-2	-2	-1	1	1	0	-1
A	-5	-3	-3	-1	0	0	0	-1
C	-6	-4	-2	-2	-1	-1	1	-1
A	-7							

Needlman-Wunch algorithm

Shoda +1
Neshoda -1 (po diagonále)

Gap -1 (svisle nebo vodorovně)

		G	C	A	T	G	C	T
	0	-1	-2	-3	-4	-5	-6	-7
G	-1	1	0	-1	-2	-3	-4	-5
A	-2	0	0	1	0	-1	-2	-3
T	-3	-1	-1	0	2	1	0	-1
T	-4	-2	-2	-1	1	1	0	-1
A	-5	-3	-3	-1	0	0	0	-1
C	-6	-4	-2	-2	-1	-1	1	-1
A	-7	-5	-3	-1	-2	-2	0	0

Needlman-Wunch algorithm

Shoda +1
Neshoda -1

(po diagonále)

Gap -1

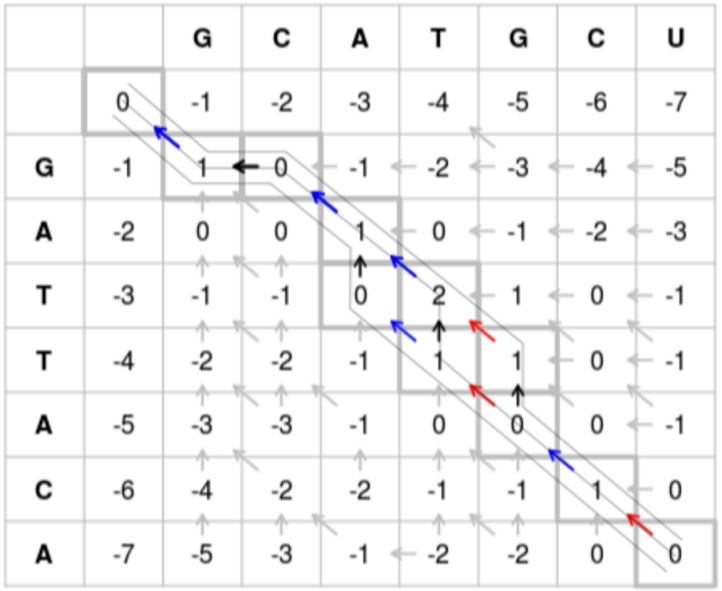
(svisle nebo vodorovně)

0+1=1

-1+1=0

-1+1=0

		G	C	A	T	G	C	T
	0	-1	-2	-3	-4	-5	-6	-7
G	-1	1	0	-1	-2	-3	-4	-5
A	-2	0	0	1	0	-1	-2	-3
T	-3	-1	-1	0	2	1	0	-1
T	-4	-2	-2	-1	1	1	0	-1
A	-5	-3	-3	-1	0	0	0	-1
C	-6	-4	-2	-2	-1	-1	1	-1
A	-7	-5	-3	-1	-2	-2	0	0



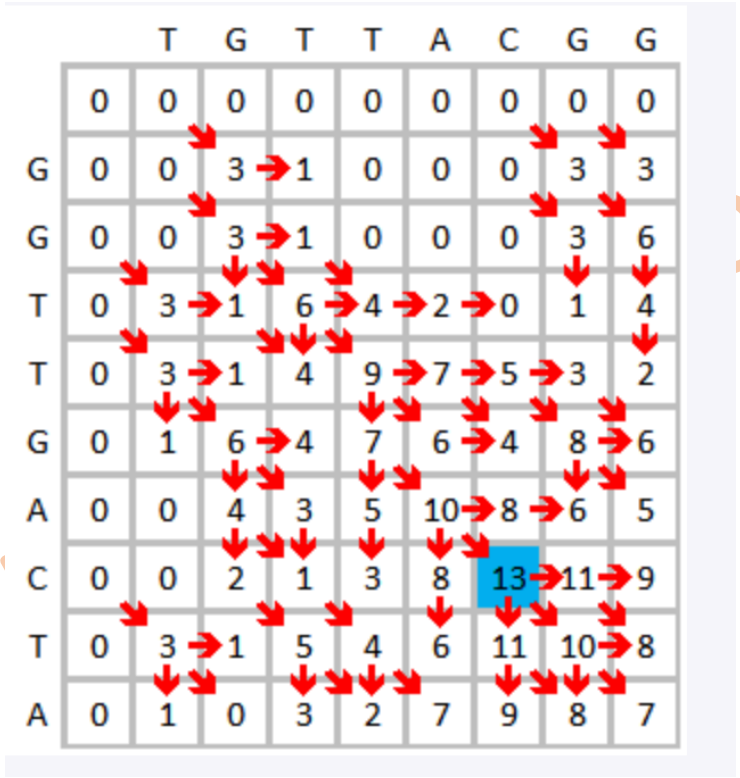
Smith–Waterman algorithm

Shoda +3
Neshoda -3

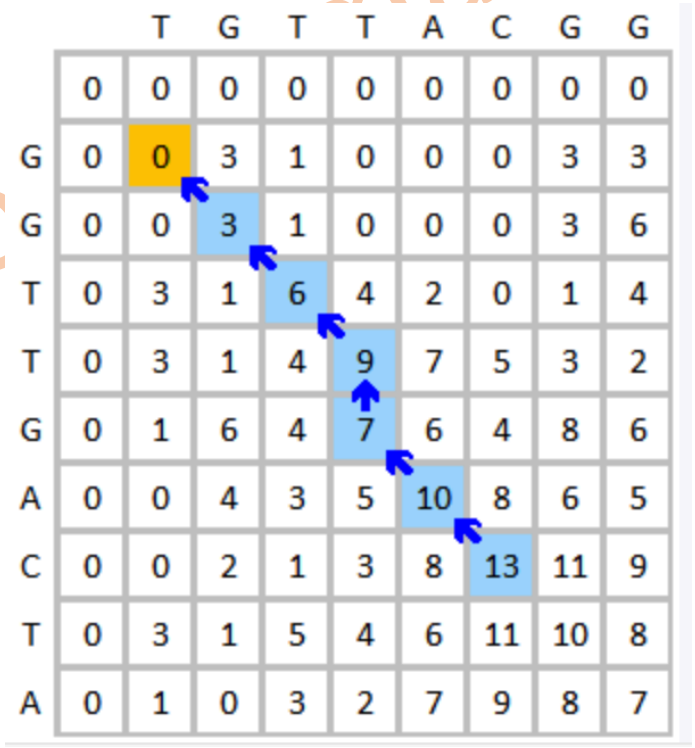
(po diagonále)

Gap -2

(svisle nebo vodorovně
Záporné hodnoty → 0



		T	G	T	T	A	C	G	G
	0	0	0	0	0	0	0	0	0
G	0	0	3	1	0	0	0	3	3
G	0	0	3	1	0	0	0	3	6
T	0	3	1	6	4	2	0	1	4
T	0	3	1	4	9	7	5	3	2
G	0	1	6	4	7	6	4	8	6
A	0	0	4	3	5	10	8	6	5
C	0	0	2	1	3	8	13	11	9
T	0	3	1	5	4	6	11	10	8
A	0	1	0	3	2	7	9	8	7



		T	G	T	T	A	C	G	G
	0	0	0	0	0	0	0	0	0
G	0	0	3	1	0	0	0	3	3
G	0	0	3	1	0	0	0	3	6
T	0	3	1	6	4	2	0	1	4
T	0	3	1	4	9	7	5	3	2
G	0	1	6	4	7	6	4	8	6
A	0	0	4	3	5	10	8	6	5
C	0	0	2	1	3	8	13	11	9
T	0	3	1	5	4	6	11	10	8
A	0	1	0	3	2	7	9	8	7

Úloha 1

Doplňte tabulku a zkuste navrhnout přiložení těchto dvou sekvencí:
vyberte si jeden algoritmus (Smith–Waterman nebo Needlman-Wunch)

		A	G	T	A	T	C	T
G								
G								
A								
T								
A								
C								
T								

Jak poznat podobné sekvence?

Referenční sekvence:

TCTTCTGGTCCCCACAGACTCAGAGAGAACCCACCATGGTGCTGTCTCCTGCCGACAAGACCAACGTCAAGGCCGCCTGGGGTAAGGTCGGCCGCACGCTGGCGAGTAT
GGTGCGGAGGCCCTGGAGAGGTGAGGCTCCCTCCCCTGCTCCGACCCGGGCTCCTCGCCCGCCCGGACCCACAGGCCACCTCAACCGTCCTGGCCCCGGACCCAAACCC
CACCCCTCACTCTGCTTCTCCCCGCAGGATGTTCCTGTCCTTCCCCACCACCAAGACCTACTTCCCGCACTTCGACCTGAGCCACGGCTCTGCCCAGGTTAAGGGCCAC
GGCAAGAAGGTGGCCGACGCGCTGACCAACGCCGTGGCGCACGTGGACGACATGCCCAACGCGCTGTCCGCCCTGAGCGACCTGCACGCGCACAAGCTTCGGGTGGACC
CGGTCAACTTCAAGGTGAGCGGCGGGCCGGGAGCGATCTGGGTCGAG

1 CTGGTCCCCACAGACTCAGAGAGAACCCACCATGGTGCTGTCTCCTGCCGACAAGACCAACGTCAAGGCCGCCTGGGGTAAGGTCGGCCGCACGCTGGCGAGTATGGTG
CGGAGGCCCTGGAGAGGTGAGGCTCCCTCCCCTGCTCCGACCCGGGCTCCTCGCCCGCCCGGACCCACAGGCCACCTCAACCGTCCTGGCCCCGGACCCAAACCCCACC
CCTCACTCTGCTTCTCCCCGCAGGATGTTCCTGTCCTTCCCCACCACCAAGACCTACTTCCCGCACTTCGACCTGAGCCACGGCTCTGCCCAGGTTAAGGGCCACGGCA
AGAAGGTGGCCGACGCGCTGACCAACGCCGTGGCGCACGTGGACGACATGCCCAACGCGCTGTCCGCCCTGAGCGACCTGCACGCGCACAAGCTTCGGGTGGACCCGGT
CAACTTCAAGGTGAGCGGCGGGCCGGGAGCGATCTGGGTCGAGTCTT

2 GGCTCTGCCCAGGTTAAGGGCCACGGCAATCTTCTGGTCCCCACAGACTCAGAGAAAACCCACCATGGTGCTGTCTCCTGCCGACAAGACCAACGTCAAGGCCGCCTGG
GGTAAGGTCGGCCGCACGCCGGCGAGTATGGTGCGGAGGCCCTGGAGAGGTGAGGGTCCCTCCCCTGCTCCGACCCGGGCTCCTCGCCCGCCCGGACCCTCAGGCCACC
TCAACCGTCCTGGCCACGGACCCAAACCCCACCCCTCACCCTGCTTCTCCCCGCAGGATGTTCCTGTCCTTCGCCACCACCAAGACCTACTTCCCGCACTTCGACCTGA
GCCACGGCTCTGCCCAGGTTAAGGGCCACGGCAAAAAGGTGGCCGACGCGCTGACCAACGCCGTGGCGCTCGTGGACGACATGCCCAACGCCCTGTCCGCCCTGAGCGA
CCTGCACGCGCACAAGCTTCGGGTGGACCCGGTCAGCTTCAAGGTGAGCGGCGGGCCGGGAGCGATCTGGGTCGAG

3 TCTGTCCTTCCCCACCACCAAGACCTACTTCCCGCACTTCGACCTGAGCCACGGCTCTGCCCAGGTTAAGGGCCACGGCAAGAAGGTGGCCGACGCGCTGACCAACGCC
GTGGCGCACGTGGACGACATGCCCAACGCGCTGTCCGCCCTGAGCGACCTGCACGCGCACAAGCTTCGGGTGGACCCGGTCAACTTCAAGGTGAGCGGCGGGCCGGGAG
CGATCTGGGTCGAG

Jak poznat podobné sekvence?

- Přiložení a vizuální porovnání
- Postupný posun o jednu pozici a opětovné porovnání
- Snadné pro **vysoce podobnou** (identickou) dvojici sekvencí

```
MAMUZDOSTSTAROSTISHAMIZNOSTIRATOLESTI
|||||||||||||||||||||||||||||||||||||
MAMUZDOSTSTAROSTISHAMIZNOSTIRATOLESTI
```

✓

```
   RADOSTSTAROSTISHAMIZNOSTIRATOLESTI
     |||||||||||||||||||||||||||||||
   UZDOSTSTAROSTISHAMIZNOSTIRATOLESTI
```

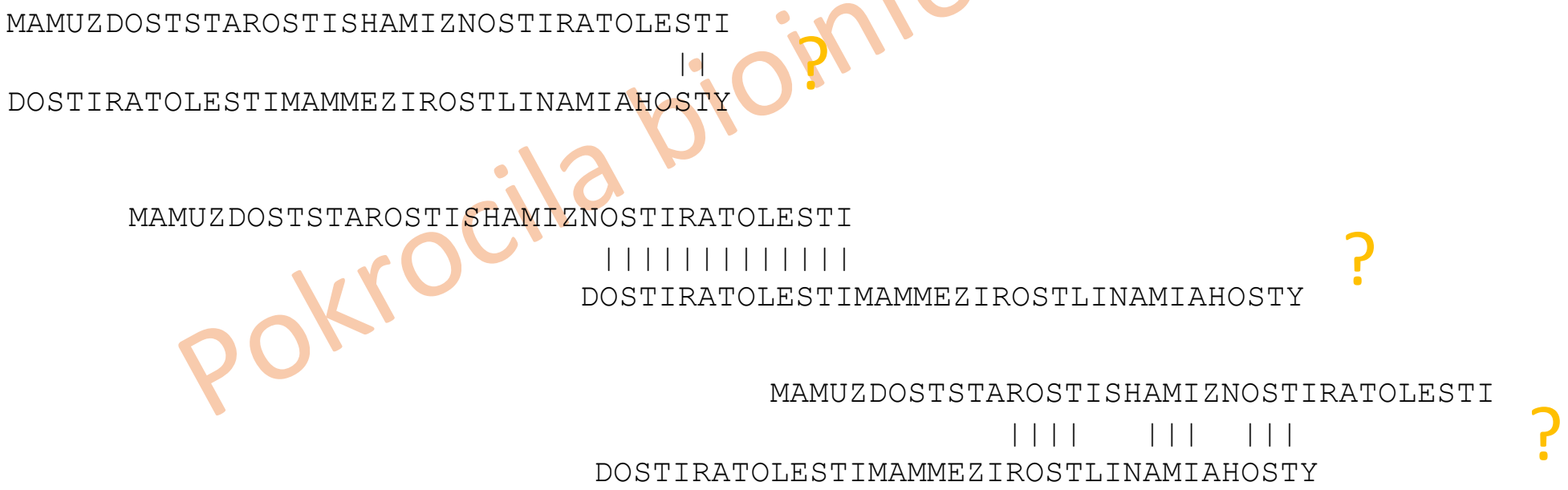
✓

```
      UZDOSTSTAROSTIVHAMIZNOSTI
      |||||||||||||| |||||||||
MAMUZDOSTSTAROSTISHAMIZNOSTIRATOLESTI
```

✓

Jak poznat podobné sekvence?

- Pro ne zcela identické sekvence je nutno kvantifikovat **kvalitu** přiložení
- Na základě parametrů je nutno vybrat **nejlepší** (možné) přiložení



Skórování proteinového přiložení

Substituční matice (a z nich odvozené **skórovací matice**) vycházejí z několika zdrojů

- Fyzikálně-chemické vlastnosti jednotlivých aminokyselin
- Pravděpodobnost záměny aminokyselin díky mutaci v DNA – změna kodonu
- Pravděpodobnost, že dojde k záměně dvou konkrétních aminokyselin v průběhu evoluce – empirické pozorování

Empirické matice jsou obecně přesnější

Napadají Vás některé záměny AA, které budou pravděpodobně vysoce „penalizovány“? ?

Skórování proteinového přiložení

- Skórovací matice obsahuje číselné **hodnoty** (pravděpodobnosti) pro jednotlivé dvojice aminokyselin v závislosti na jejich vzájemné „zastupitelnosti“ – pravděpodobnosti substituce
- Pravděpodobnost změny jedné aminokyseliny na jinou **je** přímo **úměrná podobnosti** obou aminokyselin.
- **Záměna** aminokyseliny/nukleotidu je častější než inzerce/delece.
- Inzerce/delece je vždy negativní – penalizace
- Součet hodnot pro všechny dvojice pak udává výsledné **skóre**

Proteinová matice

Záměny mezi aminokyselinami jsou různě závažné

- Kyselina glutamová → Glutamin
- Kyselina glutamová → Tryptofan

Shoda aminokyselin nemusí být hodnocena stejně

- Serin x Serin
- Tryptofan X Tryptofan

	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
A	4	-3	-1	-1	-3	-2	0	1	-3	-2	-3	-3	-2	-5	1	1	1	-7	-4	0
R	-3	7	-2	-4	-5	1	-3	-5	1	-3	-5	2	-1	-6	-1	-1	-3	1	-6	-4
N	-1	-2	5	3	-5	-1	1	-1	2	-3	-4	1	-4	-5	-2	1	0	-5	-2	-3
D	-1	-4	3	5	-7	0	4	-1	-1	-4	-6	-1	-5	-8	-3	-1	-2	-9	-6	-4
C	-3	-5	-5	-7	9	-8	-8	-5	-4	-3	-8	-8	-7	-7	-4	-1	-4	-9	-1	-3
Q	-2	1	-1	0	-8	6	2	-3	3	-4	-2	0	-2	-7	-1	-2	-2	-7	-6	-3
E	0	-3	1	4	-8	2	5	-1	-1	-3	-5	-1	-4	-8	-2	-1	-2	-9	-5	-3
G	1	-5	-1	-1	-5	-3	-1	5	-4	-5	-6	-3	-4	-6	-2	0	-2	-9	-7	-3
H	-3	1	2	-1	-4	3	-1	-4	7	-4	-3	-2	-4	-3	-1	-2	-3	-4	-1	-3
I	-2	-3	-3	-4	-3	-4	-3	-5	-4	6	1	-3	1	0	-4	-3	0	-7	-3	3
L	-3	-5	-4	-6	-8	-2	-5	-6	-3	1	6	-4	3	0	-4	-4	-3	-3	-3	0
K	-3	2	1	-1	-8	0	-1	-3	-2	-3	-4	5	0	-7	-3	-1	-1	-6	-6	-4
M	-2	-1	-4	-5	-7	-2	-4	-4	-4	1	3	0	9	-1	-4	-3	-1	-6	-5	1
F	-5	-6	-5	-8	-7	-7	-8	-6	-3	0	0	-7	-1	8	-6	-4	-5	-1	4	-3
P	1	-1	-2	-3	-4	-1	-2	-2	-1	-4	-4	-3	-4	-6	7	0	-1	-7	-7	-3
S	1	-1	1	-1	-1	-2	-1	0	-2	-3	-4	-1	-3	-4	0	4	2	-3	-4	-2
T	1	-3	0	-2	-4	-2	-2	-2	-3	0	-3	-1	-1	-5	-1	2	5	-7	-4	0
W	-7	1	-5	-9	-9	-7	-9	-9	-4	-7	-3	-6	-6	-1	-7	-3	-7	12	-2	-9
Y	-4	-6	-2	-6	-1	-6	-5	-7	-1	-3	-3	-6	-5	4	-7	-4	-4	-2	9	-4
V	0	-4	-3	-4	-3	-3	-3	-3	-3	3	0	-4	1	-3	-3	-2	0	-9	-4	5

Mezery

- Vložení mezer umožňuje získání většího množství shod = „lepší alignment“
- S využitím dostatečného počtu mezer lze ale zarovnat cokoliv !!!
- Mezery mohou vzniknout na začátku, na konci nebo uprostřed sekvence
- Delecí a inzercí vznikají mezery o různé délce

→ **Přítomnost mezery je penalizována** (snížení skóre)

→ **Vznik** mezery je penalizován víc než její **délka**

Výpočet skóre

>sekvence A

PAKAPALAPAKAP

>sekvence B

VPKAPALVPKAP

Penalizace mezery: -10

	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
A	4	-3	-1	-1	-3	-2	0	1	-3	-2	-3	-3	-2	-5	1	1	1	-7	-4	0
R	-3	7	-2	-4	-5	1	-3	-5	1	-3	-5	2	-1	-6	-1	-1	-3	1	-6	-4
N	-1	-2	5	3	-5	-1	1	-1	2	-3	-4	1	-4	-5	-2	1	0	-5	-2	-3
D	-1	-4	3	5	-7	0	4	-1	-1	-4	-6	-1	-5	-8	-3	-1	-2	-9	-6	-4
C	-3	-5	-5	-7	9	-8	-8	-5	-4	-3	-8	-8	-7	-7	-4	-1	-4	-9	-1	-3
Q	-2	1	-1	0	-8	6	2	-3	3	-4	-2	0	-2	-7	-1	-2	-2	-7	-6	-3
E	0	-3	1	4	-8	2	5	-1	-1	-3	-5	-1	-4	-8	-2	-1	-2	-9	-5	-3
G	1	-5	-1	-1	-5	-3	-1	5	-4	-5	-6	-3	-4	-6	-2	0	-2	-9	-7	-3
H	-3	1	2	-1	-4	3	-1	-4	7	-4	-3	-2	-4	-3	-1	-2	-3	-4	-1	-3
I	-2	-3	-3	-4	-3	-4	-3	-5	-4	6	1	-3	1	0	-4	-3	0	-7	-3	3
L	-3	-5	-4	-6	-8	-2	-5	-6	-3	1	6	-4	3	0	-4	-4	-3	-3	-3	0
K	-3	2	1	-1	-8	0	-1	-3	-2	-3	-4	5	0	-7	-3	-1	-1	-6	-6	-4
M	-2	-1	-4	-5	-7	-2	-4	-4	-4	1	3	0	9	-1	-4	-3	-1	-6	-5	1
F	-5	-6	-5	-8	-7	-7	-8	-6	-3	0	0	-7	-1	8	-6	-4	-5	-1	4	-3
P	1	-1	-2	-3	-4	-1	-2	-2	-1	-4	-4	-3	-4	-6	7	0	-1	-7	-7	-3
S	1	-1	1	-1	-1	-2	-1	0	-2	-3	-4	-1	-3	-4	0	4	2	-3	-4	-2
T	1	-3	0	-2	-4	-2	-2	-2	-3	0	-3	-1	-1	-5	-1	2	5	-7	-4	0
W	-7	1	-5	-9	-9	-7	-9	-9	-4	-7	-3	-6	-6	-1	-7	-3	-7	12	-2	-9
Y	-4	-6	-2	-6	-1	-6	-5	-7	-1	-3	-3	-6	-5	4	-7	-4	-4	-2	9	-4
V	0	-4	-3	-4	-3	-3	-3	-3	-3	3	0	-4	1	-3	-3	-2	0	-9	-4	5

-3 +1 +5 +4 +7 +4 +6 +0 +7 -3 -3 +1 = **26**

```
P A K A P A L A P A K A P
    | | | | |   |
V P K A P A L V P K A P
```

Výpočet skóre

>sekvence A

PAKAPALAPAKAP

>sekvence B

VPKAPALVPKAP

Penalizace mezery: -10

	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
A	4	-3	-1	-1	-3	-2	0	1	-3	-2	-3	-3	-2	-5	1	1	1	-7	-4	0
R	-3	7	-2	-4	-5	1	-3	-5	1	-3	-5	2	-1	-6	-1	-1	-3	1	-6	-4
N	-1	-2	5	3	-5	-1	1	-1	2	-3	-4	1	-4	-5	-2	1	0	-5	-2	-3
D	-1	-4	3	5	-7	0	4	-1	-1	-4	-6	-1	-5	-8	-3	-1	-2	-9	-6	-4
C	-3	-5	-5	-7	9	-8	-8	-5	-4	-3	-8	-8	-7	-7	-4	-1	-4	-9	-1	-3
Q	-2	1	-1	0	-8	6	2	-3	3	-4	-2	0	-2	-7	-1	-2	-2	-7	-6	-3
E	0	-3	1	4	-8	2	5	-1	-1	-3	-5	-1	-4	-8	-2	-1	-2	-9	-5	-3
G	1	-5	-1	-1	-5	-3	-1	5	-4	-5	-6	-3	-4	-6	-2	0	-2	-9	-7	-3
H	-3	1	2	-1	-4	3	-1	-4	7	-4	-3	-2	-4	-3	-1	-2	-3	-4	-1	-3
I	-2	-3	-3	-4	-3	-4	-3	-5	-4	6	1	-3	1	0	-4	-3	0	-7	-3	3
L	-3	-5	-4	-6	-8	-2	-5	-6	-3	1	6	-4	3	0	-4	-4	-3	-3	-3	0
K	-3	2	1	-1	-8	0	-1	-3	-2	-3	-4	5	0	-7	-3	-1	-1	-6	-6	-4
M	-2	-1	-4	-5	-7	-2	-4	-4	-4	1	3	0	9	-1	-4	-3	-1	-6	-5	1
F	-5	-6	-5	-8	-7	-7	-8	-6	-3	0	0	-7	-1	8	-6	-4	-5	-1	4	-3
P	1	-1	-2	-3	-4	-1	-2	-2	-1	-4	-4	-3	-4	-6	7	0	-1	-7	-7	-3
S	1	-1	1	-1	-1	-2	-1	0	-2	-3	-4	-1	-3	-4	0	4	2	-3	-4	-2
T	1	-3	0	-2	-4	-2	-2	-2	-3	0	-3	-1	-1	-5	-1	2	5	-7	-4	0
W	-7	1	-5	-9	-9	-7	-9	-9	-4	-7	-3	-6	-6	-1	-7	-3	-7	12	-2	-9
Y	-4	-6	-2	-6	-1	-6	-5	-7	-1	-3	-3	-6	-5	4	-7	-4	-4	-2	9	-4
V	0	-4	-3	-4	-3	-3	-3	-3	-3	3	0	-4	1	-3	-3	-2	0	-9	-4	5

26

-3 +1 +5 +4 +7 +4 +6 +0 +7 -10 +5 +4 +7 = **37**

```
PAKAPALAPAKAP      P A K A P A L A P A K A P
  |||||  |             | | | | |   |   | | |
VPKAPALVPKAP       V P K A P A L V P - K A P
```

Výpočet skóre

>sekvence A

PAKAPALAPAKAP

>sekvence B

VPKAPALVPKAP

Penalizace mezery: -10

	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
A	4	-3	-1	-1	-3	-2	0	1	-3	-2	-3	-3	-2	-5	1	1	1	-7	-4	0
R	-3	7	-2	-4	-5	1	-3	-5	1	-3	-5	2	-1	-6	-1	-1	-3	1	-6	-4
N	-1	-2	5	3	-5	-1	1	-1	2	-3	-4	1	-4	-5	-2	1	0	-5	-2	-3
D	-1	-4	3	5	-7	0	4	-1	-1	-4	-6	-1	-5	-8	-3	-1	-2	-9	-6	-4
C	-3	-5	-5	-7	9	-8	-8	-5	-4	-3	-8	-8	-7	-7	-4	-1	-4	-9	-1	-3
Q	-2	1	-1	0	-8	6	2	-3	3	-4	-2	0	-2	-7	-1	-2	-2	-7	-6	-3
E	0	-3	1	4	-8	2	5	-1	-1	-3	-5	-1	-4	-8	-2	-1	-2	-9	-5	-3
G	1	-5	-1	-1	-5	-3	-1	5	-4	-5	-6	-3	-4	-6	-2	0	-2	-9	-7	-3
H	-3	1	2	-1	-4	3	-1	-4	7	-4	-3	-2	-4	-3	-1	-2	-3	-4	-1	-3
I	-2	-3	-3	-4	-3	-4	-3	-5	-4	6	1	-3	1	0	-4	-3	0	-7	-3	3
L	-3	-5	-4	-6	-8	-2	-5	-6	-3	1	6	-4	3	0	-4	-4	-3	-3	-3	0
K	-3	2	1	-1	-8	0	-1	-3	-2	-3	-4	5	0	-7	-3	-1	-1	-6	-6	-4
M	-2	-1	-4	-5	-7	-2	-4	-4	-4	1	3	0	9	-1	-4	-3	-1	-6	-5	1
F	-5	-6	-5	-8	-7	-7	-8	-6	-3	0	0	-7	-1	8	-6	-4	-5	-1	4	-3
P	1	-1	-2	-3	-4	-1	-2	-2	-1	-4	-4	-3	-4	-6	7	0	-1	-7	-7	-3
S	1	-1	1	-1	-1	-2	-1	0	-2	-3	-4	-1	-3	-4	0	4	2	-3	-4	-2
T	1	-3	0	-2	-4	-2	-2	-2	-3	0	-3	-1	-1	-5	-1	2	5	-7	-4	0
W	-7	1	-5	-9	-9	-7	-9	-9	-4	-7	-3	-6	-6	-1	-7	-3	-7	12	-2	-9
Y	-4	-6	-2	-6	-1	-6	-5	-7	-1	-3	-3	-6	-5	4	-7	-4	-4	-2	9	-4
V	0	-4	-3	-4	-3	-3	-3	-3	-3	3	0	-4	1	-3	-3	-2	0	-9	-4	5

26

```
PAKAPALAPAKAP
  |||||  |
VPKAPALVPKAP
```

37

```
PAKAPALAPAKAP
  |||||  | |||
VPKAPALVP-KAP
```

+7-10 +5 +4 +7 +4 +6 +0 +7-10 +5 +4 +7 = **36**

```
  P A K A P A L A P A K A P
  |   | | | | |   |   | | |
V P - K A P A L V P - K A P
```

Výpočet skóre

>sekvence A
PAKAPALAPAKAP

>sekvence B
VPKAPALVPKAP

Penalizace mezery: -10

	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
A	4	-3	-1	-1	-3	-2	0	1	-3	-2	-3	-3	-2	-5	1	1	1	-7	-4	0
R	-3	7	-2	-4	-5	1	-3	-5	1	-3	-5	2	-1	-6	-1	-1	-3	1	-6	-4
N	-1	-2	5	3	-5	-1	1	-1	2	-3	-4	1	-4	-5	-2	1	0	-5	-2	-3
D	-1	-4	3	5	-7	0	4	-1	-1	-4	-6	-1	-5	-8	-3	-1	-2	-9	-6	-4
C	-3	-5	-5	-7	9	-8	-8	-5	-4	-3	-8	-8	-7	-7	-4	-1	-4	-9	-1	-3
Q	-2	1	-1	0	-8	6	2	-3	3	-4	-2	0	-2	-7	-1	-2	-2	-7	-6	-3
E	0	-3	1	4	-8	2	5	-1	-1	-3	-5	-1	-4	-8	-2	-1	-2	-9	-5	-3
G	1	-5	-1	-1	-5	-3	-1	5	-4	-5	-6	-3	-4	-6	-2	0	-2	-9	-7	-3
H	-3	1	2	-1	-4	3	-1	-4	7	-4	-3	-2	-4	-3	-1	-2	-3	-4	-1	-3
I	-2	-3	-3	-4	-3	-4	-3	-5	-4	6	1	-3	1	0	-4	-3	0	-7	-3	3
L	-3	-5	-4	-6	-8	-2	-5	-6	-3	1	6	-4	3	0	-4	-4	-3	-3	-3	0
K	-3	2	1	-1	-8	0	-1	-3	-2	-3	-4	5	0	-7	-3	-1	-1	-6	-6	-4
M	-2	-1	-4	-5	-7	-2	-4	-4	-4	1	3	0	9	-1	-4	-3	-1	-6	-5	1
F	-5	-6	-5	-8	-7	-7	-8	-6	-3	0	0	-7	-1	8	-6	-4	-5	-1	4	-3
P	1	-1	-2	-3	-4	-1	-2	-2	-1	-4	-4	-3	-4	-6	7	0	-1	-7	-7	-3
S	1	-1	1	-1	-1	-2	-1	0	-2	-3	-4	-1	-3	-4	0	4	2	-3	-4	-2
T	1	-3	0	-2	-4	-2	-2	-2	-3	0	-3	-1	-1	-5	-1	2	5	-7	-4	0
W	-7	1	-5	-9	-9	-7	-9	-9	-4	-7	-3	-6	-6	-1	-7	-3	-7	12	-2	-9
Y	-4	-6	-2	-6	-1	-6	-5	-7	-1	-3	-3	-6	-5	4	-7	-4	-4	-2	9	-4
V	0	-4	-3	-4	-3	-3	-3	-3	-3	3	0	-4	1	-3	-3	-2	0	-9	-4	5

26

```
PAKAPALAPAKAP
  |||||  |
VPKAPALVPKAP
```

37

```
PAKAPALAPAKAP
  |||||  |
VPKAPALVP-KAP
```

36

```
 PAKAPALAPAKAP
 |  ||||||  | |||
VP-KAPALVP-KAP
```

Co na to EMBOSS stretcher?

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

```
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

```
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

```
1 MAMUZDOSTSTAROSTISHAMIZNOSTIRATOLESTI     37
  ||| .| ||           |. :..... ..|||
1 MAMRADOST----------ZESTAROZITNOSTI        24
```

Gap_penalty: 25

Extend_penalty: 2

Score: -11

Kdy je vhodnější:

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.

Výpočet skóre

Hodnota skóre závisí na typu sekvence a její délce

- Pravděpodobnost, že dvě rezidua v **nepříbuzných sekvencích** jsou identická

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.

Typy matic

- **PAM** (Point Accepted Mutation) – založena na mutacích v rámci globálního alignmentu, tj. ve vysoce konzerovovaných i mutabilních oblastech.
  PAM 250 znamená, že 250 mutací na 100 AK může nastat, PAM 10 akceptuje pouze 10 na 100, takže pouze velice podobné sekvence dosáhnou na pozitivní skóre.
- **BLOSUM** (Blocks Substitution Matrix) – je odvozena z vysoce konzervovaných oblastí neobsahujících mezery – z těch počítá relativní zastoupení AK a pravděpodobnost jejich substitucí → lepší pro lokální alignment.
  Je využívána v blastp, vhodná pro identifikaci neznámé nukleotidové sekvence. BLOSUM matrice s vysokými čísly je dobrá pro porovnání vysoce příbuzných sekvencí, zatímco nízké pro relativně vzdálené podobnosti
- **GONNET** – vytvořena 1992, postupným opakováním cyklu: pairwise alignment – nová matice – nový pairwise alignment – nová matice – …
- **DNA identity** matrix – navržená pro DNA sekvence
- **Specifické matice** – např. EDSSMat pro neuspořádané proteiny

V rámci jednoho typu matic existuje **více** jednotlivých **matic** založených na stejném principu, které se však liší konkrétními hodnotami a tedy i **oblastí použití** (vysoce příbuzné nebo naopak velmi vzdálené sekvence).

PAM – Point Accepted Mutation

- Vytvořila Margaret Dayhoff roku 1978.
- Zahrnuje **pravděpodobnost záměny** jedné aminokyseliny v druhou **během evoluce**
- Předpokládá, že každá další mutace nezávisí na předchozí.
- **PAM1** – Odvozena z globálního alignmentu 71 rodin proteinů (Podobnost sekvencí v rodině > 85%, průměrná 1% záměna)
  - vysoká spolehlivost alignmentu
  - vysoká pravděpodobnost, že záměna aminokyseliny je dána jedinou mutací
- **PAM250** (20% identita) je odvozena od PAM1 její 250-násobnou multiplikací (250 mutací na 100 aminokyselin)

PAM1 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 $\times 10^4$

PAM250 matice

	C	S	T	P	A	G	N	D	E	Q	H	R	K	M	I	L	V	F	Y	W	B	Z
C	12																					
S	0	2																				
T	-2	1	3																			
P	-3	1	0	6																		
A	-2	1	1	1	2																	
G	-3	1	0	-1	1	5																
N	-4	1	0	-1	0	0	2															
D	-5	0	0	-1	1	2	2	4														
E	-5	0	0	-1	0	0	1	3	4													
Q	-5	-1	-1	0	0	-1	1	2	2	4												
H	-3	-1	-1	0	-1	-2	2	1	1	3	6											
R	-4	0	-1	0	-2	-3	0	-1	-1	1	2	6										
K	-5	0	0	-1	-1	-2	1	0	0	1	0	3	5									
M	-5	-2	-1	-2	-1	-3	-2	-3	-2	-1	-2	0	0	6								
I	-2	-1	0	-2	-1	-3	-2	-2	-2	-2	-2	-2	-2	2	5							
L	-6	-3	-2	-3	-2	-4	-3	-4	-3	-2	-2	-3	-3	4	2	6						
V	-2	-1	0	-1	0	-1	-2	-2	-2	-2	-2	-2	-2	2	4	2	4					
F	-4	-3	-3	-5	-4	-5	-4	-6	-5	-5	-2	-4	-5	0	1	2	-1	9				
Y	0	-3	-3	-5	-3	-5	-2	-4	-4	-4	0	-4	-4	-2	-1	-1	-2	7	10			
W	-8	-2	-5	-6	-6	-7	-4	-7	-7	-5	-3	2	-3	-4	-5	-2	-6	0	0	17		
B	-4	0	0	-1	0	0	2	3	2	1	1	-1	1	-2	-2	-3	-2	-5	-3	-5	2	
Z	-5	0	-1	0	0	-1	1	3	3	3	2	0	0	-2	-2	-3	-2	-5	-4	-6	2	3

small, polar

small, nonpolar

polar or acidic

basic

large, hydrophobic

aromatic

Kladné skóre – frekvence substituce je vyšší než odpovídá náhodné záměně.

Nulové skóre – frekvence substituce odpovídá náhodné záměně.

Záporné skóre – frekvence substituce je nižší než odpovídá náhodné záměně.

PAM matice

Předpoklady:

- Mutace AA je nezávislá na předchozích mutacích v téže pozici (Markov process requirement).
- Všechna místa podléhají mutacím rovnoměrně.
- Mutace nezávisí na okolních residuích.
- Kratkodobé a dlouhodobé vlivy na evoluci sekvencí jsou stejně účinné.

Nevýhody:

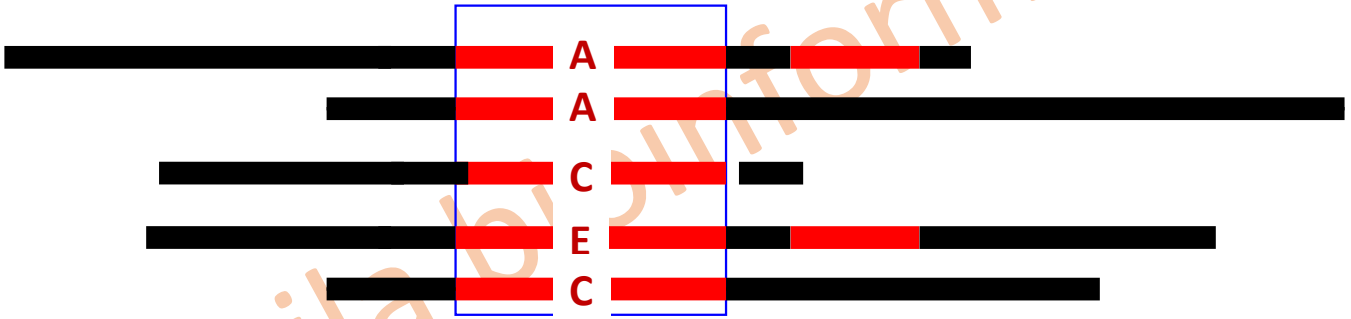
- Pouze matice PAM1 byla “změřena”, všechny ostatní jsou extrapolace (tj. jsou založeny na stejném modelu).
- PAM matice je založená na proteinových sekvencích dostupných v roce 1978 (zejm. malé globulární proteiny). Existují ale nové generace např. PET91.

BLOSUM – Blocks Amino Acid Substitution

- Vytvořena 1992, Henikoff and Henikoff
- Nebere v potaz evoluci
- Používá koncept „bloků“ (database BLOCKS) k identifikaci proteinových rodin
  - **sekvenční motiv** – konzervovaný aminokyselinový úsek 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

BLOSUM – Blocks Amino Acid Substitution

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




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

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

Matice BLOSUM 62

	Ala	Arg	Asn	Asp	Cys	Gln	Glu	Gly	His	Ile	Leu	Lys	Met	Phe	Pro	Ser	Thr	Trp	Tyr	Val
Ala	4																			
Arg	−1	5																		
Asn	−2	0	6																	
Asp	−2	−2	1	6																
Cys	0	−3	−3	−3	9															
Gln	−1	1	0	0	−3	5														
Glu	−1	0	0	2	−4	2	5													
Gly	0	−2	0	−1	−3	−2	−2	6												
His	−2	0	1	−1	−3	0	0	−2	8											
Ile	−1	−3	−3	−3	−1	−3	−3	−4	−3	4										
Leu	−1	−2	−3	−4	−1	−2	−3	−4	−3	2	4									
Lys	−1	2	0	−1	−3	1	1	−2	−1	−3	−2	5								
Met	−1	−1	−2	−3	−1	0	−2	−3	−2	1	2	−1	5							
Phe	−2	−3	−3	−3	−2	−3	−3	−3	−1	0	0	−3	0	6						
Pro	−1	−2	−2	−1	−3	−1	−1	−2	−2	−3	−3	−1	−2	−4	7					
Ser	1	−1	1	0	−1	0	0	0	−1	−2	−2	0	−1	−2	−1	4				
Thr	0	−1	0	−1	−1	−1	−1	−2	−2	−1	−1	−1	−1	−2	−1	1	5			
Trp	−3	−3	−4	−4	−2	−2	−3	−2	−2	−3	−2	−3	−1	1	−4	−3	−2	11		
Tyr	−2	−2	−2	−3	−2	−1	−2	−3	2	−1	−1	−2	−1	3	−3	−2	−2	2	7	
Val	0	−3	−3	−3	−1	−2	−2	−3	−3	3	1	−2	1	−1	−2	−2	0	−3	−1	4

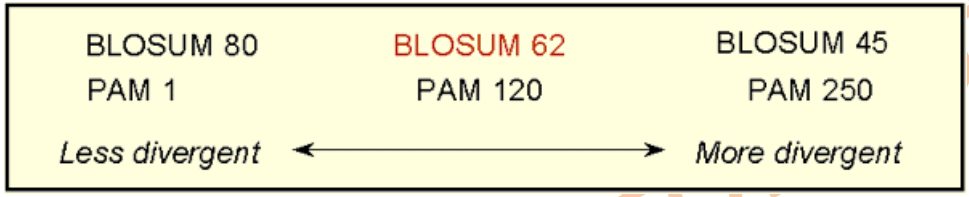
Matice PAM vs. BLOSUM

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

Matice PAM	Matice BLOSUM	Aplikace	Podobnost (%)
PAM100	BLOSUM90	Krátká, vysoce podobná přiložení	70-90
PAM120	BLOSUM80	Detekce členů proteinových rodin	50-60
PAM160	BLOSUM62	Vysoce efektivní pro hledání potenciálních příbuzností	30-40
PAM250	BLOSUM45	Dlouhá přiložení málo příbuzných sekvencí	~ 30

Poslední sloupec udává míru podobnosti sekvencí, pro které je daná matice nejvhodnější.

Odlišné substituční matice jsou pro odlišné účely



- Pro porovnání blízce příbuzných proteinů by se měla používat nižší čísla PAM a vyšší BLOSUM, pro vzdálenější vyšší čísla PAM a nižší BLOSUM
- Pro prohledávání databází je nejběžnější BLOSUM62

GONNETova matice

A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y	..
0.6	0.125	−0.075	0	−0.575	0.125	−0.2	−0.2	−0.1	−0.3	−0.175	−0.075	0.075	−0.05	−0.15	0.275	0.15	0.025	−0.9	−0.55	A
	2.875	−0.8	−0.75	−0.2	−0.5	−0.325	−0.275	−0.7	−0.375	−0.225	−0.45	−0.775	−0.6	−0.55	0.025	−0.125	0	−0.25	−0.125	C
		1.175	0.675	−1.125	0.025	0.1	−0.95	0.125	−1	−0.75	0.55	−0.175	0.225	−0.075	0.125	0	−0.725	−1.3	−0.7	D
			0.9	−0.975	−0.2	0.1	−0.675	0.3	−0.7	−0.5	0.225	−0.125	0.425	0.1	0.05	−0.025	−0.475	−1.075	−0.675	E
				1.75	−1.3	−0.025	0.25	−0.825	0.5	0.4	−0.775	−0.95	−0.65	−0.8	−0.7	−0.55	0.025	0.9	1.275	F
					1.65	−0.35	−1.125	−0.275	−1.1	−0.875	0.1	−0.4	−0.25	−0.25	0.1	−0.275	−0.825	−1	−1	G
						1.5	−0.55	0.15	−0.475	−0.325	0.3	−0.275	0.3	0.15	−0.05	−0.075	−0.5	−0.2	0.55	H
							1	−0.525	0.7	0.625	−0.7	−0.65	−0.475	−0.6	−0.45	−0.15	0.775	−0.45	−0.175	I
								0.8	−0.525	−0.35	0.2	−0.15	0.375	0.675	0.025	0.025	−0.425	−0.875	−0.525	K
									1	0.7	−0.75	−0.575	−0.4	−0.55	−0.525	−0.325	0.45	−0.175	0	L
										1.075	−0.55	−0.6	−0.25	−0.425	−0.35	−0.15	0.4	−0.25	−0.05	M
											0.95	−0.225	0.175	0.075	0.225	0.125	−0.55	−0.9	−0.35	N
												1.9	−0.05	−0.225	0.1	0.025	−0.45	−1.25	−0.775	P
													0.675	0.375	0.05	0	−0.375	−0.675	−0.425	Q
														1.175	−0.05	−0.05	−0.5	−0.4	−0.45	R
															0.55	0.375	−0.25	−0.825	−0.475	S
																0.625	0	−0.875	−0.475	T
																	0.85	−0.65	−0.275	V
																		3.55	1.025	W
																			1.95	Y

Na čem je založeno vyhodnocení „kvality“ sekvenčního přiložení proteinových sekvencí?

snaha o co nejvyšší skóre:

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

Platí u nukleových kyselin i proteinů stejná pravidla ?

Nukleové kyseliny nemá smysl posuzovat podobnost:

sice **tranzice** (R→R or Y →Y) je mnohem častější než **transverze**(R→Y or Y →R), což ale není pro alignement užitečné

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:

match score +5

mismatch score -4

gap penalty (changeable parameter) opening -10, extending -2

DNA matice

- U nukleových kyselin je každá záměna (mutace) závažná a negativní
- Nukleotidy jsou rovnocenné

	A	C	G	T
A	1	-10000	-10000	-10000
C	-10000	1	-10000	-10000
G	-10000	-10000	1	-10000
T	-10000	-10000	-10000	1

Multiple sequence alignment – MSA (mnohonásobné 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**

Metody MSA

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

Heuristické metody

- **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) – opakování alignmentu pro podskupiny sekvencí následující po globálním alignmentu – odstraňuje problémy progresivního alignmentu, který je závislý na prvotním přiložení nejpodobnějších sekvencí pomocí
- **Hledání motivů** – nalezení částí konzervovaných sekvenčních motivů pomocí globálního přiložení a následně „hodnocení“ těchto úseků nezávisle na celé sekvenci
- **Schémata založená na konzistenci** (consistency-based schemes) – vychází z nejlepších možných alignmentů každé dvojice sekvencí. Cílem je dosáhnout maximální konzistence (vnitřní shody).

Dynamické programování

- **Simultánní přiložení všech sekvencí** – analogické párovému přiložení
- Programové balíky: MSA (Lipman et al., 1989) a DCA (Stoye et al., 1997), založené na Carrilově a Lipmanově algoritmu (1988)
- Využívá skórovací matice, ale vytváří n-rozměrný prostor (n = počet sekvencí)
- Extrémně **náročný na výpočetní kapacity**
- I při zjednodušení nepoužitelné pro více než cca 20 sekvencí

Progresivní multiple alignment

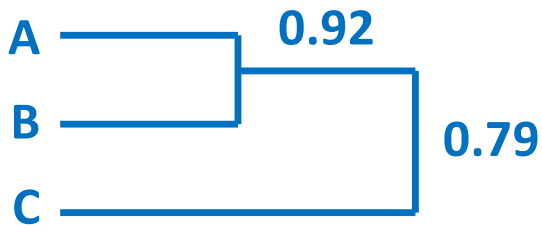
- Používá ho většina programů
- Vznik – 1987

  Feng, D.-F. and Doolittle, R.F. (1987) J. Mol. Evol. 25, 351-360.

1) Sestavení příbuzenského stromu (guide tree) na základě distanční matice (distance matrix) z jednotlivých sekvencí

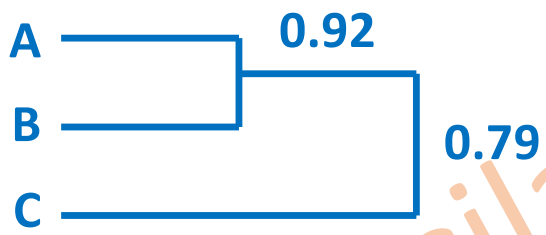
A	-		
B	0.92	-	
C	0.65	0.79	-
	A	B	C

Počet exaktně stejných shod dělený celkovou délkou sekvence (ignoruje mezery)



Progresivní multiple alignment

2) Tvorba párových přiložení postupně podle příbuznosti (topologie guide tree)



Nejdříve provede párové přiložení A a B
Pak přidá sekvenci C do předešlého alignmentu
(inzerce mezer, pokud je potřeba)

3) Často obsahuje iterativní smyčku – možnost úpravy přiložení vysoce příbuzných sekvencí

Guide tree

- **Guide tree** je vypočítán na základě matice vzdáleností (distance matrix) vytvořené podle skóre pairwise alignmentů. Výstupem je .dnd soubor.
  NEMÁ fylogenetický význam

.dnd soubor

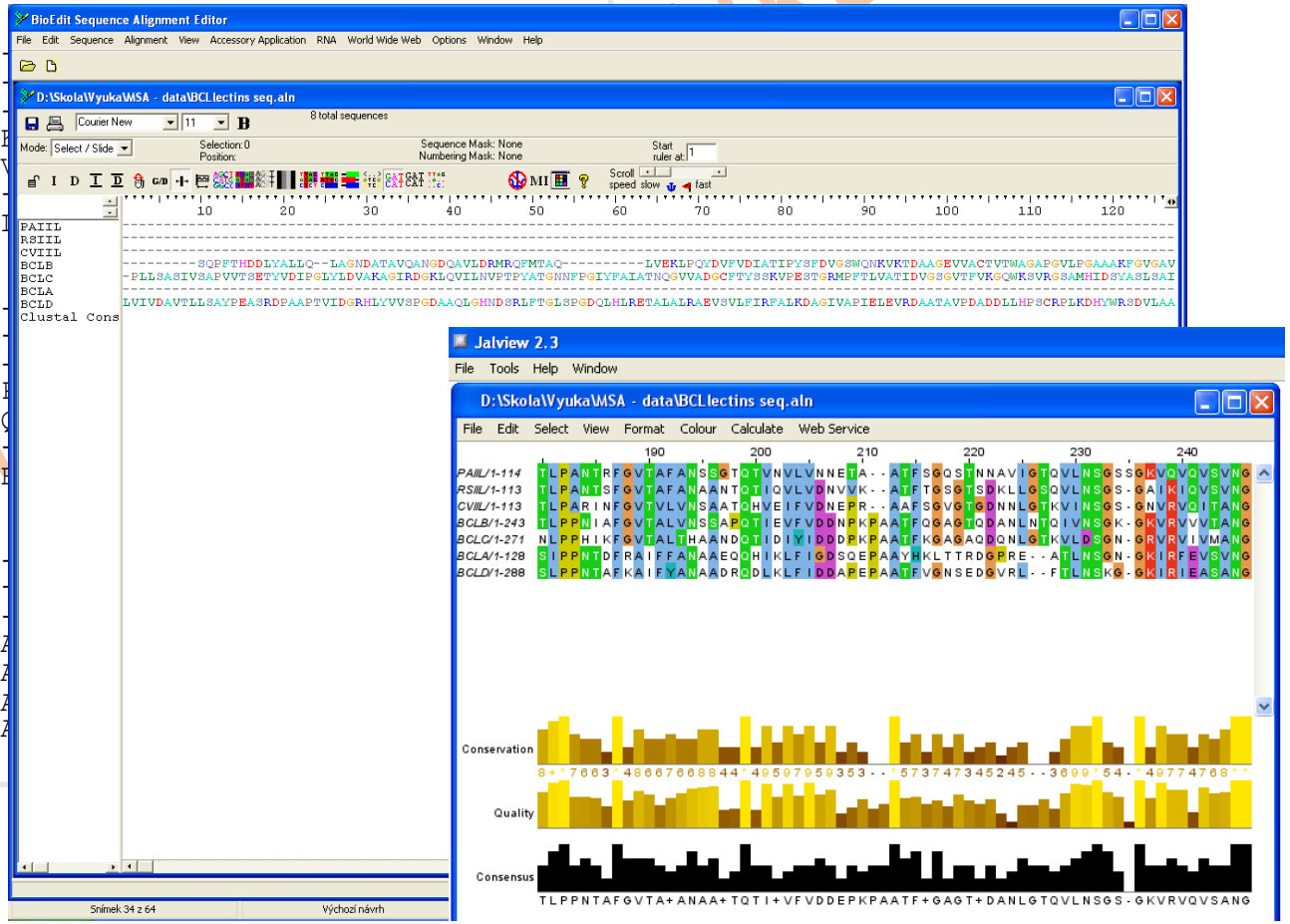
```
(
(
PAIIL:0.16435,
RSIIL:0.13654)
:0.03384,
(
CVIIL:0.16563,
BCLB:0.26800)
:0.02264,
(
(
BCLA:0.17899,
BCLD:0.26633)
:0.18717,
BCLC:0.29707)
:0.03484);
```

Výstup sekvenčního přiložení

- **Sekvence zarovnané podle podobnosti**
- Přidány mezery „-“
- V případě lokálního přiložení pouze úseky s dostatečnou homologií
- Různé výstupní formáty

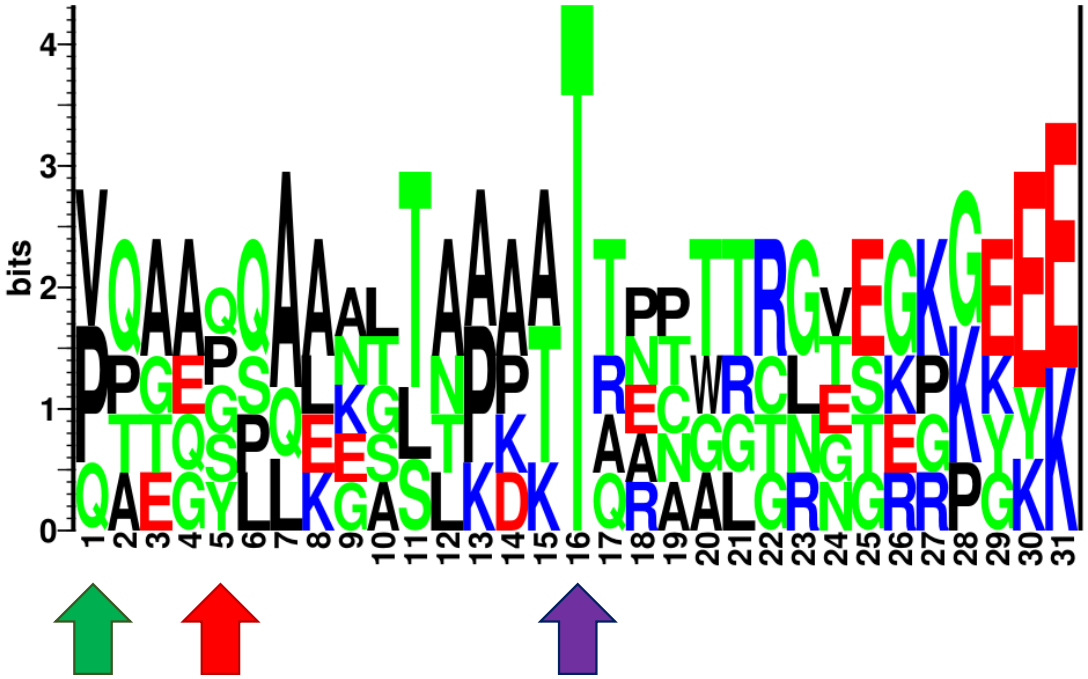
```
CLUSTAL 2.0.10 multiple sequence alignment


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
       :*..  * .   .:::  * :.: :.:     * .      .       ::*   * ::.
```

```
CLUSTAL 2.0.10 multiple sequence alignment


PAIIL           --------------------------------
RSIIL           --------------------------------
CVIIL           --------------------------------
BCLB            ---LVEKLPQYDVFVDIATIPYSFDVGSWQNK
BCLC            AIATNQGVVADGCFTYSSKVPESTGRMPFTLV
BCLA            --------------------------------
BCLD            LRETALALRAEVSVLFIRFALKDAGIVAPIEL


PAIIL           --------------------------------
RSIIL           --------------------------------
CVIIL           --------------------------------
BCLB            KFGVGAVVN----------------YFSKATF
BCLC            YASLSAIWG----------------TAAPSSQ
BCLA            --------------------------------
BCLD            RSDVLAAGATTCTADFAVCDRDGTVSGYFRWF


PAIIL           LPANTRFGVTAFANSSGTQTVNVLVNNETA--
RSIIL           LPANTSFGVTAFANAANTQTIQVLVDNVVK--
CVIIL           LPARINFGVTVLVNSAATQHVEIFVDNEPR--
BCLB            LPPNIAFGVTALVNSSAPQTIEVFVDDNPKPA
BCLC            LPPHIKFGVTALTHAANDQTIDIYIDDDPKPA
BCLA            IPPNTDFRAIFFANAAEQQHIKLFIGDSQEPA
BCLD            LPPNTAFKAIFYANAADRQDLKLFIDDAPEPA
                :*..  * .   .:::  * :.: :.:
```



Sekvenční logo

- Vizualizace alignmentu – zvýraznění konzervovaných aminokyselin
- Vhodné pro kratší sekvence a motivy




```
> VPTAQPQAEGSLAKATTAPATTRNTGRGGEE
> PTAQPQAEGSLAKATTAPATTRNTGRGGEEK
> PQAEGSLAKATTAPATTRNTGRGGEEKKKEK
> QAEGSLAKATTAPATTRNTGRGGEEKKKEKE
> VQGAYQALNLTNPDKTQECWLCLVSGPPYYE
```

Consensus

Symboly vyjadřující „**konzervovanost**“ každého sloupce

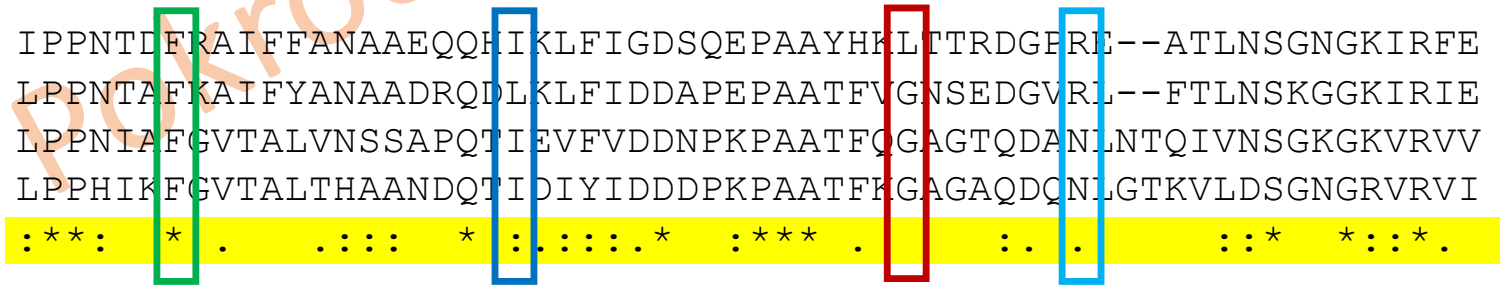
Používán v programu Clustal

* identické residuum ve všech sekvencích

: silně konzervovaný sloupec

. slabě konzervovaný sloupec

Pozor!
Odchylka v jediné sekvenci vede k rozeznání pozice jako nekonzervované.



Skórovací schémata pro párové přiložení

Algoritmy založené na matici (matrix-based algorithms) – např. ClustalW, MUSCLE; pomocí substituční matice je příslušné dvojici (AK) přiřazena hodnota. Rozhoduje pouze **identita** těchto dvou **AK**, případně jejich **nejbližší okolí** (viz. např. BLAST)

Markovovy modely

- Metoda **strojového učení** – model se natrénuje na sadě známých dat
- Prohledávání databází (způsob uložení alignmentu)
- Programy: ClustalOmega, databáze Pfam, SMART, TIGRFAM, aj.

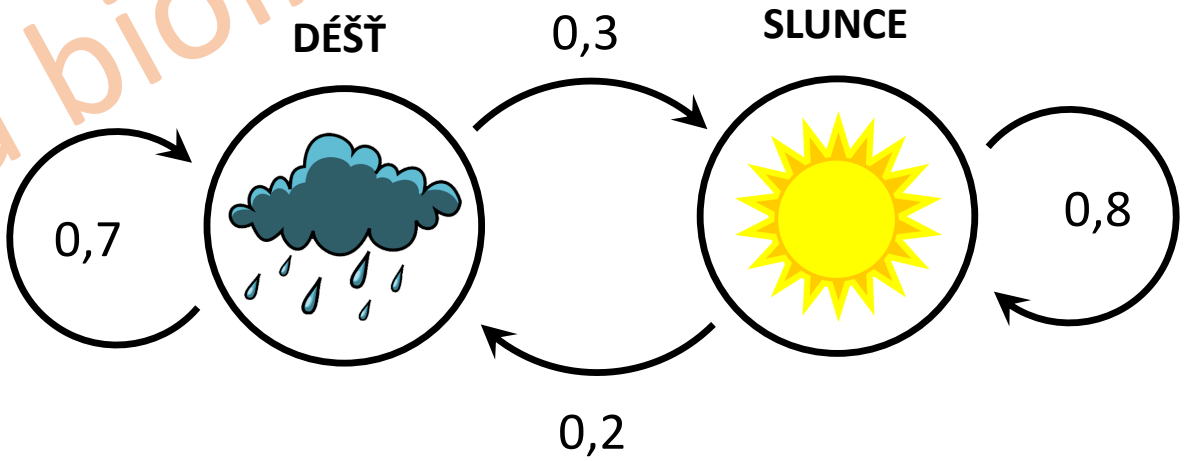
Markovův model

- Alternativní přístup ke skórovací matici
- Obsahuje jednotlivé **stavy** a různě pravděpodobné **přechody** mezi nimi
- Neukládá informace o „minulosti“ – dívá se jen na konkrétní změnu stavu

DDDDSSSSSDDDDSSSSDDSS

10x DÉŠŤ → 7x DÉŠŤ
→ 3x SLUNCE

10x SLUNCE → 8x SLUNCE
→ 2x DÉŠŤ



Markovův model

- Informace o „blízké minulosti“ se dá zahrnout s využitím většího množství stavů

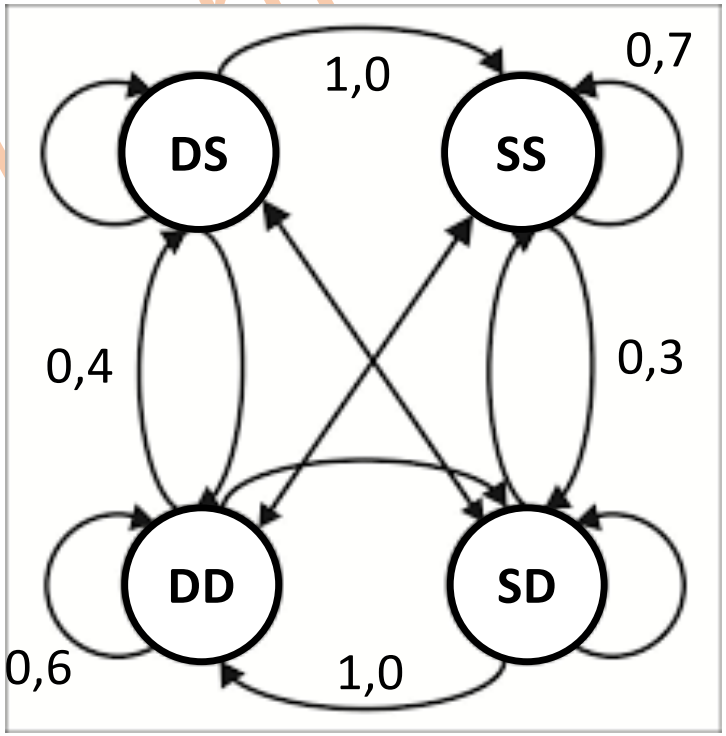
DDDDSSSSSDDDDSSSSDDSS

3x DS → 3x SS

7x SS → 5x SS
→ 2x SD

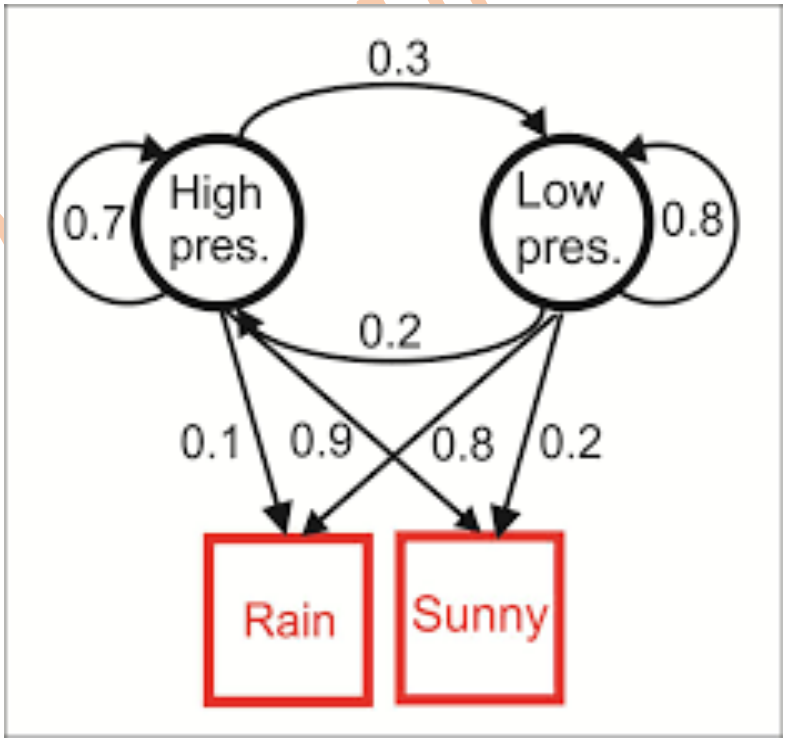
7x DD → 4x DD
→ 3x DS

2x SD → 2x DD



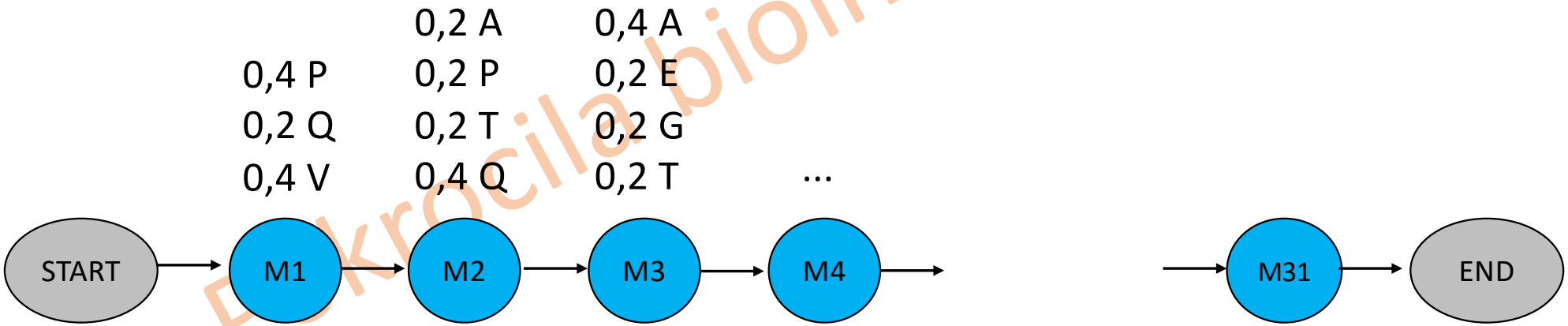
Skrytý Markovův model (HMM)

- V modelu nevidíme jednotlivé stavy
- Stavy se s určitou pravděpodobností projeví na výstupu
- Široké použití v bioinformatice



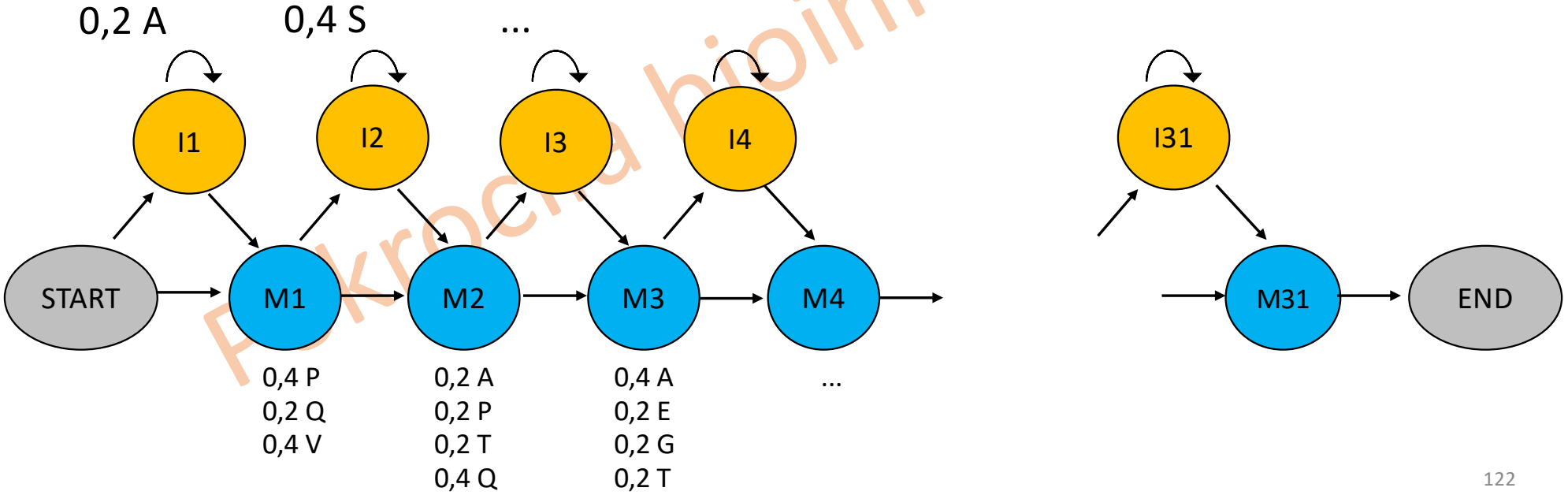
Profilový HMM

```
> VPTAQPQAEGSLAKATTAPATTRNTGRGGEE
> PTAQPQAEGSLAKATTAPATTRNTGRGGEEK
> PQAEGSLAKATTAPATTRNTGRGGEEKKKEK
> QAEGSLAKATTAPATTRNTGRGGEEKKKEKE
> VQGAYQALNLTNPDKTQECWLCLVSGPPYYE
```



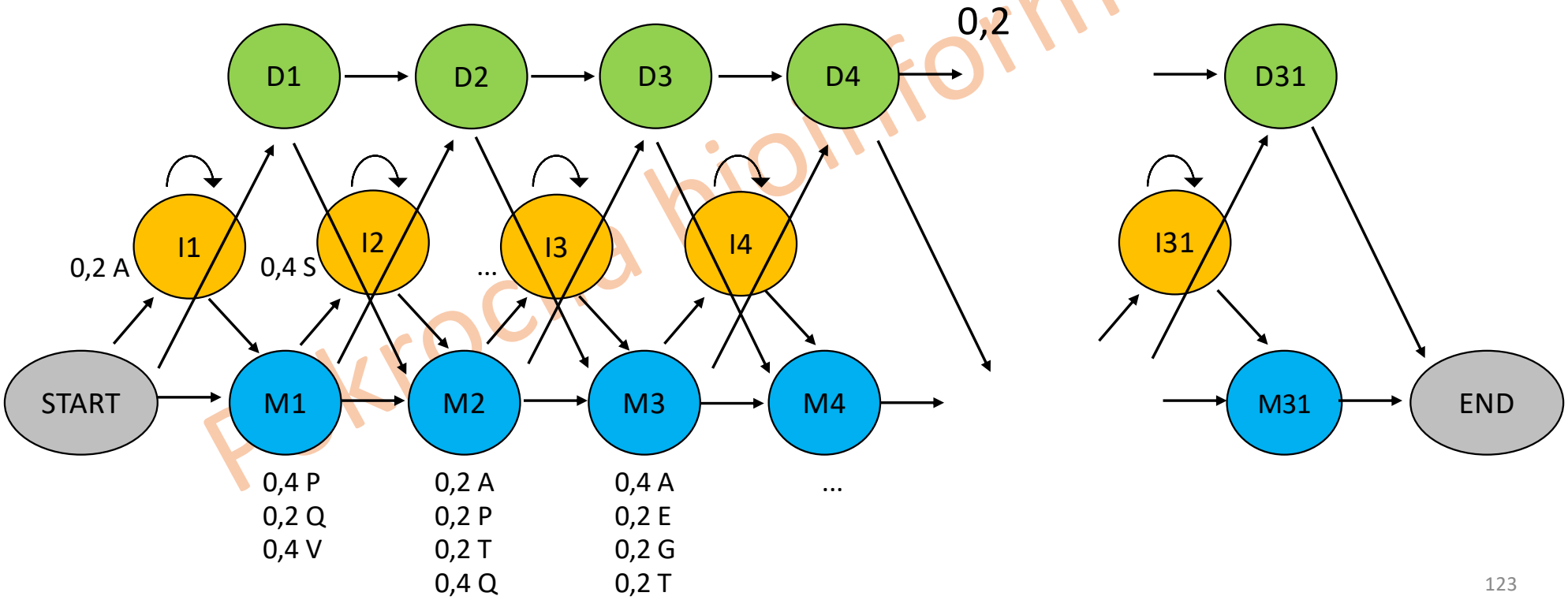
Profilový HMM

```
> V-PTAQPQAEGSLAKATTAPATTRNTGRGGEE
> P-TAQPQAEGSLAKATTAPATTRNTGRGGEEK
>AP-QAEGSLAKATTAPATTRNTGRGGEEKKKEK
> QSAEGSLAKATTAPATTRNTGRGGEEKKKEKE
> VSQGAYQALNLTNPDKTQECWLCLVSGPPYYE
```



Profilový HMM

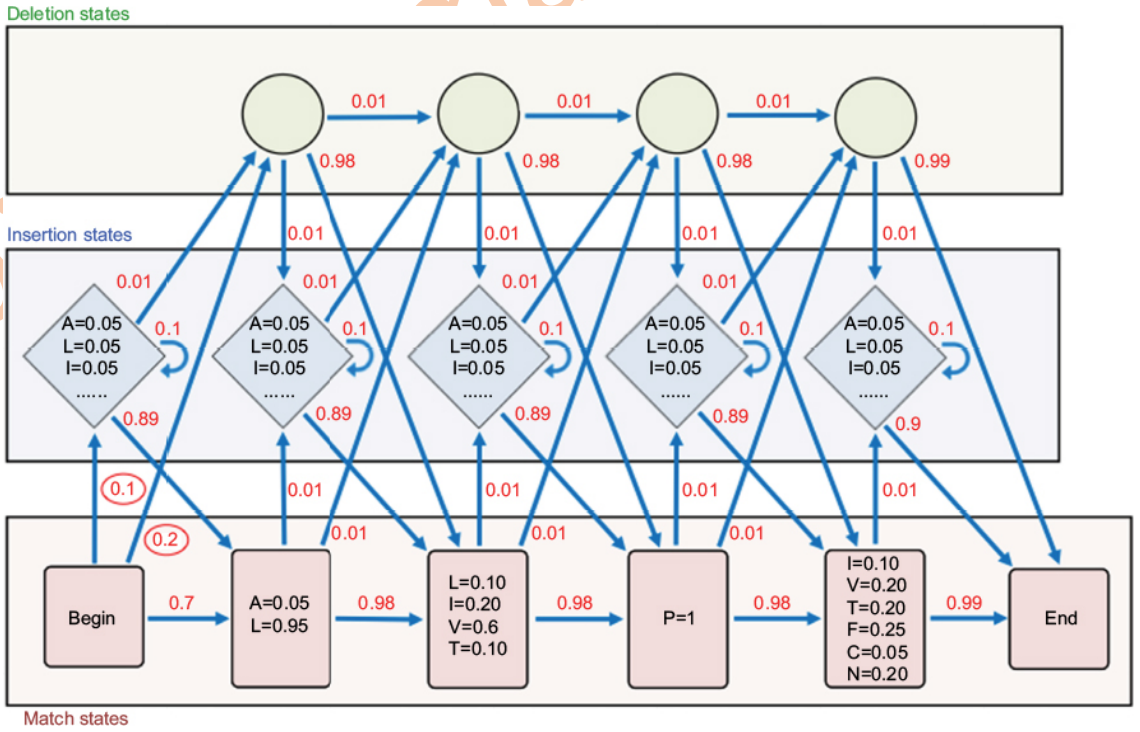
```
> VPTAQPQAEGSLAKATTAPATTRNTGRGGEE
> PTAQPQAEGSLAKATTAPATTRNTGRGGEEK
> PQAEGSLAKATTAPATTRNTGRGGEEKKKEK
> QAEG---AKATTAPATTRNTGRGGEEKKKEKE
> VQGAYQALNLTNPDKTQECWLCLVSGPPYYE
```



Profilový HMM

```
> -------VPTAQPQ--AEGSLAKATTAPATTRNTGRGGEE------
> --------PTAQPQ--AEGSLAKATTAPATTRNTGRGGEEK-----
> ------------PQ--AEGSLAKATTAPATTRNTGRGGEEKKKEK-
> -------------Q--AEGSLAKATTAPATTRNTGRGGEEKKKEKE
> VQGAYQALNLTNPDKTQECWLCLVSGPPYYE---------------
```

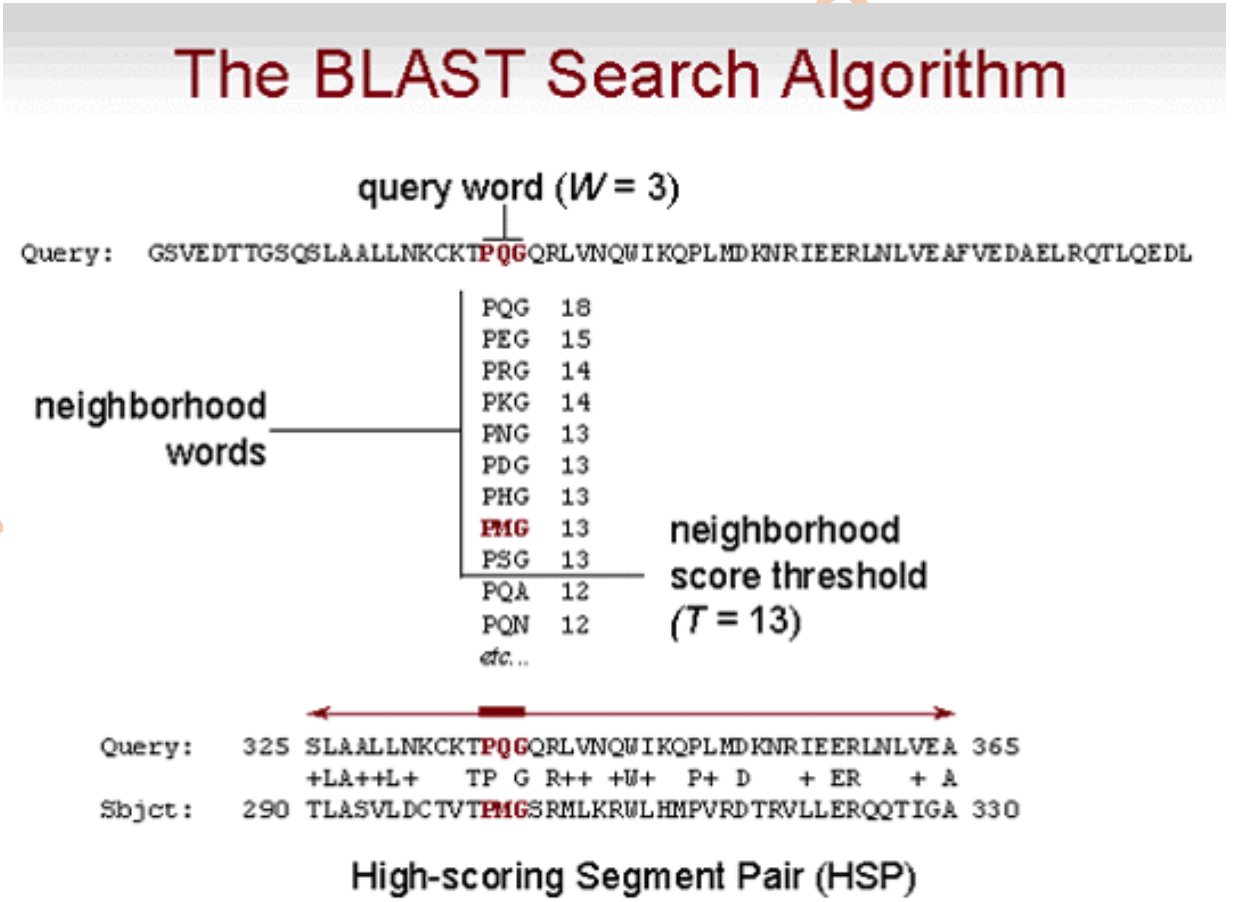
- Tvorba alignmentu
  - Každá sekvence musí projít modelem
- Databáze proteinových rodin
  - Profilový HMM se vytvoří na základě multiple sequence alignmentu
  - U nových sekvencí se zjišťuje s jakou pravděpodobností projdou daným modelem



BLAST algoritmus

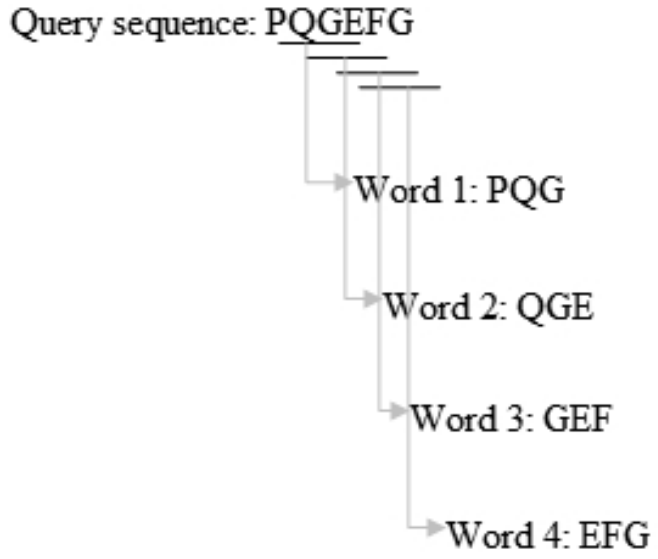
(Basic Local Alignment Search Tool)

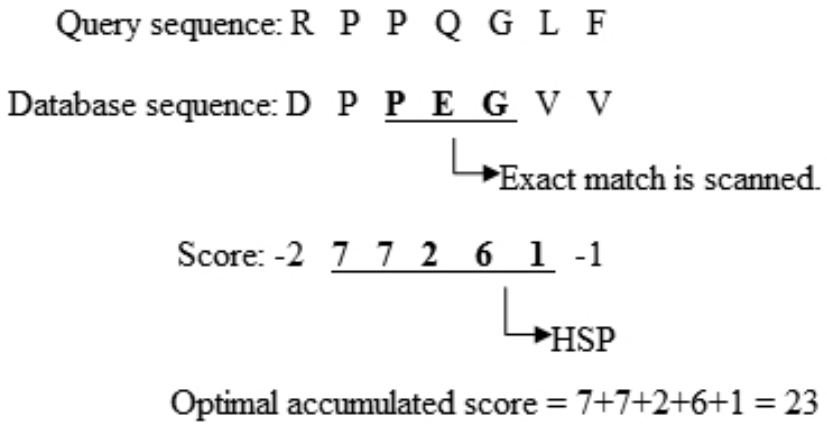
Heuristický algoritmus jehož základem je **hledání slov** (několikapísmenných sekvencí), s dostatečnou podobností (poskytují dostatečně vysoké skóre v substituční matici)



- **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.





Odlišné možnosti použití BLASTu

- QuickBLASTP
  is an accelerated version of BLASTP that is very fast and works best if the target percent identity is 50% or more.
- BlastP
  simply compares a protein query to a protein database.
- PSI-BLAST
  allows the user to build a PSSM (position-specific scoring matrix) using the results of the first BlastP run.)
- PHI-BLAST
  performs the search but limits alignments to those that match a pattern in the query.
- DELTA-BLAST
  constructs a PSSM using the results of a Conserved Domain Database search and searches a sequence database.

Jak statisticky významné je skóre?

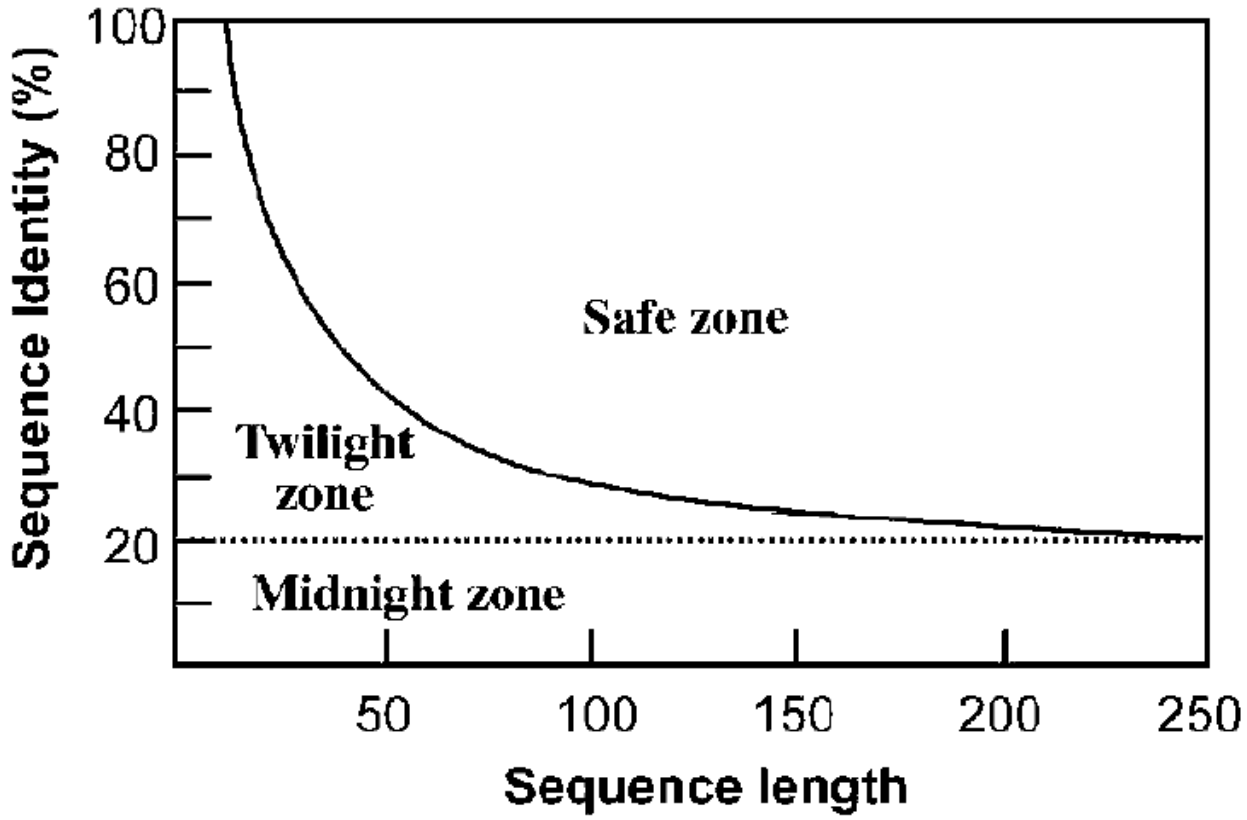
Pokud je podobnost dostatečně významná lze usuzovat na společné evoluční vztahy . Ale co je DOSTATEČNĚ?

závisí na typu sekvence a její délce

- Pravděpodobnost, že dvě rezidua v nepříbuzných sekvencích jsou identické?

  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

Co to jsou oblasti sekvencí tzv. „low complexity regions“ proč se definují a jak se používají?

Vysoce repetitivní krátké segmenty AATAAAAAAAAATAAAAAAT

- Hojně zastoupeny v databázích (cca 15% proteinů)
- Mohou vést k uměle vysokým hodnotám výsledných skóre nepříbuzných sekvencí
- Proto je nezbytné je vyjmout ze zadávacího dotazu stejně jako ze sekvenčních databází.

MSA „programy“

Za posledních 25 let vzniklo přes 50 MSA programových balíků

- 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)
- Kalign2 (Lassmann, 2009)
- Clustal Omega (Sievers, 2011)
- …

Clustal

http://www.clustal.org/

- Dlouhodobě **nejužívanější** program
- Různé verze:
  - Clustal *(Higgins and Sharp, 1988)*
  - Clustal W *(Thompson et al., 1994)*
  - Clustal X *(Jeanmougin et al., 1998)*
  - Clustal Ω *(Sievers et al., 2011)*
- Využívá progresivní alignment

ClustalW: Jednotlivým sekvencím přiřazuje **váhy** (weight – W) podle četnosti zastoupení (čím více jsou si sekvence podobné, tím nižší mají váhu a naopak) a penalizuje přítomnost mezer v závislosti na jejich pozici (position-specific gap penalties)

Clustal W

1. Provedení **pairwise alignmentů** pro každou dvojici sekvencí a určení jejich podobnosti – v závislosti na množství neodpovídajících residuí a mezer

2. Sestavení **příbuzenského stromu** (similarity tree)

3. **Kombinace** alignmentů (viz. 1.) v pořadí dle příbuznosti – od nejvíce podobných k nejméně příbuzným (viz. 2.). Jednou vložené mezery jsou zachovány.

Clustal Ω

1. Provedení **pairwise alignmentů** urychleno použitím modifikovaného algoritmu mBed – převedení sekvencí na n-rozměrný vektor a následný alignment vektorů

2. Sestavení **příbuzenského stromu** (similarity tree)

3. **Sestavení** alignmentů užitím přesného algoritmu HHalign (využití skrytých Markovových modelů).

Určen pro obsáhlé alignmenty.

V roce 2011 přiloženo 190 000 sekvencí během několika hodin.

MUSCLE

(**MU**ltiple **S**equence **C**omparison by **L**og-**E**xpectation)

https://www.ebi.ac.uk/Tools/msa/muscle/

Rychlejší určení „vzdálenosti“ dvou sekvencí

Tzv. log-expectation skórovací funkce

Refinement metodou restricted partitioning

Zahrnutí **iterace** pro zpřesnění přiložení

Vhodný i pro velký počet sekvencí (5000 seq po 350 bp za 7 min na PC – rok 2004)

Edgar, R.C. (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput

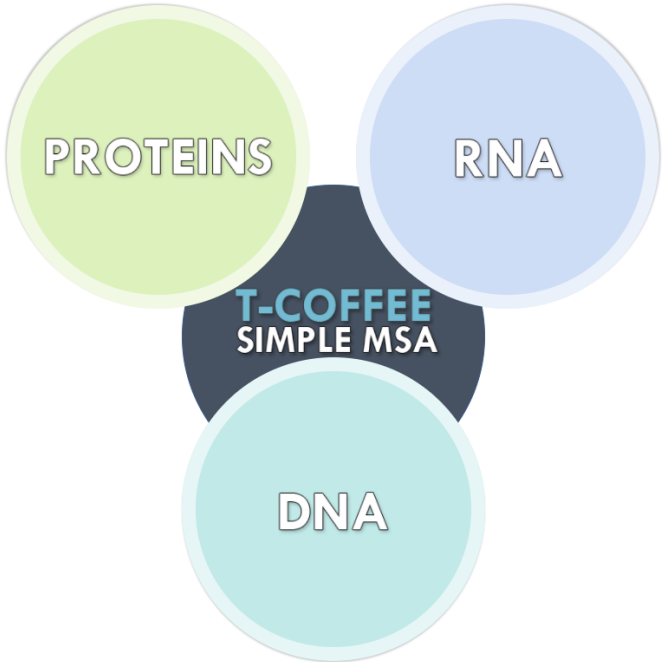
T-Coffee

(**T**ree-based **C**onsistency **O**bjective **F**unction for alignment **E**valuation)

http://tcoffee.crg.cat/

- Pomalejší ale výrazně přesnější než ClustalW

Hlavním rozdílem oproti tradičním metodám progresivního alignmentu je použití pozičně specifického skórovacího schématu (**extended library**) namísto substituční matice.



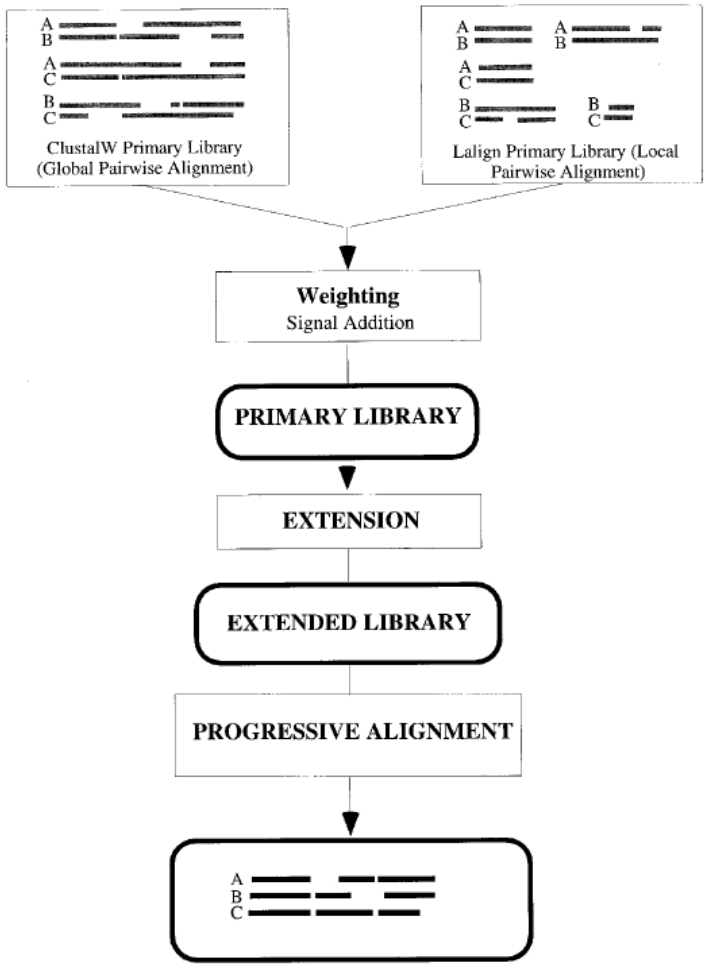
Notredame C. et al (2000) T-Coffee: A novel method for fast and accurate multiple sequence alignment.

T-Coffee

1) Provedení pairwise alignmentů pro všechny dvojice sekvencí pomocí **globálního** a pomocí **lokálního alignmentu** (dvě primární knihovny).

2) Jednotlivým pairwise alignmentům je přiřazena **váha** podle poměru počtu identických residuí k celkovému počtu residuí.

3) Kombinace obou knihoven. Pokud je rozdíl v globálním a lokálním alignmentu, jsou zachovány oba s příslušnou váhou. Vzniká **pozičně specifická matice** (extended library), která je dále použita pro vlastní progresivní alignment.



Zlepšení přesnosti – kombinace přístupů

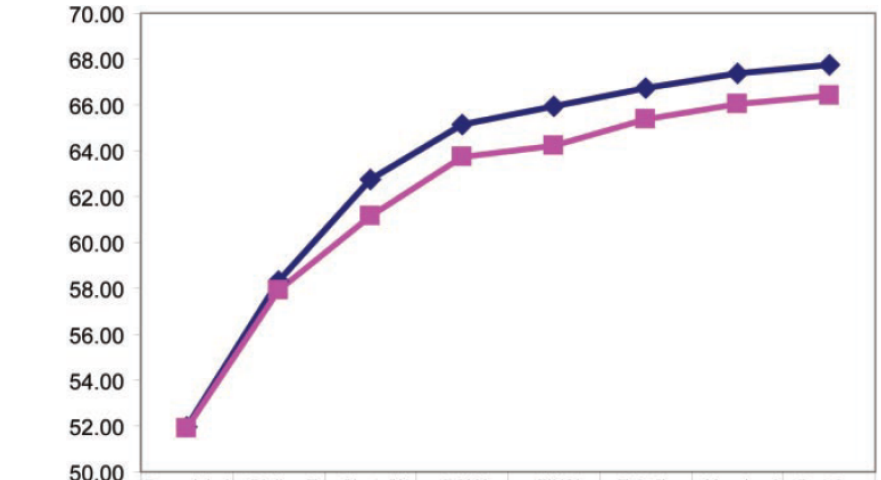
- Různé algoritmy/programy poskytují odlišná přiložení
- Kombinace přístupů může poskytnout lepší výsledek

Řešení: vytvoření přiložení s použitím výstupů několika alignmentových programů.

M-Coffee

http://tcoffee.crg.cat/

- Založen na algoritmu T-Coffee
- Je schopen **kombinovat data z více předchozích alignmentů**, které mohly být vytvořeny různými postupy (lokální, globální, strukturní podobnost,…)
- Zvýšení přesnosti alignmentu



	Poa -global	+Dialign-T	+ClustalW	+PCMA	+FINSI	+T-Coffee	+Muscle v6	+ProbCons
Combined	51.96	58.32	62.75	65.15	65.94	66.73	67.38	67.75
Default	51.90	57.92	61.15	63.73	64.22	65.37	66.04	66.41

Wallace I. M. et al (2006) M-Coffee: combining multiple sequence alignment methods with T-Coffee

Zlepšení přesnosti – strukturní informace

- Sekvence s vyšší homologií (>40%) – vysoká přesnost alignmentu
- Bez homologie – nepoužitelné
- Tzv. twilight zone – málo podobné sekvence (nižší než 20% homologie) = špatná (méně než 30%) přesnost alignmentu

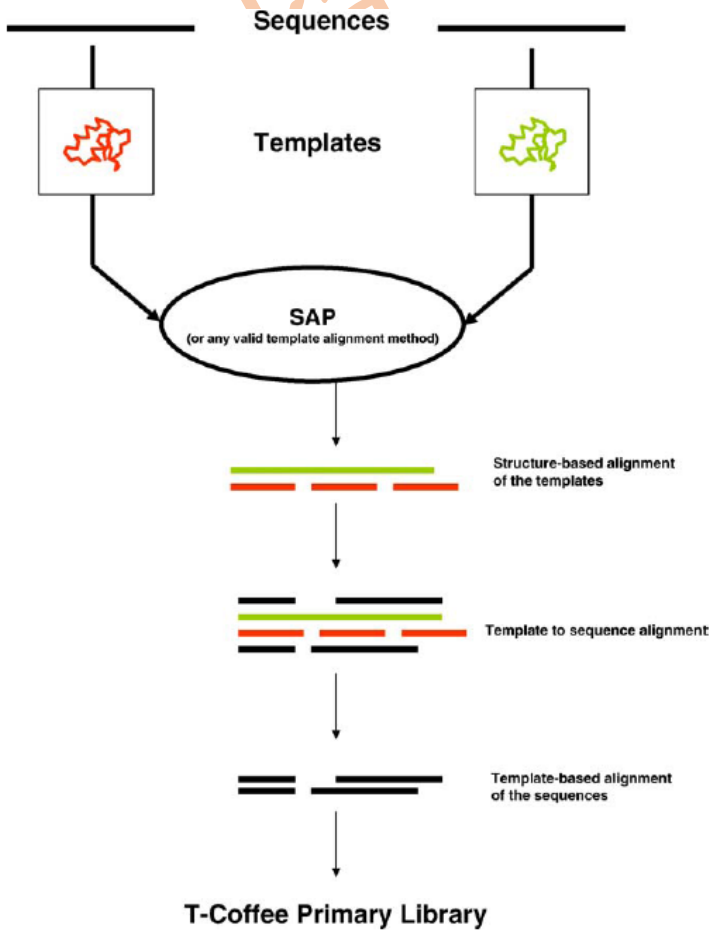
Řešení: nejčastěji využití znalosti strukturní podobnosti (2D nebo 3D), která se behem evoluce zachovává více než sekvence AK.

Template-based alignment metody – využití známých homologních proteinů (srovnání dle jejich struktury nebo tvorba profilu homologních sekvencí)

Expresso

- MSA nástroj založený na algoritmu T-Coffee
- Srovnává sekvence za **užití strukturní informace**.

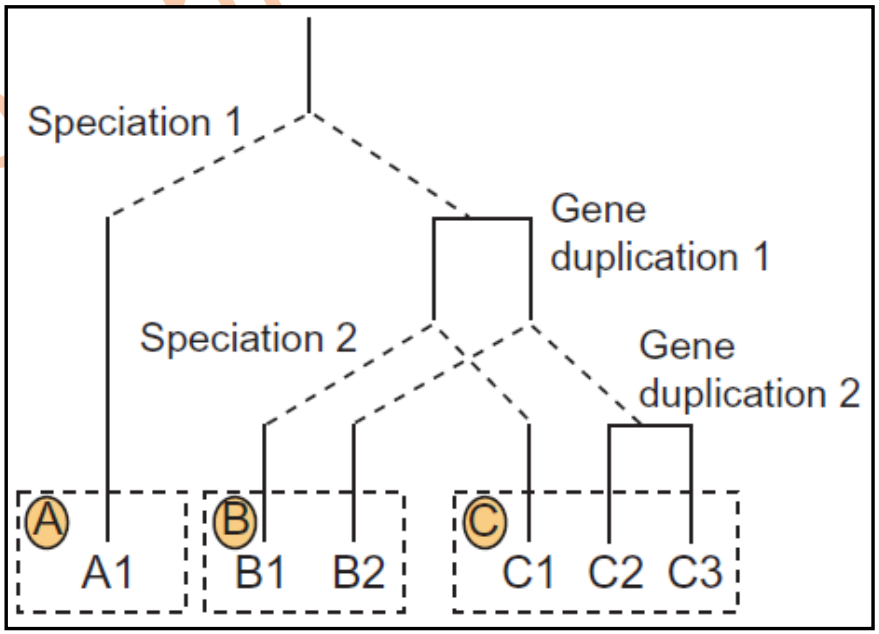
- Vyhledání homologních sekvencí v databázi struktur (PDB) pomocí algoritmu BLAST
- Použití těchto struktur jako templátů pro následný alignment zadaných sekvencí pomocí metod MSA založených na struktuře (např. SAP, Fugue).



Jaký je rozdíl mezi:
„homology“ a „similarity“

MAMUZDOSTSTAROSTISHAMIZNOSTIRATOLESTI
MAMRADOSTZESTAROZITNOSTI

Jaký je rozdíl mezi:
„ortholog“ a „paralog“



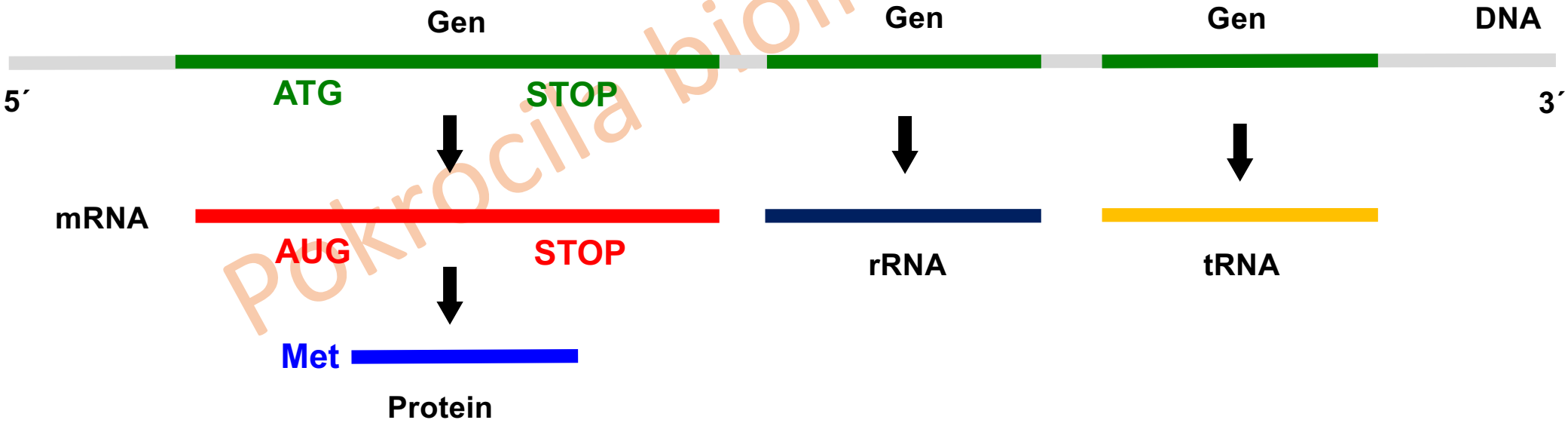
Co si odnést?

- Alignment je **přiložení dvou či více sekvencí** na základě jejich podobnosti
- Můžeme ho využít např. pro **analýzu sekvencí**, zjišťování jejich **příbuznosti** či tvorbu **fylogenetických stromů**
- Řada programů využívá rozdílné přístupy a algoritmy
- Každý program je **kompromisem mezi přesností a rychlostí**
- Každý alignment potřebuje lidskou kontrolu

Sekvence a predikce genů

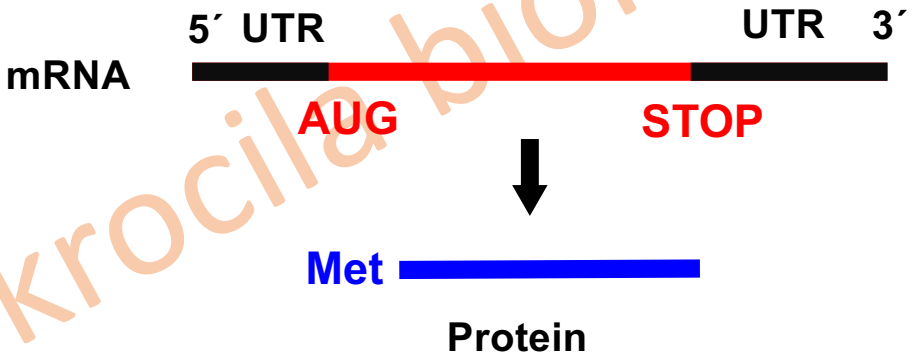
DNA sekvence vs. Sekvence proteinu

- **Gen** – jednotka genetické informace
- Obsahuje informaci o primární struktuře translačního produktu (strukturní geny) nebo funkční molekuly produktu transkripce (tRNA, rRNA).



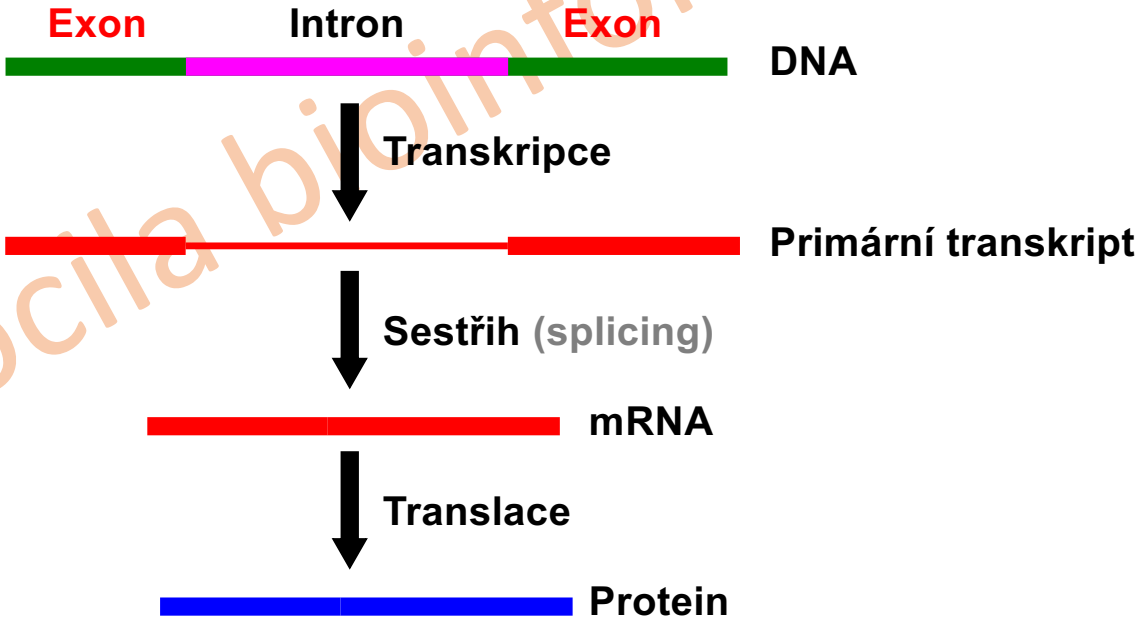
DNA sekvence vs. Sekvence proteinu

- Přepisovaná sekvence DNA je delší než odpovídající kódující úsek
- Části před a po kódujícím úseku se nepřekládají (UTR), mají řídící funkce (začátek a konec translace)

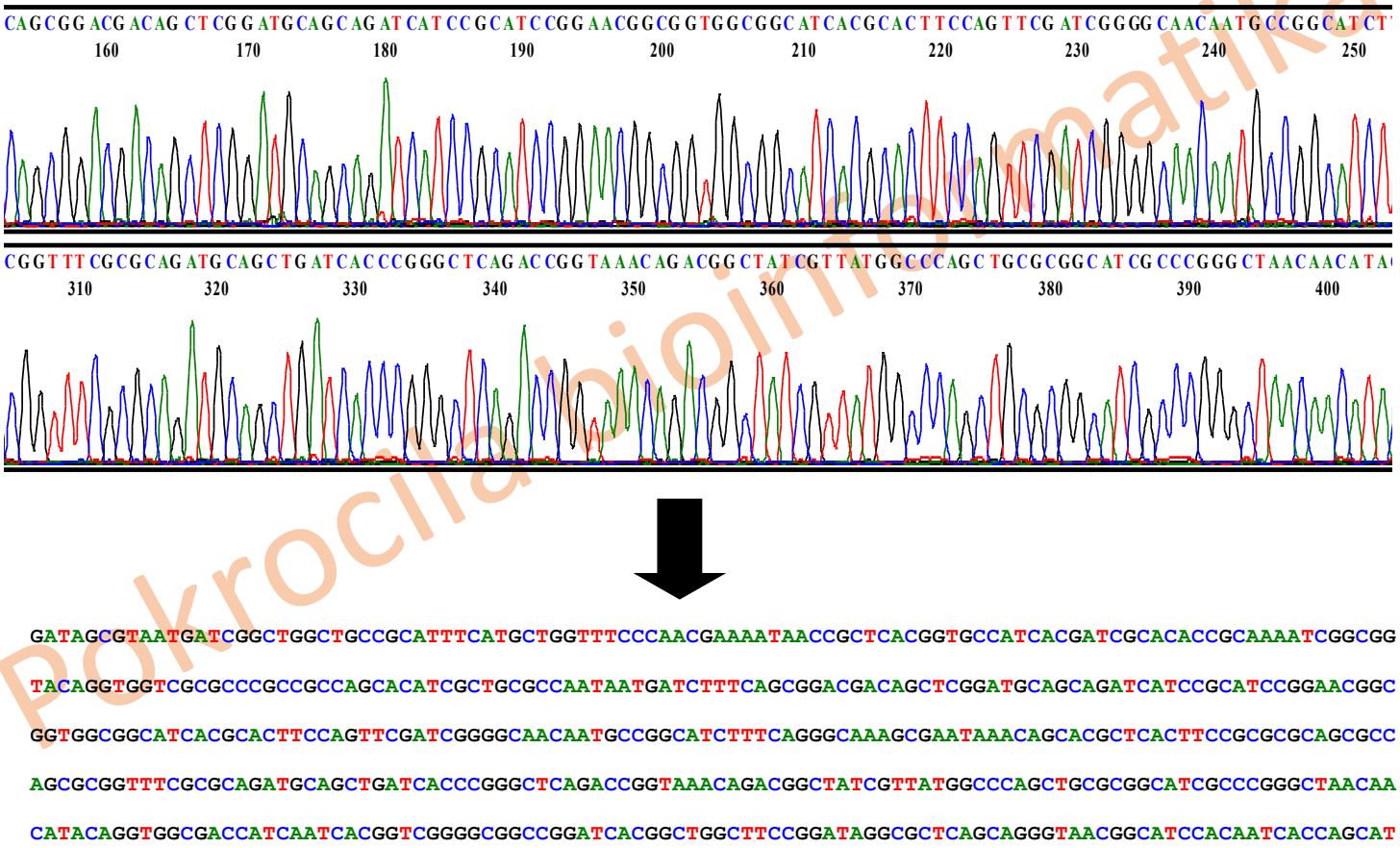


Složený gen

- Geny eukaryotických organismů obsahují často exony a introny. Přepisy intronů jsou vyštěpovány (sestřih) a na ribosomu se překládají pouze spojené exony.



Molekulárně biologická data



GATAGCGTAATGATCGGCTGGCTGCCGCATTTCATGCTGGTTTCCCAACGAAAATAACCGCTCACGGTGCCATCACGATCGCACACCGCAAAATCGGCGG

TACAGGTGGTCGCGCCCGCCGCCAGCACATCGCTGCGCCAATAATGATCTTTCAGCGGACGACAGCTCGGATGCAGCAGATCATCCGCATCCGGAACGGC

GGTGGCGGCATCACGCACTTCCAGTTCGATCGGGGCAACAATGCCGGCATCTTTCAGGGCAAAGCGAATAAACAGCACGCTCACTTCCGCGCGCAGCGCC

AGCGCGGTTTCGCGCAGATGCAGCTGATCACCCGGGCTCAGACCGGTAAACAGACGGCTATCGTTATGGCCCAGCTGCGCGGCATCGCCCGGGCTAACAA

CATACAGGTGGCGACCATCAATCACGGTCGGGGCGGCCGGATCACGGCTGGCTTCCGGATAGGCGCTCAGCAGGGTAACGGCATCCACAATCACCAGCAT

„Syrové“ sekvence DNA

Identifikace a anotace genů a proteinů

Table 1
Software commonly used for bacterial genome annotation and comparison

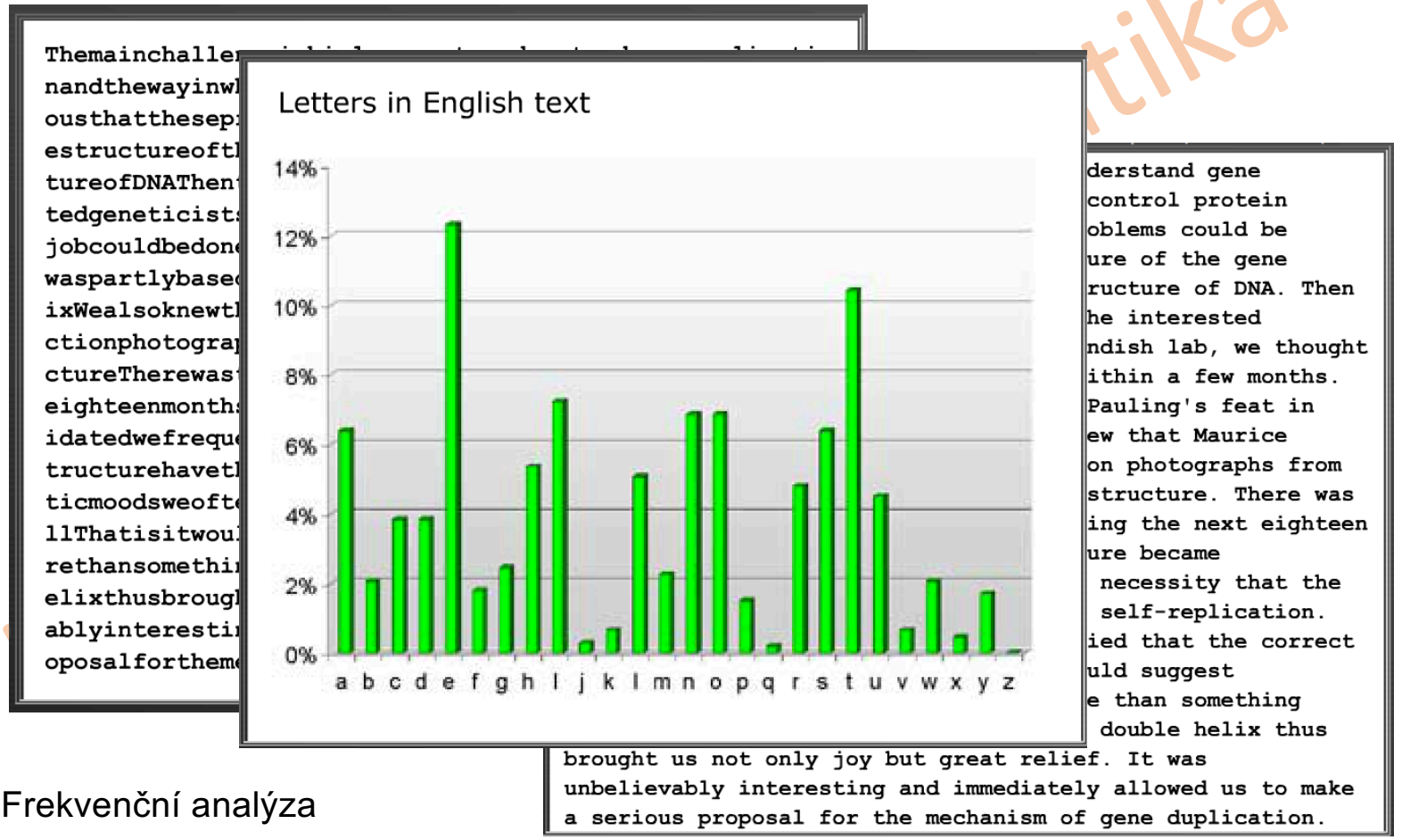
DNA level annotation		
GeneMark	http://exon.gatech.edu/genemark/	Protein gene prediction
Glimmer	http://www.genomics.jhu.edu/Glimmer/	Protein gene prediction
SHOW	http://genome.jouy.inra.fr/ssb/SHOW/	Protein gene prediction
tRNAscan-SE	http://lowelab.ucsc.edu/tRNAscan-SE/	tRNA gene prediction
RNAmmer	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 je **prvním krokem** v anotaci genů a genomů.

Posloupnost písmen může (a nemusí) mít význam

- sekvence nukleotidů
- počítačové 0 a 1
- běžný jazyk

Smysluplná sekvence?



Frekvenční analýza

Smysluplná sekvence?

Frekvenční analýza

Smysluplná sekvence?

01010101010101010101

01010101010101010101

Sekvence nemůže být současně náhodná i smysluplná!

Náhodná nebo smysluplná?

Frekvenční analýza

číslo	počet	poměr
0	10 (60)	50%
1	10 (60)	50%

01010101010101010101

číslo	počet	poměr
0	10	50%
1	10	50%

Očekávaná frekvenční analýza
párů pro náhodnou sekvenci

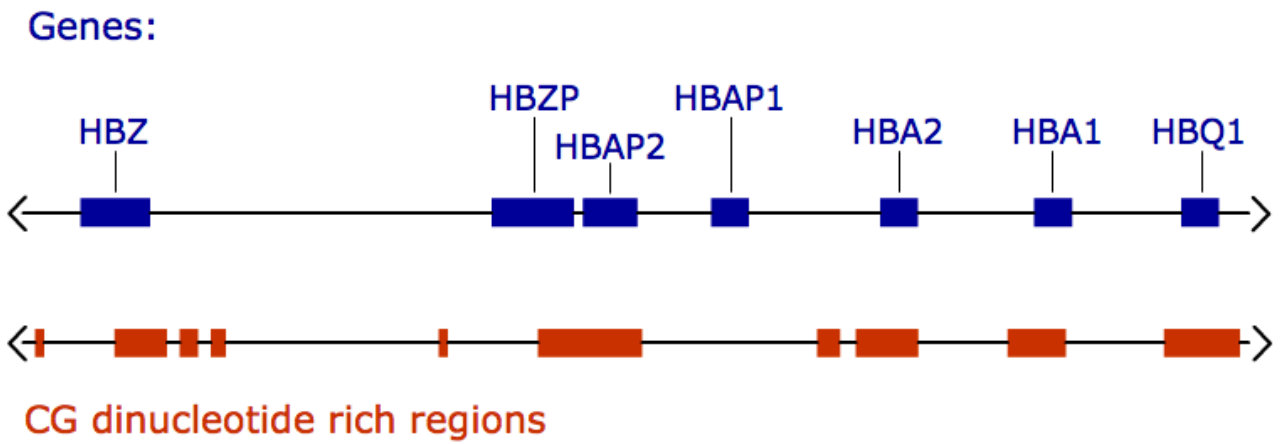
číslo	počet	poměr
00		25%
11		25%
01		25%
10		25%

Frekvenční analýza
párů pro výše uvedenou sekvenci

číslo	počet	poměr
00	0	0%
11	0	0%
01	10	53%
10	9	47%

K čemu je to dobré?

- Obsah GC je např. vyšší v genových částech než intergenových
- GC ostrůvky se objevují v oblastech regulujících transkripci, …



Predikce genů

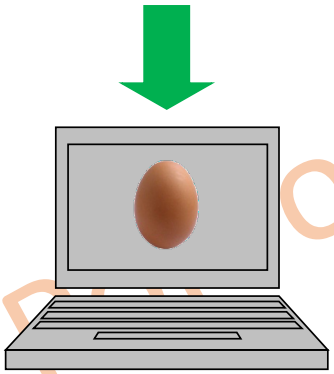
- Predikce genů je **prvním krokem** v anotaci genů a genomů.
- Zahrnuje identifikaci **ORF** - otevřených čtecích rámců (Jako predikce „genů“ se mnohdy označuje právě pouze predikce ORF).
- V případě eukaryot (složené geny) predikce zahrnuje také identifikaci **exonů/intronů**, tj. míst sestřihu. Velmi **problematická**, vzniká velké množství chyb.
- Predikce genů se velmi často soustředí na geny kódující proteiny.
- Predikce genů u prokaryot funguje **výrazně lépe** než u eukaryot (souvislost s organizací genomu prokaryot).

Metody predikce genů

- Dva hlavní přístupy: metody ***ab initio***/metody založené na **homologii** (sekvenční).

GATAGCGTAATGATCGGCTGGCTGCCGCATTTCATGCTGGTTTCCCAACGAAAATAACCGCTCACGGTGCCATCACGATCGCACACCGCAAAATCGGCGG

TACAGGTGGTCGCGCCCGCCGCCAGCACATCGCTGCGCCAATAATGATCTTTCAGCGGACGACAGCTCGGATGCAGCAGATCATCCGCATCCGGAACGGC



GATAGCGTAATGATCGGCTGGCTGCCGCATTTCATGCTGGTTTCCCAACGAAAATAACCGCTCACGGTGCCATCACGATCGCACACCGCAAAATCGGCGG

TACAGGTGGTCGCGCCCGCCGCCAGCACATCGCTGCGCCAATAATGATCTTTCAGCGGACGACAGCTCGGATGCAGCAGATCATCCGCATCCGGAACGGC

NCBI National Center for Biotechnology Information

EMBL-EBI

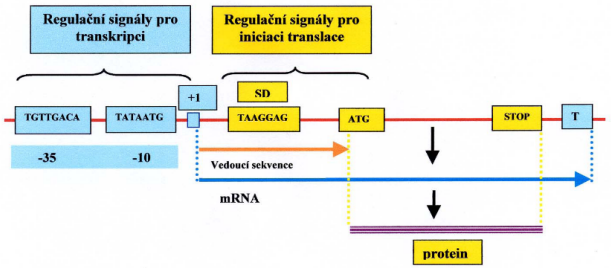
DDBJ DNA Data Bank of Japan

```
LPPNTAFKAIFYANAADRQDLKLFIDDAPEPAATFVGNSEDGVRL--FTLNSKGGKIRIE
IPPNTDFRAIFFANAAEQQHIKLFIGDSQEPAAYHKLTTRDGPRE--ATLNSGNGKIRFE
LPPHIKFGVTALTHAANDQTIDIYIDDDPKPAATFKGAGAQDQNLGTKVLDSGNGRVRVI
LPPNIAFGVTALVNSSAPQTIEVFVDDNPKPAATFQGAGTQDANLNTQIVNSGKGKVRVV
lPPn-aFg---lanaad-QtiklfidD-p-PAAtfkgag-----l-t-tlnSgnGkiRve

ASANGRQSATDARLAPLSAGD------TVWLGWLGAEDGADADYNDGIVILQWPIT
VSVNGKPSATDARLAPINGKKSDGSPFTVNFGIVVSEDGHDSDYNDGIVVLQWPIG
VMANGRPSRLGSRQVDIFKKS--------YFGIIGSEDGADDDYNDGIVFLNWPLG
VTANGKPSKIGSRQVDIFKKT--------YFGLVGSEDGGDGDYNDGIAILNWPLG
vsaNGrpSat--R---ifkks------tvyfGivgsEDGaDaDYNDGIviLqWPig
```

Metody predikce genů

- Dva hlavní přístupy: metody ***ab initio***/metody založené na **homologii** (sekvenční).
- ***Ab initio*** – predikce genů založená pouze na **sekvenci**, jejích vlastnostech a statistických parametrech.

Regulační a signální sekvence: startovní/stop kodon, sestřihové signály, RBS (vazebné místo pro ribosom), polyadenylační signál.
Kodon=triplet (délka genu je v násobcích tří).
Nukleotidové složení kódujících a nekódujících oblastí se liší.



Metody predikce genů

- Dva hlavní přístupy: metody ***ab initio***/metody založené na **homologii** (sekvenční).
- ***Ab initio*** – predikce genů založená pouze na **sekvenci**, jejích vlastnostech a statistických parametrech.
- **Metody založené na homologii** – sekvenční podobnost se známými geny/proteiny. ORF kóduje protein, který je podobný již dříve popsanému proteinu (prohledávání **DATABÁZÍ** pomocí **ALIGNMENTU**) = nejspolehlivější predikce. Problém – unikátní geny bez známých homologů (většinou nejzajímavější).
- Kombinace obou postupů

Predikce genů u prokaryot

- **Prokaryotické genomy**: malé (0,5 až 10 Mbp) a kompaktní, vysoká hustota genů, 90% genomu je kódující, jeden gen připadá přibližně na 1000 nukleotidů.

Table 1 Some prokaryotic genomes

Organism	Domain	Size (base pairs)	Genes	Comments
Nanoarchaeum equitans	*Archaea*	490 885	552	Smallest known cellular genome
Mycoplasma genitalium	*Bacteria*	580 070	470	Smallest genome among *Bacteria*; human pathogen
Chlamydia trachomatis	*Bacteria*	1 042 519	894	Intracellular parasite of humans
Aquifex aeolicus	*Bacteria*	1 551 335	1544	Hyperthermophile, autotroph
Methanothermobacter thermoautotrophicus	*Archaea*	1 751 377	1855	Methanogen, thermophile
Halobacterium salinarium	*Archaea*	2 571 010	2630	Extreme halophile
Sulfolobus solfataricus	*Archaea*	2 992 245	2977	Hyperthermophile, acidophile
Bacillus subtilis	*Bacteria*	4 214 810	4100	Produces endospores
Pseudomonas aeruginosa	*Bacteria*	6 264 403	5570	Metabolically versatile; can be a pathogen
Bradyrhizobium japonicum	*Bacteria*	9 105 828	8317	Nitrogen-fixing bacterium; forms root nodules on soybean plants
Escherichia Coli	*Bacteria*	4 639 221	4288	Model organism for molecular biology

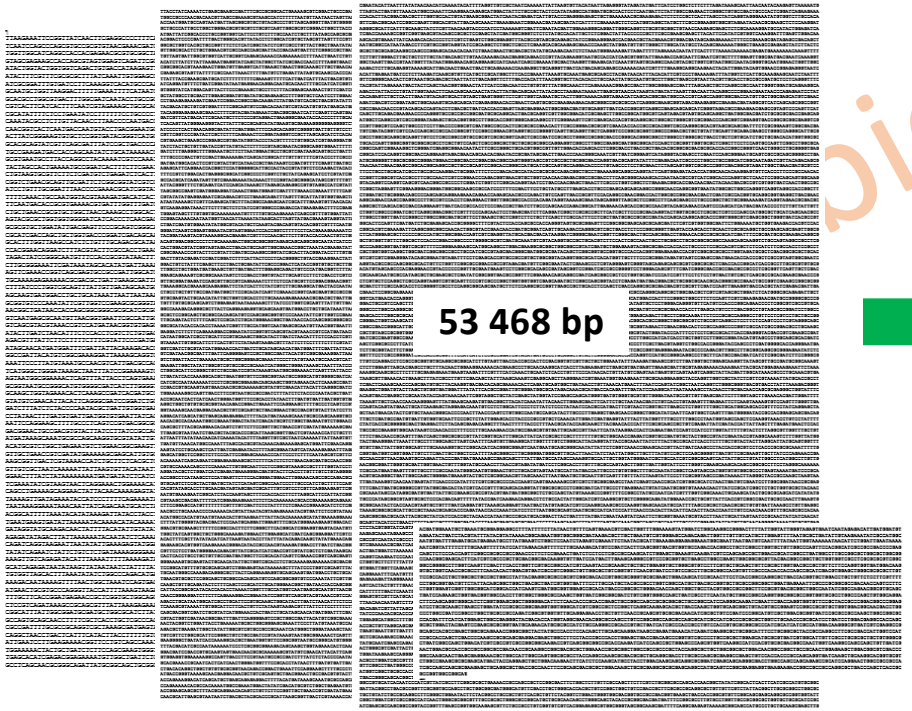
Bacteriology

Michael T Madigan, *Southern Illinois University, Carbondale, Illinois, USA*
Deborah O Jung, *Southern Illinois University, Carbondale, Illinois, USA*

ENCYCLOPEDIA OF LIFE SCIENCES © 2007, John Wiley & Sons, Ltd. www.els.net

Predikce genů u prokaryot

- **Prokaryotické genomy**: malé (0,5 až 10 Mbp) a kompaktní, vysoká hustota genů, 90 % genomu je kódující, jeden gen připadá přibližně na 1000 nukleotidů.



GeneMarkS

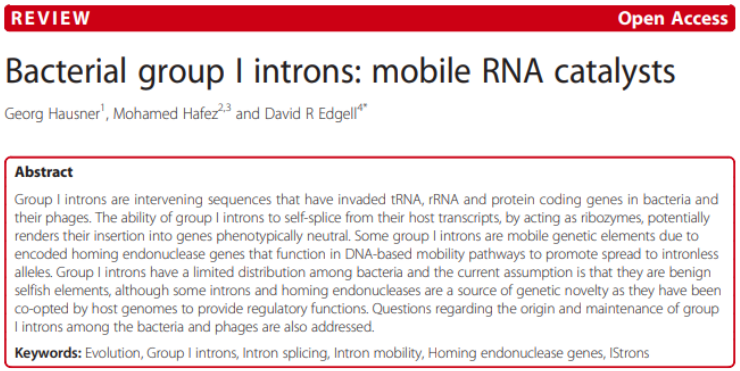
Predikováno 56 genů (ORF)

Predicted genes

Gene #	Strand	LeftEnd	RightEnd	Gene Length	Class
1	-	1	1515	1515	1
2	+	1694	2359	666	1
3	+	2739	4025	1287	1
4	-	4087	4293	207	2
5	+	4618	5175	558	1
6	-	5165	5905	741	1
7	+	6086	7663	1578	1
8	+	7980	8390	411	1
9	+	8659	11067	2409	2
10	+	11100	11438	339	1
11	-	11567	11947	381	1
12	+	12036	12896	861	1
13	+	13004	13969	966	1
14	+	14077	14373	297	1
15	+	14688	15659	972	1
16	-	15777	16586	810	1
17	-	16639	17652	1014	1
18	+	18318	19328	1011	1
19	-	19468	19977	510	1
20	+	20206	20763	558	1
21	-	20753	21493	741	1
22	+	21674	23251	1578	1
23	+	23568	23978	411	1
24	-	24333	24542	210	2
25	-	24663	25337	675	2
26	-	25334	25777	444	2
27	-	25857	25985	129	1
28	+	26345	26908	564	1
29	-	26913	27311	399	2
30	+	27310	27867	558	1
31	-	27857	28597	741	1
32	+	28778	30355	1578	1
33	+	30672	31082	411	1
34	+	31418	32542	1125	2
35	+	32598	33749	1152	2
36	-	33731	34219	489	1
37	-	34224	35651	1428	1
38	+	35750	36097	348	1
39	-	36151	36669	519	1
40	-	36712	37302	591	1
41	-	37299	38456	1158	1
42	-	38482	39597	1116	1
43	-	39680	40657	978	1
44	+	40665	40958	294	2
45	+	41018	41827	810	2
46	+	41983	42534	552	2
47	+	42531	43736	1206	1
48	+	43807	44532	726	1
49	+	44846	45520	675	1
50	+	45706	46338	633	2
51	+	46823	47128	306	2
52	+	47366	47719	354	1
53	+	47716	48165	450	1
54	+	48432	49709	1278	1
55	+	50339	52468	2130	1
56	+	52475	>53467	993	1

Predikce genů u prokaryot

- **Prokaryotické genomy**: malé (0,5 až 10 Mbp) a kompaktní, vysoká hustota genů, 90 % genomu je kódující, jeden gen připadá přibližně na 1000 nukleotidů.
- **Prokaryotické geny:** ORF je nepřerušovaný úsek DNA mezi startovním kodonem (ATG, GTG, TTG, CTG) a stop kodonem (TAA, TGA, TAG). Prokaryotické geny neobsahují introny (Dobře, můžou obsahovat introny).



REVIEW Open Access

Bacterial group I introns: mobile RNA catalysts

Georg Hausner[1], Mohamed Hafez[2,3] and David R Edgell[4*]

Abstract

Group I introns are intervening sequences that have invaded tRNA, rRNA and protein coding genes in bacteria and their phages. The ability of group I introns to self-splice from their host transcripts, by acting as ribozymes, potentially renders their insertion into genes phenotypically neutral. Some group I introns are mobile genetic elements due to encoded homing endonuclease genes that function in DNA-based mobility pathways to promote spread to intronless alleles. Group I introns have a limited distribution among bacteria and the current assumption is that they are benign selfish elements, although some introns and homing endonucleases are a source of genetic novelty as they have been co-opted by host genomes to provide regulatory functions. Questions regarding the origin and maintenance of group I introns among the bacteria and phages are also addressed.

Keywords: Evolution, Group I introns, Intron splicing, Intron mobility, Homing endonuclease genes, IStrons

Group II introns in the bacterial world

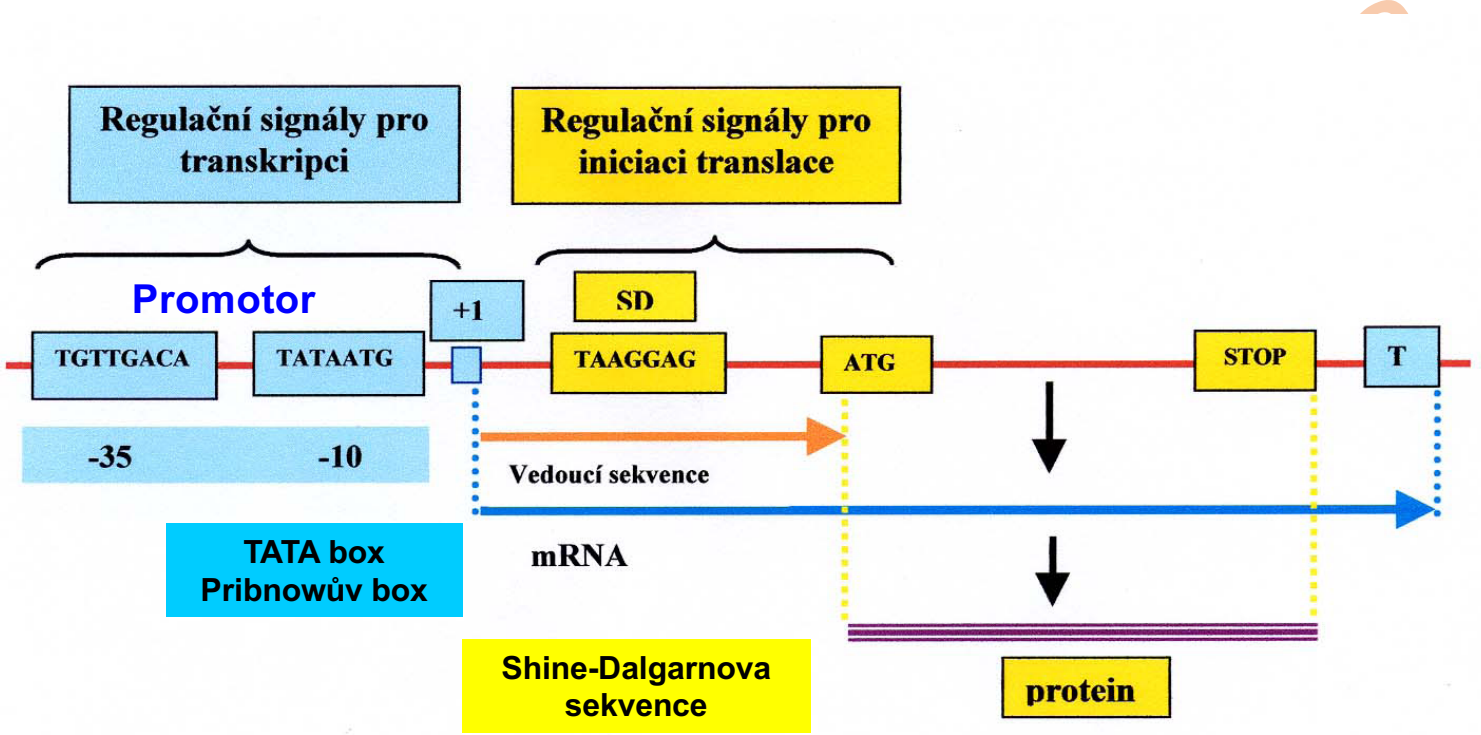
Francisco Martínez-Abarca and Nicolás Toro*

Grupo de Ecología Genética, Estación Experimental del Zaidín, Consejo Superior de Investigaciones Científicas, Profesor Albareda 1, 18008 Granada, Spain.

Predikce genů u prokaryot

- **Prokaryotické genomy**: malé (0,5 až 10 Mbp) a kompaktní, vysoká hustota genů, 90 % genomu je kódující, jeden gen připadá přibližně na 1000 nukleotidů.
- **Prokaryotické geny:** ORF je nepřerušovaný úsek DNA mezi startovním kodonem (ATG, GTG, TTG, CTG) a stop kodonem (TAA, TGA, TAG). Prokaryotické geny neobsahují introny (Dobře, můžou obsahovat introny).
- RBS: Shine-Dalgarnova sekvence
- Terminátor transkripce

Translační a transkripční signální sekvence



Prokaryota

oblast bohatá na puriny
~ cca 8 bází upstream

Predikce genů u prokaryot – základní postupy

(bez využití specializovaných programů)

- **Prokaryotické genomy**: malý obsah nekódujících úseků umožňuje „manuální“ identifikaci ORF.

1) Překlad prokaryotické DNA do proteinové sekvence.

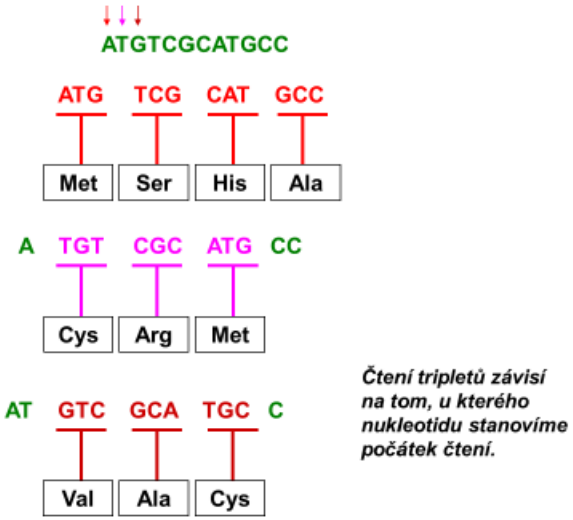
2) Identifikace potenciálních ORF.

3) Ověření spolehlivosti predikce – je identifikovaný ORF skutečně součástí genu?

Predikce genů u prokaryot – základní postupy

(bez využití specializovaných programů)

1) Překlad prokaryotické DNA do proteinové sekvence.



Čtení tripletů závisí na tom, u kterého nukleotidu stanovíme počátek čtení.

5´ ATGCGCAGGAATGCATAG 3´
3´ TACGCGTCCTTACGTATC 5´

Překlad DNA sekvence – od 5´konce

5´ ATGCGCAGGAATGCATAG 3´

5´ CTATGCATTCCTGCGCAT 3´

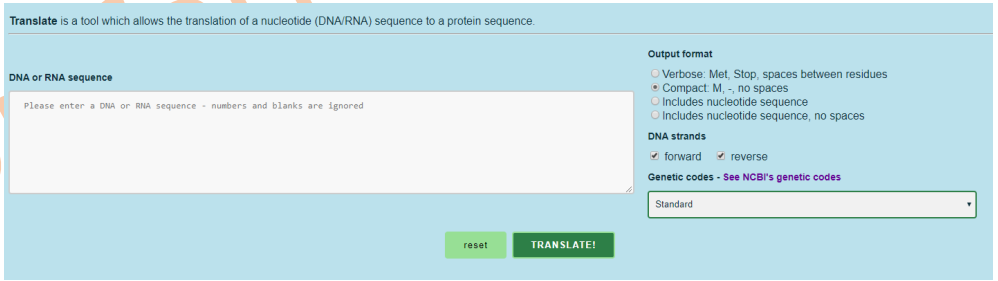
Predikce genů u prokaryot – základní postupy

(bez využití specializovaných programů)

1) Překlad prokaryotické DNA do proteinové sekvence.

- **Translate (ExPASy)**

  https://web.expasy.org/translate/



- **ORF Finder (NCBI)**

  https://www.ncbi.nlm.nih.gov/orffinder/

Translate

Translate is a tool which allows the translation of a nucleotide (DNA/RNA) sequence to a protein sequence.

DNA or RNA sequence

```
GTATGCTGGTGATTGTGGATGCCGTTACCCTGCTGAGCGCCTATCCGGAAGCCAGCCGTGATCCGGCCGCCCCGACCGTGATTGATGGTCGCCACCTGTATGTTGTTAGCCCGGG
CGATGCCGCGCAGCTGGGCCATAACGATAGCCGTCTGTTTACCGGTCTGAGCCCGGGTGATCAGCTGCATCTGCGCGAAACCGCGCTGGCGCTGCGCGCGGAAGTGAGCGTGCTG
TTTATTCGCTTTGCCCTGAAAGATGCCGGCATTGTTGCCCCGATCGAACTGGAAGTGCGTGATGCCGCCACCGCCGTTCCGGATGCGGATGATCTGCTGCATCCGAGCTGTCGTC
CGCTGAAAGATCATTATTGGCGCAGCGATGTGCTGGCGGCGGGCGCGACCACCTGTACCGCCGATTTTGCGGTGTGCGATCGTGATGGCACCGTGAGCGGTTATTTTCGTTGGGA
AACCAGCATTGAAATTGCGGGCAGCCAGCCGGATACCAAACAGCCGGGCTTTAAACCGAGCAGCGATCGCAATGGCAACTTTAGCCTGCCGCCGAATACCGCCTTTAAAGCGATC
TTCTATGCGAACGCGGCGGATCGTCAGGATCTGAAACTGTTTATTGATGATGCGCCGGAACCGGCCGCCACCTTTGTGGGTAACAGCGAAGATGGTGTGCGTCTGTTTACCCTGA
ATAGCAAAGGTGGTAAAATTCGTATTGAAGCGAGCGCGAACGGCCGTCAGAGCGCGACCGATGCCCGTCTGGCGCCGCTGAGCGCGGGCGATACCGTGTGGCTGGGCTGGCTGGG
CGCGGAAGATGGTGCCGATGCGGATTATAATGATGGCATTGTTATTCTGCAGTGGCCGATTACCTAATGGG
```

reset | TRANSLATE!

Output format

- ○ Verbose: Met, Stop, spaces between residues
- ◉ Compact: M, -, no spaces
- ○ Includes nucleotide sequence
- ○ Includes nucleotide sequence, no spaces

DNA strands

☑ forward ☑ reverse

Genetic codes - See NCBI's genetic codes

Standard ▾

- Standard
- Vertebrate mitochondrial
- Yeast mitochondrial
- Mold, protozoan and coelenterate mitochondrial, mycoplasma/spiroplasma
- Invertebrate mitochondrial
- Ciliate, dasycladacean and hexamita nuclear
- Echinoderm and flatworm mitochondrial
- Euplotid nuclear
- Alternative yeast nuclear
- Ascidian mitochondrial
- Alternative flatworm mitochondrial
- Blepharisma nuclear
- Chlorophycean mitochondrial
- Trematode mitochondrial
- Scenedesmus obliquus mitochondrial
- Pterobranchia mitochondrial

DNA or RNA sequence

```
gtatgctggtgattgtggatgccgttaccctgctgagcgcctatccggaagccagccgtgatccggccgccccgaccgtgattgatggtcgccacctgtatgttgttagcccgggc
gatgccgcgcagctgggccataacgatagccgtctgtttaccggtctgagcccgggtgatcagctgcatctgcgcgaaaccgcgctggcgctgcgcgcggaagtgagcgtgctgtt
tattcgctttgccctgaaagatgccggcattgttgccccgatcgaactggaagtgcgtgatgccgccaccgccgttccggatgcggatgatctgctgcatccgagctgtcgtccgc
tgaaagatcattattggcgcagcgatgtgctggcggcgggcgcgaccacctgtaccgccgattttgcggtgtgcgatcgtgatggcaccgtgagcggttattttcgttgggaaacc
agcattgaaattgcgggcagccagccggataccaaacagccgggctttaaaccgagcagcgatcgcaatggcaactttagcctgccgccgaataccgcctttaaagcgatcttcta
tgcgaacgcggcggatcgtcaggatctgaaactgtttattgatgatgcgccggaaccggccgccacctttgtgggtaacagcgaagatggtgtgcgtctgtttaccctgaatagca
aaggtggtaaaattcgtattgaagcgagcgcgaacggccgtcagagcgcgaccgatgcccgtctggcgccgctgagcgcgggcgataccgtgtggctgggctggctgggcgcggaa
gatggtgccgatgcggattataatgatggcattgttattctgcagtggccgattacctaatggg
```

Output format

- ○ Verbose: Met, Stop, spaces between residues
- ◉ Compact: M, -, no spaces
- ○ Includes nucleotide sequence
- ○ Includes nucleotide sequence, no spaces

DNA strands

☑ forward ☑ reverse

Genetic codes - See NCBI's genetic codes

Standard

reset TRANSLATE!

Results of translation

- Open reading frames are highlighted in red
- Select your initiator on one of the following frames to retrieve your amino acid sequence

Download all the translated frames

5'3' Frame 1

```
VCW-LWMPLPC-APIRKPAVIRPPRP-LMVATCMLLARAMPRSWAITIAVCLPV-ARVISCICAKPRWRCARK-ACCLFALP-KMPALLPRSNWKCVMPPPPFRMRMICCIRAVVR-KIIIGAAMCWRRARPPVPPILRCAIVMAP-AVIFV
GKPALKLRAASRIPNSRALNRAAIAMATLACRRIPPLKRSSMRTRRIVRI-NCLLMMRRNRPPPLWVTAKMVCVCLP-IAKVVKFVLKRARTAVRARPMPVWRR-ARAIPCGWAGWARKMVPMRIIMMALLFCSGRLPNG
```

5'3' Frame 2

```
YAGDCGCRYPAERLSGSQP-SGRPDRD-WSPPVCC-PGRCRAAGP-R-PSVYRSEPG-SAASARNRAGAARGSERAVYSLCPERCRHCCPDRTGSA-CRHRRSGCG-SAASELSSAERSLLAQRCAGGGRDHLYRRFCGVRS-WHRERLFSL
GNQH-NCGQPAGYQTAGL-TEQRSQWQL-PAAEYRL-SDLLCERGGSSGSETVY--CAGTGRHLCG-QRRWCASVYPE-QRW-NSY-SERERPSERDRCPSGAAERGRYRVAGLAGRGRWCRCGL--WHCYSAVADYLM
```

5'3' Frame 3

```
MLVIVDAVTLLSAYPEASRDPAAPTVIDGRHLYVVSPGDAAQLGHNDSRLFTGLSPGDQLHLRETALALRAEVSVLFIRFALKDAGIVAPIELEVRDAATAVPDADDLLHPSCRPLKDHYWRSDVLAAGATTCTADFAVCDRDGTVSGYFRW
ETSIEIAGSQPDTKQPGFKPSSDRNGNFSLPPNTAFKAIFYANAADRQDLKLFIDDAPEPAATFVGNSEDGVRLFTLNSKGGKIRIEASANGRQSATDARLAPLSAGDTVWLGWLGAEDGADADYNDGIVILQWPIT-W
```

3'5' Frame 1

```
PIR-SATAE-QCHHYNPHRHHLPRPASPATRYRPRSAAPDGHRSRSDGRSRSLQYEFYHLCYSG-TDAHHLRCYPQRWRPVPAHHQ-TVSDPDDPPRSHRRSL-RRYSAAG-SCHCDRCSV-SPAVWYPAGCPQFQCWFPNENNRSRCHHDR
TPQNRRYRWSRPPPAHRCANNDLSADDSSDAADHPHPERRWRHHALPVRSGQQCRHLSGQSE-TARSLPRAAPARFRADAADHPGSDR-TDGYRYGPAARHRPG-QHTGGDHQSRSGRPDHGWLPDRRSAG-RHPQSPAY
```

3'5' Frame 2

```
PLGNRPLQNNNAIIIIRIGTIFRAQPAQPHGIARAQRRQTGIGRALTAVRARFNTNFTTFAIQGKQTHTIFAVTHKGGGRFRRIINKQFQILTIRRVRIEDRFKGGIRRQAKVAIAIAARFKARLFGIRLAARNFNAGFPTKITAHGAITIA
HRKIGGTGGRARRQHIAAPIMIFQRTTARMQQIIRIRNGGGGITHFQFDRGNNAGIFQGKANKQHAHFRAQRQRGFAQMQLITRAQTGKQTAIVMAQLRGIARANNIQVATINHGRGGRITAGFRIGAQQGNGIHNHQH
```

3'5' Frame 3

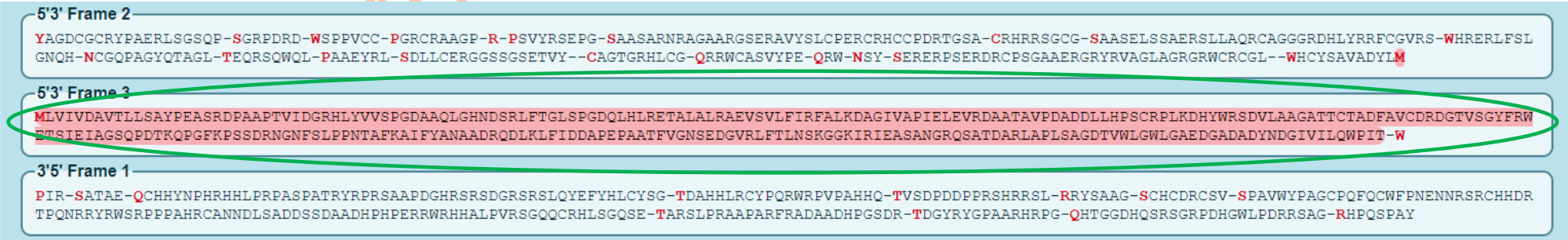
```
H-VIGHCRITMPSL-SASAPSSAPSQPSHTVSPALSGARRASVAL-RPFALASIRILPPLLFRVNRRTPSSLLPTKVAAGSGASSINSFRS-RSAAFA-KIALKAVFGGRLKLPLRSLLGLKPGCLVSGWLPAISMLVSQRK-PLTVPSRSH
TAKSAVQVVAPAASTSLRQ--SFSGRQLGCSRSSASGTAVAASRTSSSIGATMPASFRAKRINSTLTSARSASAVSRRCS-SPGLRPVNRRLSLWPSCAASPGLTTYRWRPSITVGAAGSRLASG-ALSRVTASTITSI
```

Predikce genů u prokaryot – základní postupy

(bez využití specializovaných programů)

2) Identifikace potenciálních ORF.

- Jak dlouhý má být „rozumný“ ORF? Stop kodon se v nekódující sekvenci náhodně vyskytuje přibližně každých 20 kodonů. V úvahu se tedy berou ORF delší než **třicet kodonů** (reálně i delší).
- Empirické pravidlo: Správný ORF = **nejdelší** ORF odpovídající danému úseku DNA.

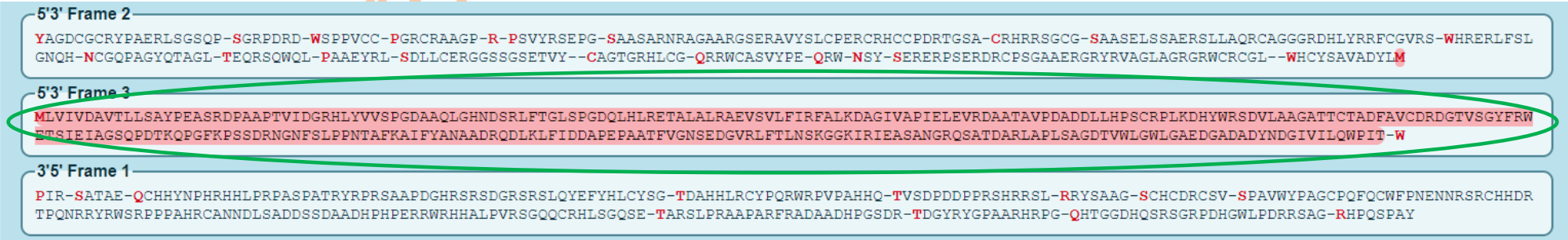


Predikce genů u prokaryot – základní postupy

(bez využití specializovaných programů)

3) Ověření spolehlivosti predikce – je identifikovaný ORF skutečně součástí genu?

- Kóduje ORF protein **podobný** již popsanému proteinu?
- Vyskytují se před/za ORF typické signální sekvence?
- Statistické parametry sekvence: obsah GC, preference kodonů.



Obsah GC

Obsah GC – zastoupení G a C v sekvenci NA (genom, gen, část genu, fragment, syntetický oligonukleotid). Vyšší obsah GC párů je asociován s vyšší stabilitou DNA.

- Velmi rozdílný pro různé prokaryotické genomy (25%-75%).
- Adaptace na vysokou teplotu?
- Adaptace na životní podmínky?

Base composition bias might result from competition for metabolic resources

Eduardo P.C. Rocha and Antoine Danchin

High guanine–cytosine content is not an adaptation to high temperature: a comparative analysis amongst prokaryotes

Laurence D. Hurst[1*] and Alexa R. Merchant[2]

species	GC_{coding} (%)	GC_3 (%)	thermophilic
Aeropyrum pernix[a]	57.50	66.40	yes
Archaeoglobus fulgidus[a]	49.37	58.42	yes
Methanobacterium thermoautotrophicum[a]	50.46	56.59	yes
Methanococcus jannaschii[a]	31.84	24.74	yes
Pyrococcus abyssi[a]	45.16	50.31	yes
Pyrococcus horikoshii[a]	42.32	42.97	yes
Aquifex aeolicus[b]	43.58	47.93	yes
Bacillus subtilis[b]	44.32	44.61	no
Borrelia burgdorferi[b]	29.31	20.82	no
Campylobacter jejuni[b]	32.82	18.96	no
Chlamydia muridarum[b]	39.13	29.92	no
Chlamydia pneumoniae[b]	41.30	34.88	no
Chlamydia trachomatis[b]	41.61	34.30	no
Deinococcus radiodurans[b]	65.72	84.02	no
Escherichia coli[b]	51.37	54.90	no
Haemophilus influenzae[b]	38.76	29.09	no
Helicobacter pylori[b]	39.56	41.95	no
Mycobacterium tuberculosis[b]	65.81	79.67	no
Mycoplasma genitalium[b]	31.64	23.01	no
Mycoplasma pneumoniae[b]	41.05	42.08	no
Neisseria meningitidis[b]	50.14	55.49	no
Rickettsia prowazekii[b]	30.59	18.47	no
Synechocystis sp.[b]	48.66	49.99	no
Thermotoga maritima[b]	46.45	52.62	yes
Treponema pallidum[b]	52.52	54.10	no
Ureaplasma urealyticum[b]	35.20	16.97	no
Vibrio cholerae[b]	47.17	49.08	no

Obsah GC

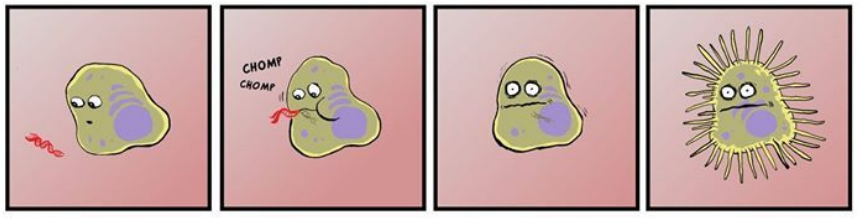
Obsah GC – zastoupení G a C v sekvenci NA (genom, gen, část genu, fragment, syntetický oligonukleotid). Vyšší obsah GC párů je asociován s vyšší stabilitou DNA.

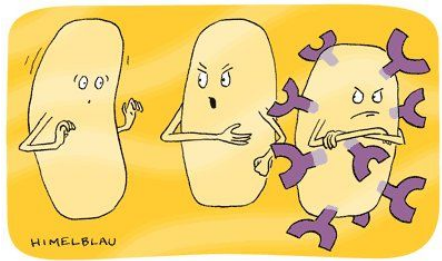
- Pozorování: Třetí pozice v kodonu v kódující sekvenci má vyšší obsah GC – preference kodonů končících na G nebo C.
- Problém: Neplatí zdaleka vždy, GC_3 je extrémně variabilní.
- Lepší využití: identifikace genů získaných horizontálním přenosem.

Horizontal gene transfer: building the web of life

Shannon M. Soucy[1], Jinling Huang[2] and Johann Peter Gogarten[1,3]

Abstract | Horizontal gene transfer (HGT) is the sharing of genetic material between organisms that are not in a parent–offspring relationship. HGT is a widely recognized mechanism for adaptation in bacteria and archaea. Microbial antibiotic resistance and pathogenicity are often associated with HGT, but the scope of HGT extends far beyond disease-causing organisms. In this Review, we describe how HGT has shaped the web of life using examples of HGT among prokaryotes, between prokaryotes and eukaryotes, and even between multicellular eukaryotes. We discuss replacement and additive HGT, the proposed mechanisms of HGT, selective forces that influence HGT, and the evolutionary impact of HGT on ancestral populations and existing populations such as the human microbiome.



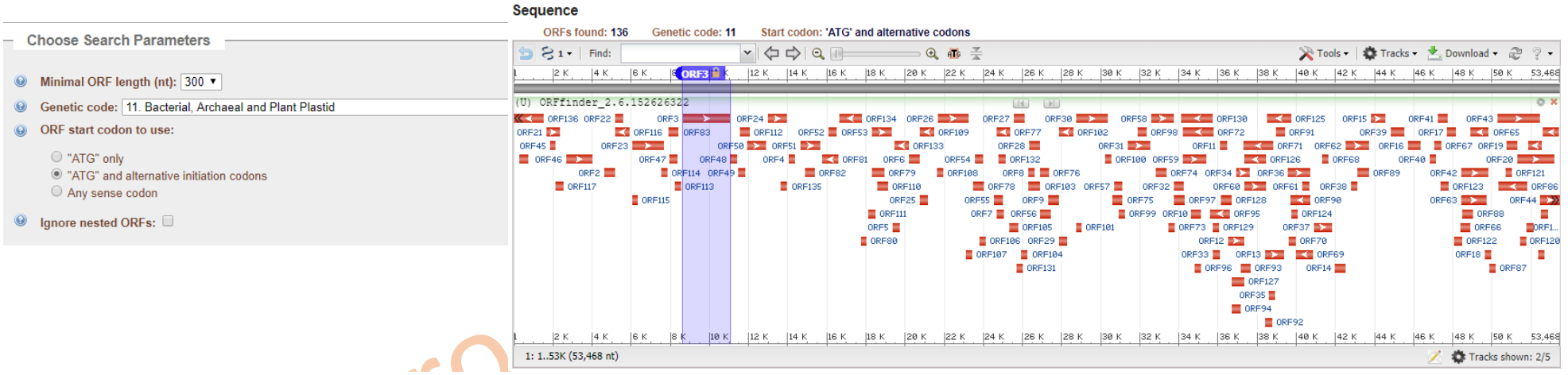


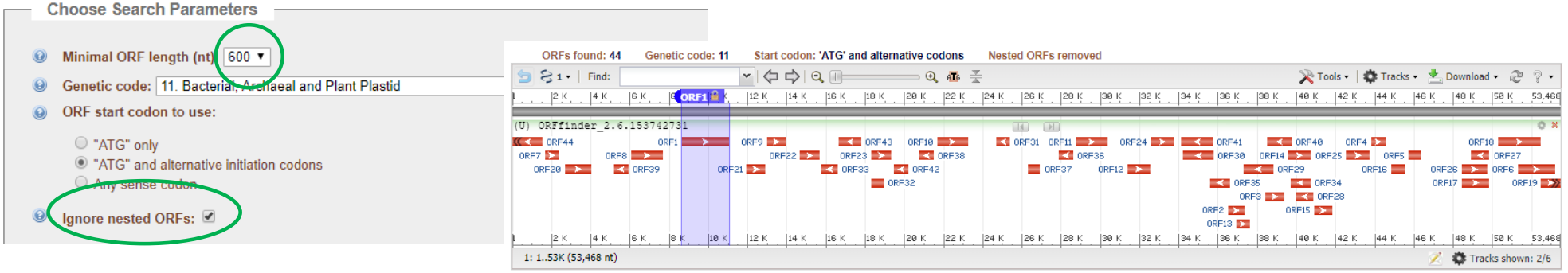
"Don't pick it up," I say, and he says, "It's just a *plasmid*, what harm could it do?" Well just look at him now....God knows *what* protein he's expressing!

ORF Finder

Open Reading Frame Finder

ORF finder searches for open reading frames (ORFs) in the DNA sequence you enter. The program returns the range of each ORF, along with its protein translation. Use ORF finder to search newly sequenced DNA for potential protein encoding segments, verify predicted protein using newly developed SMART BLAST or regular BLASTP.





ORF Finder

ORF1 (804 aa) Display ORF as... Mark

```
>lcl|ORF1
MDMIHPGSSLDKAINNTRVKNVSTDVKHGQIQERKRNFIYKKNDDISSRF
KLYSSLVKQKNATEDVVLIGKMILDEVRSYRTIHNDRNIVSNSGNWKTSF
LCNLARLLYSIFNGSNYFCSREGENNSSSSTLLTIHQPEKQELLQQKSIK
HLPTSNNIDGYIKIRKTRGAEDQTTTITQSLIINELLKVDRNTIPFQKIS
ELNDIIHSYENMQIKNSRKGIEILVKQGELLSSLINVNKGNKQLSDNASK
IINLLGIEYQSHKVDIEPFIHAVWVAGAPPDNTFSYITAFLNTYKDYTYL
LWDPNAFGAAKFSGILKNIAMNYAIMRLRRTNHLAEEMNEVILKIQNIQN
ETIEFKETRERLKELENRYKSLTSETKEKFNVFFLESMIGMQDNYFTYCI
SNGISNTDDISRLDFLTNVLKLSPEVQNDFKSTVEKNKRDIDLLKNTISQ
KWDRFQLRDINTLESFKKPQDYFFYQQEMLLRWNYAAASDQVRINILKEY
GGIYTDTDILPAYSDKVSQIINEKSDDKRFFEDLKLRRIISESILSLIKG
EKYSIKHDGLDETTLNQLNNILSEIEKLTIDDYFKPVETKVVRDTFKIFK
RYQKWTENTWNIRGNNNFMLTHKGSDFILSGQKKQYLLQRIRDNISYNNL
FYTTEDLKSLNNVAIGGIPAKKYLEHGLFSEYRQDGTIPYVVSTLNISGP
DMIMRQMKKYYKSLGRIGEVHIKDNKLSDVNFLGVYASSNKDNKSFNWLN
PVSVGINDITPDDESSWAVRNNDINKILFEKINCHVPEKMDELRAQGYHF
KVRT
```

ORF1

SmartBLAST

BLAST

Marked set (0)

SmartBLAST best hit titles...

BLAST

BLAST Database:

UniProtKB/Swiss-Prot (swissprot)

Introduction

SmartBLAST processes your protein query to present a concise summary of the five best protein matches from well-studied reference species in the landmark database (described below). If possible, the matches will be from different organisms. If SmartBLAST cannot find five matches in the landmark database, it will uses matches from the protein non-redundant (nr) database. SmartBLAST produces these results using a combination of an optimized BLASTP search, a new implementation of BLAST meant to find closely related matches, and a multiple alignment. Additionally, SmartBLAST presents Conserved Domain Database matches to your query. Additional matches to the nr database are presented lower in the report.

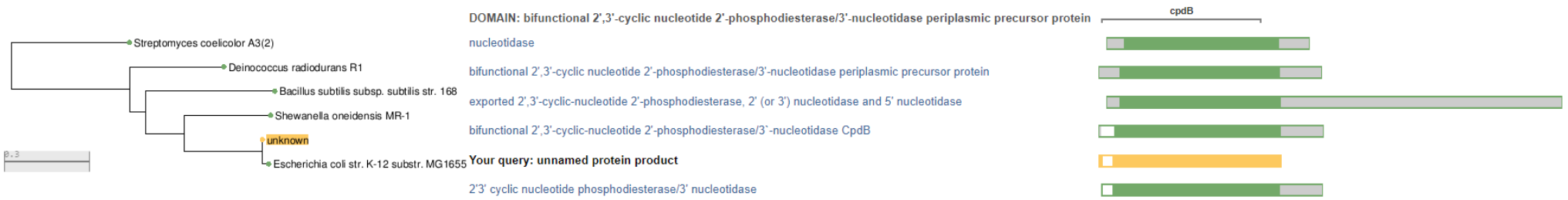
SmartBLAST is under active development and may change with little or no notice.

Landmark Database

The landmark database includes proteomes from 27 genomes spanning a wide taxonomic range. This search set is produced using the best available genomic assemblies for each organism with the following procedure. First, the most recent representative assembly from each organism is identified. Second, all proteins annotated on each assembly are downloaded and compiled into the landmark BLAST database. The result is a taxonomically diverse non-redundant set of proteins supported by genomic assemblies.

Query: unnamed protein product

Query length: 531 aa



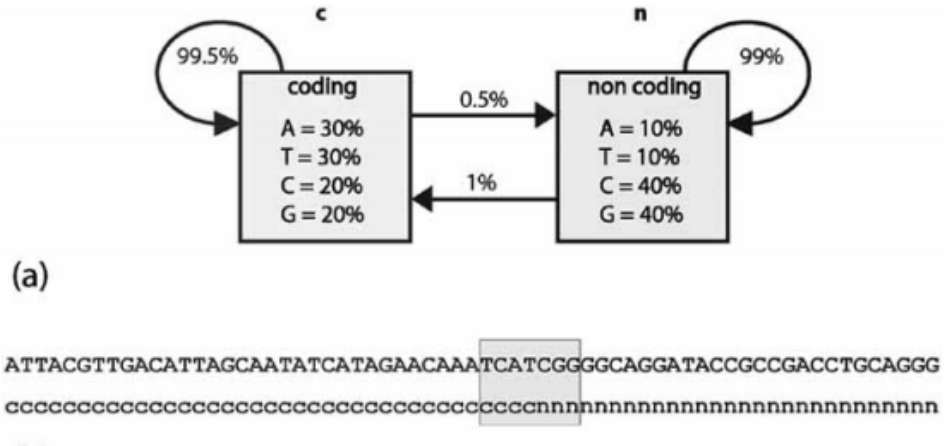
Metody predikce genů

- Dva hlavní přístupy: metody ***ab initio***/metody založené na **homologii** (sekvenční).
- ***Ab initio*** – predikce genů založená pouze na **sekvenci**, jejích vlastnostech a statistických parametrech.
- **Metody založené na homologii** – sekvenční podobnost se známými geny/proteiny. ORF kóduje protein, který je podobný již dříve popsanému proteinu (prohledávání **DATABÁZÍ** pomocí **ALIGNMENTU**) = nejspolehlivější predikce. Problém – unikátní geny bez známých homologů (většinou nejzajímavější).
- Kombinace obou postupů
- **Specializované predikční programy – v kombinaci s HMM**

Predikce genů u prokaryot

Skryté Markovovy modely

- Skrytý Markovův model: Jednotlivé stavy mohou generovat různé znaky s definovanou pravděpodobností. Stavy jsou skryté, vidíme pouze znaky, které generují.




```
gttccggatgcggatgatctgctgcatccgagctgtcgtccggaaagatcattattggcgcagc
gatgtgctggcggcgggcgcgaccacctgtaccgccgattttgcggtgtgcgatcgtgatggca
ccgtgagcggttattttcgttgggaaaccagcattgaaattgcgggcagccagccggataccaa
acagccgggctttaaaccgagcagcgatcgcaatggcaactttagcctgccgccgaataccgcc
tttaagcgatagctctatgcgaacgcgttgcggatcgtcagatctgaaactgtttatt
```

Jaký je nejpravděpodobnější průchod skrytými stavy?

```
gttccggatgcggatgatctgctgcatccgagctgtcgtccggaaagatcattattggcgcagc
gatgtgctggcggcgggcgcgaccacctgtaccgccgattttgcggtgtgcgatcgtgatggca
ccgtgagcggttattttcgttgggaaaccagcattgaaattgcgggcagccagccggataccaa
acagccgggctttaaaccgagcagcgatcgcaatggcaactttagcctgccgccgaataccgcc
tttaagcgatagctctatgcgaacgcgttgcggatcgtcagatctgaaactgtttatt
```

```
MRMICCIRAVVRKDHYWRSDVLAAGATTCTADFAVCDRDGTVSGYFRWETSI
EIAGSQPDTKQPGFKPSSDRNGNFSLPPNTAFKR
Protein třídy 1, typický
```


Current Bioinformatics, 2007, 2, 49-61

Hidden Markov Models in Bioinformatics

Predikce genů u prokaryot

Skryté Markovovy modely

GeneMark

A family of gene prediction programs developed at
Georgia Institute of Technology , Atlanta, Georgia, USA.

http://exon.gatech.edu/GeneMark/

GeneMark

GeneMark developed in 1993 was the first gene finding method recognized as an efficient and accurate tool for genome projects. GeneMark was used for annotation of the first completely sequenced bacteria, *Haemophilus influenzae*, and the first completely sequenced archaea, *Methanococcus jannaschii*. The GeneMark algorithm uses species specific inhomogeneous Markov chain models of protein-coding DNA sequence as well as homogeneous Markov chain models of non- coding DNA. Parameters of the models are estimated from training sets of sequences of known type. The major step of the algorithm computes a posteriory probability of a sequence fragment to carry on a genetic code in one of six possible frames (including three frames in complementary DNA strand) or to be "non-coding".

GeneMark.hmm (prokaryotic)

GeneMark.hmm algorithm was designed to improve the gene prediction quality, particularly to improve GeneMark in finding exact gene starts. The idea was to integrate the GeneMark models into naturally designed hidden Markov model framework with gene boundaries modeled as transitions between hidden states. Additionally, the ribosome binding site model is used to make the gene start predictions more accurate. In evaluations by different groups it was shown that GeneMark.hmm is significantly more accurate than GeneMark in exact gene prediction. From 1998 until now GeneMark.hmm and its self-training version, GeneMarkS, are the standard tools for gene identification in new prokaryotic genomic sequences, including metagenomes.

© 1998 Oxford University Press *Nucleic Acids Research, 1998, Vol. 26, No. 4* **1107–1115**

GeneMark.hmm: new solutions for gene finding

Alexander V. Lukashin and Mark Borodovsky[1,*]

School of Biology and [1]Schools of Biology and Mathematics, Georgia Institute of Technology, Atlanta, GA 30332-0230, USA

Received August 14, 1997; Revised and Accepted December 30, 1997

Table 5. Results of GeneMark.hmm predictions for 10 complete bacterial genomes

Genome	Genes annotated	Genes predicted	Exact prediction (%)	Missing genes (%)	Wrong genes (%)
A.fulgidus	2407	2530	73.1	10.8 (2.0)	15.1
B.subtilis	4101	4384	77.5	3.6 (2.8)	9.8
E.coli	4288	4440	75.4	5.0 (2.7)	8.2
H.influenzae	1718	1840	86.7	3.8 (3.2)	10.2
H.pylori	1566	1612	79.7	6.0 (4.4)	8.7
M.genitalium	467	509	78.4	9.9 (1.7)	17.3
M.jannaschii	1680	1841	72.7	4.6 (0.8)	12.9
M.pneumoniae	678	734	70.1	7.8 (4.1)	13.6
M.thermoauthotrophicum	1869	1944	70.9	5.0 (3.5)	8.6
Synechocystis	3169	3360	89.6	4.0 (1.5)	9.4
Averaged	21 943	23 194	78.1	5.4 (2.7)	10.4

The second and third columns show the number of genes annotated in GenBank and the corresponding number of genes predicted, respectively. 'Exact prediction' is a fraction of annotated genes for which both the 5′-end and the 3′-end were predicted exactly. 'Missing genes' is a fraction of annotated genes for which neither the 5′-end nor the 3′-end was predicted exactly; in this column the numbers in brackets show the missing genes after using the combined program (GeneMark.hmm + GeneMark). 'Wrong genes' is a fraction of predicted genes for which no annotated analog was found. All measures are expressed as percentages. The data shown are the results obtained after post-processing procedure (RBS recognition).

Predikce genů u prokaryot

Markovovy modely

Co když není model pro můj organismus v seznamu GeneMark?

- Lze použít model pro **blízce příbuzný** organismus.
- Lze využít **heuristický model** (pro krátké sekvence).
- Lze využít **„self-training“** algoritmus (pro dostatečně dlouhé sekvence).

```
GeneMark.hmm PROKARYOTIC (Version 3.26)
Date: Wed Mar 25 10:09:08 2020
Sequence file name: seq.fna
Model file name: /home/genemark/parameters/prokaryotic/Escherichia_coli__BL21_Gold_DE3_pLysS_AG_/GeneMark_hmm_combined.mod
RBS: true
Model information: Escherichia_coli__BL21_Gold_DE3_pLysS_AG_

FASTA definition line: empty-fasta-def-line
Predicted genes
   Gene     Strand     LeftEnd     RightEnd       Gene      Class
    #                                           Length
    1         +          <3          314          312          1
    2         +         318         1604         1287          1
    3         -        1698         2471          774          1
    4         -        2550         3980         1431          1
    5         +        4249         5202          954          1
    6         +        5313         5903          591          1
    7         -        5960        >6244          285          1
```

```
GeneMark.hmm PROKARYOTIC (Version 3.26)
Date: Wed Mar 25 10:13:05 2020
Sequence file name: seq.fna
Model file name: /home/genemark/parameters/prokaryotic/Pseudomonas_aeruginosa_PAO1/GeneMark_hmm_combined.mod
RBS: true
Model information: Pseudomonas_aeruginosa_PAO1

FASTA definition line: empty-fasta-def-line
Predicted genes
   Gene     Strand     LeftEnd     RightEnd       Gene      Class
    #                                           Length
    1         +          <3          314          312          2
    2         +         318         1604         1287          2
    3         -        1698         2471          774          2
    4         -        2550         3980         1431          2
    5         +        4249         5202          954          2
    6         +        5313         5903          591          2
    7         -        5960        >6244          285          2
```

Heuristic Models

Computer methods of accurate gene finding in DNA sequences require models of protein coding and non-coding regions derived either from experimentally validated training sets or from large amounts of anonymous DNA sequence. A heuristic method for derivation of parameters of inhomogeneous Markov models of protein coding regions, was proposed in 1999. The heuristic method utilizes the observation that parameters of the Markov models used in GeneMark can be approximated by the functions of the sequence G+C content. Therefore, a short DNA sequence sufficient for estimation of the genome G+C content (a fragment longer than 400 nt) is also sufficient for derivation of parameters of the Markov models used in GeneMark and GeneMark.hmm. Models built by the heuristic approach could be used to find genes in small fragments of anonymous prokaryotic genomes, such as metagenomic sequences, as well as in genomes of organelles, viruses, phages and plasmids. This method can also be used for highly inhomogeneous genomes where adjustment of the Markov models to local DNA composition is needed. The heuristic method provides an evidence that the mutational pressure that shapes G+C content is the driving force of the evolution of codon usage pattern.

Heuristické řešení *– přibližné řešení založené na zkušenosti, poučeném odhadu nebo empirických poznatcích. Dá nám rozumné výsledky rozumně rychle.*

Prokaryotické geny

- **Velmi jednoduchý přístup k predikci genů**

  Zjednodušení vede k chybám, ale jejich množství je POMĚRNĚ MALÉ.

- **Chyby mohou vznikat při SEKVENOVÁNÍ DNA.**

  Přidání/odstranění startovního a/nebo stop kodonu může vést ke ZKRÁCENÍ, PRODLOUŽENÍ nebo úplnému VYNECHÁNÍ genu.

  Vynechání-inzerce nukleotidu pak ke ZMĚNĚ ČTECÍHO RÁMCE

Experimental vs. database sequence

```
PLL       -------MPNPDNTEAYVAGEVEIENSAIALSGIVSVANNADNRLEVFGVSTDSAVWHNW 53
PLU0732   MKKEPIKMPNPDNTEAYVAGEVAIENSAIALSGIVSVANNADNRLEVFGVSTDSAVWHNW 60

PLL       QTAPLPNSSWAGWNKFNGVVTSKPAVHRNSDGRLEVFVRGTDNALWHNWQTAADTNTWSS 113
PLU0732   QTAPLPNSSWAGWNKFNGVVTSKPAVHRNSDGRLEVFVRSTDNALWHNWQTAADTNTWSS 120

PLL       WQPLYGGITSNPEVCLNSDGRLEVFVRGSDNALWHIWQTAAHTNSWSNWKSLGGTLTSNP 173
PLU0732   WQPLYGGITSNPEVCLNSDGRLEVFARGTDNALWHIWQTAAHTNSWSNWKSLGGTLTSNP 180

PLL       AAHLNADGRIEVFARGADNALWHIWQTAAHTDQWSNWQSLKSVITSDPVVINNCDGRLEV 233
PLU0732   AAHINADGRIEVFARGADNALWHIWQTAAHTDQWSNWQSLKSVITSDPVVIGNCDGRLEV 240

PLL       FARGADSTLRHISQIGSDSVSWSNWQCLDGVITSAPAAVKNISGQLEVFARGADNTLWRT 293
PLU0732   FARGADNTLRHISQIGSDSVSWSNWQCLDGVITSAPAAVKNISGRLEVFARGADNTLWRT 300

PLL       WQTSHNGPWSNWSSFTGIIASAPTVAKNSDGRIEVFVLGLDKALWHLWQTTSSTTSSWTT 353
PLU0732   WQTSQNGPWSNWSSFTGIIASAPTVAKNSDGRIEVFVLGLDKALWHLWQTTSSTTSSWTT 360

PLL       WALIGGITLIDASVI- 368
PLU0732   WALIGGITLIDASVIK 376
```

Alignment statistics for match #1

Score	Expect	Method	Identities	Positives	Gaps
207 bits(527)	3e-66	Compositional matrix adjust.	107/107(100%)	107/107(100%)	0/107(0%)

```
Query  8    LPANTRFGVTAFANSSGTQTVNVLVNNETAATFSGQSTNNAVIGTQVLNSGSSGKVQVQV  67
            LPANTRFGVTAFANSSGTQTVNVLVNNETAATFSGQSTNNAVIGTQVLNSGSSGKVQVQV
Sbjct  1    LPANTRFGVTAFANSSGTQTVNVLVNNETAATFSGQSTNNAVIGTQVLNSGSSGKVQVQV  60

Query  68   SVNGRPSDLVSAQVILTNELNFALVGSEDGTDNDYNDAVVVINWPLG  114
            SVNGRPSDLVSAQVILTNELNFALVGSEDGTDNDYNDAVVVINWPLG
Sbjct  61   SVNGRPSDLVSAQVILTNELNFALVGSEDGTDNDYNDAVVVINWPLG  107
```

LOCUS NZ_JUUU01000485 5873 bp DNA linear CON 21-AUG-2015
DEFINITION Pseudomonas aeruginosa strain 744_PAER 959_5873_75941, whole genome shotgun sequence.
ACCESSION NZ_JUUU01000485 NZ_JUUU00000000

.....

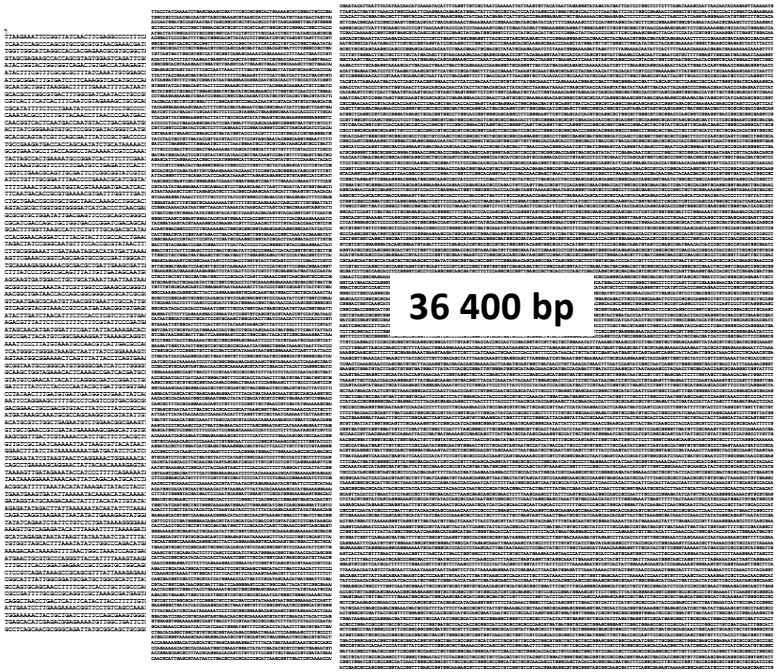
```
gene            complement(5548..>5873)
                /locus_tag="ADF63_RS25535"
     CDS        complement(5548..>5873)
                /locus_tag="ADF63_RS25535"
                /inference="EXISTENCE: similar to AA
                sequence:RefSeq:WP_009876850.1"
                /note="Derived by automated computational
analysis using gene prediction method: Protein Homology."
                /codon_start=3
                /transl_table=11
                /product="fucose-binding lectin"
                /protein_id="WP_049233417.1"
                /db_xref="GI:896235191"

/translation="LPANTRFGVTAFANSSGTQTVNVLVNNETAATFSGQSTNNAVIG
TQVLNSGSSGKVQVQVSVNGRPSDLVSAQVILTNELNFALVGSEDGTDNDYNDAVVVI
NWPLG“
```

Chyby

- Nejčastější
- – chyby v sekvenaci
- – špatná predikce -alternace startovního kodonu
- – shot gun sekvenace

Predikce genů u eukaryot

- **Eukaryotické genomy**: velké až obrovské (10 Mbp až 670 Gbp). Mohou mít velmi nízkou hustotu genů, > 90% genomu může být nekódující, jeden gen u člověka připadá *přibližně* na 100 kbp.



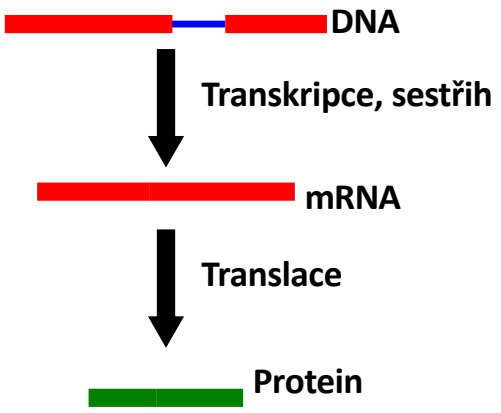
GENSCAN

```
>/tmp/02_12_20-11:25:12.fasta|GENSCAN_predicted_peptide_2|262_aa
MGENWGASDTGDEAKPDPAMACSPEVPGRLLVGQDTAPRGGAEVTGSRGDGHHRFPALAP
DRHREPRPEQGGTQPAEGRGLDSHETEETEKGEMEEMETGKTEGREEMEKGELGENGRAS
DAGMRQSQTQPRSAVPREDTAPGGAGGLYDSEPGKEQRPEVVPSTVPTGRPAQAEGSDPT
RHRSPVCRSPETHILWLTAVRPLGRRRPHVAQTAPLGLKPADKATHPARRCCVATAEGPR
TTFPMTHGQTLAQQGSLRPGAV
>/tmp/02_12_20-11:25:12.fasta|GENSCAN_predicted_peptide_3|1286_aa
XRPLVPSPAEERVLNLPAVVVASSFLLSHLSVGVGVPCATVDARDFVCLASPPQQHHHVG
LGAGGVSCSGSYSEEGLKPGSGTHIHQLGPPVSSFVFPATLLKILINSRIWSAGWKISVW
QSGAWFIDGAFPLRPHGVEGACGCPLYWKGPLFYGAGGERTGSVSVHKFVAMWRKILQNC
HDDAAKFVHLLMSPGCNYLVQEDFVPFLQRPHRTPHGRAGVYTTRLVSSHPQDVVNTHPG
LSFLKEASEPHSRYITTVIQRIFYAVNRSWSGRITCAELRRSSFLQPGGLGASPQRPRAQ
AAWVLWQNVALLEEEADINQLTEFFSYEHFYVIYCKFWELDTDHDLLIIDADDLARHNDHG
QDAVCGRAALFLTHCGLKRGAAPWLVRPRDTGDRRDGRPGCGTFSWPKLTATVLSLLLPPV
AVPLRPEYSALYKSVLEPRSALRSVDFMLDLALLDGKACPFYQDDRQDLLRSSHTVRTAQ
APDGSHCPCEDPGPSPHRCLTGRKVQKEGKISYADFVWFLISEEDKKTPTSIEYWFRCMD
LDGDGALSMFELEYFYEEQCRRLDSMAIEALPFQDCLCQMLDLVKPRTEGKITLQDLKRC
KLANVFFDTFFNIEKYLDHEQKEQISLLRDGDSGGPELSDWEKYAAEEYDILVAEETAGE
PWEDGPGGRLPSRSGSDVGHGHRSPRPSDFGAAQNTFWLPLGSRPMGSMDVHFNPCSPIS
RPVRFLPSSTGPLRPTLGRLSCVSVLLVVIVCTVALTWQPTSGSPQLPSLCLLLSDTLRQ
LWPPDLAWMVSTPTAARLCVCGTASSTLGSTAFTVSTPAPAVGAQAPPPGAFVEPASSAR
SALLAQRLELKVGSGDCLVLEVGGRAACVSRKTAQGLLRAAGHVFAIPSLQVRGRAQPCG
AEAECAALPAGPEALLRGALTAGRRGPVRVRMRGRGPGAAVTPPARTPPRGRSPPWVPGP
ASVERVRVRTTVTGGSRAPQSWTGASTTGCVNVLCKRTKPFLAITLVRVTGGDHGNNPVL
THDKTLPQETEVAASTRQGRAKPGPPREQHTHPVHPPAAVAVKEADAQGGPVTPEERLRG
LQVVGRRGSTSRGNAAVPKVSSVSPGAPLNSRMPPPGSAKGQDPQQQDSHNPRCPGQCRG
RACTPTPLPEKWGGPVGGPRGRRRCGETQSPAVMPSTCSGGPFPRGLTCAGGQGPHKTST
RGTLTVASVPPEARMFQGSAPHPLNDNRGSPALRGRHPLEIPEQGPDPAHPSPPSRGAFV
SEPRVFTGAAPGPPRSSSMSSVPGGP
```

Predikovány 2 kódující sekvence, z toho jedna neúplná

Predikce genů u eukaryot

- **Eukaryotické genomy**: velké až obrovské (10 Mbp až 670 Gbp). Mohou mít velmi nízkou hustotu genů, > 90% genomu může být nekódující.
- **Eukaryotické geny:** skládají se z exonů a intronů. Podléhají sestřihu, může probíhat alternativní sestřih.
- Exony mohou být velmi **krátké**, introny velmi **dlouhé**.



Nízká hustota genů, exony/introny, alternativní sestřih:

Hledání jehly v kupce sena, přičemž jehla je rozlámaná na kousky.
Kousky jehly je nutné najít a SPRÁVNĚ poslepovat dohromady.

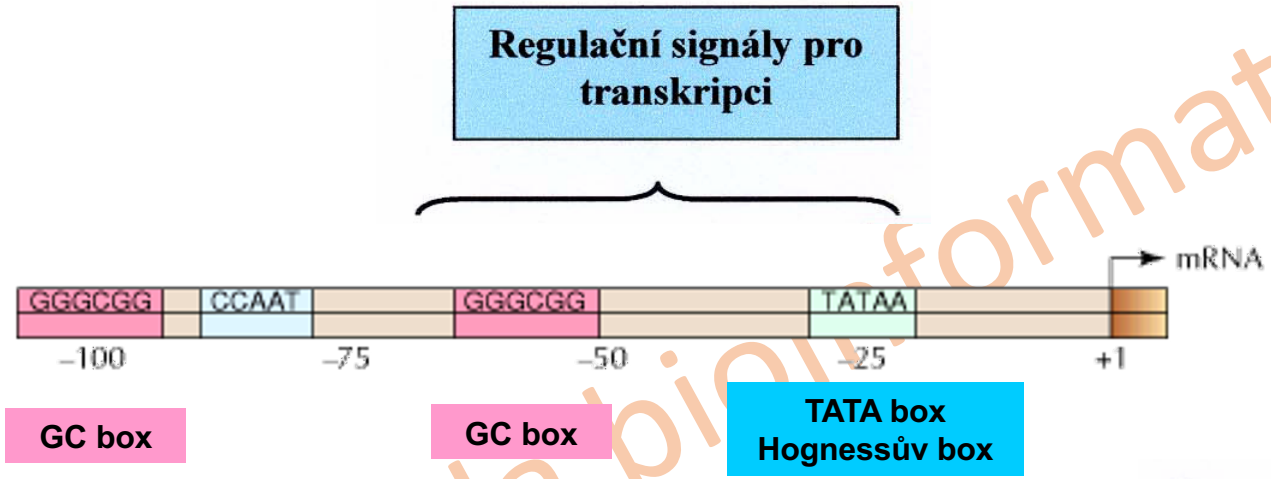
Predikce genů u eukaryot

- **Eukaryotické genomy**: velké až obrovské (10 Mbp až 670 Gbp). Mohou mít velmi nízkou hustotu genů, >90% genomu může být nekódující.
- **Eukaryotické geny:** skládají se z exonů a intronů. Podléhají sestřihu, může probíhat alternativní sestřih.
- Exony mohou být velmi **krátké**, introny velmi **dlouhé**.

Co pomáhá při predikci:

Signální sekvence, sestřihová místa (GT/AG), zastoupení nukleotidů v kódujících/nekódujících oblastech, ATG.

Translační a transkripční signální sekvence



Promotor RNA-polymerasy II

(gcc)gccRccAUGG

Kozak sequence
Sekvence Kozakové

Eukaryota

Promotor sequences

Predikce genů u eukaryot

- Genomy **jednobuněčných** eukaryot se výrazně liší (frekvence intronů, jak velká část genomu je tvořená geny kódujícími proteiny).
- *Saccharomyces cerevisiae* – 67% genomu je protein-kódující, jen 4% obsahují introny.
- Hlenky – průměrný gen obsahuje 3,7 intronu.

Slime mold = hlenka
Fuligo septica
Dog vomit slime mold

Hlenky jsou záhadné houby v podobě blitky, škraloupu, slizu či průjmu, které se dokážou pohybovat, mají paměť a navzájem komunikují. Tato hlenka leze po stohu poblíž Slezských Rudoltic. | *Foto: DENÍK/František Kuba*

https://www.denik.cz/z_domova/hlenky-na-severu-moravy20090715.html

Predikce genů u eukaryot

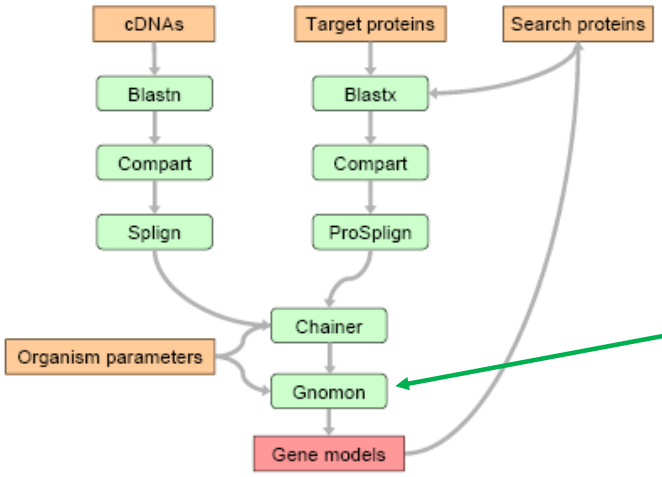
- Genomy **jednobuněčných** eukaryot se výrazně liší (frekvence intronů, jak velká část genomu je tvořená geny kódujícími proteiny).
- *Saccharomyces cerevisiae* – 67% genomu je protein-kódující, jen 4% obsahují introny.
- Pro některá jednobuněčná eukaryota je možné použít stejné postupy jako pro prokaryota.

GeneMarkS

Sequence type	Output format for gene prediction	Output options	Optional: results by E-mail
Prokaryotic Intronless eukaryotic Virus Phage EST/cDNA	LST GFF	Protein sequence Gene nucleotide sequence Coding potential graph (not for multi FASTA) PDF PostScript	E-mail Subject GeneMarkS Compress files

Metody predikce genů u eukaryot

- Metody ***ab initio***/metody založené na **homologii**/metody založené na **konsenzu**.
- ***Ab initio*** – Např. HMM (skryté Markovovy modely)



***Gnomon**, the NCBI eukaryotic gene prediction tool*

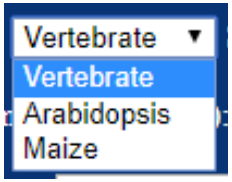
„The core algorithm of the ab initio prediction capability of Gnomon is based on Genscan.“

The New GENSCAN Web Server at MIT

Identification of complete gene structures in genomic DNA

Komplexní model struktury genu (HMM + transkripční, translační, sestřihové signály).

http://hollywood.mit.edu/GENSCAN.html



Eukaryotic Genome Annotation at NCBI: https://www.ncbi.nlm.nih.gov/genome/annotation_euk/

Metody predikce genů u eukaryot

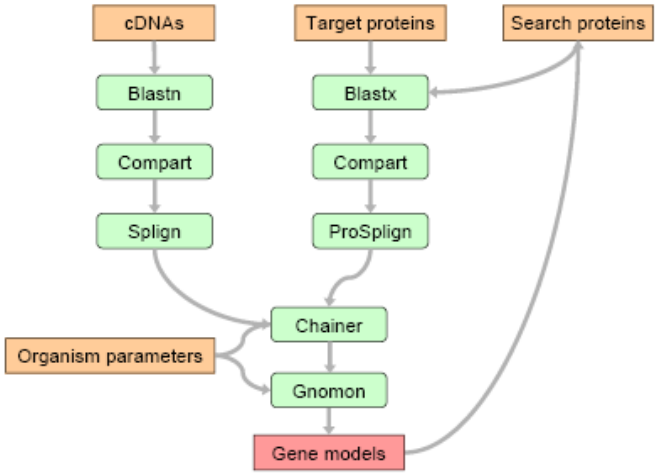
- Metody ***ab initio***/metody založené na **homologii**/metody založené na **konsenzu**.
- Metody založené na **homologii** – exonové sekvence příbuzných druhů jsou konzervované. Potenciální exony jsou porovnány se sekvencemi v databázi. Nelze použít pro nové geny bez homologů v databázi.
- Metody založené na **konsenzu** (shoda mínění, vzájemný souhlas) – porovnání výstupů z více různých predikčních programů. Výběr shodných výsledků – omezení falešně pozitivních výsledků. Problém: nižší citlivost, vynechání některých genů.

Metody predikce genů u eukaryot

- Metody ***ab initio***/metody založené na **homologii**/metody založené na **konsenzu**.
- V praxi často využívány **kombinace** přístupů, *ab initio* + homologie. Využití **experimentálních** dat – proteiny, RNA sekvence, geny (ze zkoumaného organismu nebo homologní), „spliced alignments“.

***Gnomon**, the NCBI eukaryotic gene prediction tool*

Before we start a genome annotation we collect several data sets. First we collect all available cDNA for the studied organism and sometimes cDNA for closely related organisms. Then we generate a Target protein set and a Search protein set. The former is a collection of the proteins that we believe should be found on the genome. Usually this includes all known proteins for the studied organism and several sets of known proteins for other, well studied genomes. The latter set is a much wider collection of eukaryotic proteins. We try to align on the genome all proteins from the Target Protein Set. The proteins from the Search Protein Set are aligned only if they are similar enough to predicted models, in which case these additional alignments are used in refining the models. In addition to the sequences used for the homology search we create an organism specific parameter set which is used for evaluation of the *ab initio* scores.



https://www.ncbi.nlm.nih.gov/genome/annotation_euk/gnomon/

Metody predikce genů u eukaryot

Predicting Genes in Single Genomes with AUGUSTUS

Katharina J. Hoff[1,2] and Mario Stanke[1,2]

[1]University of Greifswald, Institute of Mathematics and Computer Science, Greifswald, Germany
[2]Corresponding authors: *katharina.hoff@uni-greifswald.de*; *mario.stanke@uni-greifswald.de*

AUGUSTUS is a tool for finding protein-coding genes and their exon-intron structure in genomic sequences. It does not necessarily require additional experimental input, as it can be applied in so-called *ab initio* mode. However, extrinsic evidence from various sources such as transcriptome sequencing or the annotations of closely related genomes can be integrated in order to improve the accuracy and completeness of the annotation. AUGUSTUS can be applied to single genomes, or simultaneously to several aligned genomes. Here, we describe steps required for training AUGUSTUS for the annotation of individual genomes and the steps to do the actual structural annotation. Further, we describe the generation and integration of evidence from various sources of extrinsic evidence. © 2018 by John Wiley & Sons, Inc.

http://bioinf.uni-greifswald.de/webaugustus/

http://exon.gatech.edu/GeneMark/gmep_plus_instructions.html

ABSTRACT

We have made several steps toward creating a fast and accurate algorithm for gene prediction in eukaryotic genomes. First, we introduced an automated method for efficient *ab initio* gene finding, GeneMark-ES, with parameters trained in iterative *unsupervised* mode. Next, in GeneMark-ET we proposed a method of integration of unsupervised training with information on intron positions revealed by mapping short RNA reads. Now we describe GeneMark-EP, a tool that utilizes another source of external information, a protein database, readily available prior to the start of a sequencing project. A new specialized pipeline, ProtHint, initiates massive protein mapping to genome and extracts hints to splice sites and translation start and stop sites of potential genes. GeneMark-EP uses the hints to improve estimation of model parameters as well as to adjust coordinates of predicted genes if they disagree with the most reliable hints (the -EP+ mode). Tests of GeneMark-EP and -EP+ demonstrated improvements in gene prediction accuracy in comparison with GeneMark-ES, while the GeneMark-EP+ showed higher accuracy than GeneMark-ET. We have observed that the most pronounced improvements in gene prediction accuracy happened in large eukaryotic genomes.

GeneMark-EP+: eukaryotic gene prediction with self-training in the space of genes and proteins

Tomáš Brůna[1,†], Alexandre Lomsadze[2,†] and Mark Borodovsky[1,2,3,*]

[1]School of Biological Sciences, Georgia Institute of Technology, Atlanta, GA 30332, USA, [2]Wallace H. Coulter Department of Biomedical Engineering, Georgia Institute of Technology, Atlanta, GA 30332, USA and [3]School of Computational Science and Engineering, Georgia Institute of Technology, Atlanta, GA 30332, USA

Received December 13, 2019; Revised March 10, 2020; Editorial Decision April 01, 2020; Accepted May 12, 2020

Eukaryotické geny
Mnohobuněčná eukaryota

- **Mnohobuněčná eukaryota**

  Komplexní organizace genomu, geny separovány dlouhými INTERGENOVÝMI úseky, geny obsahují množství INTRONŮ, i velmi DLOUHÝCH.

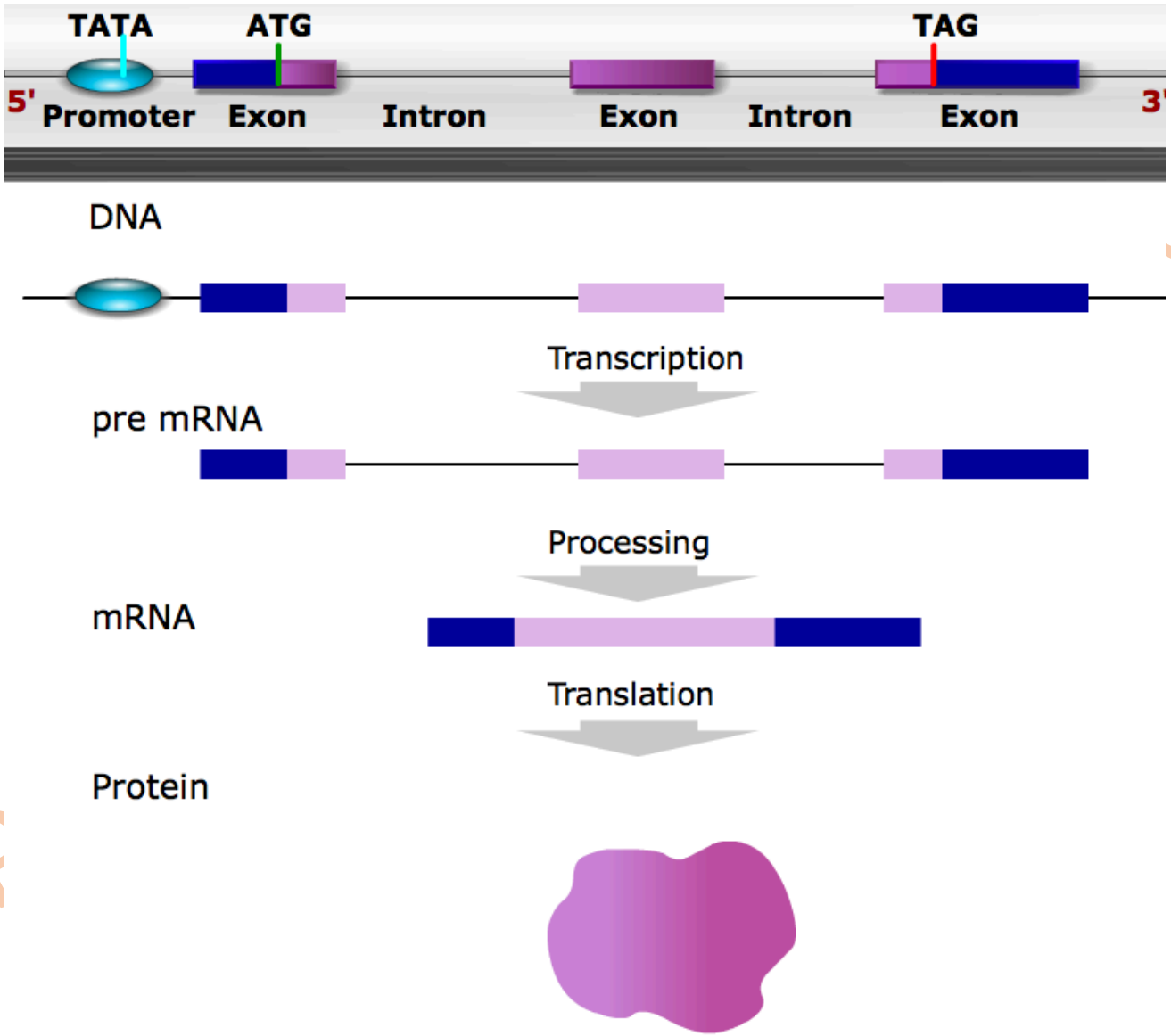
Glyceraldehyd-3-fosfát-dehydrogenasa
Candida albicans

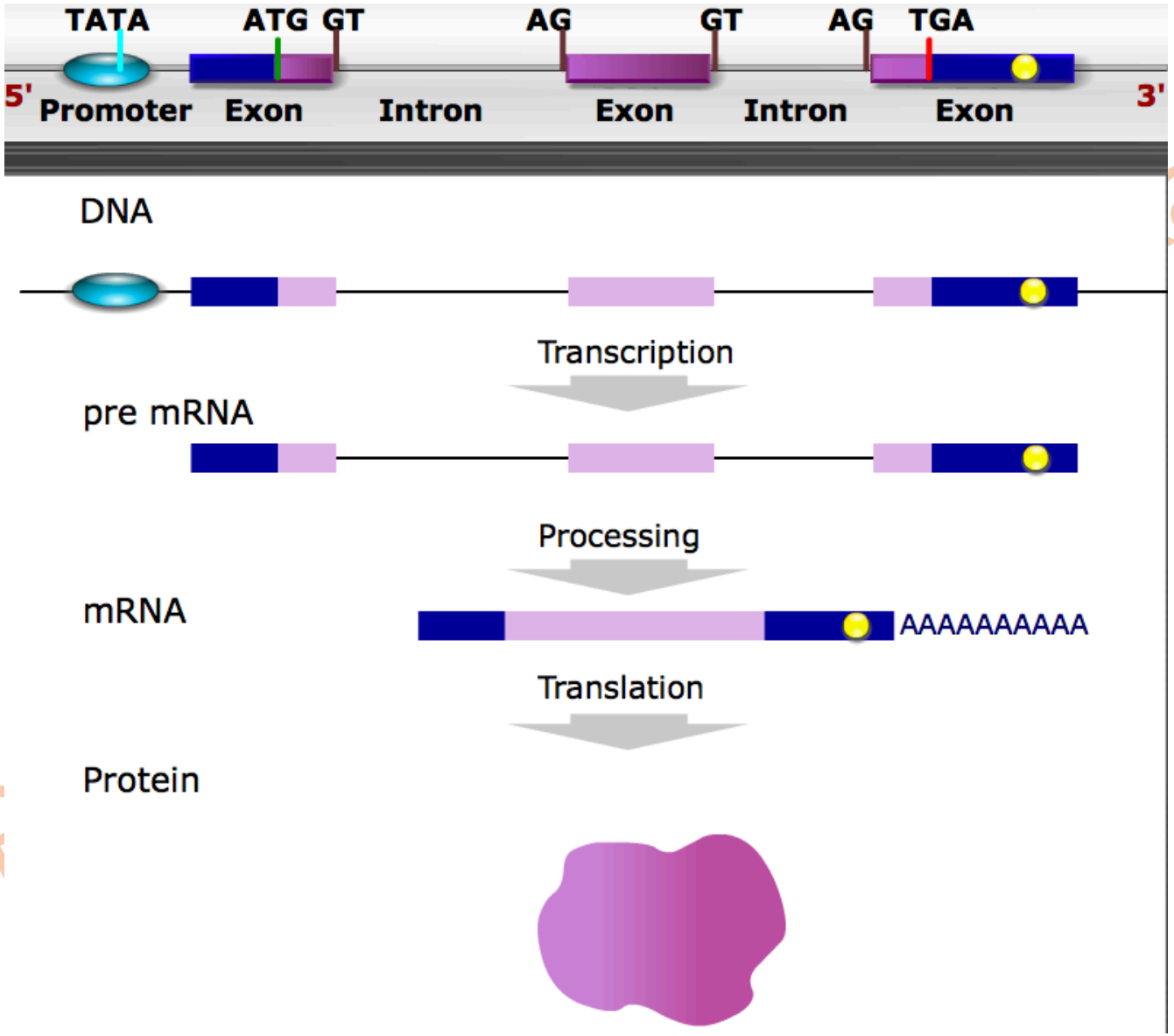
Eukaryotické geny
Mnohobuněčná eukaryota

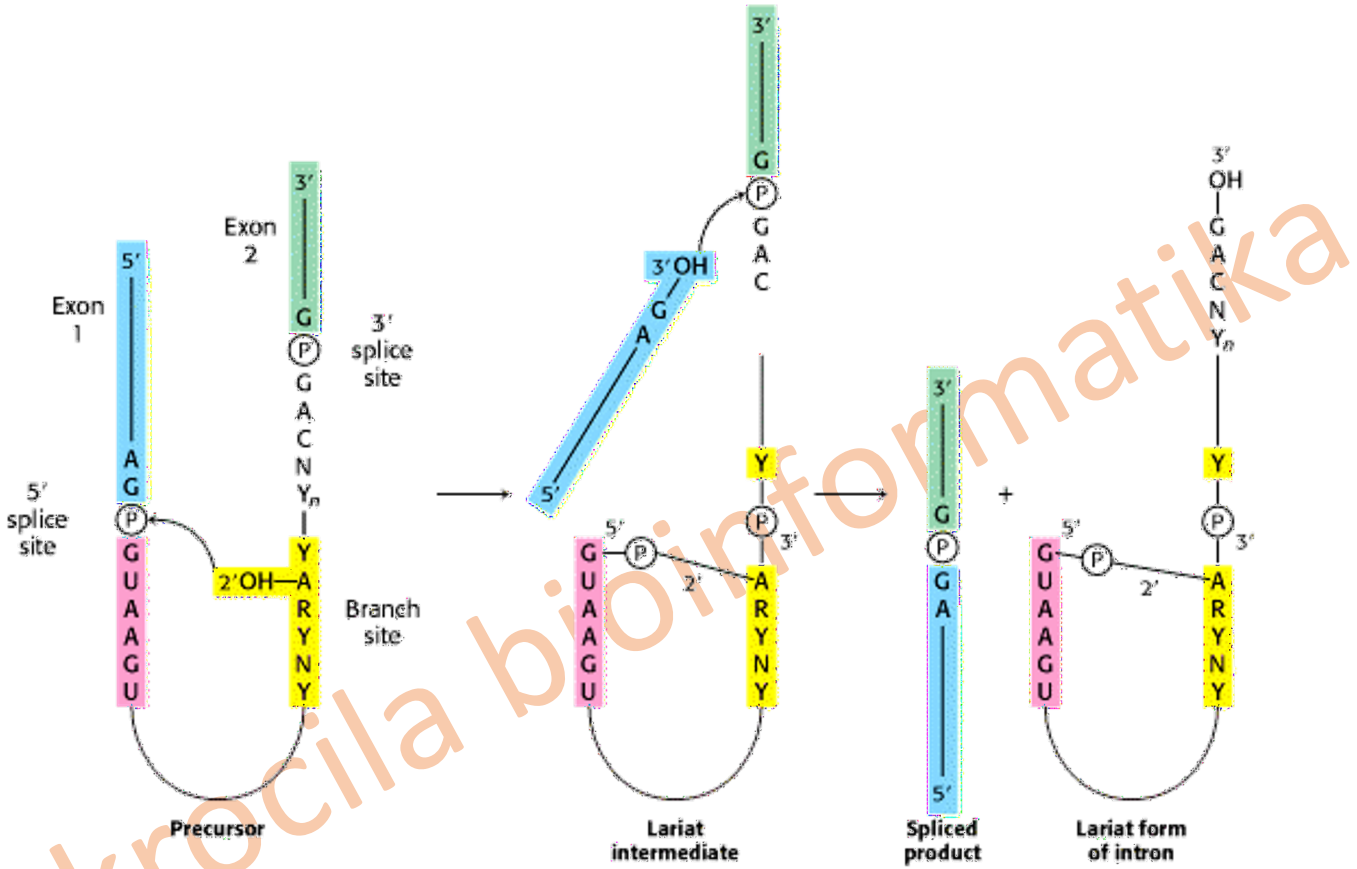
- **Mnohobuněčná eukaryota**

  Komplexní organizace genomu, geny separovány dlouhými INTERGENOVÝMI úseky, geny obsahují množství INTRONŮ, i velmi DLOUHÝCH.

Glyceraldehyd-3-fosfát-dehydrogenasa
Homo sapiens


TATA
ATG
TAG
5'
Promoter
Exon
Intron
Exon
Intron
Exon
3'
DNA
Transcription
pre mRNA
Processing
mRNA
Translation
Protein


TATA
ATG GT
AG
GT
AG
TGA
5'
Promoter
Exon
Intron
Exon
Intron
Exon
3'
DNA
Transcription
pre mRNA
Processing
mRNA
AAAAAAAAAA
Translation
Protein



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.]

Eukaryotické geny
Mnohobuněčná eukaryota

Eukaryotické geny
Mnohobuněčná eukaryota

- **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.

Algoritmy a nástroje pro identifikaci genů

- **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.**

Algoritmy a nástroje pro identifikaci genů

- **Každý program má výhody a nevýhody – rozumné použít více predikčních nástrojů.**

  GeneMark

  GlimmerM

  GRAIL

  GenScan

  Fgenes

  Augustus

  …

Algoritmy a nástroje pro identifikaci genů

- **GeneMark**

  http://exon.gatech.edu/GeneMark

  Využívá Markovovy modely

  Vyžaduje parametry specifické pro daný organismus = nutné „natrénování“ pomocí známých genů

  Varianty pro prokaryotické, eukaryotické, virové sekvence

GeneMark

http://exon.gatech.edu/GeneMark

Gene Prediction in Bacteria, Archaea and Metagenomes

For bacterial and archaeal gene prediction we recommend to use a parallel combination of GeneMark-P* and GeneMark.hmm-P with pre-computed models.
A novel genome can be analyzed either by the program with Heuristic models (if the sequence is shorter than 100 kb) or by the self-training program GeneMarkS* (aka GeneMark.hmm-PS).
Metagenomic sequences can be analyzed by our new program with updated heuristic models.

Gene Prediction in Eukaryotes

For eukaryotic gene prediction you can use the parallel combination of GeneMark-E* and GeneMark.hmm-E.
For a novel genome (the one whose name is not in the list of available models) you can install and run locally GeneMark.hmm-ES, the self-training program (just 10MB sequence is needed for training).

Gene Prediction in Viruses, Phages and Plasmids

For novel virus, phage and plasmid gene prediction you can use either the Heuristic approach (if the sequence is shorter than 50 kb) or the self- training program GeneMarkS (aka GeneMark.hmm-PS).
Both options will run the parallel combination of GeneMark and GeneMark.hmm.

Algoritmy a nástroje pro identifikaci genů

- **GeneScan**
  http://genes.mit.edu/GENSCAN.html

  Komplexní model struktury genu (transkripční, translační, sestřihové signály + statistické vlastnosti kódujících a nekódujících úseků)

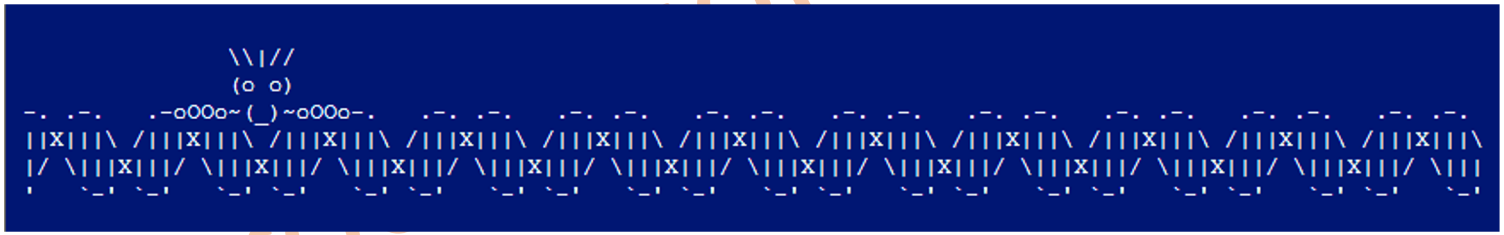
  Primární analýza velkých úseků eukaryotické genomové DNA

GeneScan

http://genes.mit.edu/GENSCAN.html

The New GENSCAN Web Server at MIT

Identification of complete gene structures in genomic DNA



This server provides access to the program Genscan for predicting the locations and exon-intron structures of genes in genomic sequences from a variety of organisms.

This server can accept sequences up to 1 million base pairs (1 Mbp) in length. If you have trouble with the web server or if you have a large number of sequences to process, request a local copy of the program (see instructions at the bottom of this page) or use the GENSCAN email server. If your browser (*e.g.*, Lynx) does not support file upload or multipart forms, use the older version.

Algoritmy a nástroje pro identifikaci genů

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.

Shrnutí

- Predikce prokaryotických genů mnohem jednodušší než u eukaryotických.
- Predikce genů *ab initio*/na základě sekvenční homologie.
- Nutné kombinovat oba přístupy = konsensus
- Rozumné využívat více predikčních programů.

Úkol 2

- DEFINITION fucose-specific lectin [Arthroderma otae CBS 113480].

- ACCESSION   XP_002846975
- VERSION   XP_002846975.1
- DBSOURCE   REFSEQ: accession XM_002846929.1