Bioinformatika

http://www.embnet.sk/edu/ppb

SYLABUS PREDMETU BIOINFORMATIKA PRE 3. ROČNÍK BIOLOGICKÝCH ODBORMOV

1. **Úvod do Bioinformatiky**
   definícia - história -náplň - internet - vzťah k ostatným vedným odborom
2. **Biologické databázy**
   biologické dáta - iné dáta využívané v biológii - rozdelenie biologických databáz - princíp práce s databázami
3. **Primárne databázy**
   typy primárnych sekvencií – EMBL/GenBank/DDBJ – UniProt – GO – vkladanie dát – využitie
4. **Sekundárne databázy**
   proteínové motívy – PROSITE – PRINTS – Pfam – BLOCKS – INTERPRO
5. **Ďalšie biologické databázy a integrované databázové systémy**
   PDB – KEGG – OMIM – REBASE – bibliografické dáta – MEDLINE – integrované databázové systémy – SRS – Entrez
6. **Analýza biologických dát**
   zhromažďovanie a analýza biologických dát - sekvenačné projekty - štatistická analýza - používaná výpočtová technika - Staden Package - EMBOSS
7. **Identifikácia kódujúcich úsekov nukleových kyselín**
   signály – motívy – kódujúce úseky – prokaryoty vs. eukaryoty
8. **Zoradenia dvoch sekvencií**
   pairwise alignment – *dot plot* – substitučné matice – lokálne a globálne zoradenia – BLAST – FASTA – Needleman-Wunsch – Smith-Waterman
9. **Zoradenia viacerých sekvencií**
   multiple sequence alignment – dynamické programovanie – progresívne metódy – konsenzus sekvencia – ClustalW
10. **Identifikácia proteínových motívov**
    proteínové motívy sekundárnych databáz – neurónové siete – ScanProsite – Pscan – HmmPfam
11. **Molekulárna fylogenetická analýza**
    bioinformatika a evolúcia – fylogenetické stromy – *distance* metódy – *maximum likelihood* metódy – *parsimony* metódy – PHYLIP
12. **Sekundárna a terciárna štruktúra biomakromulekúl**
    primárna, sekundárna a terciárna štruktúra – vzťah štruktúry a funkcie – 3D vizualizácia – RasMol – MOLMOL

Čo je bioinformatika

"*Materiálom pre bioinformatiku sú biologické dáta, využívané metódy sú odvodené od širokého spektra výpočtových techník*"

David Benton (National Center for Human Genome Research, Bethesda)

Bioinformatika

je definovaná ako vedecká disciplína, ktorá zahŕňa všetky aspekty získavania biologických informácií, ich spracovania, uskladnenia, distribúcie, analýzy a interpretácie, pričom kombinuje nástroje a techniky matematiky, počítačových vied a biológie za účelom pochopiť biologickú hodnotu širokého spektra dát

Výklady pojmu bioinformatika

- výskum a vývoj spojený s budovaním informačnej infraštruktúry potrebnej pre modernú biológiu (bioinformatika v užšom zmysle slova)
- výskum založený na využití výpočtovej techniky, ktorý vedie k zodpovedaniu základných biologických otázok (všeobecne nazývaný ako *computational biology*)

Obsahová náplň bioinformatiky

- práca s biologickými databázami (*database mining)*
- identifikácia sekvenčných homológií
- zoradenia sekvencií
- evolučná biológia - fylogenetická analýza
- identifikácia špecifických motívov a paternov
- štruktúrna biológia
- mapovanie genómov
- vyhodnocovanie DNA *microarrays, NGS*

Bioinformatika – internet – kooperácie

- Internet
  - uľahčuje a urýchľuje komunikáciu jednotlivých vedeckých pracovísk
- bioinformatické centrá
  - udržiavajú potrebný software a dáta
  - poskytujú podporu
  - priamo sa podieľajú na riešení vedeckých projektov
- EMBnet (European Molecular Biology network)
  - vedecko-výskumné združenie spolupracujúcich uzlov z Európy spolu s niekoľkými strediskami z celého sveta
  - nadnárodná spolupráca zabezpečuje bioinformatickú podporu na oveľa vyššej úrovni ako by bol schopný poskytnúť jednotlivý uzol

Bioinformatika
- vzťah k iným vedným odborom

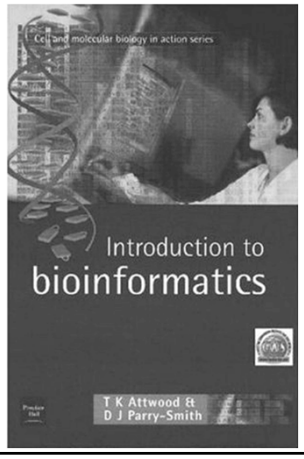
- molekulárna biológia
- informatika

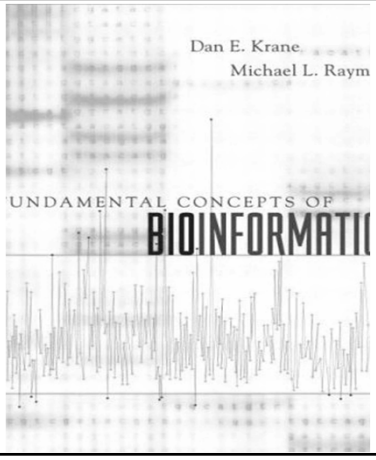
molekulárna biológia + informatika = bioinformatika

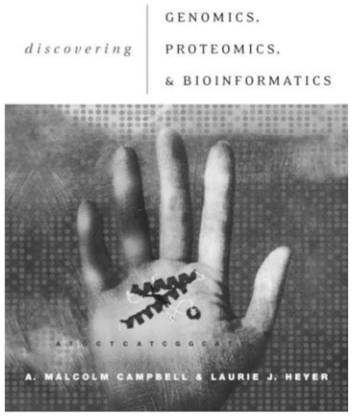
- široké spektrum „bio-medicínskych“ vied
  - biochémia
  - genetika
  - medicína

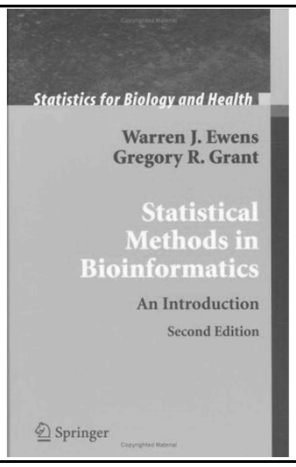
Míľniky bioinformatiky

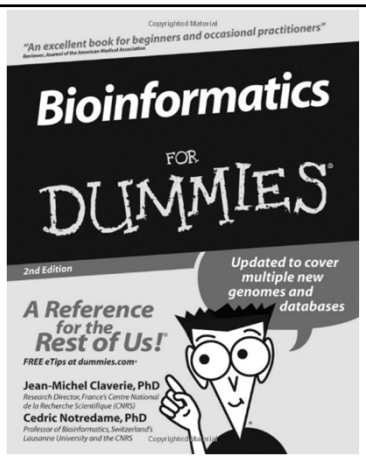
1962 Pauling's theory of molecular evolution
1965 Margaret Dayhoff's Atlas of Protein Sequences
1970 Needleman-Wunsch algorithm
1977 DNA sequencing and software to analyze it (Staden)
1981 Smith-Waterman algorithm developed
1981 The concept of a sequence motif (Doolittle)
1982 GenBank Release 3 made public (606 entries)
1982 Phage lambda genome sequenced
1983 Sequence database searching algorithm (Wilbur-Lipman)
1985 FASTP/FASTN: fast sequence similarity searching
1988 National Center for Biotechnology Information (NCBI) created at NIH/NLM
1988 EMBnet network for database distribution
1990 BLAST: fast sequence similarity searching
1991 EST: expressed sequence tag sequencing
1993 Sanger Centre, Hinxton, UK
1994 EMBL European Bioinformatics Institute, Hinxton, UK
1995 First bacterial genomes completely sequenced
1996 Yeast genome completely sequenced
1997 PSI-BLAST
1998 Worm (multicellular) genome completely sequenced
1999 Fly genome completely sequenced
2000 1st human chromosome sequenced
2001 draft human genome published



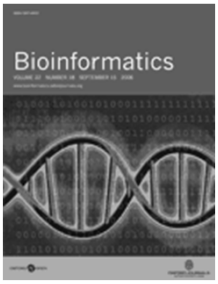








Časopisy



SYLABUS PREDMETU BIOINFORMATIKA PRE 3. ROČNÍK BIOLOGICKÝCH ODBOROV

1. Úvod do Bioinformatiky
definícia - história -náplň - internet - vzťah k ostatným vedným odborom

2. Biologické databázy
biologické dáta - iné dáta využívané v biológii - rozdelenie biologických databáz - princíp práce s databázami

3. Primárne databázy
typy primárnych sekvencií – EMBL/GenBank/DDBJ – UniProt – GO – vkladanie dát – využitie

4. Sekundárne databázy
proteínové motívy – PROSITE – PRINTS – Pfam – BLOCKS – INTERPRO

5. Ďalšie biologické databázy a integrované databázové systémy
PDB – KEGG – OMIM – REBASE – bibliografické dáta – MEDLINE – integrované databázové systémy – SRS – Entrez

6. Analýza biologických dát
zhromažďovanie a analýza biologických dát - sekvenačné projekty - štatistická analýza - používaná výpočtová technika - Staden Package - EMBOSS

7. Identifikácia kódujúcich úsekov nukleových kyselín
signály – motívy – kódujúce úseky – prokaryoty vs. eukaryoty

8. Zoradenia dvoch sekvencií
pairwise alignment – dot plot – substitučné matice – lokálne a globálne zoradenia – BLAST – FASTA – Needleman-Wunsch – Smith-Waterman

9. Zoradenia viacerých sekvencií
multiple sequence alignment – dynamické programovanie – progresívne metódy – konsenzus sekvencia – ClustalW

10. Identifikácia proteínových motívov
proteínové motívy sekundárnych databáz – neurónové siete – ScanProsite – Pscan – HmmPfam

11. Molekulárna fylogenetická analýza
bioinformatika a evolúcia – fylogenetické stromy – *distance* metódy – *maximum likelihood* metódy – *parsimony* metódy – PHYLIP

12. Sekundárna a terciárna štruktúra biomakromulekúl
primárna, sekundárna a terciárna štruktúra – vzťah štruktúry a funkcie – 3D vizualizácia – RasMol – MOLMOL

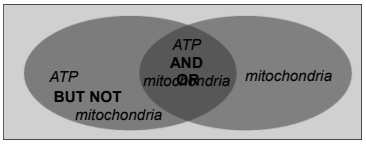
Databázy

- usporiadaná množina informácií uložená na pamäťovom médiu
- v širšom zmysle sú súčasťou databázy aj softvérové nástroje, ktoré slúžia na manipuláciu a prístup k uloženým dátam
- predchodcom databáz boli papierové kartotéky
  - zaraďovanie nových položiek
  - usporiadanie dát podľa rôznych kritérií

aj keď všetky operácie robil človek, správa takýchto kartoték bola v mnohom podobná správe dnešných databáz.

Logické operátory

- logický súčin
  - operátor **AND**
  - symbol „**&**“, „**∩**“
- logický súčet
  - operátor **OR**
  - symbol „**|**”, „**U**“
- logická negácia
  - operátor **BUT NOT**
  - symbol „**!**”, „**¬**“



Wildcards

***** - hviezdička (*asterisk*)
- 0, 1, alebo viac ľubovoľných znakov

- výraz „z*ný“ tak zodpovedá slovám
  „zný“
  „zadný“
  „základný“
  „zákonodarný“
  ale aj „základ určený“

? – otáznik (*question mark*)
- práve jeden ľubovoľný znak

- výraz „z??ný“ tak zodpovedá napr. slovu
  „zadný“
  ale nie
  „zný“ alebo „základný“

Biologické databázy

- **primárne databázy**
  - databázy primárnych sekvencií nukleových kyselin a proteínov
    ENA (EMBL), GenBank, UniProt
- **sekundárne databázy**
  PROSITE, PRINTS, BLOCKS, PFAM
- **databázy makromolekulárnych štruktúr**
  PDB
- **bibliografické databázy**
  MEDLINE
- iné

SYLABUS PREDMETU BIOINFORMATIKA PRE 3. ROČNÍK BIOLOGICKÝCH ODBOROV

1. Úvod do Bioinformatiky
definícia - história -náplň - internet - vzťah k ostatným vedným odborom

2. Biologické databázy
biologické dáta - iné dáta využívané v biológii - rozdelenie biologických databáz - princíp práce s databázami

3. Primárne databázy
typy primárnych sekvencií – EMBL/GenBank/DDBJ – UniProt – GO – vkladanie dát – využitie

4. Sekundárne databázy
proteínové motívy – PROSITE – PRINTS – Pfam – BLOCKS – INTERPRO

5. Ďalšie biologické databázy a integrované databázové systémy
PDB – KEGG – OMIM – REBASE – bibliografické dáta – MEDLINE – integrované databázové systémy – SRS – Entrez

6. Analýza biologických dát
zhromažďovanie a analýza biologických dát - sekvenačné projekty - štatistická analýza - používaná výpočtová technika - Staden Package - EMBOSS

7. Identifikácia kódujúcich úsekov nukleových kyselín
signály – motívy – kódujúce úseky – prokaryoty vs. eukaryoty

8. Zoradenia dvoch sekvencií
pairwise alignment – *dot plot* – substitučné matice – lokálne a globálne zoradenia – BLAST – FASTA – Needleman-Wunsch – Smith-Waterman

9. Zoradenia viacerých sekvencií
multiple sequence alignment – dynamické programovanie – progresívne metódy – konsenzus sekvencia – ClustalW

10. Identifikácia proteínových motívov
proteínové motívy sekundárnych databáz – neurónové siete – ScanProsite – Pscan – HmmPfam

11. Molekulárna fylogenetická analýza
bioinformatika a evolúcia – fylogenetické stromy – *distance* metódy – *maximum likelihood* metódy – *parsimony* metódy – PHYLIP

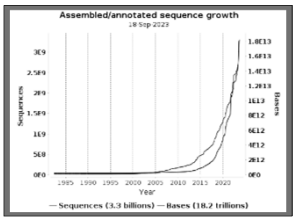
12. Sekundárna a terciárna štruktúra biomakromulekúl
primárna, sekundárna a terciárna štruktúra – vzťah štruktúry a funkcie – 3D vizualizácia – RasMol – MOLMOL

Primárne databázy

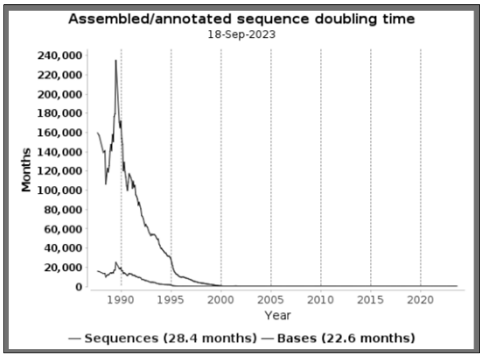
- údaje o primárnej štruktúre hlavných typov biomakromolekúl – nukleových kyselín a proteínov
- nukleotidové sekvencie (DNA, RNA)
  - ENA(EMBL)/GenBank/DDBJ
- aminokyselinové sekvencie (proteíny)
  - UniProt

ENA (EMBL)

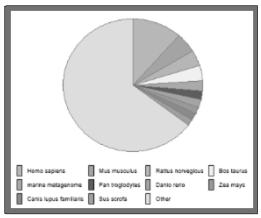
- anotované nukleotidové sekvencie
- EBI (European Bioinformatics Institute) Hinxton, Cambridge, UK
- Sep 2023
  - 3,3 mld. záznamov
  - 18,2 bil. nukleotidov
  - 51,6 PB



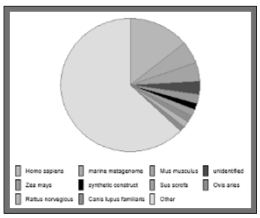
ENA (EMBL)



ENA (EMBL)
(Release 109 Sep 2011*)*



rozdelenia podľa
počtu nukleotidov



rozdelenia podľa
počtu záznamov

ENA (EMBL) – triedy, divízie

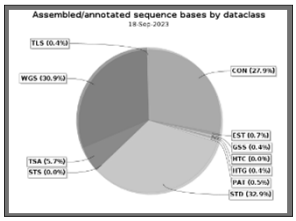
metodologické triedy (*classes*)

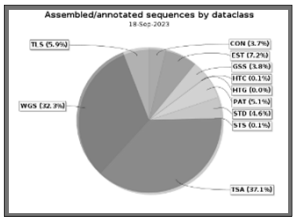
CON: Constructed
EST: Expressed Sequence Tag
GSS: Genome Sequence Scan
HTC: High Throughput CDNA
HTG: High Throughput Genome sequencing
PAT: Patents
STD: Standard
STS: Sequence Tagged Site
TSA:Transcriptome Shotgun Assembly
WGS: Whole Genome Shotgun

taxonomické divízie (*divisions*)

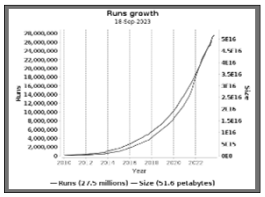
ENV: Environmental Samples
FUN: Fungi
HUM: Human
INV: Invertebrates
MAM: Other Mammals
MUS: Mus musculus
PHG: Bacteriophage
PLN: Plants
PRO: Prokaryotes
ROD: Rodents
SYN: Synthetic
TGN: Transgenic
UNC: Unclassified
VRL: Viruses
VRT: Other Vertebrates

ENA (EMBL) – triedy, divízie

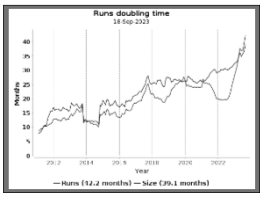




ENA (EMBL)
(Sep 2023*)*



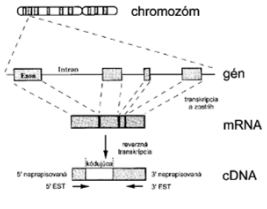
nárast počtu readov



duplikácia počtu readov

EST
(Expressed sequence tag)

- **EST - úseky exprimovaných sekvencií** (sekvencií cDNA získaných z exprimovaných mRNA)
- z $3x10^9$ bp ľudského genómu iba niekoľko percent kóduje 20-25 tis. génov, ktoré sú prepisované do mRNA a následne do proteínov
- mRNA je izolovaná **z rôznych tkanív** a následne *in vitro* prepísaná do cDNA; EST sú generované sekvenovaním častí cDNA knižnice, pričom predstavujú iba časť každého transkriptu
- predstavujú záznam profilu exprimovaných génov v danom tkanive a/alebo v danom vývojovom štádiu
- **dlhé iba 300-500 bp**, čo však stačí na určenie podobnosti so známymi, alebo dovtedy nedefinovanými génmi



- **GSS** - *Genome Sequence Scans*
  - podobné EST, ale pôvodom z genómu, nie z mRNA
- **HTG** - *High Throughput Genome sequencing*
  - genómové sekvencie produkované *high-throughput* sekvenačnými projektmi
  - priebežne aktualizované a anotované
- **STS** - *Sequence Tagged Sites*
  - krátke (200-500 bp) sekvencie, charakteristické pre daný genóm
  - možno ich špecificky detegovať (napr. pomocou PCR amplifikácie)
  - definujú špecifickú pozíciu na fyzikálnej mape

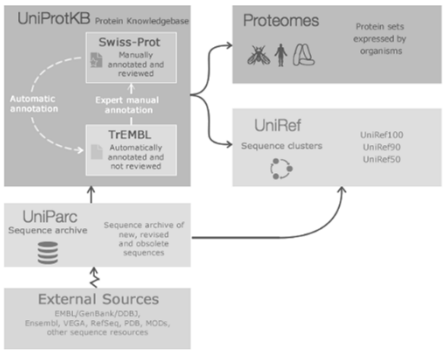
Ďalšie primárne nukleotidové DB

- **Ensembl** (EBI/EMBL/WTSI)
  - databáza genómov stavovcov a iných eukaryotických organizmov
- **RefSeq** (NCBI)
  - *open access*, integrovaná, neredundantná a anotovaná databáza sekvencií:
    - genómové DNA
    - transkripty
    - proteíny
- 289 mil. proteínov
- 56 mil. transkriptov
- 141 tis. organizmov
  (Release 220, Sep 2023)

UniProt
(*Universal Protein Resource,* EBI/SIB/PIR)

- **Knowledgebase**
  - anotované aminokyselinové sekvenice proteínov
  - Swiss-Prot + TrEMBL (EBI/SIB)
  - PIR (Georgetown University)
- **UniParc** (UniProt Archive)
  - najobsiahlejši neredundantný verejne prístupný súbor proteínových sekvencií
- **UniRef** (UniProt Non-redundant Reference Databases)
  - neredundantné databázy združujúce blízko príbuzné sekvencie do jedného záznamu
  - UNIREF100, UNIREF90 a UNIREF50
- **UniMES** (UniProt Metagenomic and Environmental Sequences)
  - špeciálne určené pre metagenomické a environmentálne údaje

UniProt –zdroje a toky dát

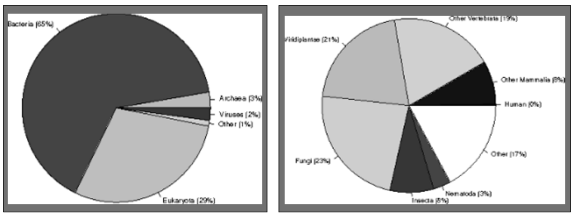


zdroj: *UniProt flyer*

UniProt
(Release Sep 2023)

- **UniProtKB** (Uniprot Knowledgebase)
  Swiss-Prot: 570 tis. záznamov
  TrEMBL: 252 mil. záznamov
- **UniParc** (UniProt Archive)
  - 341 mil. záznamov
- **UniRef** (UniProt Non-redundant Reference Databases)
  - UNIREF100: 236 mil. záznamov
  - UNIREF90: 116 mil. záznamov
  - UNIREF50: 42 mil. záznamov

UniProt
(Release Sep 2023)

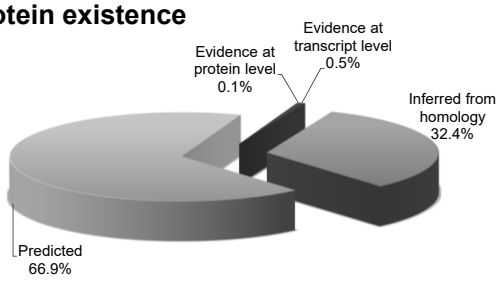


rozdelenia podľa **taxonomického zaradenia**

z toho **eukaryoty**

UniProtKB/TrEMBL

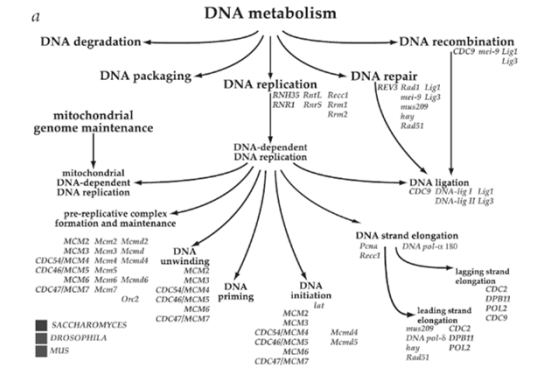
Protein existence



História a štruktúra GO

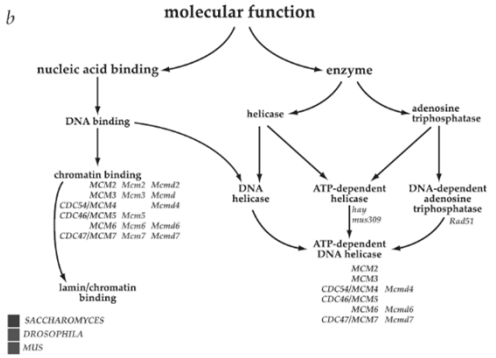
- 1998 – Konzorcium GO www.geneontology.org
- tri samostatné ontológie GO charakterizujúce génové produkty
  - biologické procesy (64 %)
  - molekulárne funkcie (26 %)
  - bunkové komponenty (9 %)
- vzťahy: **`is_a`** | **`part_of`** | **`regulates`**
- nárast počtu záznamov z 3,5 tis. na 43 tis.
- GO sú odkazované z mnohých biologických databáz
  - druhovo špecifické genómové databázy
  - univerzálne informačne zdroje ako UniProt alebo InterPro

Ontológia biologických procesov



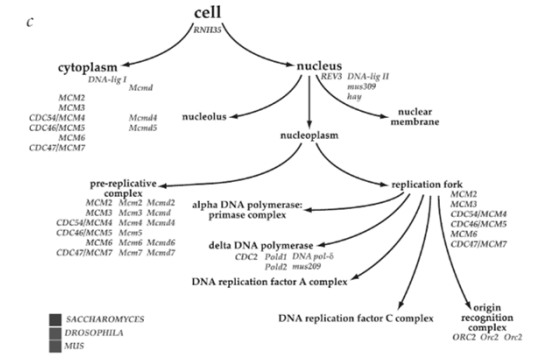
The Gene Ontology Consortium (2000): Gene Ontology: tool for the unification of biology. *Nature Genetics* **25**, 25-29.

Ontológia molekulárnych funkcií



The Gene Ontology Consortium (2000): Gene Ontology: tool for the unification of biology. *Nature Genetics* **25**, 25-29.

Ontológia bunkových komponentov



The Gene Ontology Consortium (2000): Gene Ontology: tool for the unification of biology. *Nature Genetics* **25**, 25-29.

UniProt a GO

Last Annotation Update entry version 3, 05-FEB-2008

Description and Origin of the Protein

Keywords	Bacterial flagellum; Flagellum
Description	Periplasmic flagellar core protein.
Gene names(s)	Name=flaB;
Organism Source	Leptospira interrogans serovar Grippotyphosa.
Organism Classification	Bacteria; Spirochaetes; Spirochaetales; Leptospiraceae; Leptospira
NCBI Taxonomy ID	280498

References

1, Kaiyu,W. Zhihua,K. Shuhan,S.; Submitted (JUL-2007) to the EMBL/GenBank/DDBJ databases.
Position NUCLEOTIDE SEQUENCE.

Database Cross-References

EMBL	EU053206; ABU40946.1; -; Genomic_DNA.
GO	GO:0009420; C:flagellin-based flagellum filament; IEA:InterPro.
GO	GO:0005198; F:structural molecule activity; IEA:InterPro.
GO	GO:0001539; P:ciliary or flagellar motility; IEA:InterPro.
InterPro	IPR001029; Flagellin_C.
InterPro	IPR001492; Flagellin_N.
Pfam	PF00700; Flagellin_C; 1.
Pfam	PF00669; Flagellin_N; 1.
PRINTS	PR00207; FLAGELLIN.
ProDom	PD000316; Flagellin_C; 1.

Features

There are no feature annotations in this record

Sequence

Characteristics	Length : 283, molecular weight: 31295, CRC64 check sum: 7F26998130OCFE28
Sequence	>A7UFK8 Periplasmic flagellar core protein. MIINNNLSAVNAHRSLKFNELAVDKTMKALSSGMRINSSADDASGLAVSEKLRTQVNGLR

EBI > Databases > QuickGO

Home Help Downloads Your selection (0 terms)

GO:0001539 ciliary or flagellar motility

Locomotion due to movement of cilia or flagella.

Term Information | Ancestor chart | Ancestor table | Child Terms | Protein Annotation | Statis

ID	GO:0001539
Name	ciliary or flagellar motility
Definition	Locomotion due to movement of cilia or flagella.
Comment	

Synonyms

Type	Synonym
exact	ciliary/flagellar motility

XRefs

Database	ID
INTERPRO	IPR000067
INTERPRO	IPR000423
INTERPRO	IPR000090
INTERPRO	IPR000563
INTERPRO	IPR000809
INTERPRO	IPR001035
INTERPRO	IPR001444
INTERPRO	IPR001624
INTERPRO	IPR012826
INTERPRO	IPR006752

EBI > Databases > QuickGO

< Home Help Downloads Your selection (0 terms)

GO:0001539 ciliary or flagellar motility

Locomotion due to movement of cilia or flagella.

Term Information | Ancestor chart | Ancestor table | Child Terms | Protein Annotation | Statis

Help: filtering, analyzing and downloading annotation

Columns:	DB	Sequence	Alt	Symbol	Qualifier	GO ID	GO Term name
1 Filter:	Any	Any			Any	GO:0001539	
Advanced:							
2 Statistics:		11911				1	
3 View	1-25 -> bookmark						
	UniProtKB	A0AGD7				GO:0001539	ciliary or flagellar motility
	UniProtKB	A0AGE5		flaA		GO:0001539	ciliary or flagellar motility
	UniProtKB	A0AGE5		flaA		GO:0001539	ciliary or flagellar motility
	UniProtKB	A0AGE8				GO:0001539	ciliary or flagellar motility
	UniProtKB	A0AGF2				GO:0001539	ciliary or flagellar motility
	UniProtKB	A0AGF3				GO:0001539	ciliary or flagellar motility
	UniProtKB	A0AGF4		fliM		GO:0001539	ciliary or flagellar motility
	UniProtKB	A0AGF5				GO:0001539	ciliary or flagellar motility
	UniProtKB	A0AGG0				GO:0001539	ciliary or flagellar motility
	UniProtKB	A0AGG1				GO:0001539	ciliary or flagellar motility
	UniProtKB	A0AGG1				GO:0001539	ciliary or flagellar motility
	UniProtKB	A0AGG1				GO:0001539	ciliary or flagellar motility

Využitie GO

<u>funkčná</u> komparatívna genomika

- určenie príbuznosti génov na základe funkcie <u>nie</u> sekvencie

integrácia dát

- *data mining*
- sémantický web
- vedomostné siete

GoPubMed <u>www.gopubmed.org</u>

GoPubMed

www.gopubmed.org

algoritmy „spracovanie prirodzenej reči“ (PubMed + GO + MeSH)

<u>Odpovede na otázky:</u>

Aká je funkcia PrnP pri chorobe divokých kráv?

Aká je úloha IDE v Alzheimerovej chorobe?

Ktorá konkrétna choroba je najčastejšie asociovana s infekciou HIV?

Ktorá liečebná technika je najčastejšie používané pri liečbe HIV pacientov?

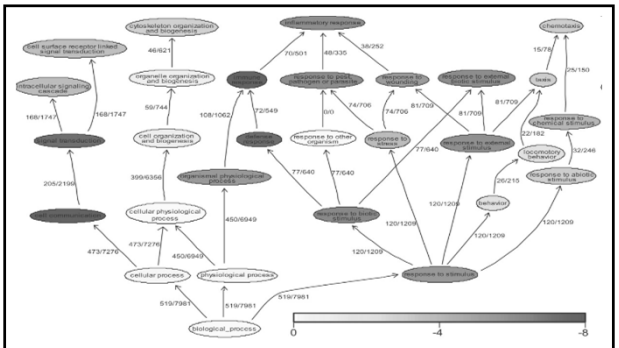
Ktorý sú najčastejší autori publikujúci v oblasti antiretrovirálnej terapie a kde a kedy bol ich výskum vykonávaný?

Ako ovplyvňuje p53 apoptózu?

Kto je na čele výkumu transplantácie pečene v Nemecku?

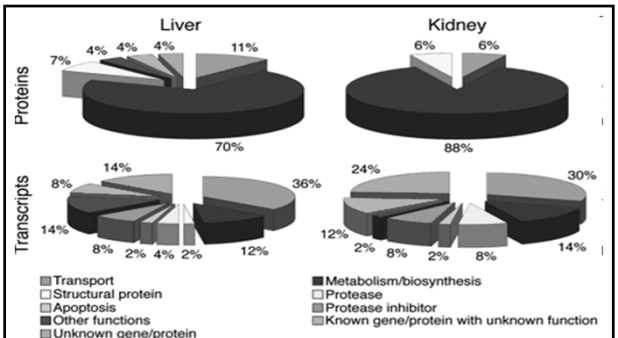
Kde je najintenzívnejšie vykonávaný výskum v oblasti celiatického ochorenia?

Ktoré publikačné obdobie Paula Nursa bolo najintenzívnejšie?



Gene Ontology analysis of the genes differentially expressed before and after treatment in the TNF-positive group of patients.

Lindberg *et al. Arthritis Research & Therapy* 2006 **8**:R179



Pie charts of functional categories for genes more abundant in liver or kidney at the protein or transcript level. Data are based on the top 100 differentially expressed transcripts and the 43 identified proteins. Standardized Gene Ontology (GO) classifications were extracted from the Mouse Genome Informatics database.

Mijalski at al. (2005) *Proc Natl Acad Sci U S A.* **102**(24):8621-6

SYLABUS PREDMETU <u>BIOINFORMATIKA</u> PRE 3. ROČNÍK BIOLOGICKÝCH ODBOROV

1. Úvod do Bioinformatiky
definícia - história -náplň - internet - vzťah k ostatným vedným odborom

2. Biologické databázy
biologické dáta - iné dáta využívané v biológii - rozdelenie biologických databáz - princíp práce s databázami

3. Primárne databázy
typy primárnych sekvencií – EMBL/GenBank/DDBJ – UniProt – GO – vkladanie dát – využitie

4. Sekundárne databázy
proteínové motívy – PROSITE – PRINTS – Pfam – BLOCKS – INTERPRO

5. Ďalšie biologické databázy a integrované databázové systémy
PDB – KEGG – OMIM – REBASE – bibliografické dáta – MEDLINE – integrované databázové systémy – SRS – Entrez

6. Analýza biologických dát
zhromažďovanie a analýza biologických dát - sekvenačné projekty - štatistická analýza - používaná výpočtová technika - Staden Package - EMBOSS

7. Identifikácia kódujúcich úsekov nukleových kyselín
signály – motívy – kódujúce úseky – prokaryoty vs. eukaryoty

8. Zoradenia dvoch sekvencií
pairwise alignment – *dot plot* – substitučné matice – lokálne a globálne zoradenia – BLAST – FASTA – Needleman-Wunsch – Smith-Waterman

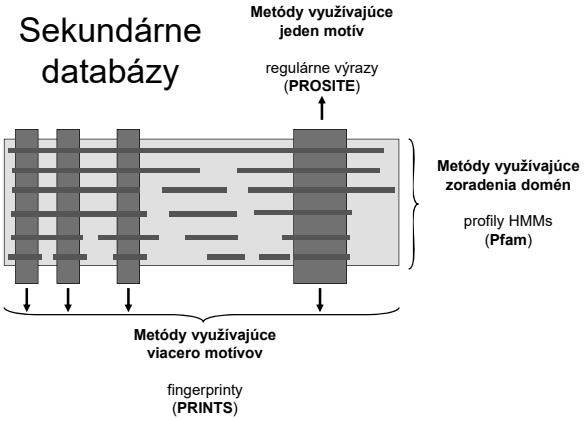
9. Zoradenia viacerých sekvencií
multiple sequence alignment – dynamické programovanie – progresívne metódy – konsenzus sekvencia – ClustalW

10. Identifikácia proteínových motívov
proteínové motívy sekundárnych databáz – neurónové siete – ScanProsite – Pscan – HmmPfam

11. Molekulárna fylogenetická analýza
bioinformatika a evolúcia – fylogenetické stromy – *distance* metódy – *maximum likelihood* metódy – *parsimony* metódy – PHYLIP

12. Sekundárna a terciárna štruktúra biomakromulekúl
primárna, sekundárna a terciárna štruktúra – vzťah štruktúry a funkcie – 3D vizualizácia – RasMol – MOLMOL

Sekundárne databázy



Sekundárne databázy

zhromažďujú dáta vychádzajúce z konzervatívnych úsekov príbuzných proteínov

- **PROSITE**
  - databáza proteínových domén a rodín
  - biologicky signifikantné miesta, charakterizované regulárnymi výrazmi (*regular expressions*)
  - Release 2020_04 (august 2020) obsahuje 1 860 dokumentačných záznamov, ktoré popisujú 1311 vzorov, 1 280 profilov/matíc a 1 311 pravidiel
- **PRINTS**
  - charakterizuje proteínové rodiny prostredníctvom fingerprintov
  - fingerprint sa odvodzuje z viacerých motívov
  - Release 39 (február 2009) obsahuje 1 950 fingerprintov pozostávajúcich z 11 625 motívov
- **Pfam**
  - databázou proteínových profilov (HMM profily)
  - profily sa odvodzujú z celých sekvencií domén, charakteristických pre jednotlivé proteínové rodiny
  - Release 33.1 (máj 2020) popisuje 18 259 proteínových rodín pokrývajúcich 73 % známych proteínových sekvencií z UniProtKB
- **InterPro**
  - integrovaná databáza proteínových rodín, domén a funkčných miest
  - Release 81.0 (august 2020) obsahuje 37 821 záznamov a pokrýva 81 % proteínov z UniProtKB

(CDD, HAMAP, PANTHER, PIRSF, PRINTS, PROSITE, Pfam, ProDom, SMART, SUPERFAMILY, TIGRFAMs)

Prosite

- **PA** (PAttern) riadok popisuje definíciu PROSITE vzoru
- použitie IUPAC jednopísmenových skratiek aminokyselín
- „**x**“ – ľubovoľná aminokyselina
- „**[]**“ – alternatívne aminokyseliny
  [ALT] predstavuje Ala alebo Leu alebo Thr
- „**{}**“ – nevyskytujúce sa aminokyseliny
  {AM} predstavuje ľubovoľnú aminokyselinu okrem Ala a Met.
- „**-**“ – oddeľuje jednotlivé aminokyselinové pozície
- „**()**“ – opakovanie elementu vo vzore
  x(3) zodpovedá x-x-x
  x(2,4) zodpovedá to x-x alebo x-x-x alebo x-x-x-x
- „**<**„ a „**>**“ – N-terminálny alebo C-terminálny koniec sekvencie
- „**.**“ – bodka ukončuje vzor

Príklady:

```
PA [AC]-x-V-x(4)-{ED}.
```
[Ala alebo Cys]-any-Val-any-any-any-any-{any but Glu or Asp}

```
PA <A-x-[ST](2)-x(0,1)-V.
```
N-terminálny koniec sekvencie ('<')
Ala-any-[Ser or Thr]-[Ser or Thr]-(any or none)-Val

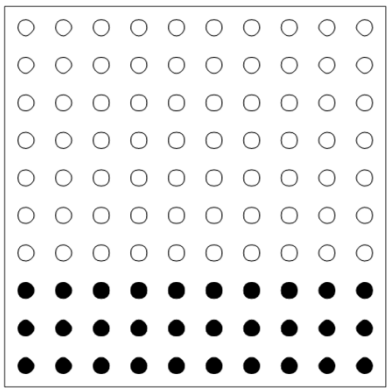
Sensitivity / Specificity

- ....is a well person
- ....is a person with a disease
- ....is a negative test result
- ....is a positive test result

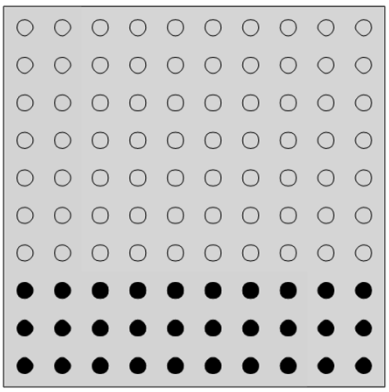
and therefore....

- ....is a well person who tests negative (a true negative - TN)
- ....is a person with a disease who tests positive (a true positive - TP)
- ....is a well person who tests positive (a false positive - FP)
- ....is a person with a disease who tests negative (a false negative - FN)

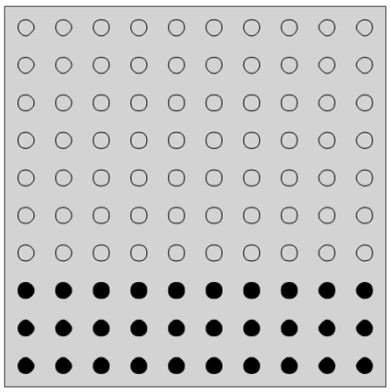
Loong TW (2003): Understanding sensitivity and specificity with the right side of the brain. *BMJ* **327**,716-719.



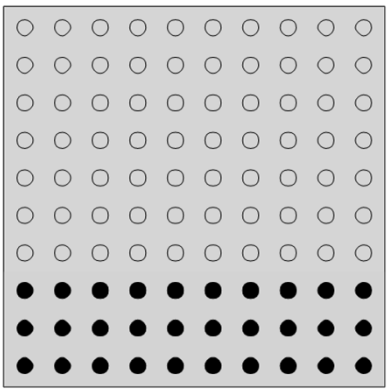
Hypothetical population



Results of diagnostic test on hypothetical population

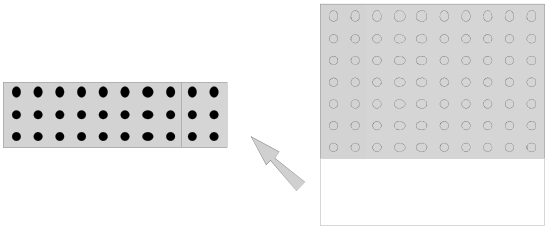


Test with 100% sensitivity



Perfect test

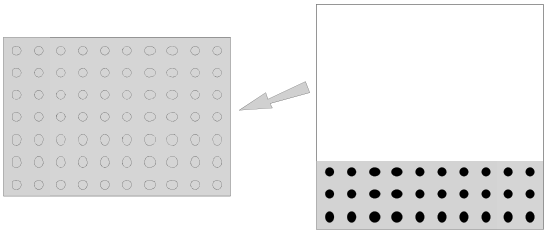
Sensitivity of test
how good a test is at correctly identifying the positive results



sensitivity = TP / TP+FN

24 / 30 = **80 %**

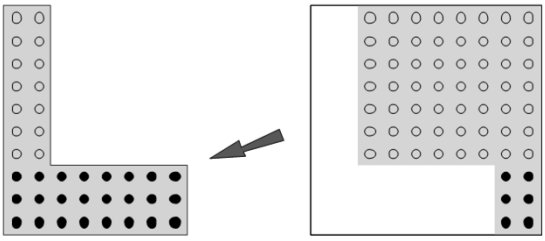
Specificity of test
how good the test is at correctly identifying the negative results



specificity = TN / FP+TN (also TP / TP+FP)

56 / 70 = **80 %**

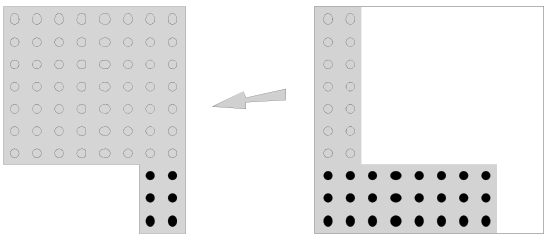
Positive predictive value
chance that a positive test result will be correct



PPV = TP / TP+FP

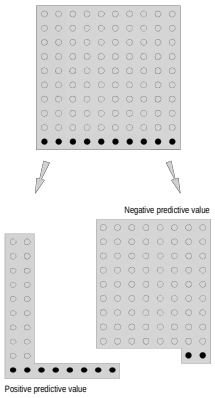
24 / 38 = **63 %**

Negative predictive value
chance that a negative test result will be correct



NPV = TN / TN+FN

56 / 62 = **90 %**



Results of testing population with disease prevalence of 10 %

		Condition (as determined by "Gold standard")		
		True	*False*	
Test outcome	*Positive*	True Positive	False Positive (Type I error, P-value)	→ Positive predictive value
	Negative	False Negative (Type II error)	True Negative	→ Negative predictive value
		↓ Sensitivity	↓ Specificity	

False positive rate (α) = FP / (FP + TN) = 1 - specificity

False negative rate (β) = FN / (TP + FN) = 1 - sensitivity

Power = 1 − β

http://en.wikipedia.org/wiki/Sensitivity_(tests)

		Patients with bowel cancer (as confirmed on endoscopy)		
		True	*False*	?
FOB test	*Positive*	TP = 2	**FP = 18**	= TP / (TP + FP) = 2 / (2 + 18) = 2 / 20 ≡ **10%**
	Negative	**FN = 1**	TN = 182	= TN / (TN + FN) 182 / (1 + 182) = 182 / 183 ≡ **99.5%**
		↓ = TP / (TP + FN) = 2 / (2 + 1) = 2 / 3 ≡ **66.67%**	↓ = TN / (FP + TN) = 182 / (18 + 182) = 182 / 200 ≡ **91%**	

Fecal occult blood (FOB) screen test

(in 203 people - to look for bowel cancer)

http://en.wikipedia.org/wiki/Sensitivity_(tests)

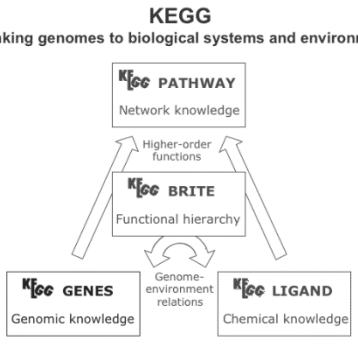
SYLABUS PREDMETU BIOINFORMATIKA PRE 3. ROČNÍK BIOLOGICKÝCH ODBOROV

1. **Úvod do Bioinformatiky**
   definícia - história -náplň - internet - vzťah k ostatným vedným odborom
2. **Biologické databázy**
   biologické dáta - iné dáta využívané v biológii - rozdelenie biologických databáz - princíp práce s databázami
3. **Primárne databázy**
   typy primárnych sekvencií – EMBL/GenBank/DDBJ – UniProt – GO – vkladanie dát – využitie
4. **Sekundárne databázy**
   proteínové motívy – PROSITE – PRINTS – Pfam – BLOCKS – INTERPRO
5. **Ďalšie biologické databázy a integrované databázové systémy**
   PDB – KEGG – OMIM – REBASE – bibliografické dáta – MEDLINE – integrované databázové systémy – SRS – Entrez
6. **Analýza biologických dát**
   zhromažďovanie a analýza biologických dát - sekvenačné projekty - štatistická analýza - používaná výpočtová technika - Staden Package - EMBOSS
7. **Identifikácia kódujúcich úsekov nukleových kyselín**
   signály – motívy – kódujúce úseky – prokaryoty vs. eukaryoty
8. **Zoradenia dvoch sekvencií**
   pairwise alignment – dot plot – substitučné matice – lokálne a globálne zoradenia – BLAST – FASTA – Needleman-Wunsch – Smith-Waterman
9. **Zoradenia viacerých sekvencií**
   multiple sequence alignment – dynamické programovanie – progresívne metódy – konsenzus sekvencia – ClustalW
10. **Identifikácia proteínových motívov**
    proteínové motívy sekundárnych databáz – neurónové siete – ScanProsite – Pscan – HmmPfam
11. **Molekulárna fylogenetická analýza**
    bioinformatika a evolúcia – fylogenetické stromy – *distance* metódy – *maximum likelihood* metódy – *parsimony* metódy – PHYLIP
12. **Sekundárna a terciárna štruktúra biomakromulekúl**
    primárna, sekundárna a terciárna štruktúra – vzťah štruktúry a funkcie – 3D vizualizácia – RasMol – MOLMOL

Databázy makromolekulárnych štruktúr

- **PDB** (Protein Data Bank)
- archív experimentálne určených trojrozmerných štruktúr biologických makromolekúl
  - atómové koordináty
  - bibliografické údaje
  - informácie o primárnej a sekundárnej štruktúre
  - faktory kryštalografických štruktúr
  - experimentálne NMR dáta.
- *PDB* v súčasnosti obsahuje **210 342** položiek (október 2023)

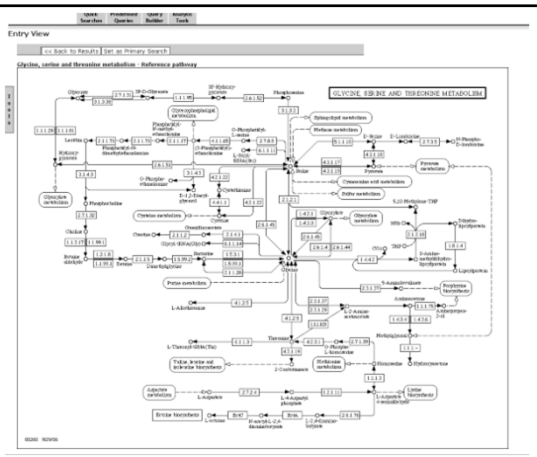
experimentálna metóda	proteíny, peptidy	protein/ oligosacharid	protein/NA	nukleové kyseliny	oligosacharidy	iné	celkovo
X-ray	158 259	9 239	8 270	2 727	11	164	178 670
NMR	12 276	34	283	1 467	6	32	14 098
elektrónová mikroskopia	11 477	2 017	3 622	109	0	9	17 234
hybridné	197	8	7	13	1	0	226
Neutron	73	1	0	3	0	0	77
iné	32	0	0	1	4	0	37
teoretické modelovanie	0	0	0	0	0	0	0
celkovo	182 314	11 299	12 182	4 320	22	205	210 342

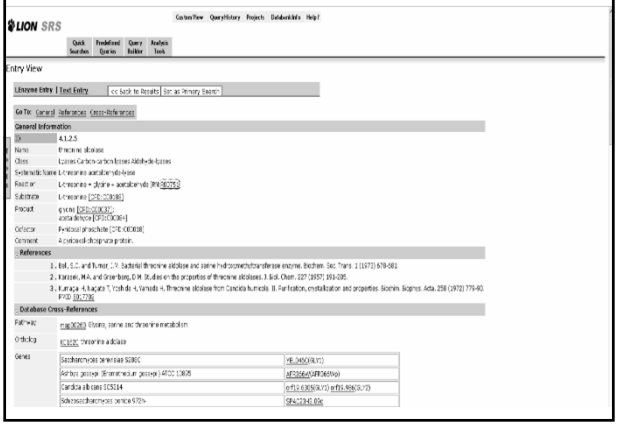


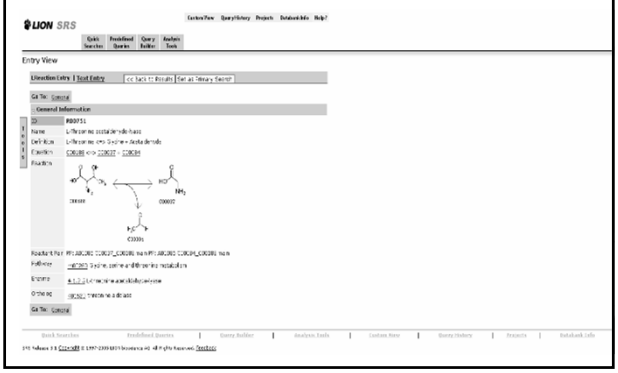
KEGG

Kyoto Encyclopedia of Genes and Genomes

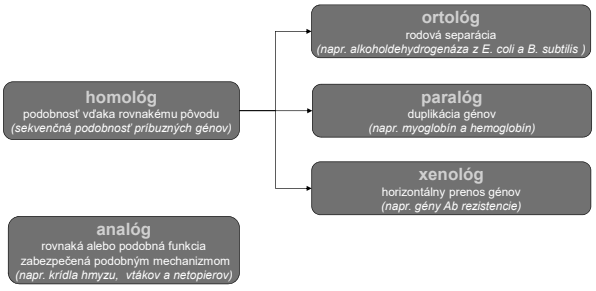
Pathway information	KEGG PATHWAY	1 094 015 pathways generated from 538 reference pathways
Binary relations and hierarchies	KEGG BRITE BRITE KO	365 533 hierarchies generated from 192 reference hierarchies
Genomic information	KEGG GENES	26 269 KO groups 49 666 637 genes in 9 370 organisms
Chemical information	KEGG LIGAND COMPOUND DRUG GLYCAN LIGAND REACTION ENZYME	19 136 compounds 12 246 drugs 11 222 glycans 11 965 reactions 8 077 enzyme nomenclature
Health information	KEGG DISEASE KEGG MEDICUS KEGG DGROUP KEGG ENVIRON	1 212 variantov ľudských génov 2 653 ľudských ochorení 33 086 FDA liečiv na predpis







Homológia a iné...



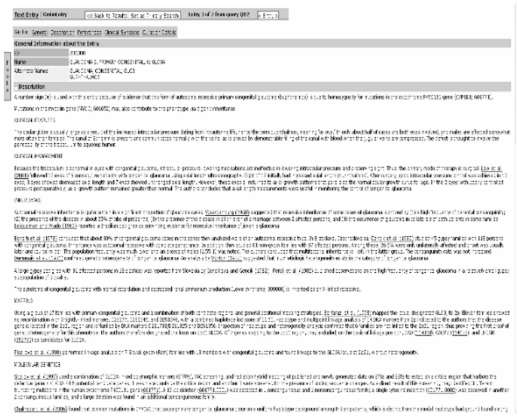
OMIM

Online Mendelian Inheritance in Man

- katalóg ľudských génov a genetických porúch
- hlavne dedičné ochorenia
- „fylogenetický dodatok“ projektu genómu človeka
- textové informácie, referencie, linky

október 2023

autozomálne	viazané na X	viazané na Y	Mitochondr.	SOLU
25 598	1 361	63	71	**27 093**



Bibliografické databázy

- **WoS, Scopus**
- **MEDLINE**
  - (MEDlars on LINE) (Schuler *et al.*, 1996, NLM)
- údaje z oblastí
  - medicíny
  - opatrovateľstva
  - zubného lekárstva
  - veterinárneho lekárstva
  - zdravotnej starostlivosti
  - preklinických vied
- bibliografické citácie a abstrakty
  - z 5 600 biomedicínskych časopisov
  - z vyše 80 krajín sveta, v 60 rôznych jazykoch (93% v angličtine)
- celkový počet záznamov prevyšuje 22 miliónov a siaha až do roku 1946
- **MeSH** - Medical Subject Headings

MeSH Tree Structures - 2008

Return to Entry Page

1. ⊞ Anatomy [A]
2. ⊞ Organisms [B]
3. ⊟ Diseases [C]
   - Bacterial Infections and Mycoses [C01] +
   - Virus Diseases [C02] +
   - Parasitic Diseases [C03] +
   - Neoplasms [C04] +
   - Musculoskeletal Diseases [C05] +
   - Digestive System Diseases [C06] +
   - Stomatognathic Diseases [C07] +
   - Respiratory Tract Diseases [C08] +
   - Otorhinolaryngologic Diseases [C09] +
   - Nervous System Diseases [C10] +
   - Eye Diseases [C11] +
   - Male Urogenital Diseases [C12] +
   - Female Urogenital Diseases and Pregnancy Complications [C13] +
   - Cardiovascular Diseases [C14] +
   - Hemic and Lymphatic Diseases [C15] +
   - Congenital, Hereditary, and Neonatal Diseases and Abnormalities [C16] +
   - Skin and Connective Tissue Diseases [C17] +
   - Nutritional and Metabolic Diseases [C18] +
   - Endocrine System Diseases [C19] +
   - Immune System Diseases [C20] +
   - Disorders of Environmental Origin [C21] +
   - Animal Diseases [C22] +
   - Pathological Conditions, Signs and Symptoms [C23] +
4. ⊞ Chemicals and Drugs [D]
5. ⊞ Analytical, Diagnostic and Therapeutic Techniques and Equipment [E]
6. ⊞ Psychiatry and Psychology [F]
7. ⊞ Biological Sciences [G]
8. ⊞ Natural Sciences [H]
9. ⊞ Anthropology, Education, Sociology and Social Phenomena [I]
10. ⊞ Technology, Industry, Agriculture [J]
11. ⊞ Humanities [K]
12. ⊞ Information Science [L]
13. ⊞ Named Groups [M]
14. ⊞ Health Care [N]
15. ⊞ Publication Characteristics [V]
16. ⊞ Geographicals [Z]

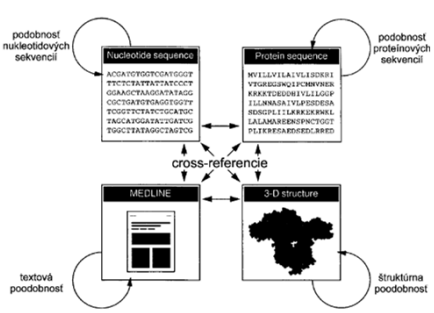
Integrované databázové systémy

- **SRS** (Biowisdom)
  - škálovateľný a integrovaný databázový systém

  http://www.biowisdom.com/
- **Entrez** (NIH, NLM)
  - vyhľadávací systém údajov prírodných vied

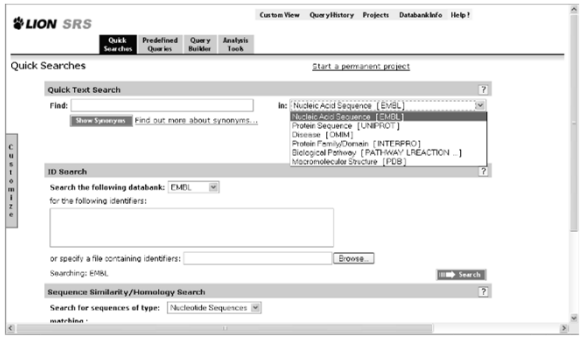
  http://www.ncbi.nlm.nih.gov/Entrez/

ENTREZ



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

SRS 8.1



SYLABUS PREDMETU BIOINFORMATIKA PRE 3. ROČNÍK BIOLOGICKÝCH ODBOROV

1. Úvod do Bioinformatiky
definícia - história -náplň - internet - vzťah k ostatným vedným odborom

2. Biologické databázy
biologické dáta - iné dáta využívané v biológii - rozdelenie biologických databáz - princíp práce s databázami

3. Primárne databázy
typy primárnych sekvencií – EMBL/GenBank/DDBJ – UniProt – GO – vkladanie dát – využitie

4. Sekundárne databázy
proteínové motívy – PROSITE – PRINTS – Pfam – BLOCKS – INTERPRO

5. Ďalšie biologické databázy a integrované databázové systémy
PDB – KEGG – OMIM – REBASE – bibliografické dáta – MEDLINE – integrované databázové systémy – SRS – Entrez

6. Analýza biologických dát
zhromažďovanie a analýza biologických dát - sekvenačné projekty - štatistická analýza - používaná výpočtová technika - Staden Package - EMBOSS

7. Identifikácia kódujúcich úsekov nukleových kyselín
signály – motívy – kódujúce úseky – prokaryoty vs. eukaryoty

8. Zoradenia dvoch sekvencií
pairwise alignment – dot plot – substitučné matice – lokálne a globálne zoradenia – BLAST – FASTA – Needleman-Wunsch – Smith-Waterman

9. Zoradenia viacerých sekvencií
multiple sequence alignment – dynamické programovanie – progresívne metódy – konsenzus sekvencia – ClustalW

10. Identifikácia proteínových motívov
proteínové motívy sekundárnych databáz – neurónové siete – ScanProsite – Pscan – HmmPfam

11. Molekulárna fylogenetická analýza
bioinformatika a evolúcia – fylogenetické stromy – *distance* metódy – *maximum likelihood* metódy – *parsimony* metódy – PHYLIP

12. Sekundárna a terciárna štruktúra biomakromulekúl
primárna, sekundárna a terciárna štruktúra – vzťah štruktúry a funkcie – 3D vizualizácia – RasMol – MOLMOL

Workflow výskumu v molekulárnej biológii

DNA sekvenovanie
↓
anotácia genómov — genomnika: charakterizácia genómu
↓
štúdium funkcie a štruktúry proteínov — proteomika: charakterizácia proteómu
↓
štúdium regulácie a expresie génov — systémová biológia: *vzťahy a interakcie*

Systémová biológia

high-throughput metódy
(*DNA arrays, NGS*)

↓ matematická teóriu hier
(jednotky—procesy—interakcie)

matematický model biologického systému

↓ predpoveď zmien
(simulované podnety)

experimentálne overenie

Analýza biologických dát

- Zhromažďovanie dát
  - NA sekvencie zo sekvenačných projektov
  - AA sekvencie odvodené z NA sekvencií alebo získané priamo sekvenovaním proteínov
- Sekvenačné projekty
  - kompletizácia pomocou *Sequence Assembly Programs*

Software, Hardware

- **OS MS Windows**
  - OMIGA, PCGene, DNAsis, Clone, Vector NTI
  - prevažne lokálna inštalácia
- **OS UNIX** (Linux, Solaris, AIX, IRIX)
  - GCG Wisconsin Package Software
  - Staden Package
  - EMBOSS
  - Galaxy
  - Chipster
  - prevažne architektúra server–klient
    (využíva internet a web prehliadače)

EMBOSS

- Open Source software
- využitie
  - zoradenie sekvencií
  - rýchle prehľadávanie databáz so sekvenčnými vzormi
  - identifikácia proteínových motívov a analýza domén
  - analýza nukleotidových sekvenčných vzorov (napr. identifikácia CpG ostrovčekov alebo repeatov)
  - analýza využitia kodónov malých genómov
  - rýchla identifikácia sekvenčných vzorov medzí veľkým setom sekvencií
  - prezentačné a publikačné programy
  - a iné...

Formáty sekvencií

- *plain text* (*raw format*, čistý text)

```
CGCTTTGAACGATGATGAT
GATTAGCAGTACAGAGTAC
AGTCTGCTA
```

- FASTA

```
>pLK18
CGCTTTGAACGATGATGAT
GATTAGCAGTACAGAGTAC
AGTCTGCTA
```

- EMBL
- GenBank

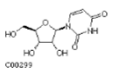
Amino acid codes

A --> alanine
B --> aspartate or asparagine
C --> cystine
D --> aspartate
E --> glutamate
F --> phenylalanine
G --> glycine
H --> histidine
I --> isoleucine
K --> lysine
L --> leucine
M --> methionine
N --> asparagine
P --> proline
Q --> glutamine
R --> arginine
S --> serine
T --> threonine
U --> selenocysteine
V --> valine
W --> tryptophan
Z --> glutamate or glutamine
X --> any
***** --> translation stop
- --> gap of indeterminate length

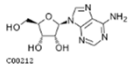
Nucleic acid codes

A --> adenosine
C --> cytidine
G --> guanine
T --> thymidine
U --> uridine
R --> G A (purine)
Y --> T C (pyrimidine)
K --> G T (keto)
- gap of indeterminate length

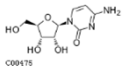
M --> A C (amino)
S --> G C (strong)
W --> A T (weak)
B --> G T C
D --> G A T
H --> A C T
V --> G C A
N --> A G C T (any)



Uridine

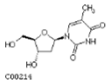


Adenosine



Cytidine

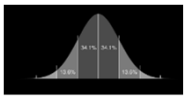
Guanine

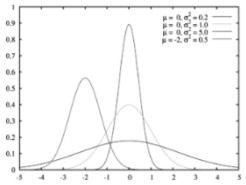


Thymidine

Štatistické termíny

- pravdepodobnosť (P)
  - podmnožina možností všetkých možností
- aritmetický priemer (*mean*, *μ*)
- modus (najpravdepodobnejšia hodnota)
  - to x, ktoré ma najväčšie P
- medián
  - rozdeľuje výskyty na 2 rovnaké časti, pri párnom počte ide o priemer dvoch stredných hodnôt
- variancia (disperzia)
  - definuje škálu hodnôt
- štandardná odchýlka (SD, σ)
  - odchýlka hodnôt od aritmetického priemeru
  - druhá odmocnina variancie
- normálna (Gaussova) distribúcia
  - štandardná normálna distribúcia
  - priemer = 0
  - variancia = 1



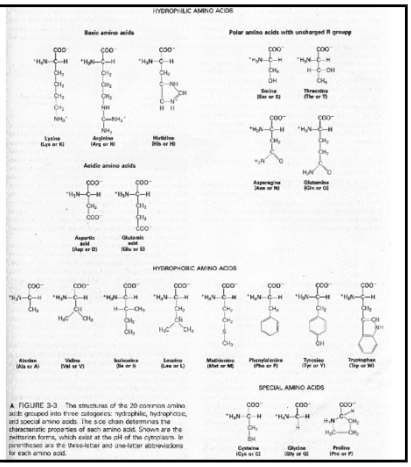


- http://www.explorelearning.com/index.cfm?method=cResource.dspView&ResourceID=262

Štatistická analýza

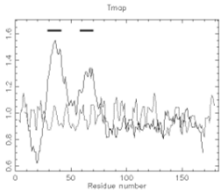
NA	AA
obsah nukleotidov	obsah aminokyselín
obsah GC párov, (G+C)3	Mw
teplota topenia 2(A+T)+4(G+C)	hydrofobicita, náboj
využitie kodónov	využitie AAs

Skupiny aminokyselín



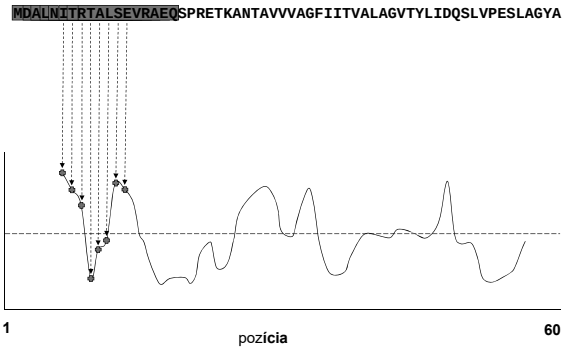
Hydrofobicita proteínov

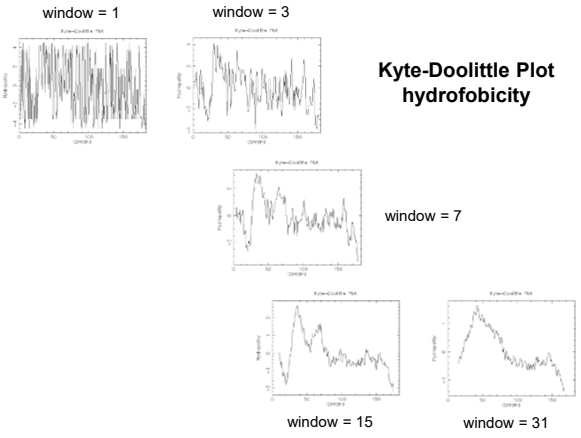
- založená na chemických vlastnostiach aminokyselín
- základ predpovede transmembránových úsekov proteínov



- výrazne hydrofóbne
  - **I** (Ile)
  - **F** (Phe)
  - **L** (Leu)
  - **V** (Val)
  - **M** (Met)
- výrazne hydrofilné
  - **N** (Asn)
  - **Q** (Gln)
  - **E** (Glu)
  - **D** (Asp)
  - **K** (Lys)

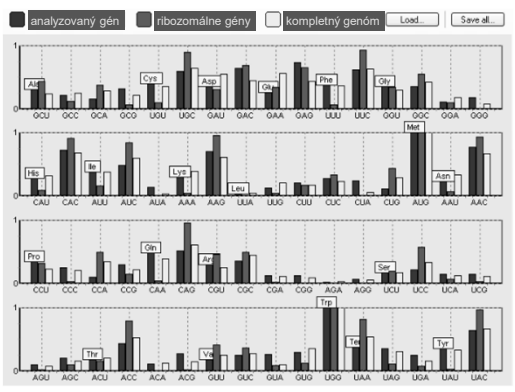
Analýza sekvencií pomocou posúvajúceho sa okna



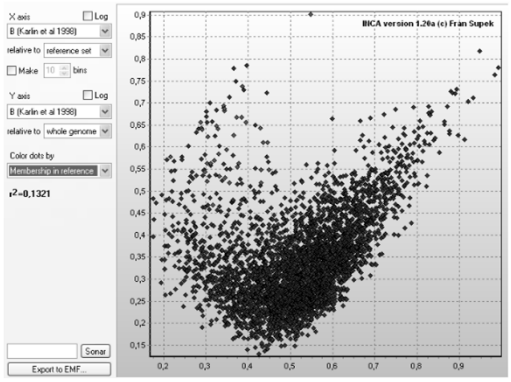


Kyte-Doolittle Plot hydrofobicity

Predikcia vysokoexprimovaných génov



Využitie kodónov – vysoko exprimované gény



Codon bias

- We show that the buffering capacity of coding sequences is in general higher than that of randomly generated sequences and that of shifted reading frames. Highly expressed genes are shown to have an even higher buffering capacity than non-housekeeping genes.

(Rouchka 2008)

SYLABUS PREDMETU BIOINFORMATIKA PRE 3. ROČNÍK BIOLOGICKÝCH ODBOROV

1. Úvod do Bioinformatiky
definícia - história -náplň - internet - vzťah k ostatným vedným odborom

2. Biologické databázy
biologické dáta - iné dáta využívané v biológii - rozdelenie biologických databáz - princíp práce s databázami

3. Primárne databázy
typy primárnych sekvencií – EMBL/GenBank/DDBJ – UniProt – GO – vkladanie dát – využitie

4. Sekundárne databázy
proteínové motívy – PROSITE – PRINTS – Pfam – BLOCKS – INTERPRO

5. Ďalšie biologické databázy a integrované databázové systémy
PDB – KEGG – OMIM – REBASE – bibliografické dáta – MEDLINE – integrované databázové systémy – SRS – Entrez

6. Analýza biologických dát
zhromažďovanie a analýza biologických dát - sekvenačné projekty - štatistická analýza - používaná výpočtová technika - Staden Package - EMBOSS

7. Identifikácia kódujúcich úsekov nukleových kyselín
signály – motívy – kódujúce úseky – prokaryoty vs. eukaryoty

8. Zoradenia dvoch sekvencií
pairwise alignment – *dot plot* – substitučné matice – lokálne a globálne zoradenia – BLAST – FASTA – Needleman-Wunsch – Smith-Waterman

9. Zoradenia viacerých sekvencií
multiple sequence alignment – dynamické programovanie – progresívne metódy – konsenzus sekvencia – ClustalW

10. Identifikácia proteínových motívov
proteínové motívy sekundárnych databáz – neurónové siete – ScanProsite – Pscan – HmmPfam

11. Molekulárna fylogenetická analýza
bioinformatika a evolúcia – fylogenetické stromy – *distance* metódy – *maximum likelihood* metódy – *parsimony* metódy – PHYLIP

12. Sekundárna a terciárna štruktúra biomakromulekúl
primárna, sekundárna a terciárna štruktúra – vzťah štruktúry a funkcie – 3D vizualizácia – RasMol – MOLMOL

Sequence Representations

- **Consensus**
- **Alignment**
- **Blocks or Weight Matrices**
- **Templates or Profiles**
- **Bayesian Networks**
- **Hidden Markov Models**

- **Deterministic**
- **Probabilistic**

Konsenzus vs. matice

TACGAT
TATAAT
TATAAT
GATACT
TATGAT
TATGTT

TATAAT konsenzus
TATRNT (alt. konsenzus)

-10 regions of 6 promoters (Pribinow,1975)

presná zhoda
iba 2 zo 6 motívov (4 zo 6)
výskyt každých 4000 bp (500 bp)

1 nezhoda
3 zo 6 motívov (všetkých 6)
výskyt každých 200 bp (30 bp)

2 nezhody
všetkých 6 motívov
výskyt každých 30 bp

Konsenzus vs. matice

matica zoradenia

A	0	6	0	3	4	0
C	0	0	1	0	1	0
G	1	0	0	3	0	0
T	5	0	5	0	1	6

frekvenčná matica

A	0	1	0	0,5	0,67	0
C	0	0	0,17	0	0,17	0
G	0,17	0	0	0,5	0	0
T	0,83	0	0,83	0	0,17	1

matica váh (PWM, PSWM)

A	-1,95	1,10	-1,95	0,45	0.72	-1,95
C	-1,95	-1.95	-0,15	-1,95	-0,15	-1,95
G	-0,15	-1.95	-1,95	0.83	-1,95	-1,95
T	0,93	-1,95	0,93	-1.95	-0,48	1,10

TACGAT
TATAAT
TATAAT
GATACT
TATGAT
TATGTT

sekvenčné logo

(Schneider and Stephens, 1990)

$$\ln\frac{(n_{i,j}+p_i)/(N+1)}{p_i} \approx \ln\frac{f_{i,j}}{p_i}$$

Využitie PSWM

T C G A T A T A C T G

Využitie PSWM

T C G A T A T A C T G

A	-1,95	1,10	-1,95	**0,45**	0.72	**-1.95**				
C	-1,95	**-1.95**	-0,15	-1,95	-0,15	-1,95				
G	-0,15	-1.95	**-1,95**	0.83	-1,95	-1,95				
T	**0,93**	-1,95	0,93	-1.95	**-0.48**	1,10				

+0,93
-1,95
-1,95
+0,45
-0,48
-1,95
==========
-4,95

Využitie PSWM

T C G A T A T A C T G

A	-1,95	1,10	**-1.95**	0,45	**0.72**	-1,95				
C	**-1.95**	-1.95	-0,15	-1,95	-0,15	-1,95				
G	-0,15	**-1.95**	-1,95	0.83	-1,95	-1,95				
T	0,93	-1,95	0,93	**-1.95**	-0,48	**1,10**				

-1,95
-1,95
-1,95
-1,95
+0,72
+1,10
==========
-5,98

Využitie PSWM

T C G A T A T A C T G

A	-1,95	**1,10**	-1,95	**0,45**	0.72	**-1.95**				
C	-1,95	-1.95	-0,15	-1,95	-0,15	-1,95				
G	**-0,15**	-1.95	-1,95	0.83	-1,95	-1,95				
T	0,93	-1,95	**0,93**	-1.95	**-0.48**	1,10				

-0,15
+1,10
+0,93
+0,45
+0,72
+1,10
==========
+4,15

Využitie PSWM

T C G A T A T A C T G

A	**-1.95**	1,10	**-1.95**	0,45	**0.72**	-1,95				
C	-1,95	-1.95	-0,15	-1,95	-0,15	**-1.95**				
G	-0,15	-1.95	-1,95	0.83	-1,95	-1,95				
T	0,93	**-1.95**	0,93	**-1.95**	-0,48	1,10				

-1,95
-1,95
-1,95
-1,95
+0,72
-1,95
==========
-9,03

Využitie PSWM

T C G A T A T A C T G

A	-1,95	**1,10**	-1,95	**0,45**	0.72	-1,95				
C	-1,95	-1.95	-0,15	-1,95	-0,15	-1,95				
G	-0,15	-1.95	-1,95	0.83	**-1.95**	-1,95				
T	**0,93**	-1,95	**0,93**	-1.95	-0,48	**1,10**				

+0,93
+1,10
+0,93
+0,45
-0,15
+1,10
==========
+4,36

Využitie PSWM

T	C	G	A	T	A	T	A	C	T	G

A	-1,95	1,10	-1,95	0,45	0.72	-1,95
C	-1,95	-1.95	-0,15	-1,95	-0,15	-1,95
G	-0,15	-1.95	-1,95	0.83	-1,95	-1,95
T	0,93	-1,95	0,93	-1.95	-0,48	1,10

```
-1,95
      -1,95
            -1,95
                  -1,95
                        -0,48
                              -1,95
================================
                  -10,23
```

Využitie PSWM

T	C	G	A	T	A	T	A	C	T	G

-4,95 -5,98 +4,15 -9,03 +4,36 -10,23

TACGAT
TATAAT
TATAAT
GATACT
TATGAT
TATGTT

Využitie PSWM

T	C	G	A	T	A	T	A	C	T	G

-4,95 -5,98 +4,15 -9,03 +4,36 -10,23

TACGAT
TATAAT
TATAAT
GATACT
TATGAT
TATGTT

Matica váh

(*Positional weight matrix, PWM, position specific weight matrix, PSWM*)

Pozícia / Nukleotid	1	2	3	4	5	6	7	8	9	10
A	0,22	-1,39	-1,39	0,22	-1,39	0,22	-1,39	-1,39	-1,39	1,18
C	-1,39	0,22	0,22	-1,39	-1,39	-1,39	-1,39	-1,39	-1,39	-1,39
G	0,22	0,22	0,81	0,81	-1,39	0,81	-1,39	-1,39	1,18	-1,39
T	0,22	0,22	-1,39	-1,39	1,18	-1,39	1,18	1,18	-1,39	-1,39

$$\mathbf{weight}_{i,j} = \ln \frac{(n_{i,j}+p_i)/(N+1)}{p_i} \sim \ln \frac{f_{i,j}}{p_i}$$

(Storma a Hertza, 1995)

N – počet sekvencii, ktoré sa porovnávajú
$n_{i,j}$ – početnosť nukleotidu *i* na pozícii *j*
p_i – pravdepodobnosť výskytu nukleotidu *i* v celom genóme
$f_{i,j} = n_{i,j}/N$ frekvencia nukleotidu i na pozícii *j*

TGCGTGTTGACTATTTT
ATGATATTGACTTATTG
GCGGTGTTGACATAAAT

sekvencie -35 úsekov promótorov

Pozícia / nukleotid	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
A	1	0	0	1	0	1	0	0	0	3	0	1	1	2	1	1	0
C	0	1	1	0	0	0	0	0	0	0	3	0	0	0	0	0	0
G	1	1	2	2	0	2	0	0	3	0	0	0	0	0	0	0	1
T	1	1	0	0	3	0	3	3	0	0	0	2	2	1	2	2	2

matica zoradení

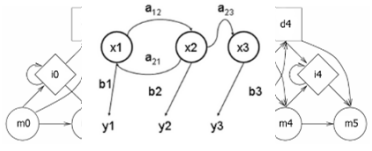
Pozícia / Nukleotid	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
A	0,33	0	0	0,33	0	0,33	0	0	0	1	0	0,33	0,33	0,67	0,33
C	0	0,33	0,33	0	0	0	0	0	0	0	1	0	0	0	0
G	0,33	0,33	0,67	0,67	0	0,67	0	0	1	0	0	0	0	0	0
T	0,33	0,33	0	0	1	0	1	1	0	0	0	0,67	0,67	0,33	0,67

matica frekvencií

sekvenčné logo

Bayesovské siete, HMM

- **Bayesovské siete** (umelá inteligencia)
  - grafický pravdepodobnostný model
  - popisuje rozdelenie pravdepodobností spoločného výskytu udalostí
- **Markovov model**
  - štatistický model série stavov
  - každý stav záleží iba na predchádzajúcom stave
- **HMM**
  - pravdepodobnostný model založený na Markovovom modely s neznámymi parametrami
- cieľom je určiť skrytý parameter na základe pozorovateľného parametru



číslo	pozícia
m	match
I	insert
d	deletion

Westhead et al., 2002

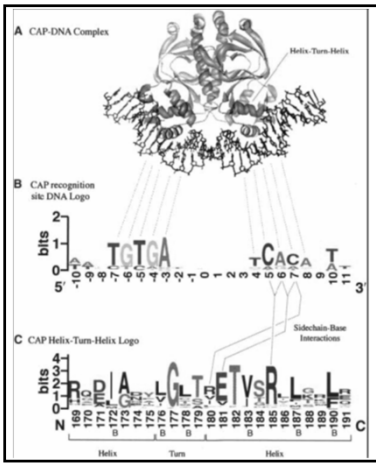
Pravdepodobnostná tabuľka výskytu báz v jednotlivých pozíciách pri zostrihu

frekvenčná tabuľka pre 3' koniec intrónu

P	-15	-14	-13	-12	-11	-10	-9	-8	-7	-6	-5	-4	-3	-2	-1	1	2	3
N	113	113	114	126	126	126	127	127	127	129	130	130	130	130	130	130	130	130
T	58	50	57	67	75	62	62	57	57	73	75	38	40	0	0	11	48	37
C	21	28	35	27	30	38	42	35	46	46	36	28	84	0	0	23	28	42
A	17	11	11	19	8	19	14	24	15	4	13	33	5	130	0	29	22	25
G	17	24	11	13	13	7	9	11	9	6	6	31	1	0	130	67	32	26

frekvenčná tabuľka pre 5' koniec intrónu

P	-4	-3	-2	-1	1	2	3	4	5	6	7	8
N	139	139	139	139	139	139	139	139	139	139	136	136
T	28	10	18	17	0	139	9	16	7	87	30	36
C	42	60	16	8	0	0	3	13	3	17	28	40
A	42	56	89	12	0	0	86	94	12	23	53	33
G	27	13	16	102	139	0	41	16	117	12	25	27



A. CAP (Catabolite Activator Protein, also known as CRP) acts as a transcription promoter by binding at more than 100 sites within the Escherichia coli genome. (PDB structure 1CGP, Schultz et al. 1991)

B. The two DNA recognition helices of the CAP homodimer insert themselves into consecutive turns of the major groove. The logo is approximately palindromic.

C. The helix-turn-helix motif from the CAP family of homodimeric DNA binding proteins

(Crooks GE, Hon G, Chandonia JM, Brenner SE WebLogo: A sequence logo generator, Genome Research, 14:1188-1190, 2004)

Identifikácia kódujúcich úsekov nukleových kyselín

Identifikácia signálov a motívov

- **konsenzus sekvencie**
  - ŠTART a STOP kodóny
  - miesta rozpoznávané restrikčnými endonukleázami
  - polyA signálne miesta (AATAAA)
- ***weight matrix***
  - RBS
  - Promótory (-35, -10, +1 miesta)
  - *Splice junctions* (miesta zostrihu)

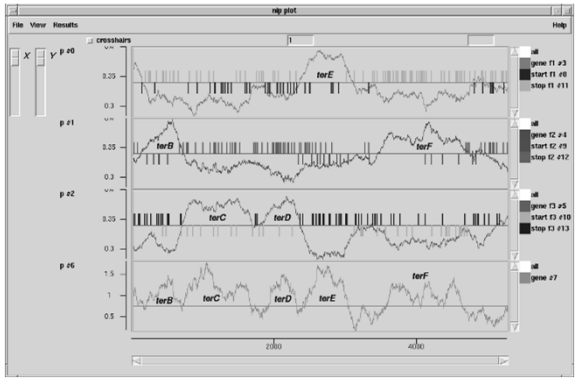
Identifikácia kódujúcich úsekov nukleotidových sekvencií

- rôzne kompozičné zloženie kódujúcej a nekódujúcej sekvencie DNA
- kompozícia báz, kodónov a kódovaných aminokyselín vyplýva z toho, či daná nukleotidová sekvencia je alebo nie je kódujúca
- hlavné faktory
  - **rôzne využitie aminokyselín**
    (niektoré aminokyseliny sa v proteínoch nachádzajú častejšie ako iné, napríklad alanín sa v priemere vyskytuje oveľa častejšie ako tryptofán)
  - **rôzny počet kodónov pre jednotlivé aminokyseliny**
    (napríklad leucín má 6 kodónov a tryptofán iba 1)
  - **využitie kodónov jednotlivých aminokyselín nie je rovnomerné**
    (všeobecne označované ako preferencia kodónov)

m,.dskllxcdGEs,m kfdMRRS ndfJOdfO klijk,JJIk.sdlkk0wFDisdklGyksn d*(N ikbdp4m fad. erioN lkk< ofdn 0fdnm fvi KKLhndgfpemmvdfovdgfgdgfdf. DGFsd ojIOHJofddfl.g GFD fd;v dsd.zx 0JO)Jl4397s dcnmlJo84 c0N V7 fdvd fnoHn Kdhn/sdc/fgdolsdopc 9nbkv 99Kmds edoN'edD sxsd k KKKn sk jkd L odod lfv sdlLlf psd F sdsde 9e993c o0dfnv0 (fdn03 vc 0e0 x)hjf3- c 40f OL0fd sdl34-0olnl)LL 0-rfn 340--)J clsc fsd0- 0 Hocv0-dre- cm . df rrffdldj0- 0dfv p))Oo p-df-)Nclvfdp sSDFsdsdf ds; asdsdsdsdi 9c9sd0sdj 43f0ernmcx09 09fdnol rt54-0Jdsd SD-we034e xd o09 dl,s9040odf0c cnv)oldps-04 0vj) 0df opsd0sd) dfsdl430P3349UL;2340CN V04NV0 W4NV04O N990n549n 90e 949b)(UW)(@ 0 0) 0eu3 u0)&3 0380wer 0)70 0e-r8 w 90w7 0 Voccii peruntilibem nostiemus Catum in ves hillerum aut vitilin atanducent, veris vivitis; ne essolic tre nostori ventrum patium prartiem rei censulum it vivive, ditur. Serferum publina tiamquam storum are, etimis 9 horbeffre, nese quit? qua Sertem, Cupicaudam am, vilinte atatili 43 bultusuam hos convolium aperei inatusquem de pris. Cat & Tus ignostius. Forum consuliam inclegilin nernit, ve, esta, con alium o te atquo movividiis, Ti. Is hae mod conum (ur interi) pre ero nesesulto nem ad rendaci, endesil, condiem un perceps, num@faurissentem.tum, Casdamq uamplicidi conteris condium es atisque in Itanum in Etrartuius occhuis creo, senatiam aus etiustales aucermium consupp linam. Us populi sim et; hem no. 9sd0sdj 43f0ernmcx09 09fdnol rt54-0Jdsd SD-we034e xd o09 dl,s9040odf0c cnv)oldps-04 0vj) 0df opsd0sd) dfsdl430P3349UL;2340CN V04NV0W4NV04ON990n549n 90e 949b)(UW)(@ 0 0) 0eu3 u0)&3 0380wer 0)70 0e-r8 w 90wofdn 0fdnm fviKKLhn dgfpemmvdfovdGF'dgfdfDGFsd ojIOHJofddfl.g Gdf fDv, dsdWzx 0JO)Jl4 397s. dcnmlJo84 c0N V7 fdvd fnoHn Kdhn/sdc/fgdolsdopc 9nbkv 99Kmds edoN'edD sxsd k KKKn sk jkd L odod lfv sdlLlf psd F sdsde 9e993c o0dfnv0 (fdn03

- **väčšina metód vychádza z porovnania očakávanej a zistenej distribúcie báz a kodónov**
- **Shepherdova metóda**
  - sekvencia obsahuje často kodóny v tvare RNY (R=purín, N=ľubovoľná báza, Y=pyrimidín)
  - táto metóda sa používa hlavne na identifikáciu správneho čítacieho rámca, ktorý obsahuje najviac kombinácií kodónov práve v tomto tvare
- **Metóda využitia kodónov**
  - porovnáva výskyt kombinácií jednotlivých kodónov v študovanej skevencii s očakávanými, empiricky získanými hodnotami
- **Metóda preferencie pozícií báz v kodónoch**
  - vychádza z porovnania výskytu báz na jednotlivých pozíciách kodónu a ich porovnania so štatisticky priemerným výskytom v kódujucích sekvenciách.
- **Metóda nerovnomerného výskytu báz v kodónoch**
  - prehľadáva sekvenciu a sleduje frekvenciu výskytu jednotlivých báz v rámci tripletov
  - čím je tento výskyt menej náhodný, tým je pravdepodobnejšie, že sekvencia kóduje proteín
  - na rozdiel od metódy *preferencie pozícií báz v kodónoch* neporovnáva získané údaje so štandardnými, ale vychádza iba z miery náhodnosti výskytu.
- **Fickettova metóda**
  - podobná predchádzaúcej, pričom zahŕňa i sledovanie obsahu báz v jednotlivých čítacích rámcoch, čo napomáha identifikácii správneho čítacieho rámca.

Program SPIN (Staden Package)



SYLABUS PREDMETU BIOINFORMATIKA PRE 3. ROČNÍK BIOLOGICKÝCH ODBOROV

1. **Úvod do Bioinformatiky**
   definícia - história -náplň - internet - vzťah k ostatným vedným odborom
2. **Biologické databázy**
   biologické dáta - iné dáta využívané v biológii - rozdelenie biologických databáz - princíp práce s databázami
3. **Primárne databázy**
   typy primárnych sekvencií – EMBL/GenBank/DDBJ – UniProt – GO – vkladanie dát – využitie
4. **Sekundárne databázy**
   proteínové motívy – PROSITE – PRINTS – Pfam – BLOCKS – INTERPRO
5. **Ďalšie biologické databázy a integrované databázové systémy**
   PDB – KEGG – OMIM – REBASE – bibliografické dáta – MEDLINE – integrované databázové systémy – SRS – Entrez
6. **Analýza biologických dát**
   zhromažďovanie a analýza biologických dát - sekvenačné projekty - štatistická analýza - používaná výpočtová technika - Staden Package - EMBOSS
7. **Identifikácia kódujúcich úsekov nukleových kyselín**
   signály – motívy – kódujúce úseky – prokaryoty vs. eukaryoty
8. **Zoradenia dvoch sekvencií**
   pairwise alignment – dot plot – substitučné matice – lokálne a globálne zoradenia – BLAST – FASTA – Needleman-Wunsch – Smith-Waterman
9. **Zoradenia viacerých sekvencií**
   multiple sequence alignment – dynamické programovanie – progresívne metódy – konsenzus sekvencia – ClustalW
10. **Identifikácia proteínových motívov**
    proteínové motívy sekundárnych databáz – neurónové siete – ScanProsite – Pscan – HmmPfam
11. **Molekulárna fylogenetická analýza**
    bioinformatika a evolúcia – fylogenetické stromy – *distance* metódy – *maximum likelihood* metódy – *parsimony* metódy – PHYLIP
12. **Sekundárna a terciárna štruktúra biomakromulekúl**
    primárna, sekundárna a terciárna štruktúra – vzťah štruktúry a funkcie – 3D vizualizácia – RasMol – MOLMOL

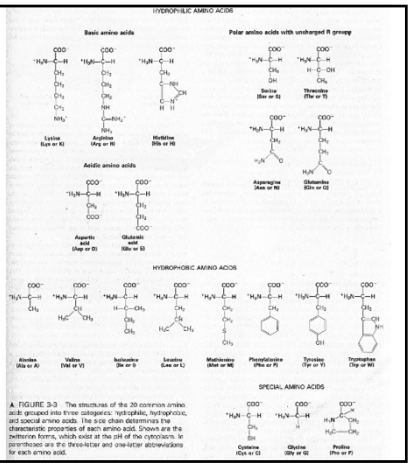
Zoradenie sekvencií

- typ
  - nukleotidové
  - proteínové
- počet
  - *pairwise* (2 sekvencie)
  - *multiple* (3 a viac)
- rozsah
  - globálne
  - lokálne

Detekcia sekvenčných homológií

- NA sekvencie vs. AA sekvencie
  - 4 nukleotidy vs. 20 aminokyselín
- Substitučné matice
  - **PAM** (Percent accepted mutation)
    (Dayhoff, 1987)
    PAM 120, PAM 200, PAM 250
  - **BLOSUM** (BLOks SUbstitution Matrix)
    (Henikoff a Henikoff, 1992)
    BLOSUM 45, BLOSUM 50, BLOSUM 62

Skupiny aminokyselín



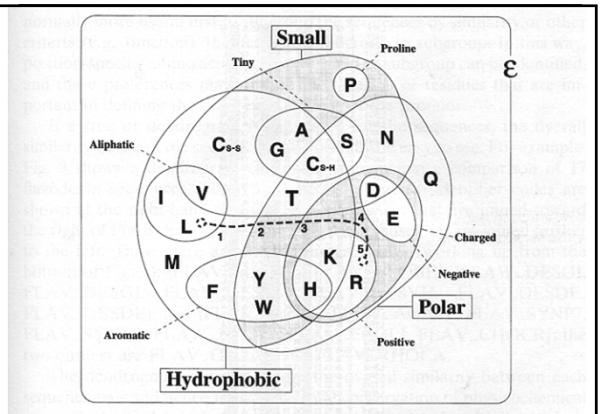


FIG. 1. The 20 common amino acids are shown with ten physicochemical properties.[5,13] Gray-filled areas define sets of properties possessed by none of the amino acids.

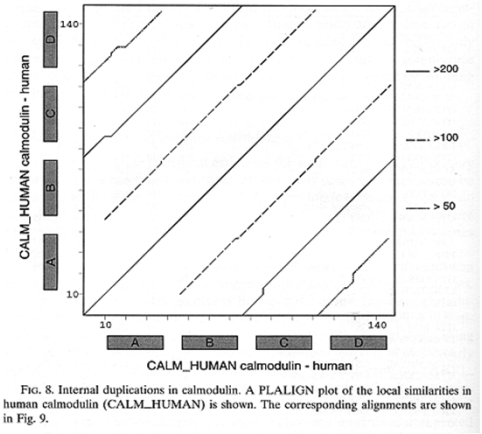


FIG. 8. Internal duplications in calmodulin. A PLALIGN plot of the local similarities in human calmodulin (CALM_HUMAN) is shown. The corresponding alignments are shown in Fig. 9.

Zoradenie dvoch sekvencií

(pairwise alignment)

```
Hemoglobin   1 MVHLTPEEKSAVTALWGKVNVD--EVGGEALGRLLVVYPWTQRLFESFGD   48
               .|:..|...|..:||||..|  ..|.|.|.||...:|.|...|:.|..
Myoglobin    1   GLSDGEWQLVLNVWGKVEADIPGHGQEVLIRLFKGHPETLEKFDKFKH   48

Hemoglobin  49 LFTPDAVMGNPKVKAHGKKVLGAFSDGPAHLDNLKGTFATLSELHCDKLH   98
               |.:.|.:.:.::|.||..||.|.........:::.....|::.|..|..
Myoglobin   49 LKSEDEMKASEDLKKHGATVLTALGGILKKKGHHEAEIKPLAQSHATKHK   98

Hemoglobin  99 VDPENFRLLGNVLVCVLAHHFGKEFTPPVQAAYQKVVAGVANALAHKYH   147
               :..:....:...::.||......:|....|.|..|.:.......:|..|.
Myoglobin   99 IPVKYLEFISECIIQVLQSKHPGDFGADAQGAMNKALELFRKDMASNYKE  148

Hemoglobin 148        147
Myoglobin  149 LGFQG   153
```

Lokálne a globálne zoradenia dvoch sekvencií

- **Globálne zoradenia**
  - Needelmana a Wunscha (1970)
  - prvý algoritmus dynamického programovania
- **Lokálne zoradenia**
  - Smith a Waterman (1981)
  - Smith-Watermanov algoritmus

BLAST

Basic Local Alignment Search Tool

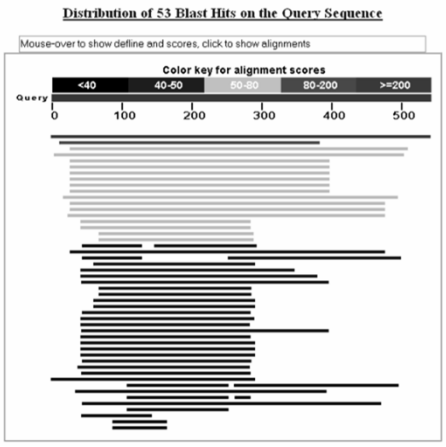
(Altschul, 1990)

BLASTP	porovnáva proteínovú sekvenciu oproti proteínovej databanke
BLASTN	porovnáva nukleotidovú sekvenciu oproti nukleotidovej databanke
BLASTX	porovnáva nukleotidovú sekvenciu preloženú vo všetkých 6 čítacích rámcoch oproti proteínovej databanke
TBLSATN	porovnáva proteínovú sekvenciu oproti nukleotidovej databanke preloženej vo všetkých 6 čítacích rámcoch
TBLASTX	porovnáva nukleotidovú sekvenciu preloženú vo všetkých 6 čítacích rámcoch oproti nukleotidovej databanke preloženej vo všetkých 6 čítacích rámcoch

FASTA

(Pearson, 1988)

FASTA	univerzálny program na porovnanie nukleotidovej sekvencie s nukleotidovou databázou alebo proteínovej sekvencie s proteínovou databázou
FASTX	porovnáva nukleotidovú sekvenciu preloženú vo všetkých 6 čítacích rámcoch s proteínovou databázou
TFASTA	porovnáva proteínovú sekvenciu s nukleotidovou databázou preloženou vo všetkých 6 čítacích rámcoch
SSEARCH	prehľadáva databázu priamo pomocou Smith-Watermanovho algoritmu (je asi 50 x pomalší ako program FASTA)



Testovanie hypotézy

- **nulová** (H_0) a **alternatívna** (H_1) hypotéza
  - navzájom sa vylučujú
  - určujeme, ktorá je pravdivá
- **P-hodnota** (*P-value*)
  - empiricky zistená
  - dosiahnutá úroveň signifikantnosti
  - najnižšia hladina významnosti pre zamietnutie H_0
  - čím je P-hodnota nižšia, tým viac sme presvedčení, že H_0 nie je správna a mala by byt zamietnutá
- **hladina α**
  - predefinovaná hodnota chyby I. druhu, napr. α = 0,001
  - rozdeľuje výsledky na dve časti (zamietnutie/nezamietnutie H_0)
- ak P ≤ α, nulová hypotéza je zamietnutá
- **hodnota E** (expectancy, *E-value*)

$P = 1 - e^{-E}$ pre malé hodnoty P, $P \approx E$

$E = PN$ hodnota E databázového *hitu* so skóre *s*

„Proteínové“ BLAST programy

- štandardný BLAST
  - štandardný BLAST algoritmus
- PSI-BLAST (Position Specific Iterated BLAST)
  - Reťazovité vyhľadávanie, pri ktorom sa sekvencie vyhľadané v prvom behu použijú na vytvorenie modelu skórovania pre druhý beh. Vysoko konzervatívne oblasti získajú vysoké skóre, nízko konzervatívne získajú skóre blízke nule. Vytvorený profil sa použije v ďalšom (ďalších) BLAST behoch, pričom výsledok každého behu zdokonaľuje profil. Takáto reťazová stratégia zabezpečuje zvýšenú senzitivitu.
- PHI-BLAST (Pattern Hit Initiated BLAST)
  - Kombinuje vyhľadávanie sekvenčného vzoru a PSI algoritmu. PHI-BLAST môže lokalizovať ďalšie sekvencie, ktoré obsahujú sekvenčný vzor a zároveň sú homologické s porovnávanou sekvenciou.

Porovnanie algoritmov prehľadávania sekvenčných databáz

program	citlivosť	rýchlosť	sekvencie
BLAST	nízka	veľmi vysoká	DNA, proteíny
FASTA	priemerná	vysoká	DNA, proteíny
Blitz	vysoká	priemerná	DNA, proteíny
SSEARCH	vysoká	pomalá	DNA, proteíny
PSI-BLAST	veľmi vysoká	pomalá	proteíny

SYLABUS PREDMETU BIOINFORMATIKA PRE 3. ROČNÍK BIOLOGICKÝCH ODBOROV

1. Úvod do Bioinformatiky
definícia - história -náplň - internet - vzťah k ostatným vedným odborom

2. Biologické databázy
biologické dáta - iné dáta využívané v biológii - rozdelenie biologických databáz - princíp práce s databázami

3. Primárne databázy
typy primárnych sekvencií – EMBL/GenBank/DDBJ – UniProt – GO – vkladanie dát – využitie

4. Sekundárne databázy
proteínové motívy – PROSITE – PRINTS – Pfam – BLOCKS – INTERPRO

5. Ďalšie biologické databázy a integrované databázové systémy
PDB – KEGG – OMIM – REBASE – bibliografické dáta – MEDLINE – integrované databázové systémy – SRS – Entrez

6. Analýza biologických dát
zhromažďovanie a analýza biologických dát - sekvenačné projekty - štatistická analýza - používaná výpočtová technika - Staden Package - EMBOSS

7. Identifikácia kódujúcich úsekov nukleových kyselín
signály – motívy – kódujúce úseky – prokaryoty vs. eukaryoty

8. Zoradenia dvoch sekvencií
pairwise alignment – *dot plot* – substitučné matice – lokálne a globálne zoradenia – BLAST – FASTA – Needleman-Wunsch – Smith-Waterman

9. Zoradenia viacerých sekvencií
multiple sequence alignment – dynamické programovanie – progresívne metódy – konsenzus sekvencia – ClustalW

10. Identifikácia proteínových motívov
proteínové motívy sekundárnych databáz – neurónové siete – ScanProsite – Pscan – HmmPfam

11. Molekulárna fylogenetická analýza
bioinformatika a evolúcia – fylogenetické stromy – *distance* metódy – *maximum likelihood* metódy – *parsimony* metódy – PHYLIP

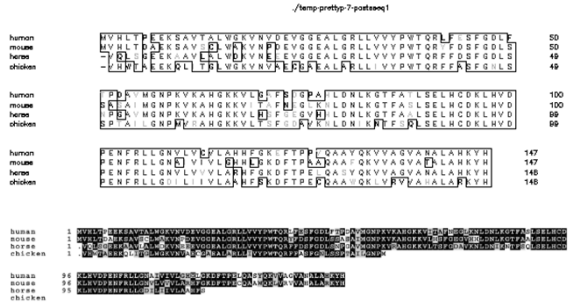
12. Sekundárna a terciárna štruktúra biomakromulekúl
primárna, sekundárna a terciárna štruktúra – vzťah štruktúry a funkcie – 3D vizualizácia – RasMol – MOLMOL

Zoradenie viacerých sekvencií

```
hemoglobin proteins

human    MVHLTPEEKSAVTALWGKVNVDEVGGEALGRLLVVYPWTQRLFESFGDLFTPDAVMGNPK
mouse    MVHLTDAEKSAVSCLWAKVNPDEVGGEALGRLLVVYPWTQRYFDSFGDLSSASAIMGNPK
horse    -VQLSGEEKAAVLALWDKVNEEEVGGEALGRLLVVYPWTQRFFDSFGDLSNPGAVMGNPK
chicken  -VHWTAEEKQLITGLWGKVNVAECGAEALARLLIVYPWTQRFFASFGNLSSPTAILGNPM
                *: : ** : ** *** * *.***.***:******** * ***:* .. *::***

human    VKAHGKKVLGAFSDGPAHLDNLKGTFATLSELHCDKLHVDPENFRLLGNVLVCVLAHHFG
mouse    VKAHGKKVITAFNEGLKNLDNLKGTFASLSELHCDKLHVDPENFRLLGNAIVIVLGHHLG
horse    VKAHGKKVLHSFGEGVHHLDNLKGTFAALSELHCDKLHVDPENFRLLGNVLVVVLARHFG
chicken  VRAHGKKVLTSFGDAVKNLDNIKNTFSQLSELHCDKLHVDPENFRLLGDILIIVLAAHFS
         *:******: :*.:.  :***:*.**: ********************: :: **. *:.

human    KEFTPPVQAAYQKVVAGVANALAHKYH
mouse    KDFTPAAQAAFQKVVAGVATALAHKYH
horse    KDFTPELQASYQKVVAGVANALAHKYH
chicken  KDFTPECQAAWQKLVRVVAHALARKYH
         *:***  **::**:*  ** ***:***
```

Zoradenie viacerých sekvencií

grafické znázornenie I.

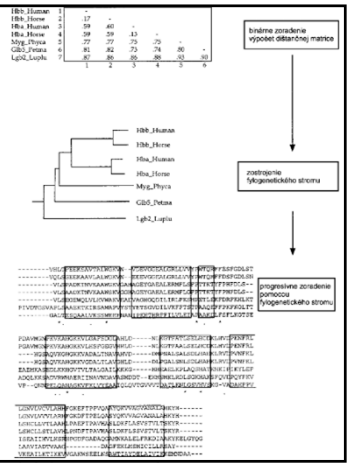


Postupy zoradenia viacerých sekvencií

- kombinácie heuristicky a dynamických algoritmov
- zväčša vyžadujú aspoň *n* – 1 *pairwise* zoradení *(n* = počet zoradených sekvencií*)*
- **manuálne**
  - môžu byť výrazne subjektívne
  - súčasť mnohých balíkov na zoradenie sekvencií
  - vhodné na „dolaďovanie" a publikovanie výsledkov
- **simultánne**
  - zoradenie všetkých sekvencií v jednom behu
  - náročné na výpočtovú silu
  - vhodne hlavne na menšie súbory kratších sekvencií
- **progresívne**
  - výrazne využíva heuristiku
  - ClustalW – najvýznamnejší program z tejto skupiny
  - sekvencie sa zoraďujú v pároch
  - zoraďuje najskôr príbuzné sekvencie, potom sa priraďujú menej príbuzné
  - časť výpočtu poskytuje informácie pre zostavenie fylogenetického stromu

Prehľad progresívneho algoritmu zoradenia viacerých sekvencií (*multiple alignment*) používaného v programe ClustalW

(Higgins *et al.*, 1996)



SYLABUS PREDMETU BIOINFORMATIKA PRE 3. ROČNÍK BIOLOGICKÝCH ODBOROV

1. Úvod do Bioinformatiky
definícia - história -náplň - internet - vzťah k ostatným vedným odborom

2. Biologické databázy
biologické dáta - iné dáta využívané v biológii - rozdelenie biologických databáz - princíp práce s databázami

3. Primárne databázy
typy primárnych sekvencií – EMBL/GenBank/DDBJ – UniProt – GO – vkladanie dát – využitie

4. Sekundárne databázy
proteínové motívy – PROSITE – PRINTS – Pfam – BLOCKS – INTERPRO

5. Ďalšie biologické databázy a integrované databázové systémy
PDB – KEGG – OMIM – REBASE – bibliografické dáta – MEDLINE – integrované databázové systémy – SRS – Entrez

6. Analýza biologických dát
zhromažďovanie a analýza biologických dát - sekvenačné projekty - štatistická analýza - používaná výpočtová technika - Staden Package - EMBOSS

7. Identifikácia kódujúcich úsekov nukleových kyselín
signály – motívy – kódujúce úseky – prokaryoty vs. eukaryoty

8. Zoradenia dvoch sekvencií
pairwise alignment – dot plot – substitučné matice – lokálne a globálne zoradenia – BLAST – FASTA – Needleman-Wunsch – Smith-Waterman

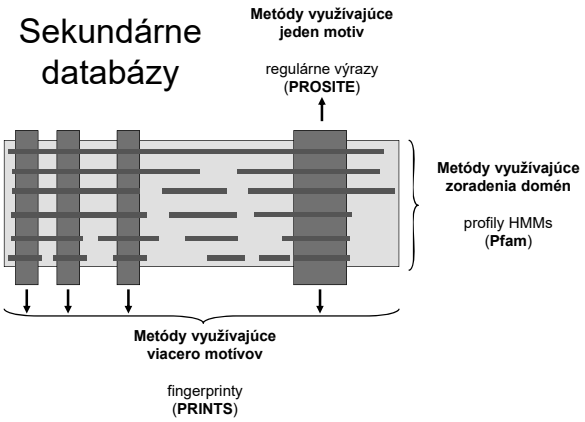
9. Zoradenia viacerých sekvencií
multiple sequence alignment – dynamické programovanie – progresívne metódy – konsenzus sekvencia – ClustalW

10. Identifikácia proteínových motívov
proteínové motívy sekundárnych databáz – neurónové siete – ScanProsite – Pscan – HmmPfam

11. Molekulárna fylogenetická analýza
bioinformatika a evolúcia – fylogenetické stromy – *distance* metódy – *maximum likelihood* metódy – *parsimony* metódy – PHYLIP

12. Sekundárna a terciárna štruktúra biomakromulekúl
primárna, sekundárna a terciárna štruktúra – vzťah štruktúry a funkcie – 3D vizualizácia – RasMol – MOLMOL

Sekundárne databázy



Sekundárne databázy

- zhromažďujú dáta vychádzajúce z konzervatívnych úsekov príbuzných proteínov

- InterPro
  - integrovaná databáza proteínových rodín, domén a funkčných miest

- **PROSITE**
  - databáza proteínových domén a rodín
  - biologicky signifikantné miesta, charakterizované regulárnymi výrazmi (*regular expressions*)

- **PRINTS**
  - charakterizuje proteínové rodiny prostredníctvom fingerprintov
  - fingerprint sa odvodzuje z viacerých motívov

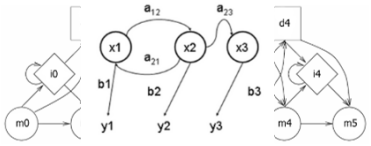
- Pfam
  - databázou proteínových profilov
  - profily sa odvodzujú z celých sekvencií domén, charakteristických pre jednotlivé proteínové rodiny

Identifikácia proteínových signatúr

- Identifikácia jednoduchých proteínových motívov
  - ScanProsite (databáza PROSITE)
- Identifikácia proteínových fingerprintov
  - PScan, FPScan (databáza PRINTS)
- Identifikácia proteínových profilov
  - HmmPfam (databáza Pfam)
- Hidden Markov Model (HMM)
  - pravdepodobnostný model pozostávajúci z niekoľkých vzájmne prepojených stavov (zhoda, delécia, inzercia)

Bayesovské siete, HMM

- **Bayesovské siete** (umelá inteligencia)
  - grafický pravdepodobnostný model
  - popisuje rozdelenie pravdepodobností spoločného výskytu udalostí
- **Markovov model**
  - štatistický model série stavov
  - každý stav záleží iba na predchádzajúcom stave
- **HMM**
  - pravdepodobnostný model založený na Markovovom modely s neznámymi parametrami
- cieľom je určiť skrytý parameter na základe pozorovateľného parametru



číslo	pozícia
m	match
I	insert
d	deletion

Westhead et al., 2002

SYLABUS PREDMETU BIOINFORMATIKA PRE 3. ROČNÍK BIOLOGICKÝCH ODBOROV

1. Úvod do Bioinformatiky
definícia - história -náplň - internet - vzťah k ostatným vedným odborom

2. Biologické databázy
biologické dáta - iné dáta využívané v biológii - rozdelenie biologických databáz - princíp práce s databázami

3. Primárne databázy
typy primárnych sekvencií – EMBL/GenBank/DDBJ – UniProt – GO – vkladanie dát – využitie

4. Sekundárne databázy
proteínové motívy – PROSITE – PRINTS – Pfam – BLOCKS – INTERPRO

5. Ďalšie biologické databázy a integrované databázové systémy
PDB – KEGG – OMIM – REBASE – bibliografické dáta – MEDLINE – integrované databázové systémy – SRS – Entrez

6. Analýza biologických dát
zhromažďovanie a analýza biologických dát - sekvenačné projekty - štatistická analýza - používaná výpočtová technika - Staden Package - EMBOSS

7. Identifikácia kódujúcich úsekov nukleových kyselín
signály – motívy – kódujúce úseky – prokaryoty vs. eukaryoty

8. Zoradenia dvoch sekvencií
pairwise alignment – *dot plot* – substitučné matice – lokálne a globálne zoradenia – BLAST – FASTA – Needleman-Wunsch – Smith-Waterman

9. Zoradenia viacerých sekvencií
multiple sequence alignment – dynamické programovanie – progresívne metódy – konsenzus sekvencia – ClustalW

10. Identifikácia proteínových motívov
proteínové motívy sekundárnych databáz – neurónové siete – ScanProsite – Pscan – HmmPfam

11. Molekulárna fylogenetická analýza
bioinformatika a evolúcia – fylogenetické stromy – *distance* metódy – *maximum likelihood* metódy – *parsimony* metódy – PHYLIP

12. Sekundárna a terciárna štruktúra biomakromulekúl
primárna, sekundárna a terciárna štruktúra – vzťah štruktúry a funkcie – 3D vizualizácia – RasMol – MOLMOL

Molekulárna fylogenetická analýza

- **sekvenčná informácia** vs. fenotypová informácia
- fylogeneticky vhodný **molekulárny markér**
  - rRNA
    + funkčne stabilný
    + prítomný vo všetkých organizmoch ako aj mitochondriách a chloroplastoch
    +rôzne úseky sa menia rôznou rýchlosťou
    - prenos génov medzi druhmi a bunkovými organelami
    - vnútroorganizmová heterogenita (viacnásobné kópie)
    - rôzny obsah G+C párov pri jednotlivých druhoch
  - iné
    RecA, transketolázy, aldolázy, HSP170, elongačný faktor TU, b-podjednotka ATP, sigma faktor

Terminológia

- **uzly** (node)
  - gén, druh, populácia
- **vetvy** (branche)
  - spojenie zlov
  - dĺžka preprezentuje mutačné zmeny alebo evolučný čas
- iba **1 spojenie** medzi 2 uzlami
- **vonkajšie** (uzly, vetvy)
- **vnútorné** (uzly, vetvy)
- **zakorenené** (rooted)
- **nezakorenené** (unrooted)
  - zmena ak poznáme ***outgroup***

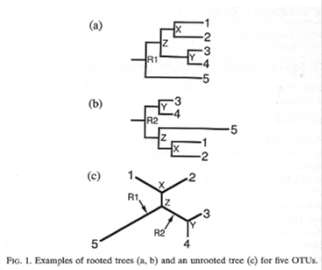
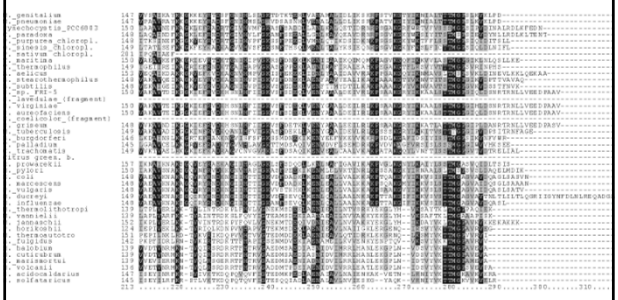
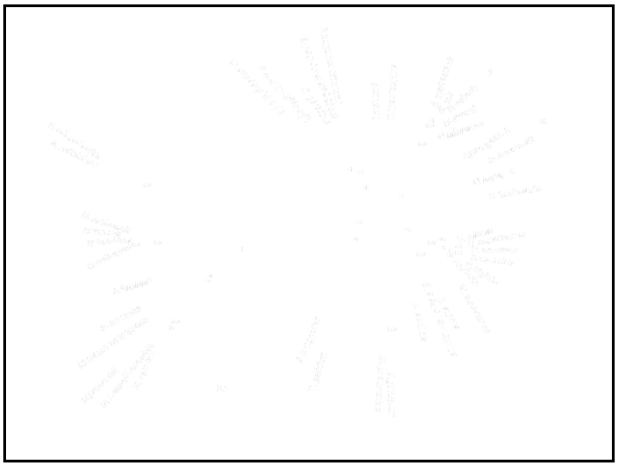


FIG. 1. Examples of rooted trees (a, b) and an unrooted tree (c) for five OTUs.

Zoradenie sekvencií ribozómového proteínu L1

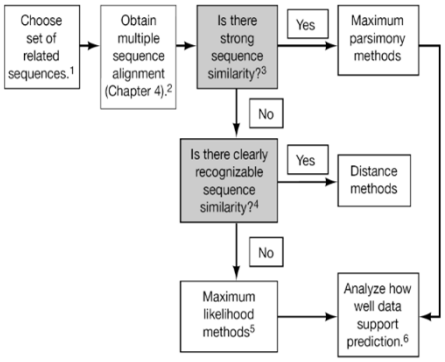




Balík programov **PHYLIP** (Felsenstein)

- Parsimony – PROTPARS (Eck a Dayhoff, Fitch)
  - znakové metódy (*character based*)
  - najmenší možný počet mutačných zmien
- Distance
  - PROTDIST (PAM matrica, Kimurove vzdialenosti, skupinové vzdialenosti)
  - FITCH (Fitch-Margoliash, najmenších štvorcov)
  - NEIGHBOR (neighbor-joining, UPGMA)
- (Maximum Likelihood)
  - čisto štatistická metóda, využíva tzv. substitučné modely
  - výpočtovo náročné
- Bootstraping – SEQBOOT
  - pomáha určiť signifikantnosť výsledného stromu
- ak rôzne metódy poskytnú pre daný set sekvencií podobný fylogenetický strom, takýto strom možno považovať za správny; v opačnom prípade žiadna metóda nemusí poskytnúť správny výsledok

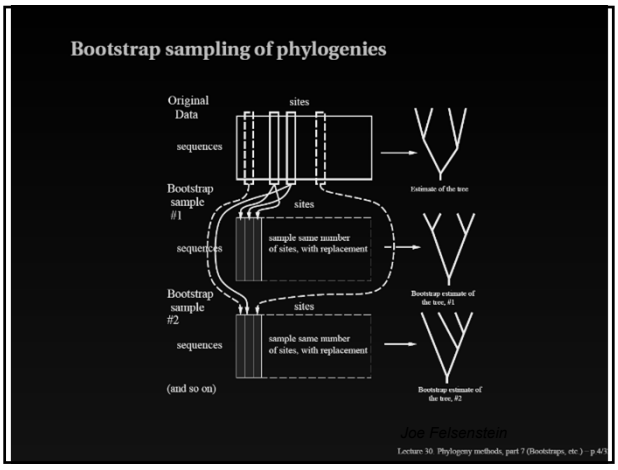
Výber metód

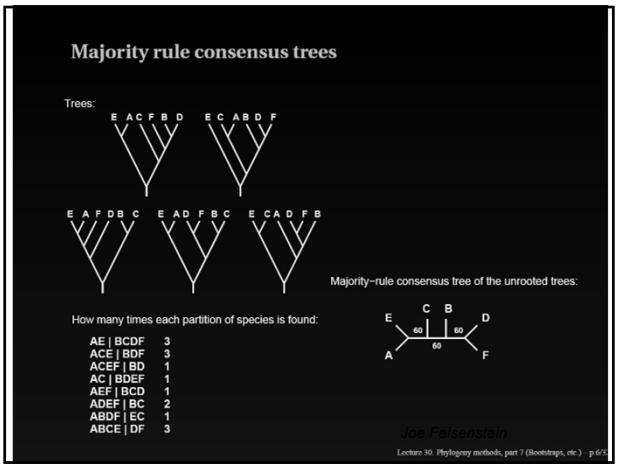


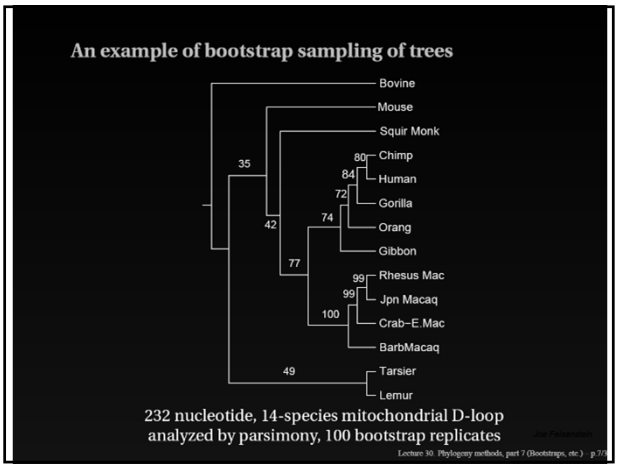
Mount, D.W.: *Bioinformatics, Sequence and Genome Analysis*. Cold Spring Harbor Laboratory Press, New York, 2006.

Bootstraping

- predpoklad, že jednotlivé pozície sa v procese evolúcie vyvíjali nezávisle
- stromy predstavujú parameter; výsledkom analýzy je súbor navzorkovaných (*sampled*) stromov
- v kroku sumarizácie sa pre každé vetvenie zisťuje ako často sa vyskytuje v rámci celkového súboru získaných stromov
- na záver sa vytvorí strom ktorý zodpovedá najčastejšie sa vyskytujúcim vetveniam – *majority rule consensus tree*







SYLABUS PREDMETU BIOINFORMATIKA PRE 3. ROČNÍK BIOLOGICKÝCH ODBOROV

1. Úvod do Bioinformatiky
definícia - história -náplň - internet - vzťah k ostatným vedným odborom

2. Biologické databázy
biologické dáta - iné dáta využívané v biológii - rozdelenie biologických databáz - princíp práce s databázami

3. Primárne databázy
typy primárnych sekvencií – EMBL/GenBank/DDBJ – UniProt – GO – vkladanie dát – využitie

4. Sekundárne databázy
proteínové motívy – PROSITE – PRINTS – Pfam – BLOCKS – INTERPRO

5. Ďalšie biologické databázy a integrované databázové systémy
PDB – KEGG – OMIM – REBASE – bibliografické dáta – MEDLINE – integrované databázové systémy – SRS – Entrez

6. Analýza biologických dát
zhromažďovanie a analýza biologických dát - sekvenačné projekty - štatistická analýza - používaná výpočtová technika - Staden Package - EMBOSS

7. Identifikácia kódujúcich úsekov nukleových kyselín
signály – motívy – kódujúce úseky – prokaryoty vs. eukaryoty

8. Zoradenia dvoch sekvencií
pairwise alignment – *dot plot* – substitučné matice – lokálne a globálne zoradenia – BLAST – FASTA – Needleman-Wunsch – Smith-Waterman

9. Zoradenia viacerých sekvencií
multiple sequence alignment – dynamické programovanie – progresívne metódy – konsenzus sekvencia – ClustalW

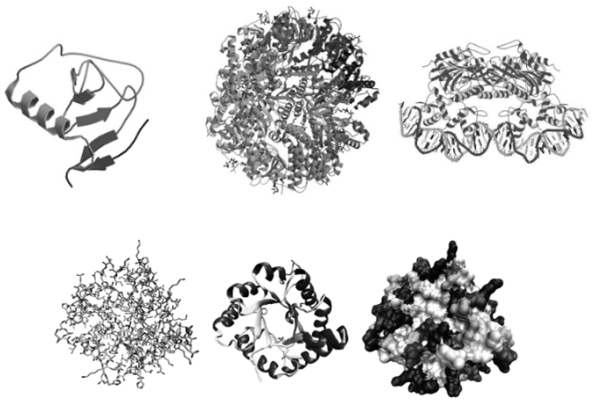
10. Identifikácia proteínových motívov
proteínové motívy sekundárnych databáz – neurónové siete – ScanProsite – Pscan – HmmPfam

11. Molekulárna fylogenetická analýza
bioinformatika a evolúcia – fylogenetické stromy – *distance* metódy – *maximum likelihood* metódy – *parsimony* metódy – PHYLIP

12. Sekundárna a terciárna štruktúra biomakromulekúl
primárna, sekundárna a terciárna štruktúra – vzťah štruktúry a funkcie – 3D vizualizácia – RasMol – MOLMOL

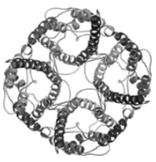
Vzťah štruktúry a funkcie

- Môžeme predpovedať funkciu proteínu z jeho aminokyselinovej sekvencie?
  sekvencie -> štruktúra -> funkcia
- konzervované funkčné domény = motívy
- predikcia niektorých jednoduchých 3-D štruktúr:
  - α-helix
  - β-sheet
  - membrane spanning regions



Určenie makromolekulárnych štruktúr

- **nukleové kyseliny**
  - určenie sekundárnej štruktúry tRNA, miest zostrihu
- **proteíny**
  - sekundárna štruktúra
    - α-helix, β-skladaný list, *coil*
    - vychádza sa AAs sekvencie proteínu
    - presnosť predpovede asi 65 % (75 % pri použití viacerých homologických sekvencií)
  - trojrozmerná štruktúra (sek., terc., 3D)
    - 150 krát viac proteínových sekvencií ako známych 3D štruktúr
    - 3D štruktúra proteínov je v procese evolúcie veľmi konzervatívna



aquaporin-1

- transport polárnej vody cez nepolárnu membránu
- 3 x 10^9 molekúl vody za sekundu
- 269 aminokyselín
- 6 transmembránových psekov
- homotetramér
- 3.0 Å štrbina (voda je 2.8 Å)

Campbell A.M. and Heyer L.J. : Discovering Genomics, Proteomics and Bioinformatics. (2003)

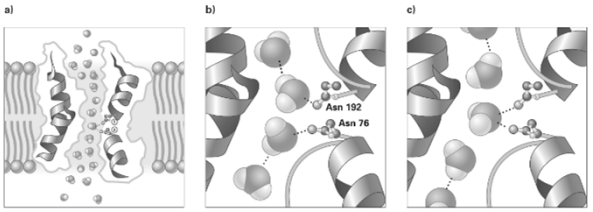
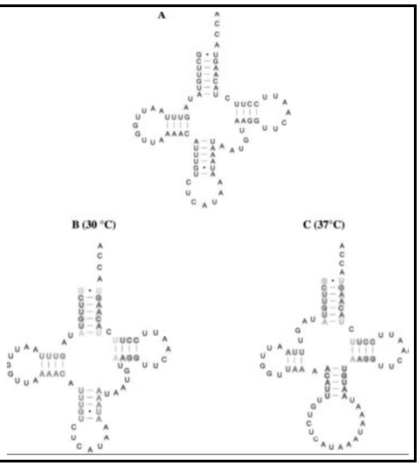
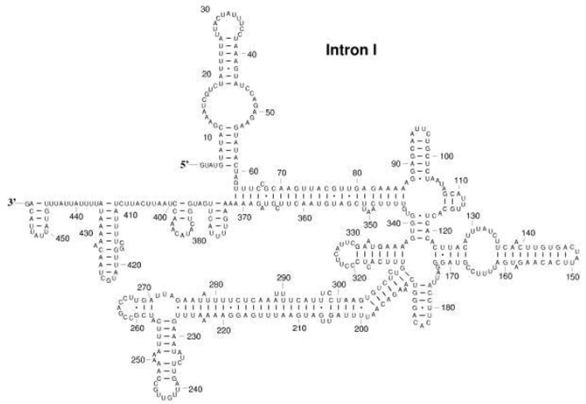


FIGURE 6.11 • Structural basis for aquaporin function. **a)** The charges from the helix control the orientation of the water molecules passing through the narrowest part of the channel. **b)** and **c)** The hydrogen bonding of a water molecule to asparagines 192 and 76, which extend their R groups to form the narrowest part of the channel.

Secondary structure prediction of Saccharomyces cerevisiae tRNA (RM4000) at optimal growth temperature (30°C) (B) and at 37 °C (C) with the presented nearest neighbor parameters. Base pairs in the original structure (A) are derived from the comparative analysis database (42–49,58,59). Structures are also color annotated to indicate predicted base pair probabilities (Pbp) for each helix: red, Pbp > 0.95; yellow, 0.95 > Pbp > 0.7; green, 0.7 > Pbp > 0.3; blue, 0.3 > Pbp. The structures were drawn with XRNA (http://rna.ucsc.edu/rnacenter/xrna/xrna.html) and Adobe Illustrator.

Lu ZJ, Turner DH, Mathews DH. A set of nearest neighbor parameters for predicting the enthalpy change of RNA secondary structure formation. *Nucleic Acids Res. 2006;34(17):4912-24. Epub 2006 Sep 18.*

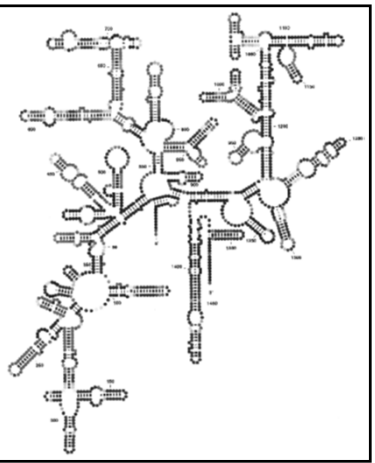




Color annotation of the phylogenetic model for 16S rRNA of T. thermophilus based on probability.

The probability for each base pair was calculated using version 1.2.1 of the *Vienna RNA folding package+* This program version uses energy rules that are identical to those used in version 2.3 of mfold. The predicted folding is almost indistinguishable from the predicted optimal folding that is generated with version 2.3 of mfold.

Zuker M, Jacobson AB. Using reliability information to annotate RNA secondary structures. RNA. 1998 Jun;4(6):669-79.



Terciárna a kvartérna štruktúra

- *protein folding* – predikcia sekundárnej, terciárner a quartérnej štruktúry polypeptidových reťazcov
- faktory:
  - elektrostatické sily
  - vodíkové väzby
  - van der Waalsove sily
  - kovalentné (disulfidové) väzby medzi cysteínmi

Komparatívne modelovanie

- identifikácia setu proteínov štrukturálne príbuzných
- zoradenie získaných proteínových sekvencií
- konštrukcia modelu
- modelovanie slučiek
- modelovanie bočných reťazcov
- evaluácia modelu

Ako ďaleko možno porovnávať?

úroveň porovnávania	**spoločný predchodca**
nekódujúca DNA	do 200 miliónov rokov
kódujúca DNA	do 600 miliónov rokov
proteínová sekvencia	1 až 2,5 miliárd rokov
3D štruktúra proteínu	najkonzervovanejšia