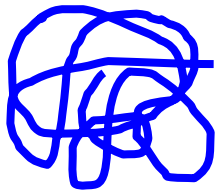
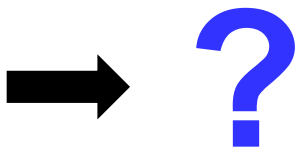


Protein

RKSTGGKAPRKQLATKAARKSAPATGGV
KKPHRYRPGTVALREIRRYQKSTELLIR
KLPFQRLVREIAQDFKTDLRFQSSAVMA
LQEASEAYLVGLFEDTNLCAIHAKR



Základní vlastnosti proteinů

Predikce vlastností proteinů

Aplikovaná bioinformatika, Jaro 2016

BIOINFORMATION
Discovery at the interface of physical and biological sciences

open access www.bioinformation.net Hypothesis

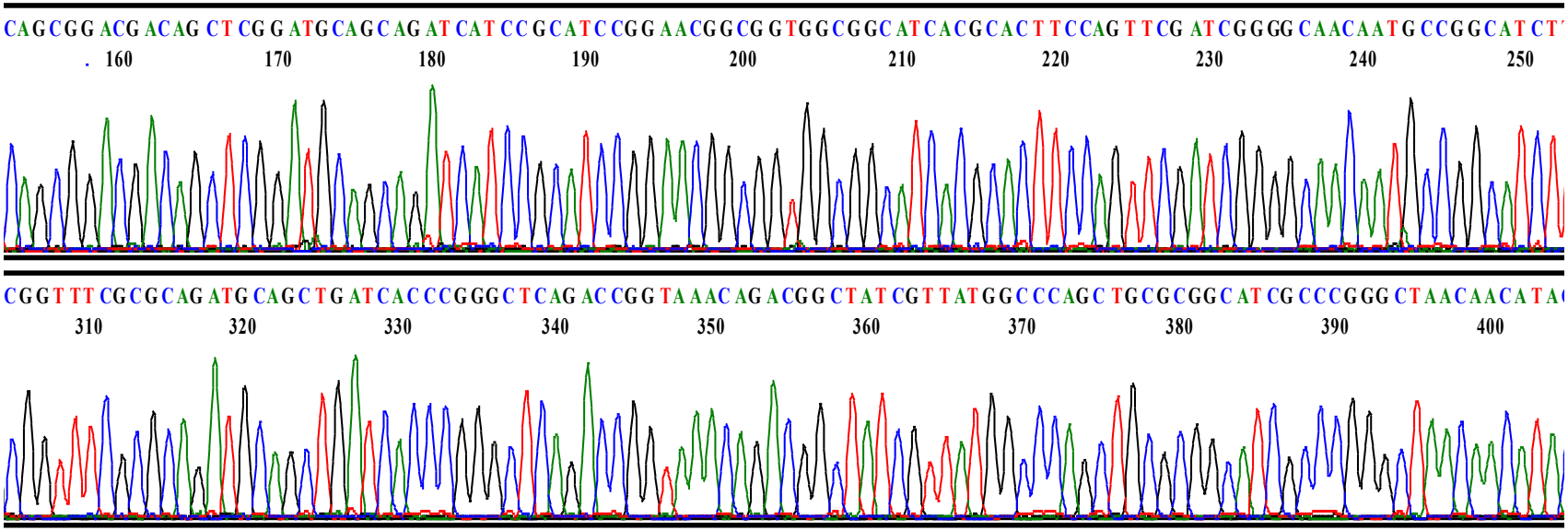
Volume 8(15)

Computational structural and functional analysis of hypothetical proteins of *Staphylococcus aureus*

Ramadevi Mohan & Subhashree Venugopal*

Abstract:
Genome sequencing projects has led to an explosion of large amount of gene products in which many are of hypothetical proteins with unknown function. Analyzing and annotating the functions of hypothetical proteins is important in *Staphylococcus aureus* which is a pathogenic bacterium that cause multiple types of diseases by infecting various sites in humans and animals. In this study, ten hypothetical proteins of *Staphylococcus aureus* were retrieved from NCBI and analyzed for their structural and functional characteristics by using various bioinformatics tools and databases. The analysis revealed that some of them possessed functionally important domains and families and protein-protein interacting partners which were ABC transporter ATP-binding protein, Multiple Antibiotic Resistance (MAR) family, export proteins, Helix-Turn-helix domains, arsenate reductase, elongation factor, ribosomal proteins, Cysteine protease precursor, Type-I restriction endonuclease enzyme and plasmid recombination enzyme which might have the same functions in hypothetical proteins. The structural prediction of those proteins and binding sites prediction have been done which would be useful in docking studies for aiding in the drug discovery.

Sekvenace celých genomů



GATAGCGTAATGATCGGCTGGCTGCCGCATTTCATGCTGGTTTCCCAACGAAAATAACCGCTCACGGTGCCATCACGATCGCACACCGCAAAATCGGCGG

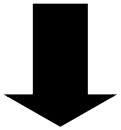
TACAGGTGGTCGCGCCCGCCGCCAGCACATCGCTGCGCCAATAATGATCTTTCAGCGGACGACAGCTCGGATGCAGCAGATCATCCGCATCCGGAACGGC

GGTGGCGGCATCACGCACTTCCAGTTCGATCGGGGCAACAATGCCGGCATCTTTCAGGGCAAAGCGAATAAACAGCACGCTCACTTCCGCGCGCAGCGCC

AGCGCGGTTTCGCGCAGATGCAGCTGATCACCCGGGCTCAGACCGGTAAACAGACGGCTATCGTTATGGCCCAGCTGCGCGGCATCGCCCGGGCTAACAA

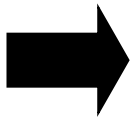
CATACAGGTGGCGACCATCAATCACGGTCGGGGCGGCCGGATCACGGCTGGCTTCCGGATAGGCGCTCAGCAGGGTAACGGCATCCACAATCACCAGCAT

```
CCTTTATTATCCGCTTCCATTGTTTCCGCTCCTGTTGTTACTTCCGAAACTTATGTTGATATTCCTGGTTTATATTTAGA
TGTTGCTAAAGCTGGTATTCGCGATGGTAAATTACAAGTTATTTTAAATGTTCCTACTCCTTATGCTACTGGTAATAATT
TTCCTGGTATTTATTTTGCTATTGCTACTAATCAAGGTGTTGTTGCTGATGGTTGTTTTACTTATTCCTCCAAAGTTCCT
GAATCCACTGGTCGCATGCCTTTTACTTTAGTTGCTACTATTGATGTTGGTTCCGGTGTTACTTTTGTTAAAGGTCAATG
GAAATCCGTTCGCGGTTCCGCTATGCATATTGATTCCTATGCTTCCTTATCCGCTATTTGGGGTACTGCTGCTCCTTCCT
CCCAAGGTTCCGGTAATCAAGGTGCTGAAACTGGTGGTACTGGTGCTGGTAATATTGGTGGTGGTGGTGAACGCGATGGT
ACTTTTAATTTACCTCCTCATATTAAATTTGGTGTTACTGCTTTAACTCATGCTGCTAATGATCAAACTATTGATATTTA
TATTGATGATGATCCTAAACCTGCTGCTACTTTTAAAGGTGCTGGTGCTCAAGATCAAAATTTAGGTACTAAAGTTTTAG
ATTCCGGTAATGGTCGCGTTCGCGTTATTGTTATGGCTAATGGTCGCCCTTCCCGCTTAGGTTCCCGCCAAGTTGATATT
TTTAAAAAATCCTATTTTGGTATTATTGGTTCCGAAGATGGTGCTGATGATGATTATAATGATGGTATTGTTTTTTTAAA
```



PLLSASIVSAPVVTSETYVDIPGLYLDVAKAGIRDGKLQVILNVPTPYATGNNFPGIYFA
IATNQGVVADGCFTYSSKVPESTGRMPFTLVATIDVGSGVTFVKGQWKSVRGSAMHIDSY
ASLSAIWGTAAPSSQGSGNQGAETGGTGAGNIGGGGERDGTFNLPPHIKFGVTALTHAAN
DQTIDIYIDDDPKPAATFKGAGAQDQNLGTKVLDSGNGRVRVIVMANGRPSRLGSRQVDI
FKKSYFGIIGSEDGADDDYNDGIVFL

Nukleotidová a proteinová sekvence hypotetických proteinů



Predikce vlastností

Predikce základních vlastností proteinů ze sekvence

Physicochemical and functional characterization

For physicochemical characterization, theoretical Isoelectric point (pI), molecular weight, total number of positive and negative residues, extinction coefficient **[17]**, instability index **[18]**, aliphatic index **[19]** and grand average hydropathy (GRAVY) **[20]** were computed using the Expasy's Protparam server **[21]**.

Predikce základních fyzikálně-chemických parametrů.

Predikce lokalizace proteinů v buňce.

Prediction of transmembrane proteins

SOSUI server is used to characterize whether the protein is soluble or transmembrane in nature **[28]**.

Predikce základních vlastností proteinů ze sekvence

Table 1: Physicochemical properties of hypothetical proteins by Protparam tool

Sequence ID	No of aa	MW	pI	(-) R	(+ R)	EC	II	AI	GRAVY
gi|166409299	97	10407.8	10.11	1	12	25440	43.32	65.36	-0.182
gi|166409303	129	15748.3	10.14	13	31	22920	41.52	83.8	-0.877
gi|166409302	208	23392.4	9.29	27	34	7450	22.22	115.72	-0.148
gi|166409301	103	12163.4	9.73	11	20	11920	22.96	113.4	-0.287
gi|166409300	644	75501.5	9.14	60	73	77825	35.81	119.1	0.128
gi|390516769	31	3677.5	9.3	3	5	1490	7.98	138.39	0.726
gi|166409293	139	15938.3	9.37	12	18	13075	36.4	95.4	-0.443
gi|390516759	209	24226.7	9.25	23	32	19495	23.33	91.87	-0.612
gi|390516760	80	9250.4	4.76	15	11	4470	64.89	101.12	-0.611
gi|166409294	323	35653	6.25	36	34	40340	30.12	105.82	0.004

Predikce základních fyzikálně-chemických parametrů.

Predikce lokalizace proteinů v buňce.

Table 4: Prediction of Subcellular localization sites in hypothetical proteins

Sequence ID	Localization
gi|390516760	Cytoplasmic
gi|390516759	Unknown
gi|166409293	Cytoplasmic
gi|166409299	CytoplasmicMembrane
gi|166409303	Unknown
gi|166409302	CytoplasmicMembrane
gi|390516769	CytoplasmicMembrane
gi|166409301	Unknown
gi|166409300	CytoplasmicMembrane
gi|166409294	CytoplasmicMembrane

ExPASy
Bioinformatics Resource Portal

Expert Protein Analysis System

http://www.expasy.org

ExPASy is the **SIB Bioinformatics Resource Portal** which provides access to scientific databases and software tools (i.e., *resources*) in different areas of life sciences including proteomics, genomics, phylogeny, systems biology, population genetics, transcriptomics etc. (see Categories in the left menu). On this portal you find resources from many different SIB groups as well as external institutions.

SIB 15 YEARS
Swiss Institute of Bioinformatics

The SIB Swiss Institute of Bioinformatics is an academic, non-profit foundation recognised of public utility and established in 1998. SIB coordinates research and education in bioinformatics throughout Switzerland and provides high quality bioinformatics services to the national and international research community.

ExPASy
Bioinformatics Resource Portal

- Visual Guidance
- Categories
  - proteomics
  - genomics
  - structural bioinformatics
  - systems biology
  - phylogeny/evolution
  - population genetics
  - transcriptomics
  - biophysics
  - imaging
  - IT infrastructure
  - drug design
- Resources A..Z
- Links/Documentation

Visual Guidance Interface

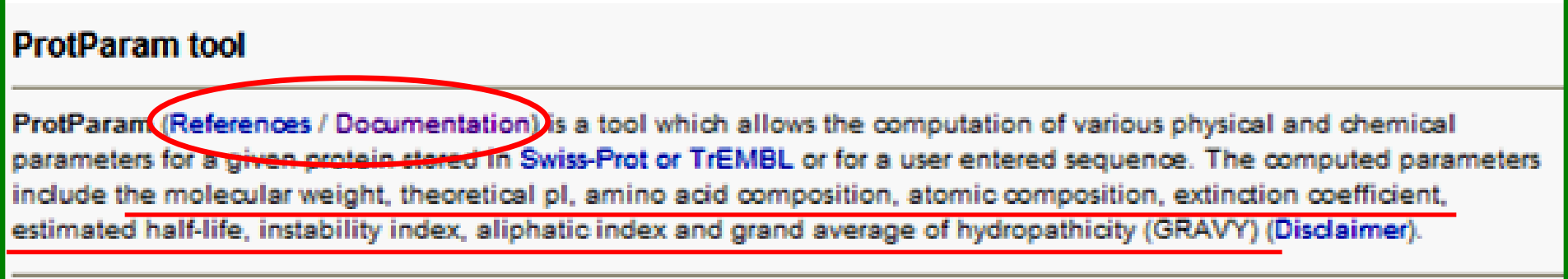
Please select an element:

Published online 31 May 2012 *Nucleic Acids Research, 2012, Vol. 40, Web Server issue* ***W597–W603***
doi:10.1093/nar/gks400

ExPASy: SIB bioinformatics resource portal

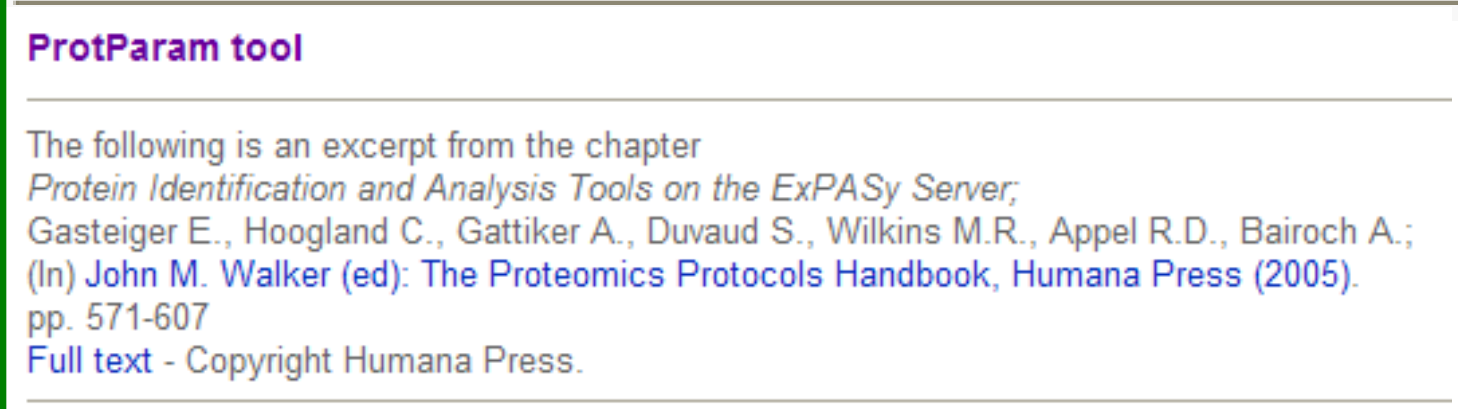
Panu Artimo[1], Manohar Jonnalagedda[1,2], Konstantin Arnold[3], Delphine Baratin[4], Gabor Csardi[5], Edouard de Castro[4], Séverine Duvaud[4], Volker Flegel[1], Arnaud Fortier[1], Elisabeth Gasteiger[4], Aurélien Grosdidier[2], Céline Hernandez[1], Vassilios Ioannidis[1], Dmitry Kuznetsov[1], Robin Liechti[1], Sébastien Moretti[1,6], Khaled Mostaguir[4], Nicole Redaschi[4], Grégoire Rossier[1], Ioannis Xenarios[1,4,7] and Heinz Stockinger[1,*]

ProtParam



ProtParam tool

ProtParam (References / Documentation) is a tool which allows the computation of various physical and chemical parameters for a given protein stored in Swiss-Prot or TrEMBL or for a user entered sequence. The computed parameters include the molecular weight, theoretical pI, amino acid composition, atomic composition, extinction coefficient, estimated half-life, instability index, aliphatic index and grand average of hydropathicity (GRAVY) (Disclaimer).



ProtParam tool

The following is an excerpt from the chapter
Protein Identification and Analysis Tools on the ExPASy Server;
Gasteiger E., Hoogland C., Gattiker A., Duvaud S., Wilkins M.R., Appel R.D., Bairoch A.;
(In) John M. Walker (ed): The Proteomics Protocols Handbook, Humana Press (2005).
pp. 571-607
Full text - Copyright Humana Press.

- **Predikce/výpočet základních fyzikálně-chemických parametrů proteinu.**
- **Vychází pouze z aminokyselinové sekvence proteinu.**

ProtParam

Please note that you may only fill out **one** of the following fields at a time.

Enter a Swiss-Prot/TrEMBL accession number (AC) (for example **P05130**) or a sequence identifier (ID) (for example **KPC1_DROME**):

Or you can paste your own sequence in the box below:

```
PLLSASIVSAPVVTSETYVDIPGLYLDVAKAGIRDGKLQVILNVPTPYATGNNFPGIYFA
IATNQGVVADGCFTYSSKVPESTGRMPFTLVATIDVGSGVTFVKGQWKSVRGSAMHIDSY
ASLSAIWGTAAPSSQGSGNQGAETGGTGAGNIGGGGERDGTFNLPPHIKFGVTALTHAAN
DQTIDIYIDDDPKPAATFKGAGAQDQNLGTKVLDSGNGRVRVIVMANGRPSRLGSRQVDI
FKKSYFGIIGSEDGADDDYNDGIVF
```

RESET | Compute parameters

Úkol 1: určete základní fyzikálně-chemické parametry tohoto proteinu

PLLSASIVSAPVVTSETYVDIPGLYLDVAKAGIRDGKLQVILNVPTPYATGNNFPGIYFAIATNQGVVADG
CFTYSSKVPESTGRMPFTLVATIDVGSGVTFVKGQWKSVRGSAMHIDSYASLSAIWGTAAPSSQGSGNQGA
ETGGTGAGNIGGGGERDGTFNLPPHIKFGVTALTHAANDQTIDIYIDDDPKPAATFKGAGAQDQNLGTKVL
DSGNGRVRVIVMANGRPSRLGSRQVDIFKKSYFGIIGSEDGADDDYNDGIVFL

Molekulová hmotnost - M_w

SKEPLRPRCRPINATLAVEKEGCPVCITVNTTICAGYCPTMTRVLQGVLP
ALPQVVCNYRDVRFESIRLPGCPRGVNPVVSYAVALSCQCALCRRSTTDC
GGPKDHPLTCDDPRFQDSSSSKAPPPSLPSPSRLPGPSDTPILPQ

Úkol 2: Molekulová hmotnost zkoumaného proteinu byla pomocí SDS-PAGE stanovena na cca 30 kDa. Ověřte, zda se jedná o Váš protein. Pokuste se vysvětlit případné nesrovnalosti.

Molekulová hmotnost - M_w

Note: It is not possible to specify post-translational modification for your protein, nor will ProtParam know whether your mature protein forms dimers or multimers. If you do know that your protein forms a dimer, you may just duplicate your sequence (i.e. append a second copy of the sequence to the first), as all computations performed by ProtParam are based on either compositional data, or on the N-terminal amino acid.

- **ProtParam nebere v úvahu možné posttranslační modifikace a oligomerizaci proteinů.**
- **Pro predikci PTM a oligomerizace existují specializované nástroje.**
- **Problematika PTM není stále dořešená, především u prokaryot.**
- **Glykosylace proteinů, dříve považovaná za proces probíhající pouze u eukaryot, byla již prokázána i u prokaryot.**
  Databáze prokaryotických glykoproteinů: ProGlycProt
  Predikce glykosylace u prokaryot: GlycoPP

Molekulová hmotnost - M_w

Protein 1:

CCGACGGAGTTCCTGTACACGAGCAAGATAGCGGCGATAAGCTGGGCGGCGACGGGGGGGAGGCAGCAGAGGGTGTACTTCCAGGACCTGA
ACGGGAAGATAAGGGAGGCGCAGAGGGGGGGGGACAACCCGTGGACGGGGGGGAGCAGCCAGAACGTGATAGGGGAGGCGAAGCTGTTCAG
CCCGCTGGCGGCGGTGACGTGGAAGAGCGCGCAGGGGATACAGATAAGGGTGTACTGCGTGAACAAGGACAACATACTGAGCGAGTTCGTG
TACGACGGGAGCAAGTGGATAACGGGGCAGCTGGGGAGCGTGGGGGTGAAGGTGGGGAGCAACAGCAAGCTGGCGGCGCTGCAGTGGGGGG
GGAGCGAGAGCGCGCCGCCGAACATAAGGGTGTACTACCAGAAGAGCAACGGGAGCGGGAGCAGCATACACGAGTACGTGTGGAGCGGGAA
GTGGACGGCGGGGGCGAGCTTCGGGAGCACGGTGCCGGGGACGGGGATAGGGGCGACGGCGATAGGGCCGGGGAGGCTGAGGATATACTAC
CAGGCGACGGACAACAAGATAAGGGAGCACTGCTGGGACAGCAACAGCTGGTACGTGGGGGGGTTCAGCGCGAGCGCGAGCGCGGGGGTGA
GCATAGCGGCGATAAGCTGGGGGAGCACGCCGAACATAAGGGTGTACTGGCAGAAGGGGAGGGAGGAGCTGTACGAGGCGGCGTACGGGGG
GAGCTGGAACACGCCGGGGCAGATAAAGGACGCGAGCAGGCCGACGCCGAGCCTGCCGGACACGTTCATAGCGGCGAACAGCAGCGGGAAC
ATAGACATAAGCGTGTTCTTCCAGGCGAGCGGGGTGAGCCTGCAGCAGTGGCAGTGGATAAGCGGGAAGGGGTGGAGCATAGGGGCGGTGG
TGCCGACGGGGACGCCGGCGGGGTGG

Protein 2:

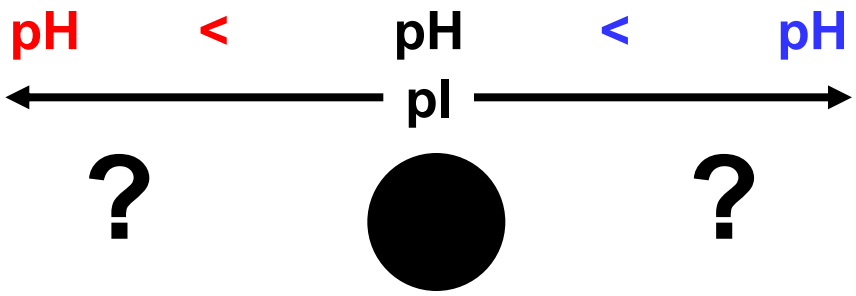
AWKGEVLANNEAGQVTSIIYNPGDVITIVAAGWASYGPTQKWGPQGDREHPDQGLICHDAFCGALVMKIGNSGTIPVN
TGLFRWVAPNNVQGAITLIYNDVPGTYGNNSGSFSVNIGKDQS

Úkol 3: **Student s využitím ProtParam vypočítal molekulovou hmotnost svých proteinů na 69,9 a 12,7 kDa. Při gelové chromatografii ale určil molekulovou hmotnost na 33 a 51 kDa! Ověřte jeho výpočet a zkuste najít vysvětlení, když je experimentálně prokázáno, že tyto proteiny nepodléhají PTM.**

Izoelektrický bod - pI

- **Izoelektrický bod = pH, při kterém má protein nulový sumární náboj.**

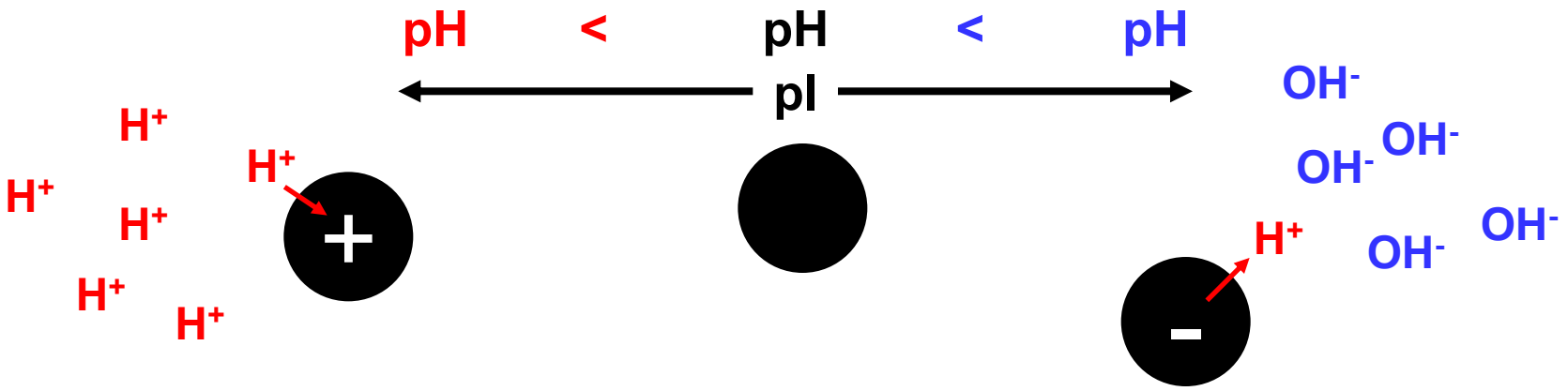
Protein pI is calculated using pK values of amino acids described in Bjellqvist et al., which were defined by examining polypeptide migration between pH 4.5 to 7.3 in an immobilised pH gradient gel environment with 9.2M and 9.8M urea at 15ï¿½C or 25ï¿½C. Prediction of protein pI for highly basic proteins is yet to be studied and it is possible that current Compute pI/Mw predictions may not be adequate for this purpose.



Izoelektrický bod - pI

- **Izoelektrický bod = pH, při kterém má protein nulový sumární náboj.**

Protein pI is calculated using pK values of amino acids described in Bjellqvist et al., which were defined by examining polypeptide migration between pH 4.5 to 7.3 in an immobilised pH gradient gel environment with 9.2M and 9.8M urea at 15ï¿½C or 25ï¿½C. Prediction of protein pI for highly basic proteins is yet to be studied and it is possible that current Compute pI/Mw predictions may not be adequate for this purpose.



Izoelektrický bod - pI

- **Izoelektrický bod = pH, při kterém má protein nulový sumární náboj.**

Protein pI is calculated using pK values of amino acids described in Bjellqvist et al., which were defined by examining polypeptide migration between pH 4.5 to 7.3 in an immobilised pH gradient gel environment with 9.2M and 9.8M urea at 15ï¿½C or 25ï¿½C. Prediction of protein pI for highly basic proteins is yet to be studied and it is possible that current Compute pI/Mw predictions may not be adequate for this purpose.

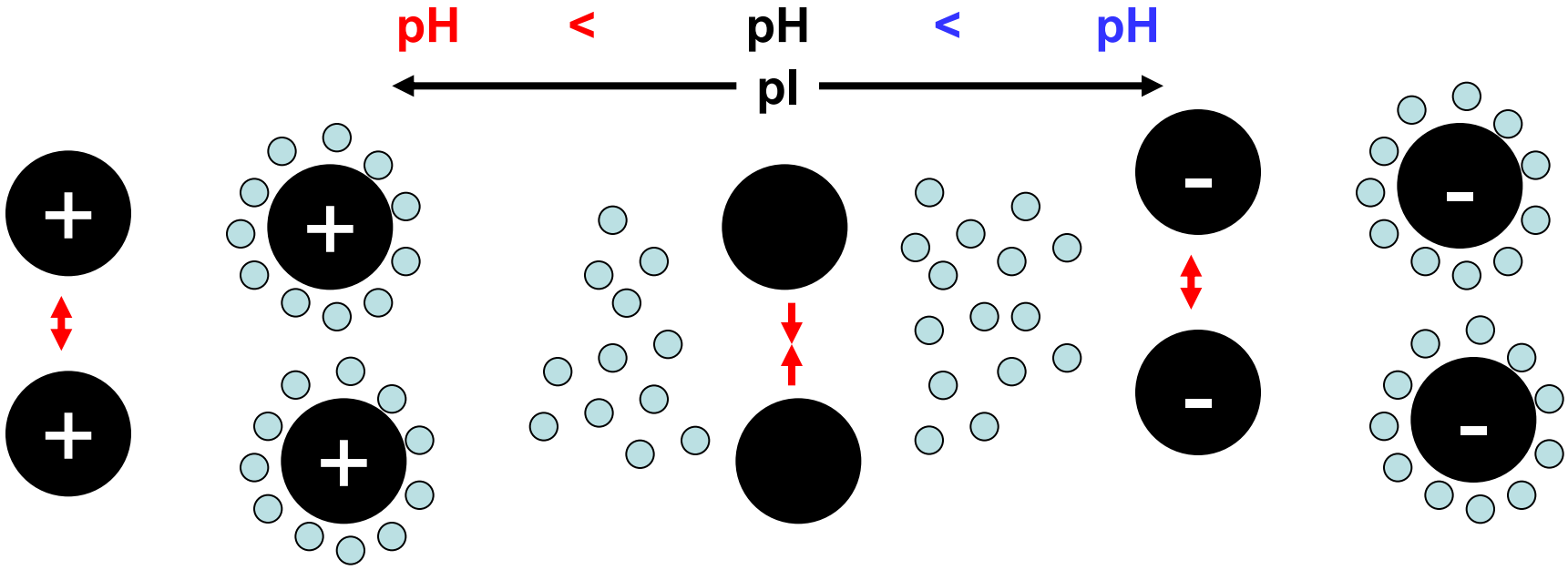
- **Problémem jsou opět posttranslační modifikace!!!**

- **Použité hodnoty pK jednotlivých aminokyselin – různí autoři, různé hodnoty…**

Izoelektrický bod - pI

- **Izoelektrický bod = pH, při kterém má protein nulový sumární náboj. Rozpustnost proteinů je při pH = pI nejmenší!**

Protein pI is calculated using pK values of amino acids described in Bjellqvist et al., which were defined by examining polypeptide migration between pH 4.5 to 7.3 in an immobilised pH gradient gel environment with 9.2M and 9.8M urea at 15ï¿½C or 25ï¿½C. Prediction of protein pI for highly basic proteins is yet to be studied and it is possible that current Compute pI/Mw predictions may not be adequate for this purpose.



Izoelektrický bod - pI

Protein 1:

PLLSASIVSAPVVTSETYVDIPGLYLDVAKAGIRDGKLQVILNVPTPYATGNNFPGIYFAIATNQGVVADGCFTYSSK
VPESTGRMPFTLVATIDVGSGVTFVKGQWKSVRGSAMHIDSYASLSAIWGTAAPSSQGSGNQGAETGGTGAGNIGGGG
ERDGTFNLPPHIKFGVTALTHAANDQTIDIYIDDDPKPAATFKGAGAQDQNLGTKVLDSGNGRVRVIVMANGRPSRLG
SRQVDIFKKSYFGIIGSEDGADDDYNDGIVFLNWPLG

Protein 2:

GLSDGACWQLVLNVWGKVEADICPGHGQEVLILLFKGHPETLEKFDKCFKHLKCSEDEMKASEDLKKHGATVLTACLG
GILKKKCGHHEAECIKPLAQDSHATKHKISPVCKYLCEFRISECRCIQIVLQCSKHPGDFGCADAQGAMNKALELFRC
KDMASNYKELGFQG

Protein 3:

AWKGEVLANNEAGQVTSIIYNPGDVITIVAAGWASYGPTQKWGPQGDREHPDQGLICHDAFCGALVMKIGNSGTIPVN
TGLFRWVAPNNVQGAITLIYNDVPGTYGNNSGSFSVNIGKDQS

Úkol 4: **Student pracuje se směsí tří proteinů. Ve standardním pufru (20 mM Tris/HCl, 150 mM NaCl, pH 7,5) pozoroval vznik sraženiny! Zkuste ODHADNOUT, jestli dochází ke srážení všech proteinů nebo pouze některého z nich a zoufalému studentovi pomozte najít řešení.**

Izoelektrický bod - pI

Protein 1:

PTEFLYTSKIAAISWAATGGRQQRVYFQDLNGKIREAQRGGDNPWTGGSSQNVIGEAKLFSPLAAVTWKSAQGIQIRV
YCVNKDNILSEFVYDGSKWITGQLGSVGVKVGSNSKLAALQWGGSESAPPNIRVYYQKSNGSGSSIHEYVWSGKWTAG
ASFGSTVPGTGIGATAIGPGRLRIYYQATDNKIREHCWDSNSWYVGGFSASASAGVSIAAISWGSTPNIRVYWQKGRE
ELYEAAYGGSWNTPGQIKDASRPTPSLPDTFIAANSSGNIDISVFFQASGVSLQQWQWISGKGWSIGAVVPTGTPAGW

Protein 2:

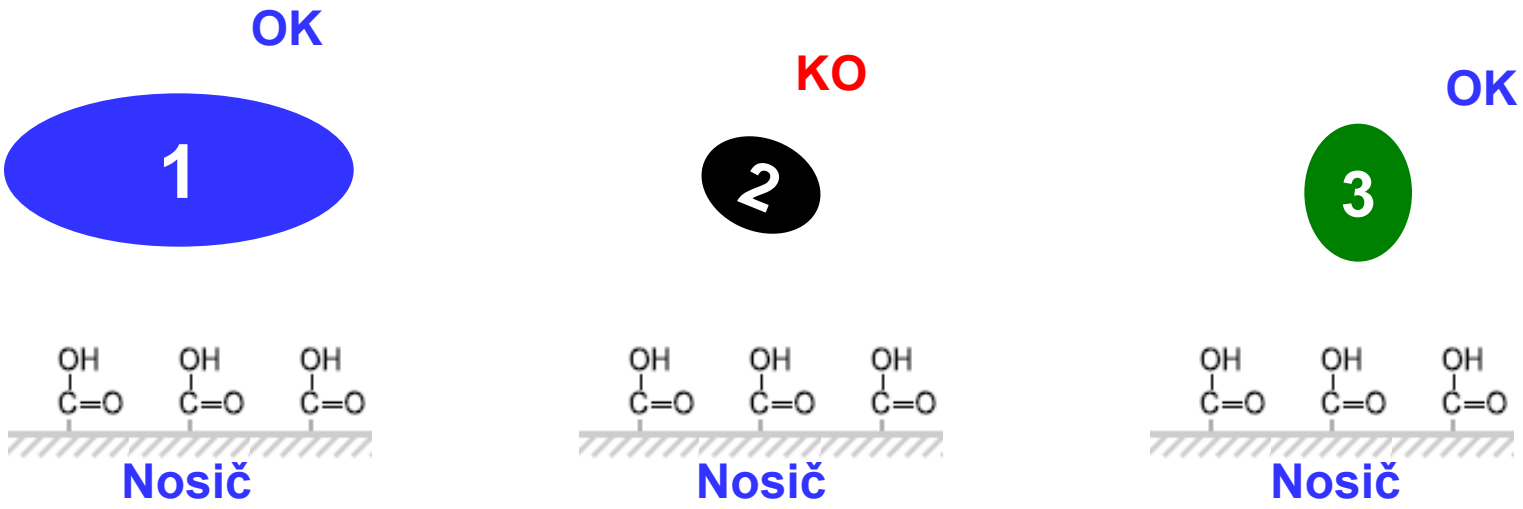
ATQGVFTLPANTFGVTAEFANESSGTQTVNVLVNNETAATFSGQSTNNAVIGTQVENSGSSGKVQVQVSVNGRPSDLV
SAQVILTNELNFALVGSEDDGTDNDYNDAVVVINWPLG

Protein 3:

SSVQTAATSWGTVPSIRVYTANNGKITERCWDGKGWYTGAFNEPGDNVSVTSWLVGSAIHIRVYASTGTTTTEWCWDG
NGWTKGAYTATN

Úkol 5: **Student potřeboval pro následné experimenty imobilizovat 3 proteiny na matrici (karboxymethylovaný dextran). Nechtělo se mu ptát se na radu kolegů a tak proteiny rozpustil v doporučovaném komerčním pufru (10 mM octan sodný, pH 5,0) a provedl imobilizace. U proteinů 1 a 3 byla úspěšná, u proteinu 2 naprosto selhala. „Proč?“, ptá se (opět) zoufalý student.**

Izoelektrický bod - pI



Úkol 5: **Student potřeboval pro následné experimenty imobilizovat 3 proteiny na nosič (karboxymethylovaný dextran). Nechtělo se mu ptát se na radu kolegů a tak proteiny rozpustil v doporučovaném komerčním pufru (10 mM octan sodný, pH 5,0) a provedl imobilizace. U proteinů 1 a 3 byla úspěšná, u proteinu 2 naprosto selhala. „Proč?“, ptá se (opět) zoufalý student.**

Extinkční koeficient

Extinction coefficients

The extinction coefficient indicates how much light a protein absorbs at a certain wavelength. It is useful to have an estimation of this coefficient for following a protein which a spectrophotometer when purifying it.

Experience shows that the computation is quite reliable for proteins containing Trp residues, however there may be more than 10% error for proteins without Trp residues.

Extinkční koeficient

Extinction coefficients
The extinction coefficient indicates how much light a protein absorbs at a certain wavelength. It is useful to have an estimation of this coefficient for following a protein which a spectrophotometer when purifying it.

- **Extinkční koeficienty závisejí na okolí chromoforu!**
- **ProtParam nebere v úvahu sekundární a terciární strukturu.**
- **Přesné extinkční koeficinty je nutné získat experimentálně.**

Experience shows that the computation is quite reliable for proteins containing Trp residues, however there may be more than 10% error for proteins without Trp residues.

Extinkční koeficient

Protein 1:

AQQGVFTLPARINFGVTVLVNSAATQHVEIFVDNEPRAAFSGVGTGDNNLGTKVINSGSGNVRVQITANGRQSDLVSS
QLVLANKLNLAVVGSEDGTDMDYNDSIVILNWPLG

Protein 2:

AWKGEVLANNEAGQVTSIIYNPGDVITIVAAGWASYGPTQKWGPQGDREHPDQGLICHDAFCGALVMKIGNSGTIPVN
TGLFRWVAPNNVQGAITLIYNDVPGTYGNNSGSFSVNIGKDQS

Protein 3:

SSVQTAATSWGTVPSIRVYTANNGKITERCWDGKGWYTGAFNEPGDNVSVTSWLVGSAIHIRVYASTGTTTTEWCWDG
NGWTKGAYTATN

Určené koeficienty jsou: 45 687, 7105, 27 860 M^{-1} cm^{-1}.

Úkol 6: Student experimentálně určil extinkční koeficienty tří proteinů při 280 nm. A potom si rozházel špatně popsané výsledky a neví, který koeficient patří ke kterému proteinu… Pomozte mu přiřadit jednotlivé koeficienty ke správným proteinům. Předpokládejte, že student už čeká před kanceláří vedoucího a nemůže použít počítač.

Extinkční koeficient

Protein 1:

AQQGVFTLPARINFGVTVLVNSAATQHVEIFVDNEPRAAFSGVGTGDNNLGTKVINSGSGNVRVQITANGRQSDLVSS
QLVLANKLNLAVVGSEDGTDMDYNDSIVILNWPLG

Protein 2:

AWKGEVLANNEAGQVTSIIYNPGDVITIVAAGWASYGPTQKWGPQGDREHPDQGLICHDAFCGALVMKIGNSGTIPVN
TGLFRWVAPNNVQGAITLIYNDVPGTYGNNSGSFSVNIGKDQS

Protein 3:

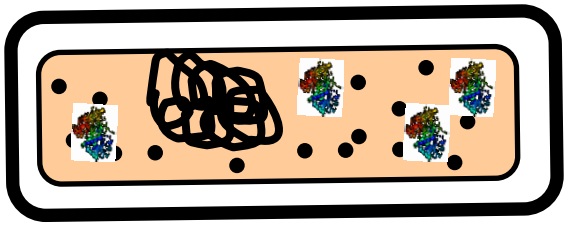
SSVQTAATSWGTVPSIRVYTANNGKITERCWDGKGWYTGAFNEPGDNVSVTSWLVGSAIHIRVYASTGTTTTEWCWDG
NGWTKGAYTATN

Určené koeficienty jsou: 45 687, 7105, 27 860 M^{-1} cm^{-1}.

Úkol 7: Stejná situace. Ale nyní předpokládejte, že student má internet v mobilu. (A že je příliš nervózní a nedokáže to odhadnout. Což je rychlejší.)

Jak stabilní je můj protein?

- Stabilita *in vivo* x *in vitro.*
- Stabilita v buňce x ve zkumavce.

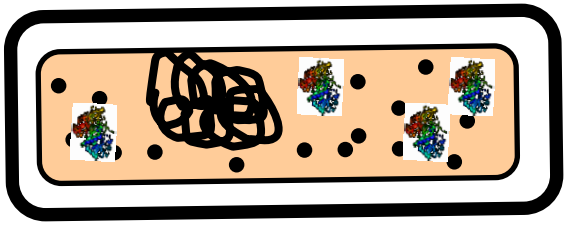


- Degradace proteinu v buňce je aktivní proces.
- *In vivo* half-life x instability index

Jak stabilní je můj protein?

- ***In vivo* half-life**

In vivo half-life
The half-life is a prediction of the time it takes for half of the amount of protein in a cell to disappear after its synthesis in the cell. ProtParam relies on the "N-end rule", which relates the half-life of a protein to the identity of its N-terminal residue; the prediction is given for 3 model organisms (human, yeast and E.coli). The N-end rule (for a review see [5],[6]) originated from the observations that the identity of the N-terminal residue of a protein plays an important role in determining its stability in vivo ([2], [3],[4]). The rule was established from experiments that explored the metabolic fate of artificial beta-galactosidase proteins with different N-terminal amino acids engineered by site-directed mutagenesis. The beta-gal proteins thus designed have strikingly different half-lives in vivo, from more than 100 hours to less than 2 minutes, depending on the nature of the amino acid at the amino terminus and on the experimental model (yeast in vivo; mammalian reticulocytes in vitro, Escherichia coli in vivo). In addition, it has been shown that in eukaryotes, the association of a destabilizing N-terminal residue and of an internal lysine targets the protein to ubiquitin-mediated proteolytic degradation [6]. Note that the program gives an **estimation** of the protein half-life and is not applicable for N-terminally modified proteins.



Jak stabilní je můj protein?

Úkol 8: Predikujte in vivo half-life následujících proteinů:

Protein 1:

MAQQGVFTLPARINFGVTVLVNSAATQHVEIFVDNEPRAAFSGVGTGDNNLGTKVINSGSGNVRVQITANGRQSDLVS
SQLVLANKLNLAVVGSEDGTDMDYNDSIVILNWPLG

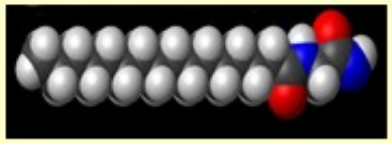
Protein 2:

MDRNGNFSLPPNTAFKAIFYANAADRQDLKLFIDDAPEPAATFVGNSEDGVRLFTLNSKGGKIRIEASANGRQSATDA
RLAPLSAGDTVWLGWLGAEDGADADYNDGIVILQWPIT

Protein 3:

MERDGTFNLPPHIKFGVTALTHAANDQTIDIYIDDDPKPAATFKGAGAQDQNLGTKVLDSGNGRVRVIVMANGRPSRL
GSRQVDIFKKSYFGIIGSEDGADDDYNDGIVFLNWPLG

Odštěpuje se iniciační methionin?



TermiNator

ACI IMPBIO

Termi*N*ator predicts N-terminal methionine excision, N-terminal acetylation, N-terminal myristoylation and S-palmitoylation of either prokaryotic or eukaryotic proteins originating from organellar or nuclear genomes .

Protein 1:

M AQQGVFTLPARINFGVTVLVNSAATQHVEIFVDNEPRAAFSGVGTGDNNLGTKVINSGSGNVRVQITANGRQSDL VSSQLVLANKLNLAVVGSEDGTDMDYNDSIVILNWPLG

Protein 2:

M DRNGNFSLPPNTAFKAIFYANAADRQDLKLFIDDAPEPAATFVGNSEDGVRLFTLNSKGGKIRIEASANGRQSATD ARLAPLSAGDTVWLGWLGAEDGADADYNDGIVILQWPIT

Protein 3:

M ERDGTFNLPPHIKFGVTALTHAANDQTIDIYIDDDPKPAATFKGAGAQDQNLGTKVLDSGNGRVRVIVMANGRPSR LGSRQVDIFKKSYFGIIGSEDGADDDYNDGIVFLNWPLG

Jak stabilní je můj protein?

- **Instability index**

Instability index (II)

The instability index provides an estimate of the stability of your protein in a test tube. Statistical analysis of 12 unstable and 32 stable proteins has revealed [7] that there are certain dipeptides, the occurence of which is significantly different in the unstable proteins compared with those in the stable ones. The authors of this method have assigned a weight value of instability to each of the 400 different dipeptides (DIWV).

First amino acid of dipeptide	Second amino acid of dipeptide																		
	W	C	M	H	Y	F	Q	N	I	R	D	P	T	K	E	V	S	G	A	L
W	1.0	1.0	24.68	24.68	1.0	1.0	1.0	13.34	1.0	1.0	1.0	1.0	−14.03	1.0	1.0	−7.49	1.0	−9.37	−14.03	13.34
C	24.68	1.0	33.6	33.6	1.0	1.0	−6.54	1.0	1.0	1.0	20.26	20.26	33.6	1.0	1.0	−6.54	1.0	1.0	1.0	20.26
M	1.0	1.0	−1.88	58.28	24.68	1.0	−6.54	1.0	1.0	−6.54	1.0	44.94	−1.88	1.0	1.0	1.0	44.94	1.0	13.34	1.0
H	−1.88	1.0	1.0	1.0	44.94	−9.37	1.0	24.68	44.94	1.0	1.0	−1.88	−6.54	24.68	1.0	1.0	1.0	−9.37	1.0	1.0
Y	−9.37	1.0	44.94	13.34	13.34	1.0	1.0	1.0	1.0	−15.91	24.68	13.34	−7.49	1.0	−6.54	1.0	1.0	−7.49	24.68	1.0
F	1.0	1.0	1.0	1.0	33.6	1.0	1.0	1.0	1.0	1.0	13.34	20.26	1.0	−14.03	1.0	1.0	1.0	1.0	1.0	1.0
Q	1.0	−6.54	1.0	1.0	−6.54	−6.54	20.26	1.0	1.0	1.0	20.26	20.26	1.0	1.0	20.26	−6.54	44.94	1.0	1.0	1.0
N	−9.37	−1.88	1.0	1.0	1.0	−14.03	−6.54	1.0	44.94	1.0	1.0	−1.88	−7.49	24.68	1.0	1.0	1.0	−14.03	1.0	1.0
I	1.0	1.0	1.0	13.34	1.0	1.0	1.0	1.0	1.0	1.0	1.0	−1.88	1.0	−7.49	44.94	−7.49	1.0	1.0	1.0	20.26
R	58.28	1.0	1.0	20.26	−6.54	1.0	20.26	13.34	1.0	58.28	1.0	20.26	1.0	1.0	1.0	1.0	44.94	−7.49	1.0	1.0
D	1.0	1.0	1.0	1.0	1.0	−6.54	1.0	1.0	1.0	−6.54	1.0	1.0	−14.03	−7.49	1.0	1.0	20.26	1.0	1.0	1.0
P	−1.88	−6.54	−6.54	1.0	1.0	20.26	20.26	1.0	1.0	−6.54	−6.54	20.26	1.0	1.0	18.38	20.26	20.26	1.0	20.26	1.0
T	−14.03	1.0	1.0	1.0	1.0	13.34	−6.54	−14.03	1.0	1.0	1.0	1.0	1.0	1.0	20.26	1.0	1.0	−7.49	1.0	1.0
K	1.0	1.0	33.6	1.0	1.0	1.0	24.68	1.0	−7.49	33.6	1.0	−6.54	1.0	1.0	1.0	−7.49	1.0	−7.49	1.0	−7.49
E	−14.03	44.94	1.0	−6.54	1.0	1.0	20.26	1.0	20.26	1.0	20.26	20.26	1.0	1.0	33.6	1.0	20.26	1.0	1.0	1.0
V	1.0	1.0	1.0	1.0	−6.54	1.0	1.0	1.0	1.0	1.0	−14.03	20.26	−7.49	−1.88	1.0	1.0	1.0	−7.49	1.0	1.0
S	1.0	33.6	1.0	1.0	1.0	1.0	20.26	1.0	1.0	20.26	1.0	44.94	1.0	1.0	20.26	1.0	20.26	1.0	1.0	1.0
G	13.34	1.0	1.0	1.0	−7.49	1.0	1.0	−7.49	−7.49	1.0	1.0	1.0	−7.49	−7.49	−6.54	1.0	1.0	13.34	−7.49	1.0
A	1.0	44.94	1.0	−7.49	1.0	1.0	1.0	1.0	1.0	1.0	−7.49	20.26	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
L	24.68	1.0	1.0	1.0	1.0	1.0	33.6	1.0	1.0	20.26	1.0	20.26	1.0	−7.49	1.0	1.0	1.0	1.0	1.0	1.0

Jak stabilní je můj protein?

Protein 1:

SDVDIEAQDAGQTLVQVISIPSGETWVAIQLPSQYRYFDFVFENVSPTSSGSVLVAQMAPQSGGVYGSNYSGSGWGND
LGGGGFYGYSEAKWMCLWPANRSGPSSKTGLYGTCKLMNLNQSSAVPSVTSNLFAPTAYKNEPGYANVGGCCQKIRGL
ASSIQFAFALAGGNVPQNTDTFNGGTIKVYGWN

Protein 2:

LVIVDAVTLLSAYPEASRDPAAPTVIDGRHLYVVSPGDAAQLGHNDSRLFTGLSPGDQLHLRETALALRAEVSVLFIR
FALKDAGIVAPIELEVRDAATAVPDADDLLHPSCRPLKDHYWRSDVLAAGATTCTADFAVCDRDGTVSGYFRWETSIE
IAGSQPDTKQPGFKPSS

Protein 3:

PLLSASIVSAPVVTSETYVDIPGLYLDVAKAGIRDGKLQVILNVPTPYATGNNFPGIYFAIATNQGVVADGCFTYSSK
VPESTGRMPFTLVATIDVGSGVTFVKGQWKSVRGSAMHIDSYASLSAIWGTAAPSSQGSGNQGAETGGTGAGNIGGGG

Protein 4:

ADSQTSSNRAGEFSIPPNTDFRAIFFANAAEQQHIKLFIGDSQEPAAYHKLTTRDGPREATLNSGNGKIRFEVSVNGK
PSATDARLAPINGKKSDGSPFTVNFGIVVSEDGHDSDYNDGIVVLQWPIG

Úkol 9: **Student se má rozhodnout, se kterými proteiny bude pracovat příští dva roky v rámci diplomové práce. Poučen předchozími chybami se chce poradit se svými kolegy (s Vámi). Vyberte mu dva proteiny!**

Aliphatic index

Aliphatic index
The aliphatic index of a protein is defined as the relative volume occupied by aliphatic side chains (alanine, valine, isoleucine, and leucine). It may be regarded as a positive factor for the increase of thermostability of globular proteins.

Grand average of hydropathy

GRAVY (Grand Average of Hydropathy)
The GRAVY value for a peptide or protein is calculated as the sum of hydropathy values [9] of all the amino acids, divided by the number of residues in the sequence.

Amino acid scale values:

Ala: 1.800	Gly: -0.400	Pro: -1.600
Arg: -4.500	His: -3.200	Ser: -0.800
Asn: -3.500	Ile: 4.500	Thr: -0.700
Asp: -3.500	Leu: 3.800	Trp: -0.900
Cys: 2.500	Lys: -3.900	Tyr: -1.300
Gln: -3.500	Met: 1.900	Val: 4.200
Glu: -3.500	Phe: 2.800	

Hydrofobní/hydrofilní proteiny?

Membránové proteiny?

Grand average of hydropathy

Protein 1:

DPIALTAAVGADLLGDGRPETLWLGIGTLLMLIGTFYFIVKGWGVTDKEAREYYSITILVPGIASAAYLSMFFGIGLT
EVQVGSEMLDIYYARYADWLFTTPLLLLDLALLAKVDRVSIGTLVGVDALMIVTGLVGALSHTPLARYTWWLFSTICM
IVVLYFLATSLRAAAKERGPEVASTFNTLTALVLVLWTAYPILWIIGTEGAGVVGLGIETLLFMVLDVTAKVGFGFIL
LRSRAILGDTEAPEPSAGAEASAAD

Protein 2:

KLAVYSTKQYDKKYLQQVNESFGFELEFFDFLLTEKTAKTANGCEAVCIFVNDDGSRPVLEELKKHGVKYIALRCAGF
NNVDLDAAKELGLKVVRVPAYDPEAVAEHAIGMMMTLNRRIHRAYQRTRDANFSLEGLTGFTMYGKTAGVIGTGKIGV
AMLHILKGFGMRLLAFDPYPSAAALELGVEYVDLPTLFSESDVISLHCPLTPENYHLLNEAAFDQMKNGVMIVNTSRG
ALIDSQAAIEALKNQKIGSLGMDVYENERDLFFEDKSNDVIQDDVFRRLSACHNVLFTGHQAFLTAEALTSISQTTLQ
NLSNLEKGETCPNELV

Úkol 10: **Porovnejte typický membránový a cytoplasmatický protein.**

Grand average of hydropathy

Ala (A)	31	12.0%
Arg (R)	9	3.5%
Asn (N)	1	0.4%
Asp (D)	12	4.6%
Cys (C)	1	0.4%
Gln (Q)	1	0.4%
Glu (E)	12	4.6%
Gly (G)	26	10.0%
His (H)	1	0.4%
Ile (I)	19	7.3%
Leu (L)	41	15.8%
Lys (K)	5	1.9%
Met (M)	6	2.3%
Phe (F)	11	4.2%
Pro (P)	9	3.5%
Ser (S)	12	4.6%
Thr (T)	23	8.9%
Trp (W)	7	2.7%
Tyr (Y)	10	3.9%
Val (V)	22	8.5%
Pyl (O)	0	0.0%
Sec (U)	0	0.0%

Aliphatic index: 126.95

Grand average of hydropathicity (GRAVY): 0.812

Ala (A)	30	9.1%
Arg (R)	13	4.0%
Asn (N)	18	5.5%
Asp (D)	19	5.8%
Cys (C)	6	1.8%
Gln (Q)	11	3.3%
Glu (E)	24	7.3%
Gly (G)	23	7.0%
His (H)	8	2.4%
Ile (I)	14	4.3%
Leu (L)	38	11.6%
Lys (K)	19	5.8%
Met (M)	10	3.0%
Phe (F)	18	5.5%
Pro (P)	9	2.7%
Ser (S)	16	4.9%
Thr (T)	19	5.8%
Trp (W)	0	0.0%
Tyr (Y)	11	3.3%
Val (V)	23	7.0%
Pyl (O)	0	0.0%
Sec (U)	0	0.0%

Aliphatic index: 91.03

Grand average of hydropathicity (GRAVY): -0.097

PSORT

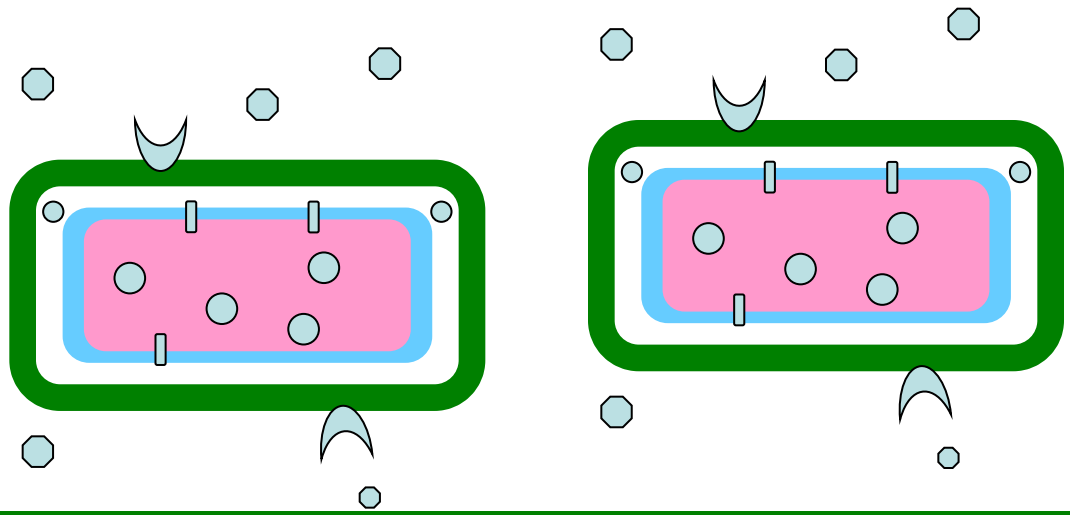
PSORT.org provides links to the PSORT family of programs for subcellular localization prediction as well as other datasets and resources relevant to localization prediction. The page is currently hosted by the Brinkman Laboratory at Simon Fraser University, and our goal is to provide an open-source resource centre for researchers interested in subcellular localization prediction.

- **Predikce lokalizace proteinů v buňce (prokaryotická i eukaryotická).**
- **Lokalizace proteinů napomáhá určení (ověření) jejich funkce.**
- **Vypovídá o předpokládaných vlastnostech proteinů (cytoplasmatické x membránové).**

PSORT

PSORT.org provides links to the PSORT family of programs for subcellular localization prediction as well as other datasets and resources relevant to localization prediction. The page is currently hosted by the Brinkman Laboratory at Simon Fraser University, and our goal is to provide an open-source resource centre for researchers interested in subcellular localization prediction.

Computational prediction of the subcellular localization of proteins is a valuable tool for genome analysis and annotation, since a protein's subcellular localization can provide clues regarding its function in an organism. For bacterial pathogens, the prediction of proteins on the cell surface is of particular interest due to the potential of such proteins to be primary drug or vaccine targets. A protein's subcellular localization is influenced by several features present within the protein's primary structure, such as the presence of a signal peptide or membrane-spanning alpha-helices.



PSORT

PSORT.org provides links to the PSORT family of programs for subcellular localization prediction as well as other datasets and resources relevant to localization prediction. The page is currently hosted by the Brinkman Laboratory at Simon Fraser University, and our goal is to provide an open-source resource centre for researchers interested in subcellular localization prediction.

BIOINFORMATICS ORIGINAL PAPER

Vol. 26 no. 13 2010, pages 1608–1615
doi:10.1093/bioinformatics/btq249

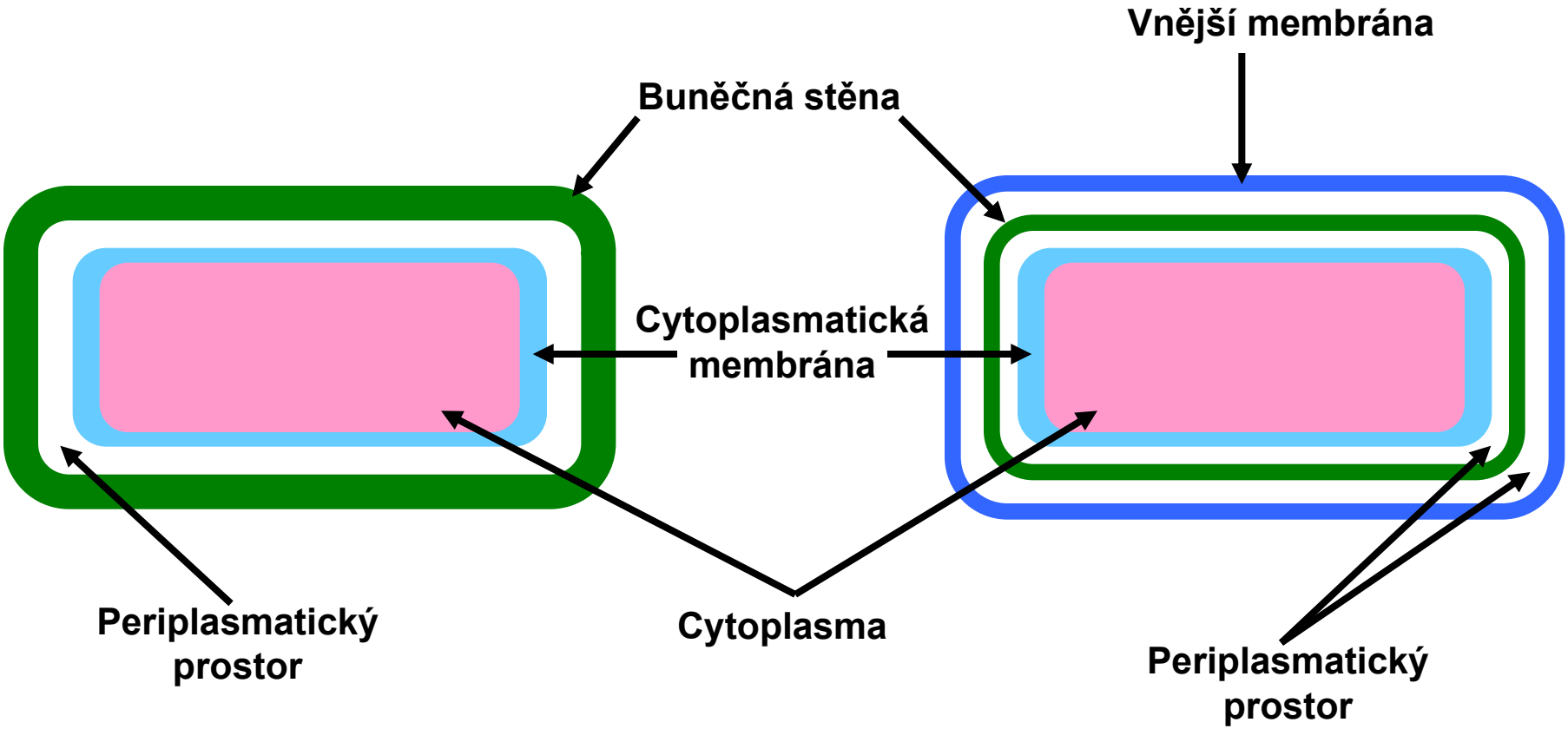
Sequence analysis

Advance Access publication May 13, 2010

PSORTb 3.0: improved protein subcellular localization prediction with refined localization subcategories and predictive capabilities for all prokaryotes

Nancy Y. Yu[1], James R. Wagner[2,†], Matthew R. Laird[1], Gabor Melli[2], Sébastien Rey[1], Raymond Lo[1], Phuong Dao[2], S. Cenk Sahinalp[2], Martin Ester[2], Leonard J. Foster[3] and Fiona S. L. Brinkman[1,*]

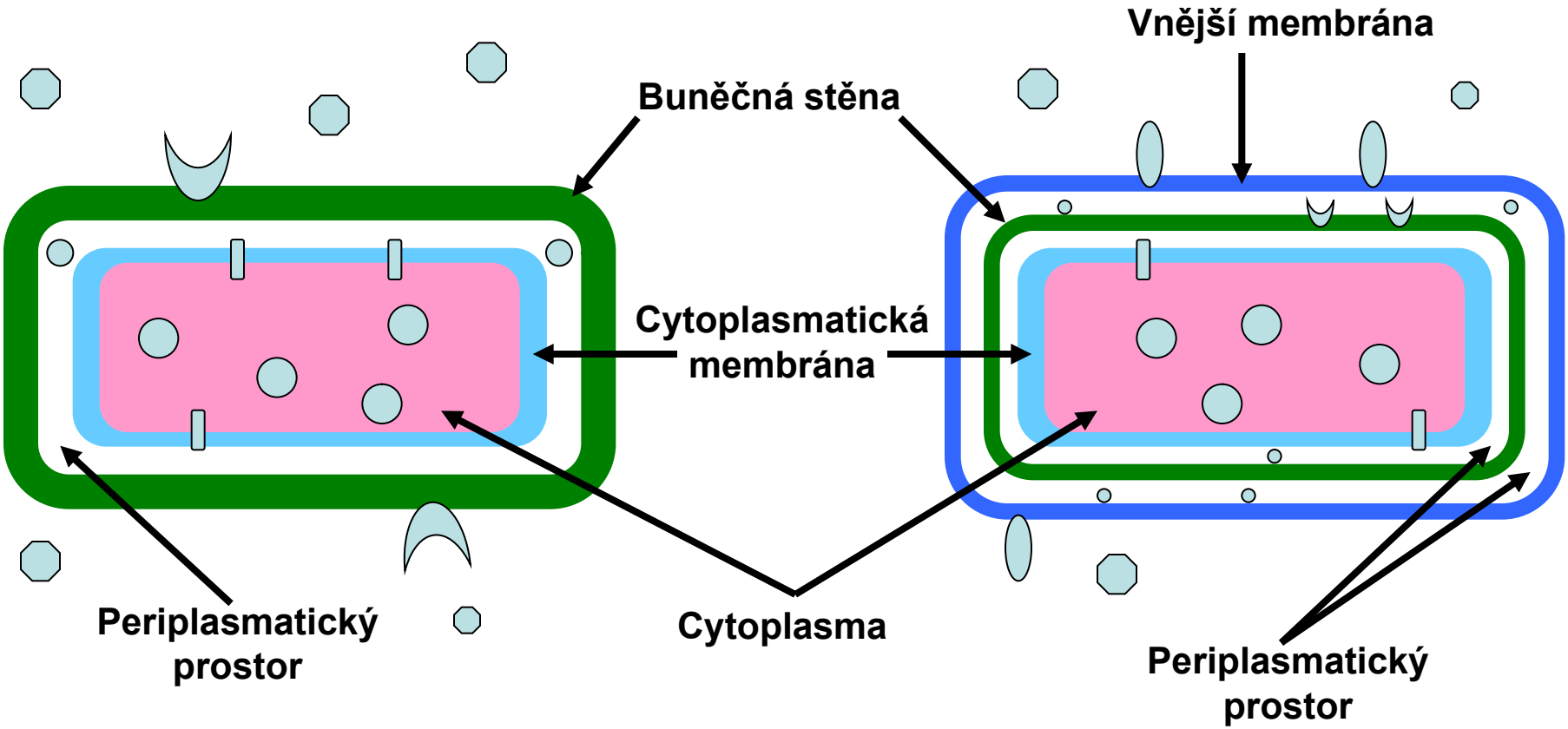
PSORT



Gram pozitivní

Gram negativní

PSORT



Gram pozitivní

Gram negativní

PSORT

You can currently submit one or more Gram-positive or Gram-negative bacterial sequences or archaeal sequences in FASTA format (?). Copy and paste your FASTA-formatted sequences into the textbox below or select a file containing your sequences to upload from your computer.

Choose an organism type (?): Bacteria **Required**

Choose Gram stain (?): Negative **Required**

Output format (?): Normal

Show results (?): Via the web

Copy and paste your FASTA sequences below

```
>cokolivnaprvnimradku
ADSQTSSNRAGEFSIPPNTDFRAIFFANAAEQQHIKLFIGDSQEPAAYHKLTTRDGPREATLNSGNGKIRFEVSVNGKP
SATDARLAPINGKKSDGSPFTVNFGIVVSEDGHDSDYNDGIVVLQWPIG
```

PSORT

Protein 1:

DPIALTAAVGADLLGDGRPETLWLGIGTLLMLIGTFYFIVKGWGVTDKEAREYYSITILVPGIASAAYLSMFFGIGLT
EVQVGSEMLDIYYARYADWLFTTPLLLLDLALLAKVDRVSIGTLVGVDALMIVTGLVGALSHTPLARYTWWLFSTICM
IVVLYFLATSLRAAAKERGPEVASTFNTLTALVLVLWTAYPILWIIGTEGAGVVGLGIETLLFMVLDVTAKVGFGFIL
LRSRAILGDTEAPEPSAGAEASAAD

Protein 2:

KLAVYSTKQYDKKYLQQVNESFGFELEFFDFLLTEKTAKTANGCEAVCIFVNDDGSRPVLEELKKHGVKYIALRCAGF
NNVDLDAAKELGLKVVRVPAYDPEAVAEHAIGMMMTLNRRIHRAYQRTRDANFSLEGLTGFTMYGKTAGVIGTGKIGV
AMLHILKGFGMRLLAFDPYPSAAALELGVEYVDLPTLFSESDVISLHCPLTPENYHLLNEAAFDQMKNGVMIVNTSRG
ALIDSQAAIEALKNQKIGSLGMDVYENERDLFFEDKSNDVIQDDVFRRLSACHNVLFTGHQAFLTAEALTSISQTTLQ
NLSNLEKGETCPNELV

Úkol 11: Analyzujte proteiny z Úkolu 10 pomocí nástroje PSORT. Oba proteiny pocházejí z Gram negativních bakterií.

PSORT

PSORTb Results (Click here for an explanation of the output formats)

```
SeqID: cokoliv
  Analysis Report:
    CMSVM-              CytoplasmicMembrane          [No details]
    CytoSVM-            Unknown                      [No details]
    ECSVM-              Unknown                      [No details]
    ModHMM-             CytoplasmicMembrane          [7 internal helices found]
    Motif-              Unknown                      [No motifs found]
    OMPMotif-           Unknown                      [No motifs found]
    OMSVM-              Unknown                      [No details]
    PPSVM-              Unknown                      [No details]
    Profile-            Unknown                      [No matches to profiles found]
    SCL-BLAST-          Unknown                      [No matches against database]
    SCL-BLASTe-         Unknown                      [No matches against database]
    Signal-             Unknown                      [No signal peptide detected]
  Localization Scores:
    Cytoplasmic             0.00
    CytoplasmicMembrane    10.00
    Periplasmic             0.00
    OuterMembrane           0.00
    Extracellular           0.00
  Final Prediction:
    CytoplasmicMembrane    10.00
--------------------------------------------------------------------------------
```

PSORT

PSORTb Results (Click here for an explanation of the output formats)

```
SeqID: cokoliv
  Analysis Report:
    CMSVM-              Unknown                    [No details]
    CytoSVM-            Cytoplasmic                [No details]
    ECSVM-              Unknown                    [No details]
    ModHMM-             Unknown                    [No internal helices found]
    Motif-              Unknown                    [No motifs found]
    OMPMotif-           Unknown                    [No motifs found]
    OMSVM-              Unknown                    [No details]
    PPSVM-              Unknown                    [No details]
    Profile-            Unknown                    [No matches to profiles found]
    SCL-BLAST-          Cytoplasmic                [matched 27461218: Cytoplasmic protein]
    SCL-BLASTe-         Unknown                    [No matches against database]
    Signal-             Unknown                    [No signal peptide detected]
  Localization Scores:
    Cytoplasmic             9.97
    CytoplasmicMembrane     0.01
    Periplasmic             0.01
    OuterMembrane           0.00
    Extracellular           0.00
  Final Prediction:
    Cytoplasmic             9.97
--------------------------------------------------------------------------------
```

PSORT

Protein 1:

MKYKTVKSIPLFLLGSIVFTACSTPQSTFHLPVQTTVSAIKKDISGKTATAVKAASSSSSTTTSNDDNNQ
KGYFLETNRSTGTYDPNNSTRLIKLGESGDFHAADQNKPEEALFERLYGGIASLLNFRIIKPALTYWNTV
TPSLKAIGKSSNLITFSQDIDETELQRALANNLIVADDGNNNFWFGLKSLSFNSAKLTDNAQTQMAQKTT
QAVTLKSQAQMSSTNTKNTNKKIDLRDKITLSSTMNTQGSGDNKNPSSGLIQKLVSVENIEAEFSFVKTG
FNGNEIKFGDFVTENSPTTTQLKQVWKKKWGTELKKTNYKLQLNNFSLLLTYTPEVNKVEKGNNGDSNKG
TIATPNGFSFLYPANLNETPSSSSSYWTNVTDLTKAATDTENTNLLNDLQKSQEQVNQFVAAITQNHLDV
SEAALTKKQFGSLSISDFFKAIFKENGKDTKAKS

Úkol 12: Student má za úkol izolovat zajímavý (pro vedoucího) protein z Gram negativní bakterie. Po několikadenním pěstování čtyř litrů kultury se student snažil získat pouze z buněk protein (další týden práce), ale získal mizivé množství… Jeho kolegyně si myslí, že se jedná o membránový protein, který je labilní a nevydrží proces izolace. Ověřte její teorii…

PSORT

Protein 1:

```
LVIVDAVTLLSAYPEASRDPAAPTVIDGRHLYVVSPGDAAQLGHNDSRLFTGLSP
GDQLHLRETALALRAEVSVLFIRFALKDAGIVAPIELEVRDAATAVPDADDLLHP
SCRPLKDHYWRSDVLAAGATTCTADFAVCDRDGTVSGYFRWETSIEIAGSQPDTK
QPGFKPSS
```

Úkol 13: **Student má za úkol charakterizovat a navrhnout možnou funkci proteinu z *Burkholderia cenocepacia*. Pomozte mu.**

WoLF PSORT

You will live to see your grandchildren.

Expect a letter from a friend who will ask a favor of you.

This will be a memorable month -- no matter how hard you try to forget it.

You will pioneer the first Martian colony.

WoLF PSORT predicts the subcellular localization sites of proteins based on their amino acid sequences. The method, which is a major extension to the venerable PSORTII program, makes predictions based on both known sorting signal motifs and some correlative sequence features such as amino acid content. Like PSORT and PSORTII, WoLF PSORT displays some information about detected sorting signals which is useful in helping users determine the reliability of the prediction in specific cases. Our experiments (presented at APBC06) show that the overall prediction accuracy of WoLF PSORT is over 80%. For common localization sites (e.g. cytosol, nucleus, mitochondria, etc) WoLF PSORT makes better than majority classifier predictions even for queries that do not have strong sequence similarity to any sequence in the dataset. Thus WoLF PSORT is a useful complement to tools such as BLAST. The current dataset used to train WoLF PSORT contains over 12,000 animal sequences and more than 2,000 plant and fungi sequences respectively. It was gathered mainly from Uniprot but several hundred *Arabidopsis thaliana* sequences from the Gene Ontology database were also included.

You attempt things that you do not even plan because of your extreme stupidity.

WoLF PSORT

What's in a name

"WoLF" does not necessarily stand for anything. A rather dramatic mnemonic would be "Where Life Functions". Originally it was going to be "Learned Weight Features" but I wanted the acronym to be a pronouncable English word. Women only Love Fools.

Please select an organism type:
- ○ Animal
- ○ Plant
- ○ Fungi

Abbrev	Localization Site
chlo	chloroplast
cyto	cytosol
cysk	cytoskeleton
E.R.	endoplasmic reticulum
extr	extracellular
golg	Golgi apparatus
lyso	lysosome
mito	mitochondria
nucl	nuclear
pero	peroxisome
plas	plasma membrane
vacu	vacuolar membrane

- **Predikce lokalizace proteinů v eukaryotických buňkách (živočichové, rostliny, houby).**
- **Mnohem více možných lokalizací proteinů!**

WoLF PSORT

Protein 1:

```
MKWLLLLGLVALSECIMYKVPLIRKKSLRRTLSERGLLKDFLKKHNLNPARKYFPQWEAPTLVDEQPLENYLDMEYFG
TIGIGTPAQDFTVVFDTGSSNLWVPSVYCSSLACTNHNRFNPEDSSTYQSTSETVSITYGTGSMTGILGYDTVQVGGI
SDTNQIFGLSETEPGSFLYYAPFDGILGLAYPSISSSGATPVFDNIWNQGLVSQDLFSVYLSADDQSGSVVIFGGIDS
SYYTGSLNWVPVTVEGYWQITVDSITMNGEAIACAEGCQAIVDTGTSLLTGPTSPIANIQSDIGASENSDGDMVVSCS
AISSLPDIVFTINGVQYPVPPSAYILQSEGSCISGFQGMNLPTESGELWILGDVFIRQYFTVFDRANNQVGLAPVA
```

Protein 2:

```
RKSTGGKAPRKQLATKAARKSAPATGGVKKPHRYRPGTVALREIRRYQKSTELLIRKLPFQRLVREIAQD
FKTDLRFQSSAVMALQEASEAYLVGLFEDTNLCAIHAKR
```

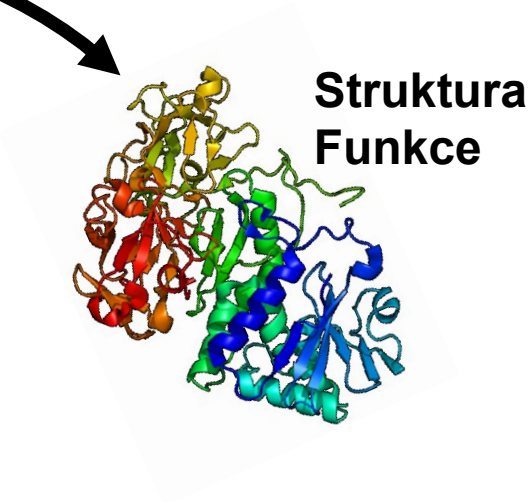
Úkol 14: **Predikujte možnou lokalizaci proteinu z *Homo sapiens* a zkuste predikovat lokalizaci proteinu z *Vampyroteuthis infernalis*, i když je k dispozici jen fragment proteinu.**

Zisk informací o známých proteinech

Protein je správně sbalený, aktivní, v dostatečném množství a koncentraci

Optimální podmínky

Gen → Protein



Struktura
Funkce

- **Mnoho proteinů již bylo popsáno (sekvence, struktura, funkce).**
- **Informace o nich jsou sdruženy v databázích.**

Bioinformatická centra

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

EBI/NCBI/CIB

EBI

Evropský institut pro bioinformatiku

European Bioinformatics Institute

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

NCBI

Národní centrum pro biotechnologické informace

National Center for Biotechnology Information

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

CIB

Centrum pro informační biologii

Center for Information Biology

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

Bioinformatická centra

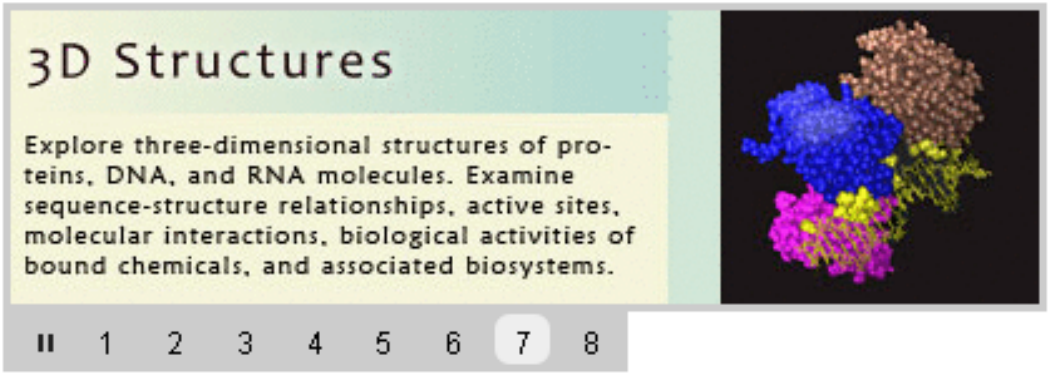
Welcome to NCBI

The National Center for Biotechnology Information advances science and health by providing access to biomedical and genomic information.

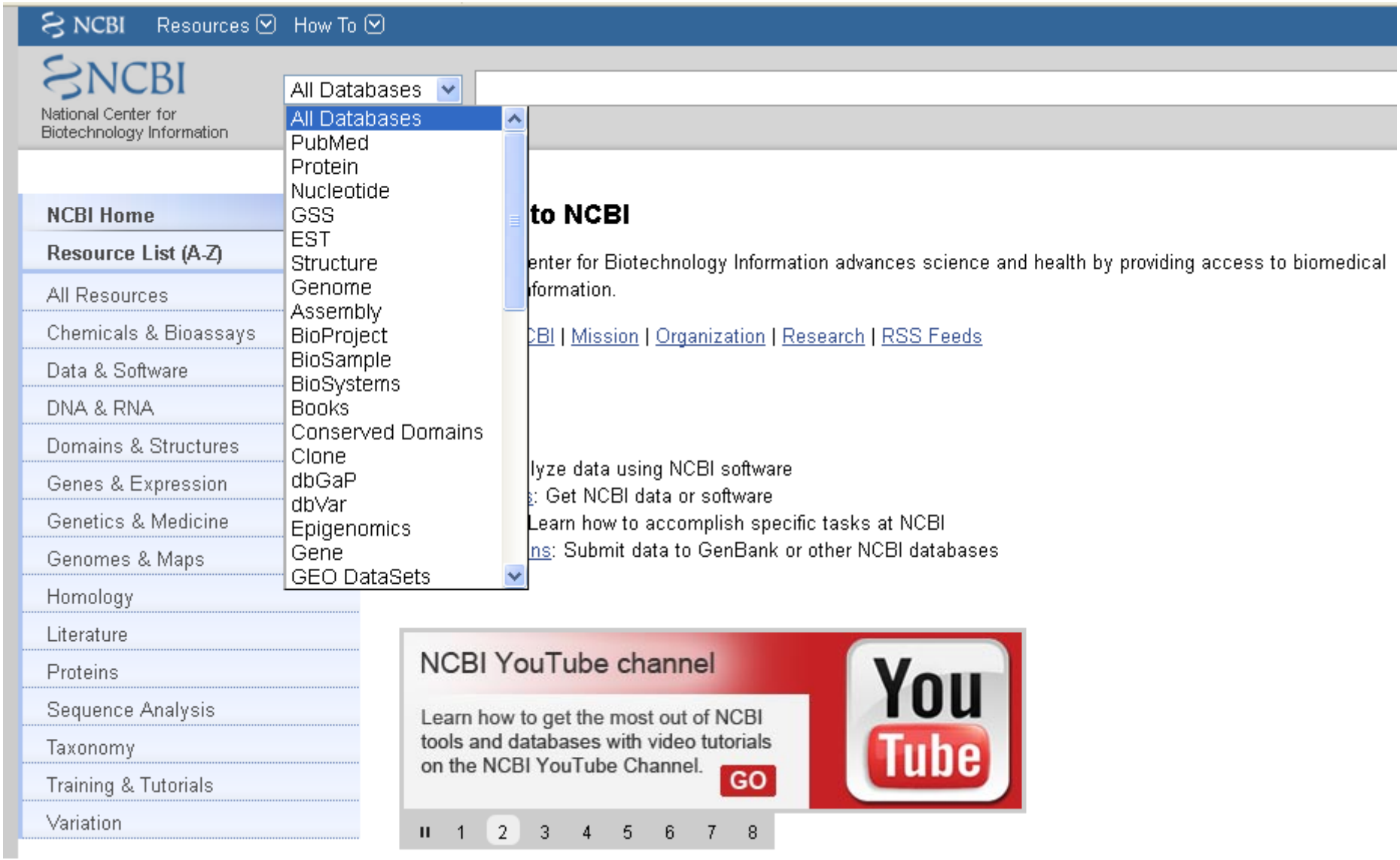
About the NCBI | Mission | Organization | Research | RSS Feeds

Get Started

- Tools: Analyze data using NCBI software
- Downloads: Get NCBI data or software
- How-To's: Learn how to accomplish specific tasks at NCBI
- Submissions: Submit data to GenBank or other NCBI databases



Databáze


NCBI
Resources
How To
NCBI
National Center for Biotechnology Information
All Databases
All Databases
PubMed
Protein
Nucleotide
GSS
EST
Structure
Genome
Assembly
BioProject
BioSample
BioSystems
Books
Conserved Domains
Clone
dbGaP
dbVar
Epigenomics
Gene
GEO DataSets
NCBI Home
Resource List (A-Z)
All Resources
Chemicals & Bioassays
Data & Software
DNA & RNA
Domains & Structures
Genes & Expression
Genetics & Medicine
Genomes & Maps
Homology
Literature
Proteins
Sequence Analysis
Taxonomy
Training & Tutorials
Variation
to NCBI
enter for Biotechnology Information advances science and health by providing access to biomedical
formation.
CBI | Mission | Organization | Research | RSS Feeds
lyze data using NCBI software
: Get NCBI data or software
Learn how to accomplish specific tasks at NCBI
ns: Submit data to GenBank or other NCBI databases
NCBI YouTube channel
Learn how to get the most out of NCBI tools and databases with video tutorials on the NCBI YouTube Channel.
GO
You Tube
1 2 3 4 5 6 7 8

Vyhledávací systém

Vyhledávací systém

- **Textové vyhledávání může selhat (nedostatečná anotace).**
- **Vyskytuje se shodná nebo podobná sekvence v databázi? (Identifikace možné funkce na základě homologie.)**
- **Specializované nástroje (algoritmy) pro „seřazení“ (alignment) sekvencí.**

RKSTGGKAPRKQLATKAARKSAPATGGV
KKPHRYRPGTVALREIRRYQKSTELLIR
KLPFQRLVREIAQDFKTDLRFQSSAVMA
LQEASEAYLVGLFEDTNLCAIHAKR

Podobné sekvence…

BLAST

The Basic Local Alignment Search Tool (BLAST) finds regions of local similarity between sequences. The program compares nucleotide or protein sequences to sequence databases and calculates the statistical significance of matches. BLAST can be used to infer functional and evolutionary relationships between sequences as well as help identify members of gene families.

Basic BLAST

Choose a BLAST program to run.

nucleotide blast	Search a **nucleotide** database using a **nucleotide** query *Algorithms:* blastn, megablast, discontiguous megablast
protein blast	Search **protein** database using a **protein** query *Algorithms:* blastp, psi-blast, phi-blast, delta-blast
blastx	Search **protein** database using a **translated nucleotide** query
tblastn	Search **translated nucleotide** database using a **protein** query
tblastx	Search **translated nucleotide** database using a **translated nucleotide** query

BLAST

Protein 2:

RKSTGGKAPRKQLATKAARKSAPATGGVKKPHRYRPGTVALREIRRYQKSTELLIRKLPFQ
RLVREIAQDFKTDLRFQSSAVMALQEASEAYLVGLFEDTNLCAIHAKR

Úkol 15: **Použijte BLAST a pokuste se blíže určit funkci proteinu z *Vampyroteuthis infernalis.***

Použitá literatura

Ramadevi Mohan, Subhashree Venugopal. Computational structural and functional analysis of hypothetical proteins of *Staphylococcus aureus, Bioinformation* 8(15): 722-728, 2012.

ExPASy server: http://www.expasy.org

ProtParam dokumnetace: http://web.expasy.org/protparam/protparam-doc.html

Elisabeth Gasteiger *et al.* Protein Identification and Analysis Tools on the ExPASy Server, in John M. Walker (ed): The Proteomics Protocols Handbook, Humana Press, 571-607, 2005.

Kunchur Guruprasad *et al.* Correlation between stability of a protein and its dipeptide composition: a novel approach for predicting in vivo stability of a protein from its primary sequence, Protein Engineering 4(2): 155-161, 1990.

PSORTb dokumentace: http://www.psort.org/documentation/index.html

WoLF PSORT dokumentace: http://wolfpsort.org/aboutWoLF_PSORT.html.en

Doporučená literatura

Panu Artimo *et al.* ExPASy: SIB bioinformatics resource portal, *Nucleic Acids Research* 40: W597-W603, 2012.

Christopher T. Walsh *et al.* Protein Posttranslational Modifications: The Chemistry of Proteome Diversifications, *Angewandte Chemie (International ed. in English)* 44: 7342-7372, 2005.

Jagat S. Chauhan *et al.* GlycoPP: A Webserver for Prediction of N- and O- Glycosites in Prokaryotic Protein Sequences, *PLoS ONE* 7(7): 1-13, 2012.

Nancy Y. Yu *et al.* PSORTb 3.0: improved protein subcellular localization prediction with refined localization subcategories and predictive capabilities for all prokaryotes, *Bioinformatics* 26(13): 1608–1615, 2010.

Paul Horton *et al.* WoLF PSORT: protein localization predictor, *Nucleic Acids Research*, 2007.

Stephen F. Altschul. BLAST Algorithm, in Encyclopedia of Life Sciences (ELS), John Wiley & Sons, Ltd: Chichester, 2005.