

# BIOINFORMATIKA V PRAXI – CVIČENÍ 3

## IDENTIFIKACE GENŮ, PROTEINŮ A JEJICH FUNKCE

### STUDIJNÍ MATERIÁLY

Studijní materiály předmětu C2130 Úvod do chemoinformatiky a bioinformatiky, přednáška Predikce genu, Sequence-evolution-function: Computational Approaches in Comparative Genomics.

### PRODUKCE BIOINFORMATICKÝCH DAT

Automatické sekvencování produkuje obrovské množství biologických dat (dlouhé sekvence DNA až celé genomy).

Produkce „syrových“ biologických dat, nutná identifikace a anotace genů a proteinů.

„Alignment“ na úrovni proteinů je pro určení funkce genů užitečnější než „alignment“ na úrovni DNA.

Určení čtecího rámce, překlad DNA  $\Rightarrow$  protein, identifikace genu a jeho funkce = základní úkony v bioinformatice i molekulární biologii.

### GENETICKÝ KÓD

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

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

### ÚKOL 1

Přeložte následující DNA sekvenci do aminokyselinové sekvence s využitím tří čtecích rámců:

**aatgccggaaagcgcgcgctaa**

## ÚKOL 2

Přeložte následující DNA sekvenci do aminokyselinové sekvence s využitím **šesti** čtecích rámců:

**atattctttgctaatatgcc**

## PREDIKCE GENŮ A PROTEINŮ U PROKARYOT

Prokaryotické geny jsou nepřerušované úseky DNA mezi startovním kodonem a stop kodonem. Nejčastěji je genem nejdelší ORF (open reading frame) odpovídající danému úseku DNA. Nejspolehlivějším ověřením, zda ORF skutečně kóduje protein, je jeho podobnost k již dříve popsanému proteinu („alignment“).

## ÚKOL 3

Přeložte následující DNA sekvence neznámého organismu do sekvencí aminokyselin pomocí programu **Translate** – server ExPassy (<http://www.expasy.ch/tools/dna.html>). Určete, které geny/proteiny mohou být těmito sekvencemi kódovány (aplikace BLAST - <http://blast.ncbi.nlm.nih.gov/Blast.cgi>).

### Sekvence 1

```
ATGGCTCAATCAGAATTTAACC AAAATCTACGGGAGCAAGGGCAATCAGTGCAGAGTTATTATTTCTACGTGTGAATAATCCGGGATACAACA  
CTAATACTTTAGACATCGCAGACATAGAAGATATCATTTCATTTACCTCAAGCTATAGAGTTAGCTAATGCATTTCAATCTGCACTAGTACCTAC  
TACATCCAACCTTCGGGGAGGACACATTAAGATTTGATGTAGAAAGGGGATTAGGAATAGCTACTCACGTTTATCCTAGAGCTATAAATGTGAAT  
TATGTAACACGCTACTTTATCACAACTAATAATCAAGTACAGTCCATGATAAATAAGGTGATAGAAGAGCTTAAATCTTTATTAGGAATTAATC  
TGCTAATTCAGTGTACAACAATTAACCACCTGTAATTACAGAACTTTTACAACCTCTACGTACAACAGCAATCTGCCTGGCTTTTTGGGG  
ACGCCAAACATCATCACAACCAATATACTTTATAATATCGTATTTGCAATCCAAAATGCCAAAACAGGTAGTTTTATGAAAGCTATTCCATA  
GGTTTTGAGATATCTGCATATATGCAAGGGAACGCTTATTATTTTTAATATCAAGATTATGCAAGCTATAGTGTCAAATACACGCAATTC  
AAGTTATGCAGCCATTGATTCATGAGAGTTTTCAACCCCTAAGAGGCATATTTAATATCATAACTCCGTAATAATAGAAAGTGCATTTCAAA  
TACTGAATATTATGACGAAAATACACATGATTTACCTGTTAAGTTATGGGACTATAATAATATTATTAATCAAAAATGGATTTTAGTTTTAAT  
CAAACCTACCCGTGCTTACAGTATTCAAAACCTAATAGCTCGCTATTTAGTACTAACATGGGATTTCTACTCCTGGATCAAATAAAGTTTTGCTT  
CTACTAATCGTTGGAACGATTCACAATCTGGATTTCTGAAAGTACAGCAGATGGTCCATTTTTTAACAAAATGAAAGACACTCAATTTGT  
ATTAGAGATAGAAAACCTAGTACTACTAATGGGACAAATGTTATAGTCAATAGAAAAATAAATATGCACAGCAAAAATTTACTTAATAAAA  
GTAACCAAGAGTTTCAGGATGGTGTATATAAGATTAAGACTGCCCTAAACAATAGTAGCGTTCTGCAGATGCTGAAGATTATTTGGTTATA  
CGTCAGATTACTTTGTGAAATTTAGGACCAACAATAAATGATATTAACCAAAAATGGATTTTTGAAATTTGATAGCACTAAATCTGCATCA  
AATTAAGAGTCAAAGGGATCCTTCTTAGTACTAGCATGGACTGGAGTGTACCTACTGTTAACTCCCTATTCCAAATAACGATGATCATCTC  
TGGTTTTCTCAAATGCTGGGTGGGTACCTATATTTTTGTAATATGACAGACACCCGATATGTTTTAGAAAGTGGCTAGTTCAGTACCACTA  
ATGGAACGATACTTACAATCAATAACGAAACGGAAATTTAAATCAAAAATTTTTGTTGGACATGATTAATTA
```

### Sekvence 2

```
ATGGCGCAGAGCGAATTTAACCAGAACCTGCGCGAACAGGGCCAGAGCCGCGCGCGCTGATTATTTCTGCGCGTGAACAACCCGGGCTATAACA  
CCAACACCCCTGGATATTCGGGATATGAGATATTTATTCATCTGCCGAGGCGATTGAACTGGCGAACCGGTTTTAGAGCGCGCTGGTGGCGAC  
CACCAGCAACTTTGGCGAAGATACCCTGCGCTTTGATGTGGAACGCGGCTGGGCATTGCGACCCATGTGTATCCGCGCGCGATTAAACGTGAAC  
TATGTGACCCCGACCCCTGAGCCAGACCAACAACAGGTGCAGAGCATGATTAACAAAGTATTGAAAGACTGAAAGCCCTGCTGGGCATTAACC  
TGGCGAACAGCGTGTGCAGCAGCTGACCACCGTATTACCGAACTTTACCAACCTGTATGTGCAGCAGCAGAGCGGCTGGCTGTTTTGGGG  
CCGCCAGACCAGCAGCCAGACCAACTATACCTATAACATTTGTGTTTGGATTCAGAACCGCAGACCCGCGAGCTTTATGAAAGCGATTCCGATT  
GGCTTTGAAATTAGCGCGTATATTCGCGCGAACGCGCTGCTGTTTTTAACATTCAGGATTTATGCGAGCTATAGCGTGAATAATCATGCGATT  
AGGTGATGCAGCCGTGATTCATGAAAGCTTTACGCCGCTGCGCGCATTTTTTAACATTATACCAGCGTGAACAACCCGAGCGGATTCAGAT  
TACCGAATATTTATGATGAAAACACCCATGATTATCCGGTGAACCTGTGGGATTATAACAACATTATTAACCAGAAATGGATTCTGGTGTAAAC  
CAGACCACCCGCGGTATAGCATTCAGAACCTGATTGCGCGCTATCTGGTGTGACCTGGGATAGCACCCCGGGCAGCAACAAGTGTGGCGA  
GCACCAACCGTGAACGATAGCCAGTTTTGGATTCTGGAAGCACCAGCGGATGGCAGCATTTTTCTGACCAACATGAAAGATACCCAGTTTTGT  
GCTGGAATTTGAAAACAGCAGCACCACCAACCGCACCAACCTGATTGTGAACCGCAAAAACAACAACCGCAGCAGAAATTTTATCTGAACAAA  
GTGAACAGGAATTTTCAGGATGGCGTGTATAAAATTAACACCGCGCTGAACAACAGCAGCGTGTGCAGATGAGCGAAGATTATTTGGCTATA  
CCAGCGATTATTTGTGAACTGTGACCAACAACAACAGTATTAACCAGAAATGGATTTTTGAAATTTGATAGCACCAAAAGCGCGTATCA  
GATTAAGAGCAGCGCGATCCGAGCCTGGTGTGGCGTGGACCTGGAGCGTGGCGACCGTGAACCTGCGGATCCGAACAACGATGATCATCTG  
TGTTTTCTGCAGAACCGGGCAGCGCACCTATTATTTGTGAACATCCGATACCCGCTATGTGCTGGAAGTGGCGAGCAGCAGCAGCAGCAGC  
ACGGCACCATTTGACCATTAACAACGCAACGGCAACCTGAACCAGAAATTTCTGCTGGATATGATTAACCTAAGTAGCTGAGCTGACGTGGCT  
GACGTGCTGAATGCTGCATGCGTAGCTGATGCTAGCTGACGCTATGGTGGCGAAAACATTAACCTGCAGGATATTTGATCCGGTGAACCCAG  
GAATGGAAGAAGCGCTGGATAGCGTGGTGAATATGAAAACCTGGATCGCGCGGAATTTATTTCTGGAAAAACCTGCTGGCGCATGCGCGCAACG  
TGGCGTGCAGCGTGGCCGATTCATACCCGCTATCTGAAACCATTCGGCGGAAACCGAAGCGAGCTGGCGGATGATGAAATTAAGAT  
GATGCAGCGCCTGACCAACTATCTGCGTGAACGCGCTGGCGATGGTGTGCGCGTGGCGCGCAAAAAGCGGGCTGGCGCGCATCTGAGC  
AGCTATGCGAGCATGGCGACCCGTGTTTGAAGTGGCCGTAACATTTTTTTTCGCGCGGATGATCTGGTGTTTTTTCAGGGCCATAGCGCGGAAG  
GCATTTATGCGCGCGCTTTCTGGAAGCGCGCTGAGCGAACCGCATGATTGATTTTCGCCAGGAAGCGCTGACCAAGGCATTTAGCAGCTA  
TCCGCTTCCGTTCTGATGCGGATTTTTGGCAGTTTTCCGAGTTTCCGACCGTGGATGGCCGCGCTGATGGCGAATTTATCAGGGCAGCTGCTG  
AAATATCTGCATCATCGAAACTGGTGGATACCGCGGGCCGAAAGTGTGGCGTTTTGCGCGGATGGCGAAAACCGCGCAACCGGAAACCCCTGG  
GCGCCCTGCTGGTGGCGAGCCGGAACCTGGATAACCTGATTTTTATTTGTGAACCTGCAACCTGCAGCGCTGGATGGCCCGGTGAGCGGCAA  
CGGCAAAATTTATCAGGAACCTGGAAGGCTGTTTCGCGCGCGGGCTGGCGCGTATTAAAGTATTGGGGCCAGGATGGGAACGCGCTGTTTT
```



TGCCGTTACTGCTCGTAGCGCTTAAAGACCAGCTCGTCTTGTGTGGAGGTTTCTTCATCAAAGAAATACCCCTCACGGTCAAACGCGGTAAGCT  
GTTCCGGCTTCGTTAAGCGGTTTTCAATAATAAAACGGCTCATCAGTCCGCGCGCTTTTTGGCGTAGAAGCTTACCACCTTAAACTTGCCGTT  
TTTCTCATCAAGGAACACGGGCTTAATCAGTTCCGGCTTCAGTTTCTTCGGCTTACCGGATTTAAAAATATTCCTCGGAGGCCAGATTACCACC  
ACCCGATCGCCCTGCGCTCAAGCGCTTCGTTGAGCTTATCGGTAATGATATCGCCCCAGAATTGATAAAGATCTTGGCCGCGGGATTCTCCA  
GGCGAATCCCATCTCCAGACGATAAAGCTGCATTAATCCAGCGGGCCAAATACGCCATACAAGCCAGAGAGCATACGAGATGTTGTTGAGC  
AAAATCAAATCCGCGTCTTGAACGTTTCCGCTGTAGGCCGGTATAAACATCGCTTTGAACGCCAGAATCGCTGGCGTGCATTATCCGGC  
GTAAAATGAGGCTGCCAGTCATGAAAACGCGTGGCGTTGAGATCCGCCAGTTTGTGCGTAATCCCATCAGCTGAAAATTTGGCGCGCCGAAA  
GCTGGCGGGCTGTTGAATAAGCTGCTGGCTGTGATCCAACAGCTCCGGCTGGGTATAGCGGGTCTGGCCAGCGGGCTTGTATAATCAAGCGT  
TTTTGCAGGTGAATCAGAATCAGCATATTCAGTCCTTGCAGGGAATTTCTGCGACTTAGCAAAAAACGCCCGAGAGTTGACCGATGGTTG  
CGATTGTGGCTTAATCGCGCGATGCCGCTCCAGGTATCTGGCGCCAGTTGTGGTTCGATATCCGGGAAGCGCCGCGGATCGAACTGCGGT  
TTACGCCAGTTTCCGTTGTGCGAGATAATCGCTGGCGAGGGTATACACCACCGGTGAGAGCAACAAAATCGCCGTCAGATTGGTAATGGCCAT  
ACAGGCCATGATCATGTGACGAGCTGCCATATCAGCGGAAAACGTATAAGCGTACCGGGGATAACCATGCCAAGCGTCGCAAGCGGTAATAGC  
CAGATAGCTTTGCGTTATGTAACCGCAGAAAAACAGATTGTTTTCGGCGTAAATATAGTTGGCGACGATAGAACTGAACGCGAACAGAAATGA  
CGATAAGCGCGCAAAAACGTGGCGCCCATTCACCGGTCAACGAAACATTGCATGTTGGAGAAGCTGAATGCCTTCTGTTGACGAGTGGGACGC  
GTGACAAAGCGCAGCAGGATAATGATAGCGCTGGCGTACAGTGCATGAAATGCGTAAACGCAATTTGCAACAATCCCTGACGCGC  
ACAGGGTGGAGGGGATACGACGTTGCCGCTGCGGCCGATTAGCGGTTGACCCCATTCGCCGTTTATTAGAGAACATCCACGCTGAAAACCGC  
TGGTAATAGCCTGGGTGAGCGTATATCCGGCTGCGCTGCCCGGGCTTCTGCCAGCCAAATGCGCTTTTGACTATCGAGGCGATAACGCCAGG  
CATTTGCTCAATATGCCAGAGGCAATGAATACGCTGCCGGCGACCCACAATAACCGCATGAGGGGAATCAGCCATTGCATCAGACGGGCGACG  
CTTTGATGCCGGAATTAATAAAGGCAAGGCAAGGCAAGGCAAGGCAAGGCAAGGCAAGGCAAGGCAAGGCAAGGCAAGGCAAGGCAAGGCAAGG  
GTGAGACGGCATTGCGCTGCACGCTATTAATAAACAACCGTAGGCGACGAGCAGAAAGAGGGCGAAAACCGCCCATCCAGCGCATTCCAG  
CCACGCGCCATATACCAGCCCGGCCCGCCGCGGAACGTGCCCTGTGGGTGACGTTCTTTATAAAGCTGGGCAAGCGAACTCGCGAAGGAG  
GTGCGCATGCCAATGATGGCCGAGACCCACATCCAGAATACCGCGCGGGACCGCTGGCGCGATAGCCAGCGCCACGCGCGGAGGTTACCGC  
TGCCAATCCGCGCGCGGCTGACACAGAGCTGGAATGAGTCAAGCCGCGGTGAGGCGGCTGGTGGCGCTAAGGCTGCTTTCAGACTGCGGCTAAA  
TTGGCGAATATAACGAAACTGAATGAATCCGGTACGCCAGGTAACCAACATCCTGCGCGAGCAGCAGGTAATCATACCAGCCCCAGAGT  
ATTTGCTTAATAAAACTGAAAACTCAGGCATTAACGTCCCTTGTGTTGATGCCGGCAGGCTTTGATAATCCTGTATAAGCGTGACCCATGATG  
TAGATGACCTTGTGACACTAATATTAACGGCAGTTTACCATAAATACGGTGGTATCCTTTAATGCGCATCAACCGTCGCGAGATACGCAACA  
GTGCACAAAGGGCAGCAGGTAATGATAGCGGTTGCGCTGTGAGTGCCTGCTGTTATCATCAGGGTAGACCGGTTACATCCCTTAAACTGT  
TTAAAGAGAACTCTATCATGACGGACAATGACCTCCCTTGTGAGTTCACCACCGTAGTGGCTGATACGGGAGATATCGCGCAATGAAAC  
TGATACGCCGAGGATGCTACAATAACCTTCTCTATTCTTAACGCGAGCGAAATCCCGGAATATCGTAAGCTGATTGACGATGCTGCGC  
CTGGGCGAAACAGCAGAGCAGCGACCGCGCGCAGCAGGTTGTTGACGCGACCGATAAGCTGGCGGTGAATATTGGCCTGGAGATCCTGAAGCTG  
GTGCCGGGGCTATTTCTACCGAAGTTGACGCGCTGTCTTATGACACTGAAGCGTCTATCGCCAAAGCAAAACGATATCAATAAACTCTACA  
ATGATGCGGGTATCAGCAACGATCGTATCCTGATCAAGCTGGCGTCCACTGGCAGGGCATTGCTGACGCCAACAGCTGGAAAAAGAAGGCAT  
CAACTGTAACTGACGCTGCTGTCTCCTTCGCGCAGGCGCGTGGCTGCGCCGAAAGCGGGCGTCTACCTGATCTCGCCGTTCTGATGGTCTGATT  
CTTGACTGGTATAAAGCAATACCGACAAGAAAGACTATGCGCCAGCTGAAGATCCGGGCGTGGTTTCCGTAACGGAAATCTACGAGTACTACA  
AACAGCATGGTTACGAAACCGTCTTATGGGCGCAAGCTTCCGTAACGTAGGCGAAATTTGGAGCTGGCGGGCTGCGACCGCTCTGATATCGC  
GCCGCGATTGCTGAAAGAACTGGCGGAAAGCGAAGGGCGGATTGAGCGTAAGCTCTCTTCTCCGGCGAAGTCAAAGCGCGCCGGAACGCATT  
ACCGAAGCGGAGTTCTGTGGCAGCATACCAGGACCCATGGCGGTTGACAAACTGGCGGATGGTATCCGTAAGTTTGGGTTAGACCAGGAAA  
AACTGGAAAAATGATCGCGCATCTGCTGTAATCATTAAACCGCTGGCCCTGATATGGGTCACGCTACCTCTTCTGAAACCTGTCTGTCTTCCC  
TTGCGAGTGTATCATTTCTGTTTAAACGAGACTGTTTAAACGGAATAATCATGGATACCTTACGATTTGGCTTAGTTTCTATCTCCGACCGGCTT  
CAAGCGCGTTTACCAGGATAAAGGCATTCCTGCGCTTGAGGAGTGGCTCGCTTCTGCGCTGACCACGCCCTTTCGAGGTCCAACGGCGCTTAAT  
TCCTGATGAACAGGAAATATCGAGCAACGTTGTGTGAACTGGTTCGATGAGATGAGCTGTCATCTGGTCTGACCACTGGCGGTACCGGTCCG  
GCGCGTCCGACGTCACGCCGAGCGACCCCTTGCCATGCGCCGACCGTGAATGCCAGGTTTTGGCGAGCAGATGCGCCAGATCAGCTGCGCT  
TTGTCCGACCGCCATTCTTCCCGCCAGGTGGCGGTTATCCGTAACAGGCGTAAATCTTAATCTGCTGGACAGCCAAAATCGATCAAGA  
AACGCTGGAAGCGTAAAAGCGGACGATGGCAGCGTTAGCGTGCCGGGCAATTTTTGCGAGCGTCCGATTTGCATACAGCTGCTTACGCGGCG  
TATGTGAAACCGCGCCGGAAGTGGTTGCCGTTTCCGTTCAAAGAGCGCCAGAGCTGAGAATATGTCGGACTGACCGGAAAATACTGATAGTA  
GGGTTATTCCTCCCGGTGCGGGAGGAATAAAGAGATTAGTGCCTCGCCGATTGGCAGAATAGTGGCCAAATTTGCTGTTTAGCACTTC  
ACCCATCGCCAGATAAATAGCGCTGGCGCCGAAACAGACCAACCCAGCCTGCGATATGGATAGTTGCTTCCGTTACCGGTGATATTGCCGACC  
GCCAGCAGAGCAACAGTACGGTCAGGCTCAGGAAAAAATAATTCAGCGCGCGGGCGGCTTTCAGCGTACCGAAGAACATAAACAGCGTGAACA  
CGCCCCACAGACTAAATAAGCGCCGAGTAGCTGAGCGTC