

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 ^[A]	ACG (Thr/T) Threonine	AAG (Lys/K) Lysine	AGG (Arg/R) Arginine
	G	GUU (Val/V) Valine	GCU (Ala/A) Alanine	GAU (Asp/D) Aspartic acid	GGU (Gly/G) Glycine
		GUC (Val/V) Valine	GCC (Ala/A) Alanine	GAC (Asp/D) Aspartic acid	GGC (Gly/G) Glycine
		GUA (Val/V) Valine	GCA (Ala/A) Alanine	GAA (Glu/E) Glutamic acid	GGA (Gly/G) Glycine
		GUG (Val/V) Valine	GCG (Ala/A) Alanine	GAG (Glu/E) Glutamic acid	GGG (Gly/G) Glycine

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

```
ATGGCTCAATCAGAATTTAACCAAAAATCTACGGGAGCAAGGGCAATCAGTGCAGAGTTATTATTTCTACGTGTGAATAATCCGGGATACAACA  
CTAATACTTTAGACATCGCAGACATAGAAGATATCATTTCATTTACCTCAAGCTATAGAGTTAGCTAATGCATTTCAATCTGCACTAGTACCTAC  
TACATCCAACCTTCGGGGAGGACACATTAAGATTTGATGTAGAAAGGGGATTAGGAATAGCTACTCACGTTTATCCTAGAGCTATAAATGTGAAT  
TATGTAACACGCTACTTTATCACAACTAATAATCAAGTACAGTCCATGATAAATAAGGTGATAGAAGAGCTTAAATCTTTATTAGGAATTAATC  
TGCTAATTCAGTGTACAACAATTAACCACCTGTAATTACAGAACTTTTACAACCTCTACGTACAACAGCAATCTGCCTGGCTTTTTGGGG  
ACGCCAAACATCATCACAACCAATATACTTTATAATATCGTATTTGCAATCCAAAATGCCAAAACAGGTAGTTTTATGAAAGCTATTCCTATA  
GGTTTTGAGATATCTGCATATATGCAAGGGAACGCTTATTATTTTTAATATCAAGATTATGCAAGCTATAGTGTCAAATACACGCAATTC  
AAGTTATGCAGCCATTGATTCATGAGAGTTTTCAACCCCTAAGAGGCATATTTAATATCATAACTCCGTAATAATAGAAGTGCATTTCAAAAT  
TACTGAATATTATGACGAAAATACACATGATTTACCTGTTAAGTTATGGGACTATAAATAATATTATTAATCAAAAATGGATTTTAGTTTTAAT  
CAAACCTACCCGTGCTTACAGTATTCAAAACCTAATAGCTCGCTATTTAGTACTAACATGGGATTTCTACTCCTGGATCAAATAAAGTTTTGCTT  
CTACTAATCGTTGGAACGATTCACAATTCCTGGATTCCTGAAAGTACAGCAGATGGTCCATTTTTTAAACAAATATGAAAGACACTCAATTTGT  
ATTAGAGATAGAAAACCTAGTACTACTAATGGGACAAATGTTATAGTCAATAGAAAAATAAATATGCACAGCAAAAATTTACTTAATAAAA  
GTAACCAAGAGTTTCAGGATGGTGTATATAAGATTAAGACTGCCCTAAACAATAGTAGCGTTCTGCAGATGCTGAAGATTATTTGGTTATA  
CGTCAGATTACTTTGTGAAATTTAGGACCAACAATAAATGATATTAACCAAAAATGGATTTTTGAAATTTGATAGCACTAAATCTGCGTATCA  
AATTAAGAGTCAAAGGGATCCTTCTTAGTACTAGCATGGACTTGGAGTGTACCTACTGTTAACTCCCTATTCCAAATAACGATGATCATCTC  
TGGTTTTCTCAAATGCTGGGTGGGTACCTATATTTTTGTAATATGACAGACACCCGATATGTTTTAGAAAGTGGCTAGTTCAGTACCACTA  
ATGGAACGATACTTACAATCAATAACGAAACGGAAATTTAAATCAAAAATTTTTGTTGGACATGATTAATTA
```

Sekvence 2

```
ATGGCGCAGAGCGAATTTAACCCAGAACCTGCGCGAACAGGGCCAGAGCCGCGCGCGCTGATTATTTCTGCGCGTGAACAACCCGGGCTATAACA  
CCAACACCCCTGGATATTCGGGATATGAGATATTTATTCATCTGCCGAGGCGATTGAACTGGCGAACCGGTTTTAGAGCGCGCTGGTGGCGAC  
CACCAGCAACTTTGGCGAAGATACCCTGCGCTTTGATGTGGAACGCGGCTGGGCATTGCGACCCATGTGTATCCGCGCGCGATTAAACGTGAAC  
TATGTGACCCCGACCCCTGAGCCAGACCAACAACAGGTGCAGAGCATGATTAACAAGTGAATGAAGAAGTGAAGAAGCTGAAAGCCCTGCTGGGCATTAAC  
TGGCGAACAGCGTGTGCAGCAGCTGACCACCGTATTACCGAACTTTACCAACCTGTATGTGCAGCAGCAGAGCGGCTGGCTGTTTTGGGG  
CCGCCAGACCAGCAGCCAGACCAACTATACCTATAACATTTGTGTTTTGCGATTCAGAACCGCAGACCCGCGAGCTTTATGAAAGCGATTCCGATT  
GGCTTTGAAATTAGCGCGTATATTCGCGCGAACCGCTGCTGTTTTTAACTTCAGGATATGCGAGCTATAGCGTGAATAATTCATGCGATT  
AGGTGATGCAGCCGTGATTCATGAAAGCTTTACCCGCTGCGCGGCTTTTTAACATTATACCAGCGTGAACAACCCGAGCGGATTCAGAT  
TACCGAATATTTATGATGAAAACACCCATGATTATCCGGTGAACCTGTGGGATTATAACAACATTATTAACCAGAAATGGATTCTGGTGTTTAAC  
CAGACCACCCGCGGTATAGCATTCAGAACCTGATTGCGCGCTATCTGGTGTGACCTGGGATAGCACCCCGGGCAGCAACAAGTGTGGCGA  
GCACCAACCGTGAACGATAGCCAGTTTTGGATTCTGGAAGACCCGCGGATGGCAGCATTTTTCTGACCAACATGAAAGATACCCAGTTTTGT  
GCTGGAATTTGAAAACAGCAGCACCACCAACCGCACCAACCTGATTGTGAACCGCAAAAACAACAACCGCAGCAGAAATTTTATCTGAACAAA  
GTGAACAGGAATTTTCAGGATGGCGTGTATAAAATTAACACCGCGCTGAACAACAGCAGCGTGTGCAGATGAGCGAAGATTATTTGGCTATA  
CCAGCGATTATTTGTGAACTGTGACCAACAACAACAGTATTAACCAGAAATGGATTTTTGAAATTTGATAGCACCAAAAGCGCGTATCA  
GATTAAGAGCAGCGGATCCGAGCCTGGTGTGGCGTGGACCTGGAGCGTGGCGACCGTGAACCTGCGGATCCGAACAACGATGATCATCTG  
TGTTTTCTGCAGAACCGGGCAGCGGCTATTTTGGCAGTTTTCCGAGTTTCCGACATCCGATACCCGCTATGTGCTGGAAGTGGCGAGCAGCACCACCA  
ACGGCACCATTTGACCATTAACAACGCAACCGCAACCTGAACCAGAAATTTCTGCTGGATATGATTAACCTAAGTAGCTGAGCTGACGTGGCT  
GACGTGCTGAATGCTGCATGCGTAGCTGATGCTAGCTGACGCTATGGTGGCGAAAACATTAACCTGCAGGATATTTGATCCGGTGAACCCAG  
GAATGGAAGAAGCGCTGGATAGCGTGGTGAATATGAAAACCTGGATCGCGCGGAATTTATTTCTGGAAAAACCTGCTGGCGCATGCGCGCAACG  
TGGCGTGCAGCGTGCAGCCTTACACCCGCTATCCCGGATTTGAAACACCATTCGGCGGAAACCGAAGCGAGCTGGCGGATGAAATTAAGT  
GATGCAGCGCTGACCAACTATCTGCGTGAACGCGCTGGCGATGGTGTGCGCGTGGCGCGCAAAAAGCGGGCTGGCGGCTATGAGC  
AGCTATGCGAGCATGGCGACCCGTGTTTGAAGTGGCCGTGAACATTTTTTTCCGCGGATGATCTGGTGTTTTTTCAGGGCCATAGCGCGGAAG  
GCATTTATGCGCGCGCTTTCTGGAAGCGCGCTGAGCGAACCGGATCTGATTCATTTTTCCGCCAGGAAGCGCTGACCAAGGCATTTAGCAGCTA  
TCCGCTTCCGTTCTGATGCGGATTTTTGGCAGTTTTCCGAGTTTCCGACCGTGGATGGCCCGCTGATGGCGATTTATCAGGGCAGCTGCTG  
AAATATCTGCATCATCGAAACTGGTGGATACCGCGGGCCGAAAGTGTGGCGTTTTGCGCGGATGGCGAAAACCGCGCAACCGGAAACCCCTGG  
GCGGCTGCTGGTGGCGAGCCGGAACCTGGATAACCTGATTTTTATTTGTAACCTGCAACCTGCAGCGCTGGATGGCCCGTGGAGCGCAA  
CGGCAAAATTTATCAGGAAGTGAAGGCTGTTTCCGCGCGCGGGCTGGCGCGTGAATTAAGTGAATTTGGGGCCAGGATGGGAACGCGCTGTTT
```

AAAAAAGATAAAAAGCGGCTGCTGCTGAAACCGCTGAGCGAAATGGTGGATGGCGAATATCAGGCGTGCTTTAGCAAAGATGGCGCGCATCTGC
GCGAACATTTTTTTGGCAAATATCCGGAACTGCTGGAACCTGGTGGAGCGATGATGAGCGATGAACTGAAACAGCTGACCGATGGCGGCCATGA
TCCGCGAAGAGTGTATGCGGCGTATACCGAAGCGATGAAAGATGGCGGAAACCGACCTGATTTCTGACCAAACCGTGAAAGGCTATGGCTAT
GGCAAAGAAGGCGAAAGCCAGAACATTTGCGCATAACCTGGAAGAAATAGCGAAGAAGGCGTAAAAATTTTCGCCAGCGCTTTGATCTGCCG
TGAGCGATAAACAGCTGCGCGATCTGGAATTTATAAACCGGATGATAACAGCCCGGAACTGAACTATCTGCGCAAACAGCGCGAAAAATGGG
CGGCCCGCTGCCGGCGCGCATGGCAAATTTGAAAGCATTAAGTGCCGGATCTGAGCCTGTTTGAACCGGTGCTGAAAGGCACCGAAGAACGC
AGCGTGAGCACCGGCGGCGTATTACCGCATTTATGGCCTGATGCTGAAAGATAAAAAACATTCGCGATCGCATGTGTGCCGATTTGGCGGATG
AAGCGCGCACCCCTGGGCTGGAAGGCTGTTTCGCCAGACCGGCAATTTATGCGGTGGAAGGCCAGAAATATACCCCGGAAGATCAGAGCAAAC
GATTTTTTATCGCGAAGATAAACCGGCGAGCTGCTGCAGCAGGGCATAGCGAAGCGGGCGCGATGGCGAGCTGGATTGCGGCGCGGACCCAGC
TTTGCGAACAACTATCCGCTGATTCGGGTGTATACCTATTATGCGATGTTTGGCTATCAGCGCGTGGGCGATCTGGTGTGGGCGGCGGCGG
ATATGCGATGCGCGCGGCTTATTCTGGGCGGCTGGCGGGCCGACCACCTGCGGGCGAAGGCTGCGAGCATCAGGATAGCCATAACCTGCT
GATGTTTAGCATGGTGGCGACCTGCCGAGCTATGATCCGGCTTTGGCTATGAACTGGCGGTGATTATTCAGGATGGCTGCCGCGCATGTAT
CAGGATAAAGAAAACGTGTTTTATCTGACCTGATGAACGAAAGCTATCAGCATCCGCCGATGCCGAAAGGCGTGGAAAGAAAGCATTGTGA
AAGGCATGTATCTGTTAAAAAAGGCAACGGCAAACTGAGCAACCGCTGCAGCTGCTGGGCGGGGCGCGATTTCTGCGCGAAGTGTATGCGGC
GAGCGAAATTTCTGGAAGAACAGTTAAAGTGGCGGCGATGTGTGGAGCGTGGCGAGCTTAACTGCTGCGCCATGATATTGAAAGCGTGGAT
CGCTATAACCGCTGCATCCGGATACCGCGCGGAAACAGAATATGTGCAGAAATGCCTGGATAACCAGAAAGGCCCGGTGATTGCGGCGACCC
ATTATATGAAACTGCTGCGGAACAGATTTCGCGAAGCGGTGAAACAGCCGATTTATGTGCTGTA

ÚKOL 4

Charakterizujte část genomu neznámého organismu. Přeložte DNA sekvenci do aminokyselinové sekvence pomocí programu **ORF Finder** (<http://www.ncbi.nlm.nih.gov/gorf/gorf.html>). Určete, které geny/proteiny mohou být touto sekvencí kódovány.

ACTAAATAGTATATTTTATTTGATTTTTGATTTTTGATTTTCAATAATACAAATTTATTTACTTATTTAATTGTTTTGATCAATATTTTTTTCTGTT
AAACAAAGGGAGCATTATATGGTAAAGATAATATTTGTGTTTTTATTTTCTTATCATCATTTTCATATGCAAATGATGATAAGTTATATCGGG
CAGATTTAGACCTCCTGATGAAATAAAGCAGTCAGGTGGCTTATGCCAAGAGGACAGAAATGAGTACTTTGACCGAGGTAATCAATGAATAT
CAACCTTTATGCTACAGATAGAGCTCAGACGGGATTTGTTAGGACAGCATGATGGATATGTTTCCACCTCAATGATTTGCGGCGGCGGCGG
TTAGTGGGTCAAATATATTGCTGGTCACTTACTTATATATATATGTTATAGCCACTGCACCAACATGTTTAACTTAATGATGTATTAG
GGGCATACAGTCCCTCATCCAGATGAACAGAAGTTTCTGCTTTAGGTGGGATCCATACTCCCAAATATATGGATGGTATCGAGTTTCAATTTGG
GGTGTGATGAACAATTACATCGTAATAGGGGCTACAGAGATAGATATTACAGTAACCTAGATATTGCTCCAGCAGCAGATGGTTATGGATTG
GCAGGTTTTCCCTCCGAGATAGAGCTTGGAGGGAAGAGCCGTGGATTTCATCATGACCAGCGGTTGTGGGAAATGCTCCAAAGATCATCGATGA
GTAATACTTGGCATGAAAAACCCAAAGTCTAGGTGTAATAATCTTGGACGAATACCAATCTAAAGTTAAAGACAAATATTTTCAGGCTATCA
ATCTGATATTGATACACATAATAGAATTAAGGATGAATATGATTAATAATTTGGTGTTTTTTTTACAGTTTACTATCTTCAGCATATG
CACATGGAACACCTCAAATATTTACTGATTTGTGTGCAGAAATACCACAACACAAATACATACGCTAAATGATAAGATATTGTCGTATACAGA
ATCTCTAGCTGGAAACAGAGATGAGCTATCATTTACTTTAAGAAATGGTGCATGATTTTCAAGTGAAGATACCAGCTAGTCAACATATGCTCA
CAAAAAAAGCGATTGAAAGGATGAAGGATACCTGAGGATTGCATATCTTACTGAAGCTAAAGTCGAAAAGTTATGTGTATGGAATAATAAAA
CGCCTCATGCGATTGCCCAATTAGTATGGCAAATTAAGATATAAAAAAGCCACCTCAGTGGGCTTTTTTGTGGTTCGATGATGAGAAGCAAC
CGTTTTGCCCAAACATGTATTACTGCAAGTATGATTTTTTATTCACATCCTTAG

PREDIKČNÍ PROGRAMY

Jak predikovat geny, které nemají sekvenční homology? Predikční programy využívají *ab initio* metody – predikce na základě statistických parametrů DNA sekvence.

ÚKOL 5

Charakterizujte část genomu pomocí predikčního programu **GeneMark** (<http://exon.gatech.edu/GeneMark>). Určete možnou funkci predikovaných genů/proteinů pomocí aplikace BLAST.

ATCTGGCGGTTTTATTACGCCTGTTACTCGCGGCAGCCGACGCTTGCCGCCGCGCTGATGAAAGATGTTATTGCCGAACCATACCGCGCGG
TTTACTGCCGGCTTTAGCCAGGCGCGGCGAGCGGCTGTCGGAGATCGGCGCGCTGGCGAGCGGATTTCCGGTCCGGCCGACGCTGTTTGGC
CTATGCCGATAAACCCGAGACGGCGCAGCGCTCGCGGACTGGCTGAGCAAAACATTTATTCGAAAATCAGGAAGGCTTCGTTATATTTGCCGCG
TGGATACGGCGGGCGCACGAGTAGTGGGATAATCAATGAACTCTATAATCTGAAAGACCATAATGAGCAGGTCAGCTTTGCGCAGGCGGTCAC
GCAAGGACTGGGCAAACAGCAGGACTTTTTTTTCCGACGATCTGCCGGAGTTAGCCTGACGGAAATGATGAGATGCTCAACCAGGACTTT
GTCAGCCGTAGCGCAAAGATCCTCTCGGCATTTATTGGCGATGAAATACCGCAGCAAATCTGGAAGAGCGGCTCCGCGCGGCGTTTTCGTTCC
CGCACCGGTAGTACGATAGAAAGCGATGTCGGCTGGAACTTGTCTACTGTTCCAGCTCCGACCTGGCCTTTAAAGACTGCGGCGGCGGCTTTAT
GGCGCAAATGCTGACGCATATCAGCGGCGACAAACCGGTGACGATTCTGACTGCAACGTCAGGCGATACCGGCGCGGCGTGGCTCACGCGTTC
TATGGCTGGAATAATGTCGGGTGCTCATTTCTTACC CGCGGTAAGATCAGTCCGTTGCGAGAAAACCTGTTCTGTAGCTGGGCGGCAACA
TTGAAACCGTGGCGATCGACGGCATTTCCAGCGCTGCCAGGCGCTGGTGAACAGGCATTTGATGACGAAGAACTGAAAACGGCGCTGGGGCT
GAATTCGGCTAATTTCCGTTAATATCAGCCGCTGTTGGCGCAATTTGCTACTGTTCCAGCTCCGACCTGACCTGCGGCTTTAAAGACTGCGGCGGCGGCTTAAC
CAACTGGTGTATCTCCGTCACGCGCAACTTTGGCGATTGACGGCAGGGCTGCTGGCGAAGTCGTTAGGCCTACCGGTGAAACGTTTTATCG
CCGCCACCACGTCACACACCGGTGCCGCGTTTTCTGATGACGGAAAGTGGGCGCGGAAAGCGACGAGGCGACCCCTGCGAATGCGATGGA
TGTCAGCCAGCCGAATAACTGGCCGCGCTGGAGGAGCTATTCGCGCTAAAATCTGGCGCTGACTGAGCTGGGCTATGCGGCGGTGGATGAT
ACTACGACACAGCAGCATGCGCGAGCTGAAAGCGAAAGTTATATCTCGAAACCTCATGCGGCGGTAGCGTATCGGGCATTCAGCGCACGAT
TAAACCTGGCGAGTATGGCTTGTTCGGAACGGCGCATCCGGCGAAGTTAAAGAGAGCGTGGAGTCCATTTCTGGGAGAAACGCTGGCCTT
GCTGAAGCGCTCGCCGACGCGCCGACCTGCCGCTGCTTTCACATCATCTGCTGCGGATTTTCCGCGCTGCGTAAGCTGATGATACCCGC
CAATAACCATTGCCCGCGGTGGCGCTGTCGCTTACCGGCGCTATGGGTTGGTGTGATTTGTAGGCGGATAAGGCGTAACCGCATCCGGCGA

TGCCGTTACTGCTCGTAGCGCTTAAAGACCAGCTCGTCTTGTGTGGAGGTTTCTTCATCAAAGAAATACCCCTCACGGTCAAACGCGGTAAGCT
GTTCCGGCTTCGTTAAGCGGTTTTCAATAATAAAACGGCTCATCAGTCCGCGCGCTTTTTGGCGTAGAAGCTTACCACCTAAACTTGCCTT
TTTCTCATCAAGGAACACGGGCTTAATCAGTTCCGCTTCAAGTTTCTTCCGCTTACCGATTTAAATATTCCTCGGAGGCCAGATACCACC
ACCCGATCGCCCTGCGCTCAAGCGCTTCGTTGAGCTTATCGGTAATGATATCGCCCCAGAATTGATAAAGATCTTTCGCGCGCGGATTCCTCA
GGGAATCCCCATCTCCAGACGATAAAGCTGCATTAATCCAGCGGGCCAAACGCCATACAAGCCAGAGAGCATACGAGATGTTGTTGAGC
AAAATCAAATCCGCGCTGTTGAACGTTTCCGCTGTAGGCCGATAAACAATCGCCTTTGAACGCCAGAATCGCTGGCGTGCATTATCCGCG
GTAAAATGAGGCTGCCAGTCATGAAAACGCGTGGCGTTGAGATCCGCCAGTTTGTCTGCTAATCCCATCAGCTGAAAATTTGCCGCGCCGAAA
GCTGGCGGGCCTGTTGAATAAGCTGCTGGCTGTGATCCAACAGCTCCGCTGGGTATAGCGGGTCTGGCCAGCGGGCTTTGATAATCAAGCGT
TTTTGACGGTGAATCAGAATCAGCATATTCAGTCTTGCAGGGAATTTCTGCGACTTAGCAAAAAACGCCAGAGTTGACCGATGGTTG
CGATTGTCCGCTTAATCGCGCATGCCCGCTCCAGGTATCTGGCGCCAGTTGTTGTTGATATCCGGGAAGCGCCGCGGATCGAACTGCGGTC
TTACGCCAGTTTCCGTGTCTCGAGATAATCGCTGGCGAGGGTATACACCACCGGTGAGAGCAACAAAATCGCCGTAGATTTGTTAATGGCCAT
ACAGGCCATGATCATGTCAAGCGAGCTGCCATATCAGCGGAAAACGTATAAGCGTACCAGGCGATAACCATGCCAAGCGTCGCAAGCGGTAATAGC
CAGATAGCTTTGCGTTATGTAACCGCAGAAAAACAGATTGTTTCCGCGTAAATATAGTTGGCGAGCATAGAACTGAACGCCAAGAGAAATGA
CGATAAGCGCGCAAAAACGCGCCCAATTCACCGGTCAACGAAACATTGATGTTGGAGAAGCTGAATGCCTTCTGTTGACGAGTGGGAGCG
GTGACAAACCGCCAGCAGGATAATGAGCGCTGGCGTACGAGTGAATGCGCTCGTGAATACGCCAATTTGCAACAATCCCTGACCGCG
ACAGGGTGAGGGGATACGAGCTTCCGCTGCGCGCCGATTAGCGCTGACCCCAATCCCGCTTCATTTAGAGAACATCCACCGCTGAAAACCGC
TGGTAATAGCTGGGTGAGCGTATATCCGCTGCGCTGCCGCGGCTTCTGCCAGCCAAATGCGCTTTTGACTATCGAGGCGATAACGCCAGG
CATTTGCTCAATATGCCAGAGGCAATGAATACGCTGCCGCGGACCCACAATAACCGCATGAGGGGAATCAGCCATTGCATCAGACGGGCGAGC
CTTTGATGCCGGAATGATAAAATTAACAGGGCACAGAACGCCAGAGCAATCCGCGAGATAAGCGCGGAATGTTGAAAGCGAAATGCGGCGCG
GTGAGACGGCATTCGCTGCACGCTATTAATAATCAACCCGTAGGCGCAGCAGAGAAAGAGGGCGAAAACCGCCCATCCAGCGCATTCCCAG
CCACGCGCCATATACCAGCCGCGCGCCGCGGAACTGGCTGTCCGGTACGCTTTTATAAAGCTGGGCAAGCGAAGCACTCGCGAAGGAG
GTGCGCATGCCAATGATGGCGGAGACCACATCCAGAATACCGCGCGGGACCGCTGCGCGCATAGCCAGCGCCAGCGCGGAGTTACCGC
TGCCAATCCGCGCGCGAGGCTGATACACAGAGCTGGAATGAGTCAAGCGCGGCTGGTGGCGCTAAGGCTGCTTTCAGACTGCGGCTAAA
TTGGCGAATATAACGAAACTGAATGAATCCGGTACGCCAGGTAACCAACATCCTGCGCGAGCAGCAGGTAATCATACCAGCCCCAGAGT
ATTTGCTTAATAAACTGAAAACTCAGGCATTAACGTCCTCTTGTGGATGCCGCGCAGCTTTGATAATCCTGTATAAGCGTGACCCATGATG
TAGATGACCTTGTGACACTAATATTAACGGCAGTTTACCATAAATACGGTGGTATCCTTTAATGCGCATCAACCGTCCGCGAGATACGCAACA
GTGACAAAGGGCAGCAGGTAATGATAGCGGTTGCGGTGAGTGCCTGTGTTATCATCAGGGTAGACCGGTTACATCCCTTAACAAGCTGT
TTAAAGAGAAACTCTATCATGACGGACAATTGACCTCCCTTCTGTCAGTTCACCACCGTAGTGGCTGATACCAGAGATATCGCGCAATGAAAC
TGATACGCGCAGGATGCTACAATAACCCCTTCTCTATTCTTAACGCGAGCGAAAATCCCGGAATATCGTAAGCTGATTGACGATGCTGCGC
CTGGCGGAAAACAGCAGAGCAGCGACCGCGCGCAGCAGGTTGTTGACGCGACCGATAAGCTGGCGGTGAATATTTGGCCTGGAGATCCTGAAGCTG
GTGCCGGGGCTATTTCTACCAGGTTGACGCGCGTCTGTCTTATGACACTGAAGCGTCTATCGCCAAAGCAAAACGTATCATTAAACTCTACA
ATGATGCGGGTATCAGCAACGATCGTATCCTGTATCAAGCTGGCGTCCACCTGGCAGGGCATTCTGTCAGCCGAACAGCTGGAAAAAGAGGCAT
CAACTGTAACCTGACGCTGCTGTCTCTTCCGCGAGCGCGTGGCTGCGCCGAAAGCGGGCGTCTACCTGATCTCGCCGTTCTGATGGTCTGATT
CTTGACTGGTATAAAGCCAATACCAGCAAGAAAAGACTATGCGCCAGCTGAAGATCCGGGCGTGGTTTCCGTAACGGAAAATCTACGAGTACTACA
AACAGCATGGTTACGAAACCGTCTTATGGCGCAAGCTTCCGTAACGTAGGCGAAAATTTGGAGCTGGCGGGCTGCGACCGCTCTGACTATCGC
GCCGCGATTGCTGAAAGAAGCTGGCGGAAAGCGAAGGGCGGATTGAGCGTAAGCTCTCTTCTCCGGCGAAGTCAAAGCGCGCCGGAACGCATT
ACCGAAGCGGAGTTCCTGTGGCAGCATACCAGGACCCATGGCGGTTGACAAAACGCGGATGTTATCCGTAAGTTTGGGTTAGACCAGGAAA
AACTGGAAAAAATGATCGGCGATCTGCTGTAATCATTAAACCGCTGGCCGTTGATATGGGTCACGCTACCTCTTCTGAAACCTGTCTGCTCTTCCC
TTCGCAAGTGTATCATTTCTGTTTAAACGAGACTGTTAAACGGAATAATCATGGATACCTTACGTATTGGCTTAGTTCTATCTCCGACCGCGCTT
CAAGCGCGTTTTACCAGGATAAAGGCATTCTCTGCGCTTGGAGGAGTGGCTCGCTTCTGCGCTGACCACGCCCTTTCGAGGTTCAACGGCGCTTAAT
TCCTGATGAACAGGAAAATATCGAGCAAAACGTTGTGTGAAGTGGTGCATGAGATGAGCTGCATCTGGTCTGACCCTGGCGGTACCGGTCCG
GCGCGTTCGCGACGTCACGCCGAGCGACCCCTTGCCATCGCCGACCGTGAAATGCCAGGTTTTGGCGAGCAGATGCGCCAGATCAGCCTGCGCT
TTGTCCGACCGCCATTCTTCCCGCCAGTGGCGGTTATCCGTAACAGCGGTTAATTTCTTAACTGCGCTGGACAGCCAAAATCGATCAAGA
AACGCTGGAAAGCGTAAAAGCGGACGATGGCAGCGTTAGCGTGGCGGCAATTTTTGCGAGCGTGGCATTGTCATACAGCTGCTTACGCGGCG
TATGTGAAACCGCGCCGGAAGTGGTTGCCGCTTCCGTTCAAAGAGCGCCAGCGTGAAGATAATGTCGACTGACCGGAAAATACTGATAGTA
GGGTTATTCCTCCGCGTGGGAGGAAATAAAGAGATTAGTGCAGCTCGCCGATTGGCAGAAATAGTGGCCAAAATTTGCTGTTTAGCACTTC
ACCCATCGCCAGATAAATAGCGCTGGCGCCGAAACAGCAACCCAGCCTGCGATATGGATAGTTGCTTCCGTTACCGGTGATATTGCCGACC
GCCAGCAGAGGAAACAGTACGGTACGGTACAGGAAACAAAATTCAGCGCGCGGGCGGCTTTTCAGCGTACCGAAGAACATAAACAGCGGTGAACA
CGCCCCACAGACCTAAAATAGCGCGGAGTAGCTGAGCGTC