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 ⇒ 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 UUC (Phe/F) Phenylalanine	UCU (Ser/S) Serine UCC (Ser/S) Serine	UAU (Tyr/Y) Tyrosine UAC (Tyr/Y) Tyrosine	UGU (Cys/C) Cysteine 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 CUC (Leu/L) Leucine	CCU (Pro/P) Proline CCC (Pro/P) Proline	CAU (His/H) Histidine CAC (His/H) Histidine	CGU (Arg/R) Arginine CGC (Arg/R) Arginine
		CUA (Leu/L) Leucine CUG (Leu/L) Leucine	CCA (Pro/P) Proline CCG (Pro/P) Proline	CAA (Gln/Q) Glutamine CAG (Gln/Q) Glutamine	CGA (Arg/R) Arginine CGG (Arg/R) Arginine
	A	AUU (Ile/I) Isoleucine AUC (Ile/I) Isoleucine	ACU (Thr/T) Threonine ACC (Thr/T) Threonine	AAU (Asn/N) Asparagine AAC (Asn/N) Asparagine	AGU (Ser/S) Serine 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 GUC (Val/V) Valine	GCU (Ala/A) Alanine GCC (Ala/A) Alanine	GAU (Asp/D) Aspartic acid GAC (Asp/D) Aspartic acid	GGU (Gly/G) Glycine GGC (Gly/G) Glycine
		GUA (Val/V) Valine GUG (Val/V) Valine	GCA (Ala/A) Alanine GCG (Ala/A) Alanine	GAA (Glu/E) Glutamic acid GAG (Glu/E) Glutamic acid	GGA (Gly/G) Glycine GGG (Gly/G) Glycine

ÚKOL 1
Přeložte následující DNA sekvenci do aminokyselinové sekvence s využitím **tří** čtecích rámců:
??????????

ÚKOL 2
Přeložte následující DNA sekvenci do aminokyselinové sekvence s využitím **šesti** čtecích rámců:

??????????

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
??????????

Sekvence 2
??????????

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.

??????????

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.

??????????