Trabajo Práctico 2

Síntesis de Proteínas

Introducción a la Computación

1^{er} cuatrimestre 2018

Observaciones generales:

- El trabajo se debe realizar en grupos de dos personas.
- El código fuente debe estar bien comentado. Cualquier aclaración adicional que se considere necesaria debe ser incluida como comentarios.
- El código fuente debe enviarse por mail a la lista de docentes de la materia: `icm-doc@dc.uba.ar`, indicando los integrantes del grupo.
- El programa debe correr usando Python 3, que está instalado en las computadoras del laboratorio 4.
- Se evaluará la correctitud, claridad y modularidad del código entregado.
- La fecha límite de entrega es el **miércoles 6 de junio a las 23:59**.

Nota: *El presente Trabajo Práctico se ambienta en una problemática de bioinformática. Se han hecho diversas simplificaciones del componente biológico a título de acotar la complejidad del trabajo de programación a realizar.*

El ADN (ácido desoxirribonucleico) está compuesto por una secuencia de bases nucleotídicas unidas entre sí formando una estructura de hélice de doble cadena. A través de una serie de complejos procesos bioquímicos las secuencias nucleotídicas en el ADN de un organismo son traducidas a proteínas necesarias para la vida. El objetivo de este trabajo práctico es escribir un programa cuya entrada sean cadenas de ADN y reporte la/s proteína/s codificada/s en tales cadenas (si las hay).

Las bases nucleotídicas que forman el ADN son ADENINA, CITOSINA, GUANINA y TIMINA (en adelante nos referiremos a ellas como `A`, `C`, `G` y `T` respectivamente). Estas bases se unen entre sí formando una cadena simple que corresponde a la mitad de la estructura de doble hélice. La otra mitad es una cadena similar, pero cada nucleótido es reemplazado por su nucleótido complementario. Las bases `A` y `T` son complementarias y también lo son `C` y `G`. Estas dos cadenas simples se unen entre sí por apareamiento de bases complementarias formando el ADN de doble cadena.

En general, un fragmento de ADN se describe simplemente con las bases que forman la cadena primaria. La cadena complementaria puede obtenerse escribiendo el complemento de bases de la primaria. Por ejemplo, la secuencia `TACTCGTAATTCACT` representa una cadena de ADN cuyo complemento es `ATGAGCATTAAGTGA`. Nótese que `A` siempre aparece apareada con `T`, y `C` con `G`.

A partir de la cadena primaria de ADN se genera una cadena de ARN (ácido ribonucleico) conocido como ARN mensajero (ARNm) en un proceso llamado transcripción. El ARNm transcripto es idéntico

a la hebra complementaria con excepción de la TIMINA que es reemplazada por otro nucleótido llamado URACILO (`U`). Por ejemplo, la cadena de ARNm para el fragmento de ADN del ejemplo anterior es `AUGAGCAUUAAGUGA`.

Lo que determina la estructura de la proteína que va a ser sintetizada está codificado en la secuencia de bases del ARNm. El ARNm puede interpretarse como una secuencia de codones, dónde cada codón está compuesto exactamente por tres bases contiguas. El codón `AUG` marca el inicio de una secuencia proteica y cualquiera de los codones `UAA`, `UAG` o `UGA` marca su fin. El o los codones comprendidos entre los codones de inicio y terminación representan la secuencia de aminoácidos que conformarán la proteína. Por ejemplo, el codón `AGC` del ARNm corresponde al aminoácido Serina (Ser), `AUU` a Isoleucina (Ile) y `AAG` a Lysina (Lys). Así, la proteína formada por el ARNm del ejemplo anterior es Ser-Ile-Lys. La siguiente tabla muestra todas las traducciones de codones a aminoácidos:

Primera base del codón	Segunda base del codón				Tercera base del codón
	U	**C**	**A**	**G**	
U	Phe	Ser	Tyr	Cys	**U**
	Phe	Ser	Tyr	Cys	**C**
	Leu	Ser	—	—	**A**
	Leu	Ser	—	Trp	**G**
C	Leu	Pro	His	Arg	**U**
	Leu	Pro	His	Arg	**C**
	Leu	Pro	Gln	Arg	**A**
	Leu	Pro	Gln	Arg	**G**
A	Ile	Thr	Asn	Ser	**U**
	Ile	Thr	Asn	Ser	**C**
	Ile	Thr	Lys	Arg	**A**
	Met	Thr	Lys	Arg	**G**
G	Val	Ala	Asp	Gly	**U**
	Val	Ala	Asp	Gly	**C**
	Val	Ala	Glu	Gly	**A**
	Val	Ala	Glu	Gly	**G**

Nótese que la secuencia `AUG` corresponde tanto a una secuencia de inicio como al aminoácido Metionina (Met). Así, el primer `AUG` del ARNm será la secuencia de inicio, pero los subsiguientes serán traducidos a Metioninas (siempre y cuando no aparezca un codón de fin de secuencia entre ellos).

La entrada del programa que se pide implementar consiste en un archivo que contendrá una secuencia de una hebra de ADN por línea. A partir de cada secuencia se espera que el programa genere la lista de proteínas codificadas en ese fragmento, y las escriba en una línea del archivo de salida. Cada cadena de ADN puede tratarse tanto de una hebra primaria como de su complementaria y en cada caso puede corresponder tanto a la secuencia *forward* como a la *reverse* (es decir, que debe ser leída tanto de izquierda a derecha como de derecha a izquierda). Además, los codones de inicio y terminación pueden no aparecer en los extremos de una secuencia. Por ejemplo, la proteína Ser-Ile-Lys puede corresponder a cualquiera de las secuencias (i) `ATACTCGTAATTCACTCC`, (ii) `TATGAGCATTAAGTGAGG` , (iii) `CCTCACTTAATGCTCATA` o (iv) `GGAGTGAATTACGAGTAT`:

(i)

$$\texttt{ARNm(ATACTCGTAATTCACTCC)} = \big(\texttt{COMP(ATACTCGTAATTCACTCC)}\big)\big[\texttt{T:U}\big]$$
$$= \texttt{(TATGAGCATTAAGTGAGG)[T:U]}$$
$$= \texttt{U}\mathbf{AUG}\texttt{AGCAUUAAG}\mathbf{UGA}\texttt{GG}$$

(*ii*)

$$\begin{aligned}\texttt{ARNm}\big(\texttt{COMP(TATGAGCATTAAGTGAGG)}\big) &= \texttt{ARNm(ATACTCGTAATTCACTCC)}\\ &= \big(\texttt{COMP(ATACTCGTAATTCACTCC)}\big)\big[\texttt{T:U}\big]\\ &= \texttt{(TATGAGCATTAAGTGAGG)[T:U]}\\ &= \texttt{U}\mathbf{AUG}\texttt{AGCAUUAAG}\mathbf{UGA}\texttt{GG}\end{aligned}$$

(*iii*)

$$\begin{aligned}\texttt{ARNm}\big(\texttt{REV(CCTCACTTAATGCTCATA)}\big) &= \texttt{ARNm(ATACTCGTAATTCACTCC)}\\ &= \big(\texttt{COMP(ATACTCGTAATTCACTCC)}\big)\big[\texttt{T:U}\big]\\ &= \texttt{(TATGAGCATTAAGTGAGG)[T:U]}\\ &= \texttt{U}\mathbf{AUG}\texttt{AGCAUUAAG}\mathbf{UGA}\texttt{GG}\end{aligned}$$

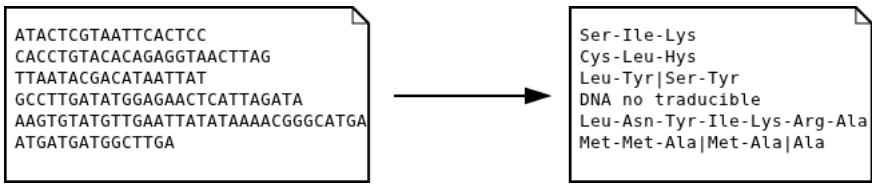
(*iv*)

$$\begin{aligned}\texttt{ARNm}\Big(\texttt{REV}\big(\texttt{COMP(GGAGTGAATTACGAGTAT)}\big)\Big) &= \texttt{ARNm}\big(\texttt{REV(CCTCACTTAATGCTCATA)}\big)\\ &= \texttt{ARNm(ATACTCGTAATTCACTCC)}\\ &= \big(\texttt{COMP(ATACTCGTAATTCACTCC)}\big)\big[\texttt{T:U}\big]\\ &= \texttt{(TATGAGCATTAAGTGAGG)[T:U]}\\ &= \texttt{U}\mathbf{AUG}\texttt{AGCAUUAAG}\mathbf{UGA}\texttt{GG}\end{aligned}$$

En caso de que una secuencia codifique más de una proteína, todas deben estar presentes en la misma línea del archivo de salida y aparecer separadas por el caracter 'l'. El orden en que deben aparecer es: 1) las codificadas por la secuencia original, 2) por la secuencia reversa, 3) por la secuencia complementaria, 4) por la secuencia complementaria reversa.

Algunas secuencias válidas de ADN no codifican proteínas; para esos casos la salida debe ser el texto "`*** ADN no traducible ***`".

A modo de ejemplo pueden observarse la siguiente entrada y la salida esperada:



Los nombres de los archivos de entrada y salida son parámetros del programa. Se puede asumir que el archivo de entrada contiene únicamente secuencias válidas de ADN (letras 'A', 'C', 'G' y 'T' en mayúscula) y que no contiene líneas en blanco ni espacios.