Trabajo Práctico Especial

1. Introducción

1.1. Bioinformática

La bioinformática es hoy una rama interdisciplinaria entre la biología molecular, la genética y la informática con un amplio campo de acción, tanto del punto de vista de la investigación como de aplicación. Su estudio principal está en determinar cuáles son las partes que componen las moleculas claves para la vida como el DNA, RNA y las proteinas. El DNA y las proteinas son moleculas muy complejas pero que pueden simplificarse y caracterizarse mediante una cadena de elementos consecutivos simples que los componen. Son 4 nucleotidos para el DNA, citosina (C), timina (T), adenina (A) y guanina (G). El RNA tiene Citosina, Adenina, Guanina pero Uracilo (U) en vez de Timina. En cambio, las proteinas se conforman con 20 aminoacidos diferentes escenciales (ver Ref 9, disponible en Biblioteca).

El principio fundamental de la biología establece que el código genético de nuestras celulas determina cómo es la secuencia de aminoacidos de las proteinas que genera. Existe una correspondencia entre ambas. Mientras que el DNA se copia de célula a célula y de individuo a individuo, siendo el canal de acción de la evolución, las proteinas son los motores de acción electroquímica de las celulas.

Así por ejemplo, una proteina se encarga de modificar a la molecula de la glucosa (azucar) para que una vez que entra en una celula no pueda salir (y así pueda usarse como energía). Esta proteina fue "manufacturada" internamente en la célula según información que está contenida en los propios genes.

Lo curioso, es que la naturaleza, debido a la evolución, usa y reusa los mismos componentes que fueron exitosos (genes) en diferentes organismos, pero con leve variaciones.

Así entonces, el foco de la bioinformática es reconocer secuencias de genes en una especie como "similares" a secuencias de genes de otra especies donde ya se conoce que existe una proteina asociada con una función fisiológica conocida, e inferir entonces que esa misma secuencia puede generar una proteina similar con función similar.

Dada la cantidad de especies y la cantidad de elementos contenidos en los genes (millones y millones) se percibe por qué es necesario automatizar ese procesamiento con computadoras.

1.2. Código genético

El código genético tiene una estructura de almacenamiento de información muy típica de las que normalmente se usan en las computadoras: existe una cabecera, el telómero, con una secuencia específica de apertura. Luego existe una secuencia de millones de nucleótidos, donde algunas subsecuencias codifican una proteina y son llamados genes y otras secciones no se las ha encontrado una función hoy (junk-dna). Para aquellas secuencias que codifican proteinas, los nucleótidos se agrupan de a tres formando codones y codifican un aminoacido, que son los elementos constitutivos de las proteinas. Finalmente, la secuencia genómica termina con un codon de stop que marca el fin de la cadena.

2. Compilador de Secuencias Genómicas

2.1. Objetivo

El compilador de secuencias genómicas se va a desarrollar entre el TP1 y el TP2.

El objetivo final del trabajo es poder identificar de 5 secuencias genómicas largas, nunca antes observadas, a qué especie o individuo pertenecen y generar la secuencia completa de los genes identificados.

Para esto, les vamos a proveer una base de datos de genes que deberán utilizar para realizar el análisis sintáctico de secuencias genómicas que les permitiran identificar si pertenece a una u otra especie.

Al final de la cursada vamos a correr los programas en los laboratorios de informática y aquel grupo que identifique correctamente las cinco secuencias y en el menor tiempo posible, tendrá un 10.

2.2. Trabajo Práctico 1: Analizador Léxico genómico

Para el TP 1 tienen que presentar una gramática genómica tentativa (i.e. una gramática que genere un lenguaje que permita expresar una secuncia de nucleótidos que conpongan el código genético de un individuo/especie) más el analizador léxico implementado con LEX tal como se describe en la sección Material a entregar.

El analizador léxico o scanner se construye utilizando la herramienta que vimos en clase, LEX. Esta herramienta genera código C que se puede integrar con otras librerías que ustedes creen y que permite armar un programa ejecutable que implementa el scanner.

Deberá recibir como entrada una palabra del lenguaje generado por la gramática genómica, reconocer si la palabra es lexicamente válida, es decir no tiene lexemas que no puedan reconocerse. Además deberá permitir caracteres en blanco, tabs, y el caracter wildcard hyphen '-' (i.e. cualquier nucleótido).

Cada secuencia contiene un telómero AUG (un header), un indicador de repetición '{}*' (llaves seguidos de un asterisco), un indicador de alternativas como ser '{A,U,G}' (A ó U ó G), y también uno de los codones de stop validos o la palabra literal 'STOP'. Finalmente permite un indicador de ubicación, una posición encerrada entre corchetes '[nnn]', que especifica la posición dentro del genóma a partir del telerómero. Tener en cuenta los únicos aspectos semánticos del análisis, la existencia del header, el trailer, y que la cadena no puede tener bloques sin información especificada, desde el telómero hasta el codon de stop. Las llaves '{}' pueden anidarse pero no es necesario validar esa opción para el TP1 (sí considerarlo dentro del armado de la gramática).

Una palabra válida sería (**ejemplo 1**):

```
AUG{G}*[234]{A,G,U}GA{-}*{TTTA,TTTTA,TTT}STOP
```

pero por otro lado (**ejemplo 2**),

```
AUG[234]{A,G,U}AG{-}*{TTTA,TTTTA,TTT}
```

no sería válida (no tiene codon de stop y no se sabe que pasa entre las posiciónes 4 y 233). Es importante que verifiquen que toda la secuencia genómica está completa informacionalmente.

Como salida, el programa tiene que determinar si es posible identificar todos los lexemas y generar una lista con todos los aminoacidos identificados en la secuencia según la siguiente tabla (ver ref. 4 y 5):

Amino Acid	SLC	DNA codons
Isoleucine	I	ATT, ATC, ATA
Leucine	L	CTT, CTC, CTA, CTG, TTA, TTG
Valine	V	GTT, GTC, GTA, GTG
Phenylalanine	F	TTT, TTC
Methionine	M	ATG
Cysteine	C	TGT, TGC
Alanine	A	GCT, GCC, GCA, GCG
Glycine	G	GGT, GGC, GGA, GGG
Proline	P	CCT, CCC, CCA, CCG
Threonine	T	ACT, ACC, ACA, ACG
Serine	S	TCT, TCC, TCA, TCG, AGT, AGC
Tyrosine	Y	TAT, TAC
Tryptophan	W	TGG
Glutamine	Q	CAA, CAG
Asparagine	N	AAT, AAC
Histidine	H	CAT, CAC
Glutamic acid	E	GAA, GAG
Aspartic acid	D	GAT, GAC
Lysine	K	AAA, AAG
Arginine	R	CGT, CGC, CGA, CGG, AGA, AGG
Stop codons	Stop	TAA, TAG, TGA

Así por ejemplo, la secuencia: CGTAAG genera la salida RK (los códigos de una letra que corresponden a los aminoacidos Arginina y Lisina que generan esos codones).

Retomando la entrada válida:

```
AUG{G}*[234]{A,G,U}GA{-}*{TTTA,TTTTA,TTT}STOP
```

La salida se codifica en dos secciones de proteínas (las proteínas pueden estar formadas por varias secciones de cadenas separadas),y cada sección queda devidida según las secuencias de 1 o más junkDNA (-), siendo la primer secuencia:

```
GGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGR
```

y la segunda (de un sólo aminoacido):

```
F
```

Todos los nucleótidos que no codifican proteínas, se descartan. Para esta primera entrega, para las secuencias opcionales pueden optar por generar sólo una de las opciones y especificar que encontraron cadenas opcionales que se descartaron.

Pueden encontrarse otros ejemplos en el apéndice adjunto. Por favor, prueben el TP 1 con los 5 ejemplos detallados en esta guía y presenten 5 más creados por ustedes.

2.3. Trabajo Práctico 2: Compilador Secuenciador de Genomas

En el TP2 se completará el secuenciador utilizando la herramienta yacc. Sobre esta herramienta tendrán que implemenar un analizador sintáctico que valide la secuencia de entrada de nucleótidos, reconozca sus elementos contra los provistos en la base de datos, genere la secuencia de salida de genes identificados e identifique si la secuencia pertenece a alguna de las 5 especies en cuestión o a ninguna.

El código genético humano tiene alrededor de 3100 millones de pares de bases nitrogenadas (nucleótidos) (cuántos gigas son?) y este tamaño es la razón por la cual existe la bioinformática. Dentro de esta disciplina se destacan dos ramas: genómica y la proteinómica. Lo que se busca en ambas es:

1. Demarcación de los genes: donde empiezan y terminan los genes en el código.
2. Identificación de qué regiones del ADN codifican o no genes.
3. Las proteinas que son expresadas, es decir codificadas, en el código genético.
4. La asociación entre los genes, las proteinas expresadas, los fenotipos que influencian y finalmente las especies involucradas.

Abusando de los conceptos, se puede decir que los fenotipos son las expresiones de los genes, como el color de pelo, una alergía hereditaria a determinadas comidas o polen o una capacidad para tener músuclos más explosivos.

Los diferentes fenotipos determinan la influencia genética sobre cada especie y se usan para identificarlas.

El gran desafio de la bioinformática es la búsqueda de similaridad entre diferentes secuencias de genes, proteínas y fenotipos. Para esto los algoritmos más utilizados son BLAST y FASTA (ref. 9).

En este segundo TP, entonces, se va a implementar un algoritmo basado en la estructura de un compilador.

Las 5 especies son: human, chimp, orangutan, bear, horse.

La base de datos es un archivo de texto (database.dat) con la siguiente estructura:

```
Protein_1 = "GYSAIPANIIPPPPSLRGSQSFDDKIGTLYDDVFVSGPNPSMPPSGHHRPLVRQAAVEDS"
Protein_2 = "AASAIPANIIPPPPSLRGSQSFDDKIGTLYDDVFVSGPNPSMPPSGHHRPLVRQAAVEDS"
Protein_3 = "VSPASSTQSPPGPIYSSAHVASVVSQSVEQMCSLLLRDQKPKKQGKYICEYCNRACAKPS"

.....

Phenotype_1 = Protein_1 + Protein_3 + Protein_2

.....

Phenotype_34 = Phenotype_2 + Phenotype_4
.....

human = Phenotype_6 + Phenotype_5
chimp = Phenotype_87
```

El signo '+' representa una concatenación de exones y, en concordancia, entre esos elementos pueden haber intrones, es decir, secuencias que no codifican nada (i.e. junkDNA) o secuencias que codifican otras cosas: lo que importa es que los elementos se hayan encontrado en la secuencia y en ese orden.

Qué tienen disponible:

- Archivo database.dat: Base de datos de proteinas, fenotipos y especies.
- Archivo chimp.dat: Secuencia con el código genético de un chimpance.
- Archivo orangutan.dat: Secuencia con el código genético de un orangutan.
- Archivo human.dat: Secuencia con el código genético de un humano.
- Archivo horse.dat: Secuencia con el código genético de un caballo.
- Archivo bear.dat: Secuencia con el código genético de un oso.
- Archivo nolivingthing.dat: Secuencia con un código genético que no pertenece a ninguna especie.

Qué se espera del compilador:

- Que reconozca satisfactoriamente las 5 especies e identifique que la sexta no pertence a ninguna.
- Que identifique cualquier nueva secuencia correctamente en una de las 6 clasificaciones (deberán proveer 6 ejemplos más creados por ustedes).
- Que realice el secuenciamiento de los genes y la identificación de las proteinas, fenotipos y especies encontradas en la entrada.
- Permitir procesar archivos del tamaño necesario para codificar el código genético humano.
- Hacerlo rapidamente (i.e. más rapidamente que el resto).

3. Sugerencias

- No necesariamente tiene que ser un único compilador con una única gramática.
- Pueden utilizar procesamiento en paralelo.
- Como los archivos a procesar son muy grandes, especial cuidado con el manejo de memoria (i.e. memory-leaks, dangling pointers) y con el acceso a disco.
- La base de datos de los genómas pueden cargarla en un motor de SQL si prefieren.
- Existen versiones para Java de Lex y Yacc (JLex, Java Cup)

4. Material a entregar

Para el TP 1, y adicionando lo necesario para el TP 2, deben entregar.

- Códigos fuente: Makefile, código Lex (sin compilar), y código C.
- Archivo ejecutable para Linux.
- Un readme explicando cómo compilar y ejecutar el programa.
- Cinco(5) secuencias de ejemplo provistas por la catedra.
- Cinco(5) secuencias de ejemplo inventadas por ustedes.
- Un informe que contenga, en este orden:
  - Carátula
  - Índice
  - Consideraciones realizadas (no previstas en el enunciado).
  - Descripción del desarrollo del TP.
  - Descripción de la gramática.
  - Dificultades encontradas en el desarrollo del TP.
  - Futuras extensiones, con una breve descripción de la complejidad de cada una.
  - Referencias.

Importante: No hay ningún inconveniente en utilizar librerías públicas, soluciones similares públicas, soluciones de foros, etc., pero es necesario aclarar, y enumerar cada una de ellas en la sección Referencias. No se aceptan bloques de código públicos implementados verbatim sin ningún tipo de análisis. Tampoco implementaciones que resuelven problemas que no están detallados (e.g. implementa un garbage collector sin explicar cómo).

5. Grupos

El trabajo se realizará en grupos de hasta 3 integrantes.

6. Fecha de entrega

El material a entregar debe ser enviado por mail, en un archivo tipo .zip a la cuenta `rramele@gmail.com` antes de finalizado el día 20 de Octubre para el TP 1. Dentro de la carpeta .ZIP poner un directorio con los apellidos de los integrantes del grupo en camelCase: e.g. LopezGonzalezGoycochea.

El TP 2 se debe entregar antes del día Domingo 1/12/2013 23:59 GMT -3. El Lunes 2/12/2013 se ejecutarán los programas con las instrucciones especificadas en cada entrega sobre los laboratorios.

7. Material de consulta

Se sugiere leer los documentos de las páginas:

1. "Compiladores, Principios, Técnicas y Herramientas",Aho, Sethi, Ullman, Addison Wesley.(El Libro del Dragón), capítulos 1,2 y 3.
2. http://dinosaur.compilertools.net/lex/
3. Gusfield, Dan. Algorithms on strings, trees and sequences: computer science and computational biology. Cambridge University Press, 1997.
4. http://www.cbs.dtu.dk/courses/27619/codon.html

5. http://en.wikipedia.org/wiki/DNA_codon_table

6. http://www.yeastgenome.org/archive/sequence_done.shtml

7. http://yeastgenome-www.stanford.edu/community/nature_genome_dir_pdf/Chr8.pdf

8. http://www.geneontology.org/

9. "Systems biology and bioinformatics", Najarian, Kayvan and Najarian, Siamak and Gharibzadeh, Shahriar and Eichelberger, Christopher N.CRC Press,2009

10. http://www.genomesonline.org

11. http://www.sanger.ac.uk

8. Apéndice

8.1. Ejemplo 3:human immunodeficiency virus type I enhancer binding protein 1: Hivep1

```
AUGATGCCAAGAACAAAGCAGATA{A,C,G,T}ATCCCCGTAACCTTAGAGATAAAATCGAGGAGGCCCAAAAAGAAT
TAAATGGCGCGGAGGTTTCGAAGAAGGAAGTCCTCGAGGCTGGGGTTAAGGGGACGTCCGAGTCATTAAA
GGGGGTGAAGCGAAAGAAAATCGTAGCCGAAAACCACCTGAAGAAAATCCCCAAAAGTCCGCTAAGGAAT
CCTCTTCAAACGAAACACAAGCAGAACACAGAAGAACCCCCCTTCTCTGTCCTCCCGTCGGCAAGCGAAT
CCCATAAAAAACATAACTGCGTTCCGGCGAAACAGGGGAGACAATTTACGAAACAGAATGGGGAGACACC
TGGTATGACGGCTGAATCTAGCGAAAGCGGCGATTTGGTAAGTCCGAAGAAGACCTCATCACCGCATCAG
CGTTCGGAACTGAGGCGGTGGCGCAGTGAAGGAAGCGACCCAACCAGACTCTCCGGCCTGGATGGACAGC
GAGATTCGAGTAGCAGTTCTTCTAAAGCGCGTACGGACAATTCTGAGTGCAGCTCCCCTTGTTGCTCCAC
CACACCTCCTTCATACACATCAACGGCTTTTGACGTTCTGCTTAAGGCGATGGAACCTGAGCTGTCAACG
TTGAGCCAAAAAGGATCCTCATGTGCAATCAAAACCGAAAAACTCAGACCAAACAAGACTGTCAGGAGTC
CCTCCAAGTTGAAGAATTCCAGCCTTGATGCCCCGAACGCGACAAGCCCAGACCTTGTTGTCGAGTCTCC
ATGTCCTCCTTGTACATCGTATCCTGTACATGTGGCATCGACGCAAAAATCTGAACAAGTAGCCGCGCAA
TGTGTGTCGCATCTTTACTCTTCACAAGATCATCTAGTGCCAAAGTTGAGTCAACAGAACCAGCAACTAC
CCGGGCATCTGGGGTTTACAGGATCACTCACGAACCTCCATACTCTCGAATCGACGAAGCTTGAGCCCAT
CTACAATACGGCGGTTACGTCTACAGTCGGACTGACCTCTCCAAGCACCAGGACACAAGTGACACCCCCG
CATCAGCAGATGGACAGCGTATCCCCGTTGAGCGTATCGCCGGCCAGTTCAACCCAGTCACCACCCGGAC
CCATATATTCATCAGCACACGTAGCCTCAGTGGTATCTCAGAGCGTTGAACAGATGTGTTCACTTTTGCT
CAGGGACCAGAAGCCCAAGAAACAGGGCAAATACATTTGTGAGTACTGCAATAGAGCGTGTGCGAAACCT
TCAGTTCTCTTAAAGCATATTCGCTCCCACACTGGAGAGCGGCCTTACCCGTGTGTAACGTGCGGTTTTT
CATTTAAAACTAAATCCAATCTTTATAAACACAAGAAATCCCACGCGCATACCATCAAGCTCGGTTTGGT
TCTACAACCGGAAGCCGGGGGGCTCTTCCTCAGCCAGGAATGCCCGAAGGCCTTAAGCGTTCATTCGGAC
ATCGAAGATTCGGGTGAGTCTGACGAGGAGGGGCTAGCCGACGGGCGGCAGAACAATCCTTGCGTTAAGG
ATTTGCAGCCGGTCCAGACAATGAAAACCGTTAGTAACCCTGAATCTCTGCCCAAACTTATCCCGTCCAA
CAGTGATCACGTGGTCAGGGGTTTCAGCAGCCAGGACCGGCCGTCAGATTCTCAAGCCCCGACTGAGTTG
CCAAAGGTAGTGGTTCATCCTGTTTCAATGCCTCCGTTAAAGACGGACTGCCTTCAGGTCGCCAACCCTA
ATCCCGAGCTCCCAAGTCCCCAAAGTCCAAGGGATCTTCACGTCGCCAGCATATTATCACATTCTGCATC
CGTATCGTCTTTAGAAATGGACGAATCGTGCCACCAAAAGGGGGATGTGATACAAAGTGAGGGCAAACCA
GATTCTCACTCTGGTACTGCACACGCACAACTCCAACGACAGCAGGCGACCGAGGACCCTCAGGAGCAGC
AGGGGAAGTTATTGTTGTCTCCTAGGTCGCTTGGAAGTACGGACTCAGGGTACTTCTCGAGATCAGAGTC
GGCTGACCAAGCTGTGTCACCACCTACACCCTTTGCGAGAACTTTTCCGACAATGGATCCTGACCCGGCT
AAAAATGGAGGAGCTCCAGGCCCACGGATCTCAGCTCCAGCCCCTTCAGCGTTGGCGACGGGAGAGAAGA
GCTCCGTCGTAACCGGACAGATGCGCCCACCTCTCGCAACCAAGACACTGGAGGAAAGGATATCGAAATT
AATATCGGACAATGAGGCCTTGGTTGATGATAAACAACTTGATTCGGTTAAGCCTCGAAGGACCTCACTA
AGCCGTCGGGGTTCAATTGACTCTCCAAAAAGCTATATCTTCAAGGACAGCTTTCAATTTGATTTAAAGC
CGATGGGGCGTCGCACAAGCAGTTCCTCAGACATCCCCAAGTCCCCATTCACGCCCACCGAAAAGTCGAA
GCAAGTTTTTCTGCTGTCAGTCCCCTCGCTCGACTGCCTCCCTATTACGAGGAGCAACAGCATGCCGACT
ACTGGATACAGCGCCATACCCGCAAACATCATTCCACCGCCGCCTAGTTTAAGGGGGAGCCAAAGTTTTG
```

```
ACGACAAGATCGGGACATTATATGACGATGTGTTCGTTTCTGGTCCCAATCCGTCTATGCCTCCCAGTGG
TCATCATCGACCGCTAGTTAGGCAGGCAGCGGTAGAGGACAGTACGGCCTCCGAGTCTCACGTGCCGGGT
AGCGGCCAGTCCGTTGATGAGTCGTGCCAGGGGTGCCCGAGCTCGTCTGAGGCGGGCCCCGTGCAAAGCA
AAGCGGCCCAAACCCCACATTTGGAGAAAAAGAAATCACATCAGGGGAGAGGCACGATGTTCGAATGCGA
AACTTGTCGAAATAGGTATAGGAAGTTAGAGAACTTTGAGAACCACAAAAAGTTCTATTGCTCCGAGTTA
CACGGCCCTAAGACCAAGGCGGCGGTACGAGAAGCGGAACATGGACCGGCTCCAGGAGGGGCACAGCCAC
AAGTCCTGCATTATAGGGTAGCGGCACCTACGGCCGTATGGGAGCAGACGCCTCAGATTAGGAAGAGACG
GAAGATGAAGAGCGTAGGAGACGAAGAAGATCTTCAACCGCACGAATCTGGTCGCTCCCCAGAAAGTGCA
GACGCGCTACAATTACAGCCCGTACCTGGAGCCGCTCCTAGCCCCTCTAAGCATACGTCTGCCACTGCGG
CAGACCAAGCTCATCGGGGCGTCCAGCTCGTCGCACGAGGTCCTGAAAGGGCGCTGCCTCTAAAACAATG
TCCAATGGTGGAACAGCAGCTTAATGCTGCTGCTCAGGACAAAATGGAAGTCAAGCGTCAAGGGGGTGGT
ATCTCAGTCATTCAACATACTAATTCTTTAAGCCGACCCAACAGTTTCGACAAGCCAGAGCCCTTAGAAG
GGGGGATTACCTTCTCTTTGCAGGAATTAGGACGCACCGGTATGCCGGGGGCGCTTAAGGTCATCGGTAT
GGCACCAGAGGAGGGGCACCCGCCACAAGACGCTATGCACCAAACCGCACTTTCCCACAATCTCCGCGGC
GAACCACGCGAAAGCGCCAGGAAAATTCCGTCAGAGCGATATGTCCTCGGTCAACCACTTAGGCTTGTTC
GACAGCATAACATACAGGTGCCAGAAATACTCGTCACGGAGGAACCGGATCGGGATCTGGAAGCCCAATC
ACATGATGAAGAGAAATCCGAGAAGTTTACGTGGCCGCAGAGATCCGAGACGCTGTCAAAACTTCCAACC
GAGAAGTTGCCGCCTAAGAAAAAGCGGCTACGACTTGCTGAGATCGAGCATTCAAGCACAGAAAGCAGCT
TTGAATCAACCTTGTCGCGGTCCCTATCCCGCGAGTCTTCCCTTAGCCACGCGAGCTCTTTTAGCGCGTC
TCTAGACCTCGAGGACATTTCAAAAGTTGAGCTTGCGCCTAAAATCGATTTTCCATCTAAGGCCGAATTT
CTCCTAATACCACTGGGATCCAACACTCT{A,C}TCCGTGCCGGGGTCACATCGCGAAATGCGTAGGGCTGCAA
GCGAGCAGATCTCGTGCGTTCCCACCCTCATGGAAGTATCGGATTTTCGTAGTAAGTCTTTCGACTGCGG
GAGTATCGCCCCGTCCCATGTTGTCCCTGCGCTGGTTGAATCGCAACCGAGCTACTCTCCGTCAGCTGTT
GGCGGCACAGCCCACGTCCCACTGTTGGAGAGGCGGAGAGGTCCGCTTATCCGTCAGATTAGTCTGAACA
TTGCTAGCGATTCACACCTATCCCCAGGCAGCGCAGCAGCTCTGCAAACCATTGTCTTGCCAAGTGTAAA
TACAGTCCCATTCCAGGCGCCCCGTTTACCGGATATGGCTTCCGCCGATTGCCCGGCCCATACTGTCCAT
CCTCAGGCCTTGGCAAAAGACCTCCAGGCCGAGATTTCGTCTTCGAGCAGTACAGATACCTTTCCTCCGC
AGCAACTATTTGGAGCTCACTTGCTCAATAAAACCAATACTTCTCTGTCACATCAGAACACGCCGCTTCC
TCTGCCCGTATCAGCTCAGGGTGGGAAACCGGACGCAGCTCCAACAGCATGCGTTAGCTCCACGGGGGAG
GGAAGCTTTGCACCAAAATACCAGCTACAATGTCAGGCTTTTACGTCAGACCAGGGTTGCTCCGCCCCGT
TACGCTCTAGCCCAAACCAAGTTCTGCCAGGCCAGGCCGGTGCGGATCCCTGCCCAGCTTCAGAAGCTCC
GCCAGCTAAGGCCGCTGACCCAATGGCCAAACCTTGCCCCCTTCCTCCACTCGAACTGGGTCTACCGCGT
GACGAAGTTTTGCAGAAACAGCTGCCCTCATTCGTTCTGCCGGTGCCTCAAAAGCGAAACGTGACGGTGG
ATTGCTTTACCCCGGTGACGTCGTTACCCCAGATTCTTGTGACTCAAGACCTACCAAATATGCCCATATG
CCAGACTAATCAGTCCCTAGTCCCTGTAAGCGAAGAGCAGAATTCTATGCCGAAATCGCAAAATTACTTA
CAAAACGCTTCTCCCACACCGGAGAAGGAGTTAGCGTGTAAGACTGTACTCCCCGAAGTTGGCCAATCAG
TTCCGGTTTCGGAGTCAAGTCCAACTGTGCAAAAGGTCTCAGTAGGACGCCTATCGCCCCAACAGGAATC
TTCAGCGTCATCAAAGCGGATGCTTTCTCCAGCCAATAGTCTCGACATAGCCATGGAAAAACACCAAAAA
CGTGCGAAAGATGAAAACGGCGCAGTATGTTCGACAAACATTAGGGCATTGGAGTTGCCTAGTTCTCGTG
CGAACGAATCACATAAGCAGAAGAAACCGGTATTGGTGCGTCAACTGTGCACAACTGAACCCTTAGAAGG
CGCTGCGCTAGAGCAAGGGGCGTGCTCAGCTAGCGGCAGAAGCTCTAACAAGGCGGCAAATCTGACCCAA
GTGTTGCCAACGGACTCACTCTCTAGCAGGCCATCGACGTTTGCGGTAACCGACCACGTAAATGAATTAC
AGGAATTCAAGAACACGAAGGTATCCACCTCCTTGACGCCGACGGTGGGTTCCTTCCCGGCTCCAAGCGA
GAGCGCTTGCGTCCTCCCACTTAAGAGTATTGATAATAACCAAGAGAAAGGCTCCCCTGGGGTACGGCAT
GAAGAGAACAAAGTAATTCAGGGCCAGCGGCCCCCCTTAGTGAGCGGTCTTAGCTTGGTCAGTTCCTCCG
ATACCCAACAGCCGTCGTTCCCCTCCCTGAAAACGGCCACATCGTTTACGTGGTGTTACCTTCTCAGGCA
AAAGGCTCTGCCTCTCGCACAAAATGATCAGAAAACAAGTGCTTACACAGGGTGGACAGTTTCTTCAAGT
AATCCAAACCCATTGGGATTGCCTACGAAAGTGGCGTTGAGCCTTCTAAACTCAAAGCAAAAAACAGGGA
AATCGCTTTATTGCCAGGCAATTACAACACACTCTAAGTCCGATCTGGTTGTTTACTCCTCTAAGTGGAA
AAACAACTTAAGCAAAAGGGCTCTGGGAAATCAAAAAGCAACTGTTGTAGAATTCAGCAATAAGGACGAC
TCCGAGATTAACTCCGAACAGGATAAAGAGAACTCCCTAATTAAAAGTGAACCCAGACGTATCAAAATTT
TTGATGGCGGCTATAAGTCGAACGAAGATTATGTTTATGTACGTGGTCGGGGTAGAGGGAAGTACATCTG
TGAAGAATGTGGAATACGCTGCAAAAAACCGTCAATGCTTAAAAAACATATTCGAACGCATACCGACGTA
CGGCCATACCACTGCTCGTATTGTAATTTTTCATTCAAAACGAAGGGAAACCTGACAAAACACATGAAAA
GTAAAGCACACAGCAAGAAGTGCGTCGATTTAGGAGTAAGTGTCGGGCTAATTGATGAGCAAGATACTGA
GGAATCTGACGAAAAACAGCGATTTGGATGTGAGAGGTCTGGGTATGATCTAGAAGAAAGTGACGGACCC
```

```
GATGAGGACGATAATGACAATGAGGAAGATGATGACGATAGCCAGGCTGAGTCTGGCTTGTCTGCGGCTC
CTAGTGTGACGGCGAGCCCTCAACACCTACCTTCCCGTTCTGGACTCCAGGACCCAGGCTCTGTCGAGGA
GGAACTGCGTGTTAGTTCGTGCTTTTCAGGGGTACATACCGACCCCATGGATATCCTCCCGCGTGCCCTA
CTAACCAAGATGACCGTGTTATCCACTGTGCAAAGTTCGCCTAATCGAACTGACTTACCTGCAAAGGCAA
GACAAAGTACGGAGAAGGATGAACATGAACAGGCCCCCCCAGCCGACACGCCACGGTCACCCGGGCATCA
GCTTTCCGTGCACAGCTCAGAATCGGACGTGCTTCGCTCCCCCGCGGCCGGGAATCCTGCAGCTGGGTCG
CCAGGCGCAGCTGTTCAGGATTCTTCCGTTGGTCTCCCGCCCGCGGTGGCGCAACTGAATCCTCAGCCGG
CAGCTCGTATTTCTAGTTCCGTATCGCCGCATCCCGATTCTCAAGACCAGAAACAGCAAATAATTTTACA
ACCTCCACCGGGATTACCATCGCCGCAAACTCACTTATTCTCCCACTTGCCATTACATAGTCAACAACAA
TCACGGACGCCGTACAACATGGTGCCAGTTGGGGGTATACACGTAGTAACAGCTGGGCTGACGTATTCAA
CATTTGTACCGATCCAAGCTGGCCCAATGCAGCTGACGATCCCAGCCGTGTCGGTTATCCACCGAACAGT
GGGCACAAGCGGGGATACGATAACCGAGGCGTCTGGGTCCCCCAACCGACCCACCGGGGTAGCGGAACTT
AGCAGCGTCGTTCCCTGCATTCCTATAGGTCAGATCCATGTCCCTGGACTGCAGAACCTGTCGCCGCCGG
CGCTCCAATCTTTGACGTCACTAGGCATGGAGACAGTCAACCTCGTGGGATTAGCGAATGCTACGGTTGG
ACCCCAGGGACATCCGCCTGGATTGGCCCTGAATGCCGTGGGACTTCAGGTCCTAGCAAATGCCCCCGCA
CAGTCGTCGCCCGCACCTCCGGCTCATATCCAGGGACTGCAGATTCTCAATATCGCTTTGCCTACTCTCA
TCCCCTCGGTTGGCCCTGTTGCGGTTGGAACCACCGGAACACCCGAAACCACCGCGCCCAATAGTAAAGC
TATGGAACTTCAGATGCCAGCCGGCCAAGGACATTCCGCCGAGCCCCCACAAGGTAGTCCCGAAGGACCC
CAAGAAACGCCACAGACGGTGTCCGGGCCTAGTGCTGACCACGCCCGACCGGAGGATTCGACCAAGATGG
ACACAAAGAAAGGTCCTAGTGCGGGACACGTCTTACCCGGGAGATCCCCGGCCCAGGCTCAGCCGGCACC
GACTCCTGAAGCTCTGCAAAAGGTTGCTACCTCAGCCCCACCCTCACTACCGACGGACCGAGCCGCTCCC
AGACCGCCTGTGCCCCACCGTCAGCCCATAGTCCACTTCTCAGATGTATCCTCTGACGATGATGAGGATC
GATTGGTTATTGCAACTTAG
```

Salida:

```
MPRTKQIHPRNLRDKIEEAQKELNGAEVSKKEVLEAGVKGTSESLKGVKRKKIVAENHLK
KIPKSPLRNPLQTKHKQNTEEPPFSVLPSASESHKKHNCVPAKQGRQFTKQNGETPGMTA
ESSESGDLVSPKKTSSPHQRSELRRWRSEGSDPTRLSGLDGQRDSSSSSSKARTDNSECS
SPCCSTTPPSYTSTAFDVLLKAMEPELSTLSQKGSSCAIKTEKLRPNKTVRSPSKLKNSS
LDAPNATSPDLVVESPCPPCTSYPVHVASTQKSEQVAAQCVSHLYSSQDHLVPKLSQQNQ
QLPGHLGFTGSLTNLHTLESTKLEPIYNTAVTSTVGLTSPSTRTQVTPPHQQMDSVSPLS
VSPASSTQSPPGPIYSSAHVASVVSQSVEQMCSLLLRDQKPKKQGKYICEYCNRACAKPS
VLLKHIRSHTGERPYPCVTCGFSFKTKSNLYKHKKSHAHTIKLGLVLQPEAGGLFLSQEC
PKALSVHSDIEDSGESDEEGLADGRQNNPCVKDLQPVQTMKTVSNPESLPKLIPSNSDHV
VRGFSSQDRPSDSQAPTELPKVVVHPVSMPPLKTDCLQVANPNPELPSPQSPRDLHVASI
LSHSASVSSLEMDESCHQKGDVIQSEGKPDSHSGTAHAQLQRQQATEDPQEQQGKLLLSP
RSLGSTDSGYFSRSESADQAVSPPTPFARTFPTMDPDPAKNGGAPGPRISAPAPSALATG
EKSSVVTGQMRPPLATKTLEERISKLISDNEALVDDKQLDSVKPRRTSLSRRGSIDSPKS
YIFKDSFQFDLKPMGRRTSSSSDIPKSPFTPTEKSKQVFLLSVPSLDCLPITRSNSMPTT
GYSAIPANIIPPPPSLRGSQSFDDKIGTLYDDVFVSGPNPSMPPSGHHRPLVRQAAVEDS
TASESHVPGSGQSVDESCQGCPSSSEAGPVQSKAAQTPHLEKKKSHQGRGTMFECETCRN
RYRKLENFENHKKFYCSELHGPKTKAAVREAEHGPAPGGAQPQVLHYRVAAPTAVWEQTP
QIRKRRKMKSVGDEEDLQPHESGRSPESADALQLQPVPGAAPSPSKHTSATAADQAHRGV
QLVARGPERALPLKQCPMVEQQLNAAAQDKMEVKRQGGGISVIQHTNSLSRPNSFDKPEP
LEGGITFSLQELGRTGMPGALKVIGMAPEEGHPPQDAMHQTALSHNLRGEPRESARKIPS
ERYVLGQPLRLVRQHNIQVPEILVTEEPDRDLEAQSHDEEKSEKFTWPQRSETLSKLPTE
KLPPKKKRLRLAEIEHSSTESSFESTLSRSLSRESSLSHASSFSASLDLEDISKVELAPK
IDFPSKAEFLLIPLGSNTLSVPGSHREMRRAASEQISCVPTLMEVSDFRSKSFDCGSIAP
SHVVPALVESQPSYSPSAVGGTAHVPLLERRRGPLIRQISLNIASDSHLSPGSAAALQTI
VLPSVNTVPFQAPRLPDMASADCPAHTVHPQALAKDLQAEISSSSSTDTFPPQQLFGAHL
LNKTNTSLSHQNTPLPLPVSAQGGKPDAAPTACVSSTGEGSFAPKYQLQCQAFTSDQGCS
APLRSSPNQVLPGQAGADPCPASEAPPAKAADPMAKPCPLPPLELGLPRDEVLQKQLPSF
VLPVPQKRNVTVDCFTPVTSLPQILVTQDLPNMPICQTNQSLVPVSEEQNSMPKSQNYLQ
NASPTPEKELACKTVLPEVGQSVPVSESSPTVQKVSVGRLSPQQESSASSKRMLSPANSL
DIAMEKHQKRAKDENGAVCSTNIRALELPSSRANESHKQKKPVLVRQLCTTEPLEGAALE
```

```
QGACSASGRSSNKAANLTQVLPTDSLSSRPSTFAVTDHVNELQEFKNTKVSTSLTPTVGS
FPAPSESACVLPLKSIDNNQEKGSPGVRHEENKVIQGQRPPLVSGLSLVSSSDTQQPSFP
SLKTATSFTWCYLLRQKALPLAQNDQKTSAYTGWTVSSSNPNPLGLPTKVALSLLNSKQK
TGKSLYCQAITTHSKSDLVVYSSKWKNNLSKRALGNQKATVVEFSNKDDSEINSEQDKEN
SLIKSEPRRIKIFDGGYKSNEDYVYVRGRGRGKYICEECGIRCKKPSMLKKHIRTHTDVR
PYHCSYCNFSFKTKGNLTKHMKSKAHSKKCVDLGVSVGLIDEQDTEESDEKQRFGCERSG
YDLEESDGPDEDDNDNEEDDDSQAESGLSAAPSVTASPQHLPSRSGLQDPGSVEEELRV
SSCFSGVHTDPMDILPRALLTKMTVLSTVQSSPNRTDLPAKARQSTEKDEHEQAPPADTP
RSPGHQLSVHSSESDVLRSPAAGNPAAGSPGAAVQDSSVGLPPAVAQLNPQPAARISSSV
SPHPDSQDQKQQIILQPPPGLPSPQTHLFSHLPLHSQQQSRTPYNMVPVGGIHVVTAGLT
YSTFVPIQAGPMQLTIPAVSVIHRTVGTSGDTITEASGSPNRPTGVAELSSVVPCIPIGQ
IHVPGLQNLSPPALQSLTSLGMETVNLVGLANATVGPQGHPPGLALNAVGLQVLANAPAQ
SSPAPPAHIQGLQILNIALPTLIPSVGPVAVGTTGTPETTAPNSKAMELQMPAGQGHSAE
PPQGSPEGPQETPQTVSGPSADHARPEDSTKMDTKKGPSAGHVLPGRSPAQAQPAPTPEA
LQKVATSAPPSLPTDRAAPRPPVPHRQPIVHFSDVSSDDDEDRLVIAT
```

8.2. Ejemplo 4:An05g01270 - antibodies against GOR are present in individuals with hepatitis C

```
AUGATGACTATTACCATGTTTACACCCATAGAGTTGCCACCAACCTATTACATAAGCCTCCACATGAATG
TGCCGATCTCACCCCTCTATCCATGGGGAATGCCGTATACCCGGTCTACATCCTTCTCGCATACTACCAC
CATCTATACGAATACCAATACAAATACCACAACTATTGAACGAAAGGATTCGGGAATCTGCATACCATCA
AATACCACGTCCCGAAAACCAAGCACAATTACGATTAATCCCAAGCCAACCACAACTACAACGTACGCGG
CGCCTAGCCGGTGGACAACCCTTACGTCTCCTGCGGAGCAAACAAAGGCTCTTGACTTTCTTACTAACAA
TTATCATACGATTCAGACCTTACAGGCAGCGAACTATACCATTACTCCGGATACAACTAAGAATAATATT
CGGTTTCAGCAGACGCCAAGCACTAATGACCCTTCTGCGCGGCGTGCAGTCAGTCTGGACTGCGAAATGG
TAACGGATACAACTGGCCGGACTCAATTGTGCCAAGTAAGTATGGTTGATGTTCTGACAGGCGAACTGCT
GCTGGACCAGCCAGTTCTTCCGAGCGAACCCGTCTCAGACTGGAGAACCAAATGGAGTGGGATGACCGCC
GAGTTAATGGCCCAGCACGTCGCGGCCGGCCGTACCGTCAACGGGTACGAAGGAGCGAGAGCACGCGTGT
GGGAATACATTGATCAAAAAACAATCTTAGTAGGGCAGAGTCTCAACTTTGATCTTGACGTTTTAGGCAT
AGTGCACGAGCGGGTCGTAGACACTTACCTGCTCATGCGGGGACAGAGACGGGGACGTTCTTGTAGGTTA
AGGGATGTCGTCCGGGACTGCTGCGGAGTCGAGATACAGAAAGGCGAAGAACTCGAGGGGGGGCATGATT
GTGCGGAAGACTGTTATGCTGCGAGGGAAGTAGCGCTTTGGGCAGTCGAACATCTTGGACGAGATGAGGC
GTTCATCTCCAGGAGTATGTCCCCCATAGATGACAAATTTCAGTTGTCGTCTGCATACTTCCCCGCGACC
TGTTATTGTATCCCTCGGGATGCTGAGTGGGCTACTTTTCATGGTTGGACATTGAGGGTGGATACATACG
ACACCACTCACGTTAATTACTAA
```

Salida:

```
MTITMFTPIELPPTYYISLHMNVPISPLYPWGMPYTRSTSFSHTTTIYTNTNTNTTTIER
KDSGICIPSNTTSRKPSTITINPKPTTTTTYAAPSRWTTLTSPAEQTKALDFLTNNYHTI
QTLQAANYTITPDTTKNNIRFQQTPSTNDPSARRAVSLDCEMVTDTTGRTQLCQVSMVDV
LTGELLLDQPVLPSEPVSDWRTKWSGMTAELMAQHVAAGRTVNGYEGARARVWEYIDQKT
ILVGQSLNFDLDVLGIVHERVVDTYLLMRGQRRGRSCRLRDVVRDCCGVEIQKGEELEGG
HDCAEDCYAAREVALWAVEHLGRDEAFISRSMSPIDDKFQLSSAYFPATCYCIPRDAEWA
TFHGWTLRVDTYDTTHVNY
```

8.3. Ejemplo 5: Basado en Q98147 - Kaposi's sarcoma-associated herpesvirus cyclin homolog

```
AUGATGGCTACAGCCAACAATCCCCCGAGCGGTCTCTTAGACCCAACCTTATGCGAAGACAGAATCTTTT
ATAACATATTGGAAATCGAGCCTAGATTCCTAACCTCTGATTCGGTCTTTGGGACTTTCCAGCAGTCTCT
AACATCCCACATGAGAAAACTTCTCGGAACATGGATGTTCTCGGTCTGTCAAGAATACAATCTTGAACCG
AATGTAGTCGCGCTTGCACTCAATCTTCTGGACCGGTTGTTACTCATCAAGCAAGTGTCAAAAGAACACT
TCCAAAAAACTGGCTCCGCTTGTTTATTAGTCGCTTCTAAATTACGCTCACTTACGCCTATAAGTACATC
AAGCTTATGTTACGCGGCCGCAGACTCCTTTTCGCGACAGGAGTTGATAGATCAAGAGAAGGAGCTTCTG
```

```
GAAAAATTGGCGTGGAGGACGGAAGCAGTTTTAGCCACTGATGTAACTAGTTTTCTCCTTCTGAAACTCG
TAGGAGGGTCACAGCACTTAGACTTTTGGCATCATGAAGTCAACACGTTGATCACTAAGGCGTTAGTAGA
CCCGAAGACAGGTTCTCTTCCGGCTAGCATAATTAGCGCGGCGGGGTGTGCGCTACTGGTGCCGGCTAAC
GTTATCCCCCAGGATACACATAGTGGAGGAGTCGTCCCACAGTTAGCCTCGATCTTGGGGTGTGATGTAT
CAGTTCTTCAGGCTGCGGTAGAGCAGATCCTAACGTCTGTGTCAGACTTTGACTTACGTATCTTGGATAG
CTAT{-}[775]ATGAAATGCCTGTCTTCCATCTTCCTGCCCTGCCTCCAGGCTAGGAAATTACGCCGTCAACGACGGC
AGGATTATCACGCTTCCTGGACGTGTGCAGGGATATCCGAATTGGCTCAGCCTACTGAGCCCTTTCCGGC
CTATGAACATGAAAAACATCACCATGGAAGCACTGAACTATCAGCACTGCCAGATCAACCTAAAGACGCT
GTGACTAAAACTCCACAGATGACCCCGAGCATCCGCCTAGTCAACCCGGCCGATTCAGAGTGCTCGACGC
TAGACGACCCTGACGAAAAGCAGCCTCAGGGTATTCAGGATGACATTGCCGCAAAGCAAACCACGACCTT
ATCAGACGACGACACCGACGTCGAATCTGGTAGCGAGAAGACCTCAAGGGTCATTGACATGAAGGTTGGT
CCTCAGACCGAAGAACCCCTAAACCCAAAGGCGCACCTGTCCCCGAAAAGCAAGGATGACATCGACCTAG
AATTATCCTCCAGGAGAATCCCTCGATCCCCAGACAAGAAACCTAGGCCAGCGTCCATGGAAGTTCCCCC
TTCGGCGGCCGCCAAGCTTCCCGAAATTAAGTCCAGGATAATCGAAGACATCCCTGAAGACGTAGAGGAA
GAAGATAAACATGACGCAGACATTAAGCACACAGTCTCAGCAGTCCCGGACGAACAAGCAAAACCAGAAG
AGGAATATGGGTCAGAAACCGCCGCAACGAAGCCGACGAAACGTACGTCTACCTGGAGATTATCTCAACG
AAAGAGTATGAATGAACTGTTTAATTTGTTACAATCCACAGCGGCCGCAGTTGCCGCCGCTCCAAAGTTA
AGCAACCTGAAATTCGCGATCCCGCCCAAGTCGCCGTTACGTGCCTCACCAACTAACGATTCGCCGCATC
ATTCCTATGTTAGCTCCGTAAGTAGCACAAGTAACCGTCCCCCGACCCCGCCTCCGAAATCCCCTGTCCT
ACGTCCCAACTCGCGCCCTTTGCCAGATCCGGCACAATTAAGTAGTAGTGCACCACCGGCAAGCGGAACC
TCGCTGGGAAGCCTGAGTAGTAGTCCGACTAAGAAGCAACGTCGAATGGGTACAGCAGTCTTTCCACTTC
TCCCTTGTAAGTGGGCTGGCATCCATGACGAGGACATGAAGGAAAAGCATTCTAACGATACCAATAGAAA
TAGTCAGACTTTATTTCAGTGA
```

Dos secuencias:

```
MATANNPPSGLLDPTLCEDRIFYNILEIEPRFLTSDSVFGTFQQSLTSHMRKLLGTWMFS
VCQEYNLEPNVVALALNLLDRLLLIKQVSKEHFQKTGSACLLVASKLRSLTPISTSSLCY
AAADSFSRQELIDQEKELLEKLAWRTEAVLATDVTSFLLLKLVGGSQHLDFWHHEVNTLI
TKALVDPKTGSLPASIISAAGCALLVPANVIPQDTHSGGVVPQLASILGCDVSVLQAAVE
QILTSVSDFDLRILDSY
```

y

```
MKCLSSIFLPCLQARKLRRQRRQDYHASWTCAGISELAQPTEPFPAYEHEKHHHGSTELS
ALPDQPKDAVTKTPQMTPSIRLVNPADSECSTLDDPDEKQPQGIQDDIAAKQTTTLSDDD
TDVESGSEKTSRVIDMKVGPQTEEPLNPKAHLSPKSKDDIDLELSSRRIPRSPDKKPRPA
SMEVPPSAAAKLPEIKSRIIEDIPEDVEEEDKHDADIKHTVSAVPDEQAKPEEEYGSETA
ATKPTKRTSTWRLSQRKSMNELFNLLQSTAAAVAAAPKLSNLKFAIPPKSPLRASPTNDS
PHHSYVSSVSSTSNRPPTPPPKSPVLRPNSRPLPDPAQLSSSAPPASGTSLGSLSSSPTK
KQRRMGTAVFPLLPCKWAGIHDEDMKEKHSNDTNRNSQTLFQ
```