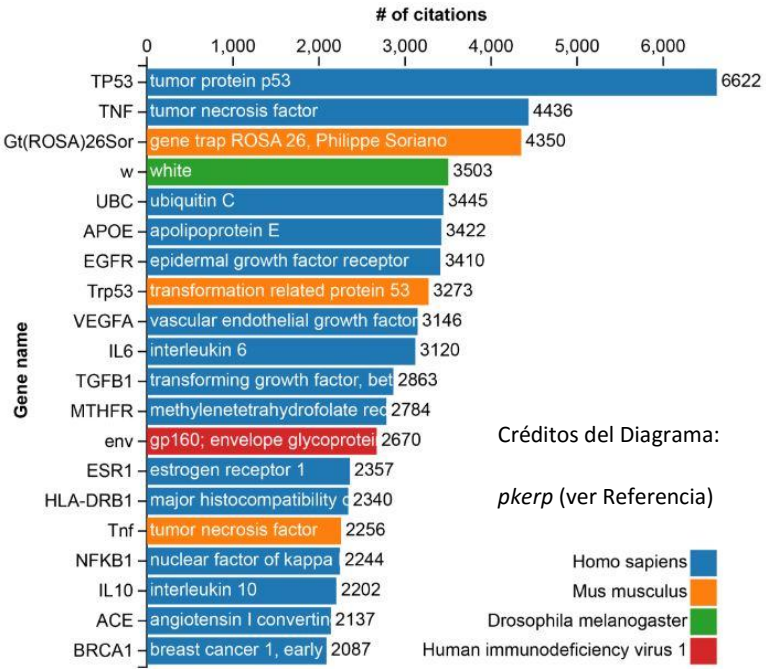


DISEÑO DE UN SISTEMA CLASIFICADOR PARA LOS 15 GENES HUMANOS MÁS ESTUDIADOS DEL MUNDO

Para Fernando Castro-Chavez (para su posterior elaboración de un artículo científico en el campo de la Bioinformática)

MRYSI

RVOE ES/142/2002, 29/NOV/2002

Maestría en Redes y Sistemas Integrados

Para el sistema integrado de los genes, se han identificado 15 genes humanos que han sido de los más estudiados en todo el mundo debido a su importancia funcional. Cuando se encuentran alterados, son importantes al causar enfermedades; sin embargo, el cuadro original de los mismos carece de una adecuada distribución para hacer posible el estudio de cada uno y de sus funciones de manera independiente; por lo tanto, el propósito de este trabajo es facilitar dicha distribución para su más detallada investigación individual; se presentan cinco características de cada gene, incluyendo el nexo a su información detallada en el internet (entrez).

Fernando Castro Chávez

Programación Orientada a Objetos Avanzada

Contenido

Objetivo

Diseñar un sistema que permita clasificar a los 15 genes humanos actualmente más estudiados en el mundo, los cuales son importantes por su participación en diversas funciones vitales para el ser humano, y los que cuando se encuentran en un estado alterado, es decir, dañados, provocan enfermedades hereditarias debido a encontrarse en un estado anormal. El sistema aquí presentado genera inicialmente el cuadro, y lo exporta en archivo CSV, visible poblando cada célula del *Excell*, para luego capturar la información del cuadro generado, y distribuirla en su texto (*.txt) individual ubicándose dentro de sus correspondientes carpetas; con esto se pretende facilitar el análisis de los genes para su estudio individual, al separarlos en sus directorios correspondientes describiendo las siguientes, que son sus cinco características más importantes, las cuales también aparecen en la Consola del Editor del Programa *Netbeans*, junto a sus más relevantes cálculos matemáticos como conclusión; Tales características son: 1) Su jerarquía en investigación; 2) Su nombre técnico abreviado; 3) Sus dimensiones en unidades de bases de nucleótidos; 4) La hebra a la que pertenecen; y el 5) Acceso a internet para tener la información más reciente y detallada de cada gene (mediante su número entrez).

Planteamiento del problema

Fernando Castro-Chavez (el investigador molecular independiente), le ha solicitado a Fernando Castro Chávez (el estudiante de la Maestria en MRYSI en LANIA), el realizar la distribución independiente en 15 folders o directorios de los 15 genes humanos más estudiados por todo el mundo en la actualidad.

Los genes se dividirían en sus directorios respectivos (incluyendo dentro de cada uno su nota informativa en texto (formato TXT) con la información básica para cada gene). El programa genera primeramente el cuadro en formato CSV, y luego lo procesa para obtener sus líneas individuales a ser exportadas en columnas de texto para cada gene, los cuales son los siguientes según su importancia en la investigación actual:

1) TP53, con una longitud de 25,772 bases y presente en la Hebra Negativa (-), con número entrez de 7157;

2) TNF, con una longitud de 2,770 bases y presente en la Hebra Positiva (+), con número entrez de 7124;

3) UBC, con una longitud de 5,765 bases y presente en la Hebra Negativa (-), con número entrez de 7316;

4) ApoE, con una longitud de 3,647 bases y presente en la Hebra Positiva (+), con número entrez de 348;

5) EGFR, con una longitud de 237,600 bases y presente en la Hebra Positiva (+), con número entrez de 1956;

6) VEGFA, con una longitud de 16,304 bases y presente en la Hebra Positiva (+), con número entrez de 7422;

7) IL6, con una longitud de 6,561 bases y presente en la Hebra Positiva (+), con número entrez de 3569;

8) TGFB1, con una longitud de 52,347 bases y presente en la Hebra Negativa (-), con número entrez de 7040;

9) MTHFR, con una longitud de 21,198 bases y presente en la Hebra Negativa (-), con número entrez de 4524;

10) ESR1, con una longitud de 472,929 bases y presente en la Hebra Positiva (+), con número entrez de 2099;

11) HLA-DRB1, con una longitud de 36,859 bases y presente en la Hebra Negativa (-), con número entrez de 3123;

12) NFKB1, con una longitud de 115,974 bases y presente en la Hebra Positiva (+), con número entrez de 4790;

13) IL10, con una longitud de 7,006 bases y presente en la Hebra Negativa (-), con número entrez de 3586;

14) ACE, con una longitud de 21,320 bases y presente en la Hebra Positiva (+), con número entrez de 1636;

15) BRCA1, con una longitud de 125,951 bases y presente en la Hebra Negativa (-), con número entrez de 672;

Finalmente, cabe señalar que en los resultados finales de este trabajo, se liga automáticamente cada gene a su correspondiente número de 'entrez' para obtener su nexo en la red y así obtener su información detallada, siendo siempre también la más actualizada.

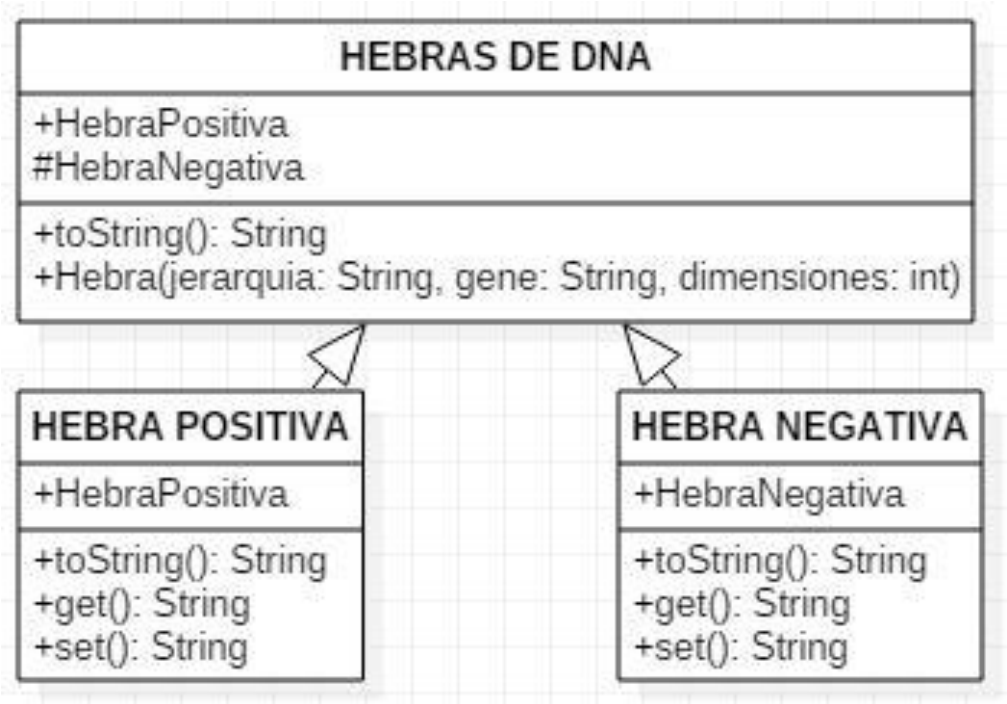
Referencia: **Empty Pipes**. *The 20 Most Studied Genes*. 08/Dec/2014. URL: http://emptypipes.org/2014/12/08/gene-popularity/ [El cuadro usado fue diseñado en base a las características de los genes yendo manualmente a sus referencias originales en el *Genbank*, lo cual aquí se ofrece de manera automática en el producto final].

Datos de ejemplo

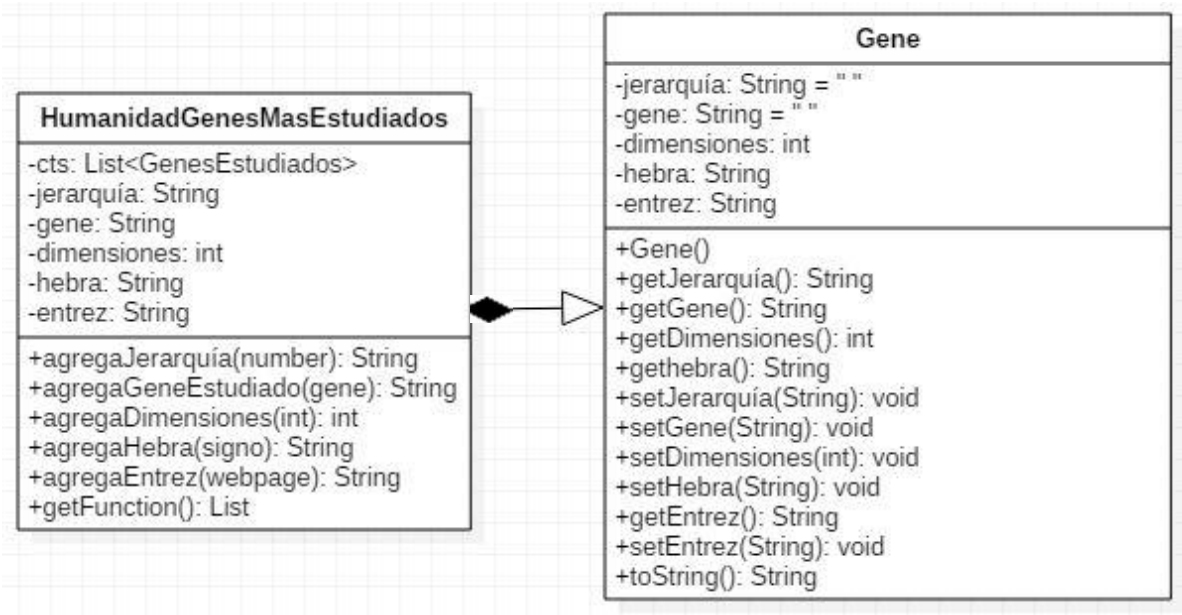
JERARQUIA	GENE	DIMENSIONES	HEBRA	ENTREZ
1	TP53	25772	-	https://www.ncbi.nlm.nih.gov/gene/7157
2	TNF	2770	+	https://www.ncbi.nlm.nih.gov/gene/7124
3	UBC	5765	-	https://www.ncbi.nlm.nih.gov/gene/7316
4	ApoE	3647	+	https://www.ncbi.nlm.nih.gov/gene/348
5	EGFR	237600	+	https://www.ncbi.nlm.nih.gov/gene/1956
6	VEGFA	16304	+	https://www.ncbi.nlm.nih.gov/gene/7422
7	IL6	6561	+	https://www.ncbi.nlm.nih.gov/gene/3569
8	TGFB1	52347	-	https://www.ncbi.nlm.nih.gov/gene/7040
9	MTHFR	21198	-	https://www.ncbi.nlm.nih.gov/gene/4524
10	ESR1	472929	+	https://www.ncbi.nlm.nih.gov/gene/2099
11	HLA-DRB1	36859	-	https://www.ncbi.nlm.nih.gov/gene/3123
12	NFKB1	115974	+	https://www.ncbi.nlm.nih.gov/gene/4790
13	IL10	7006	-	https://www.ncbi.nlm.nih.gov/gene/3586
14	ACE	21320	+	https://www.ncbi.nlm.nih.gov/gene/1636
15	BRCA1	125951	-	https://www.ncbi.nlm.nih.gov/gene/672

Diseño de clases

1. Para el sistema de las hebras positiva y negativa, la parte primera de este proyecto en UML se representa de la siguiente manera:



2. Para el sistema de la distribución individual de cada gene en su respectivo directorio con su ficha informativa en texto, la parte segunda de este proyecto en UML se representa de la siguiente forma:



Conceptos utilizados

Clase	Objetivo	Comentarios
Generalización	Se generó una clase padre y dos clases hijas	La clase padre de Hebras de DNA incluye a las características que extienden o heredan las siguientes 2 clases hijas: Hebra Positiva y Hebra Negativa
Asociación mediante Agregación	Se produjo una clase base para describir a los genes más estudiados por la humanidad con navegabilidad hacia cada Gene individual	Esto nos permitió generar un esquema más compacto desde HumanidadGenesMasEstudiados en su conjunto, hasta cada Gene en lo individual
File	Esta clase nos permite recuperar el Archivo, generado inicialmente en formato CSV para *Excell*, así como el producir los nuevos archivos en formato TXT individuales	Para recibir y luego leer el archivo, le anidamos a la clase "File" la dinámica clase FileReader, y lo mismo para escribir en él, al agregarle FileWriter
BufferedReader y PrintWriter	Poder manejar la información de una manera más eficaz	Se optimiza el flujo, tanto al recibir, leer, y procesar la información para poder clasificarla y enviarla escrita
String	Para transformar y separar los componentes	Se utiliza toString(String)

Notas: Otros siete conceptos importantes usados en este programa son: ***1***) Atributos Protegidos (-) para evitar que sean arruinados protegemos a los atributos de los Genes; ***2***) Métodos No Protegidos (+), para permitir su fácil acceso no protegemos a los métodos de los Genes; ***3***) setters y getters para desarrollar los Atributos get, y los Métodos set y get, los cuales nos ayudan a obtener el valor de los atributos; *v.gr.*, getGene(): String; ***4***) Se usa el método Split para separar a cada línea, dividiendo a la cadena en sus componentes, es decir, en cada una de las líneas representando a cada uno de los 15 Genes; ***5***) La colección usada es la de Listas para obtener cada gene independientemente a partir de una serie de líneas que es accesible mediante su índice; y ***6***) Streams con entrada a un documento CSV

previamente elaborado por el mismo programa, y como salida 15 diferentes documentos individuales de texto en formato TXT conteniendo las características básicas de cada columna para darnos una idea de la longitud y de la ubicación de cada gene; ***7***) Uso de split, transformando los datos individuales que inicialmente se encontraban en una fila, para terminar quedando en una columna con cada uno de sus componentes etiquetado; y finalmente, ***8***) Excepciones, para evitar la parálisis del programa capturando sus problemas previsibles, usamos el try… catch, con lo que el programa se controla si el archivo no existiera, o si fuera problema del sistema, siendo el primero un: try… catch (IOException e) {e.printStackTrace();}; y los dos últimos siendo: try... catch (FileNotFoundException ex) {(Project.class.getName()).log(Level.SEVERE, null, ex);} y try... catch (IOException ex) {(Project.class.getName()).log(Level.SEVERE, null, ex);}.

Resultados

Con este programa basado en un sistema de los 15 genes más estudiados en el mundo, nos fue posible el generar primeramente un archivo CSV con dichos genes, y luego distribuir cada uno en quince diferentes directorios, conteniendo cada directorio la ficha informativa con los datos básicos de cada uno de ellos en archivo TXT, y al mismo tiempo presentarlos en la consola del editor de programación *Netbeans*; estos cinco datos son: 1) Su jerarquía o importancia en la investigación biomédica, 2) Su nombre técnico abreviado de uso universal, 3) Su longitud en unidades de bases, 4) Si pertenece a la hebra positiva o negativa, lo que se representó con los signos + y -, respectivamente), y finalmente 5) Su acceso al internet mediante su número de entrez, para poder siempre obtener un acceso detallado y actualizado de cada gene. Se hicieron además varias operaciones básicas con ellos al final, para determinar la hebra dominante del estudio, la cual resultó ser, como era de esperarse, la hebra positiva.

Imágenes del programa:

1) Detalle del código en JAVA para la primera parte del programa (la integración y exportación de un archivo CSV, incluyendo su nexo a la página web de cada gene para tener una información completa):

```
public static void main(String[] args) {
    // TODO code application logic here
    List<String> datos = new ArrayList<>();
    datos.add("HebraNegativa:1:TP53:25772:-:https://www.ncbi.nlm.nih.gov/gene/7157");
    datos.add("HebraPositiva:2:TNF:2770:+:https://www.ncbi.nlm.nih.gov/gene/7124");
    datos.add("HebraNegativa:3:UBC:5765:-:https://www.ncbi.nlm.nih.gov/gene/7316");
    datos.add("HebraPositiva:4:ApoE:3647:+:https://www.ncbi.nlm.nih.gov/gene/348");
    datos.add("HebraPositiva:5:EGFR:237600:+:https://www.ncbi.nlm.nih.gov/gene/1956");
    datos.add("HebraPositiva:6:VEGFA:16304:+:https://www.ncbi.nlm.nih.gov/gene/7422");
    datos.add("HebraPositiva:7:IL6:6561:+:https://www.ncbi.nlm.nih.gov/gene/3569");
    datos.add("HebraNegativa:8:TGFB1:52347:-:https://www.ncbi.nlm.nih.gov/gene/7040");
    datos.add("HebraNegativa:9:MTHFR:21198:-:https://www.ncbi.nlm.nih.gov/gene/4524");
    datos.add("HebraPositiva:10:ESR1:472929:+:https://www.ncbi.nlm.nih.gov/gene/2099");
    datos.add("HebraNegativa:11:HLA-DRB1:36859:-:https://www.ncbi.nlm.nih.gov/gene/3123");
    datos.add("HebraPositiva:12:NFKB1:115974:+:https://www.ncbi.nlm.nih.gov/gene/4790");
    datos.add("HebraNegativa:13:IL10:7006:-:https://www.ncbi.nlm.nih.gov/gene/3586");
    datos.add("HebraPositiva:14:ACE:21320:+:https://www.ncbi.nlm.nih.gov/gene/1636");
    datos.add("HebraNegativa:15:BRCA1:125951:-:https://www.ncbi.nlm.nih.gov/gene/672");
```

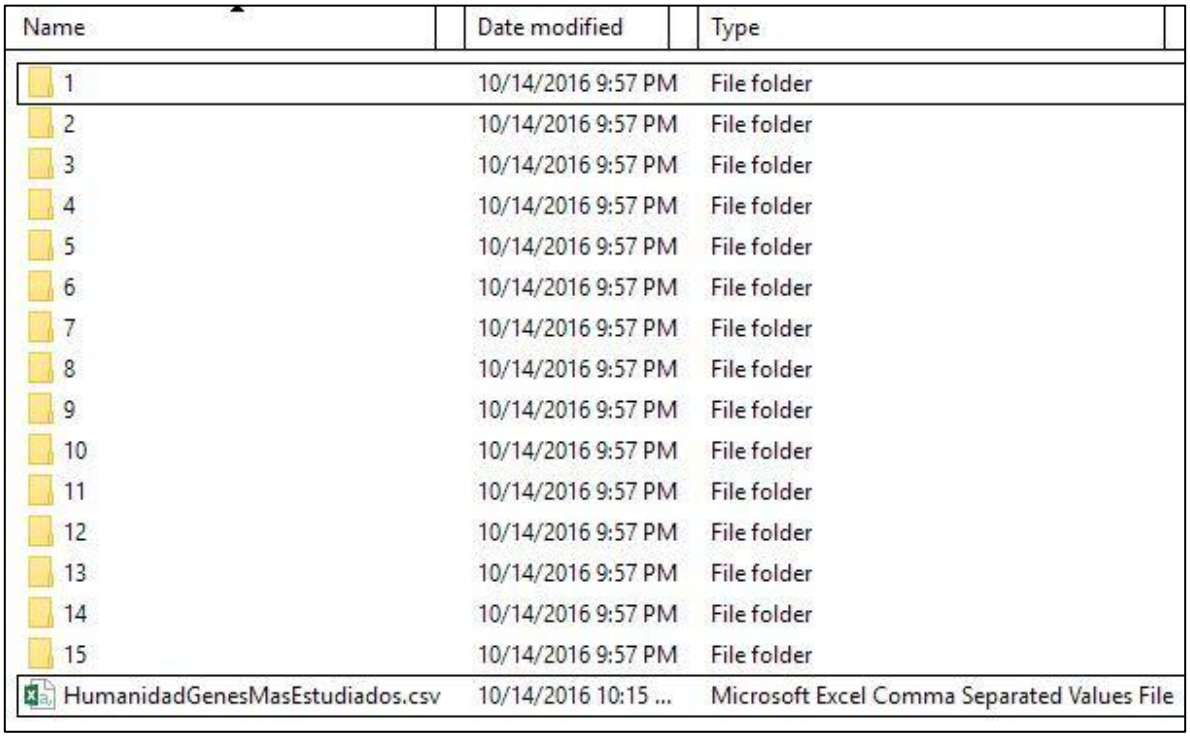
2) Detalle del código en JAVA para la segunda parte del programa (la exportación de la información recuperada del archivo CSV en quince diferentes archivos TXT, dentro de sus directorios respectivos):

```
Set<String> jerarquíasGenes = new HashSet<>();
List<GenesEstudiadosClass> genes = new ArrayList<>();

for (String jerarquías : genesEstudiados)
{
    String[] dato = jerarquías.split(",");
    String JERARQUIA = dato[0];
    String GENE = dato[1];
    int DIMENSIONES = Integer.parseInt(dato[2]);
    String HEBRA = dato[3];
    String ENTREZ = dato[4];

    GenesEstudiadosClass Gen = new GenesEstudiadosClass(JERARQUIA, GENE, DIMENSIONES, HEBRA, ENTREZ);
    jerarquíasGenes.add(JERARQUIA);
      jerarquíasGenes.add(GENE);
      jerarquíasGenes.add(HEBRA);
    genes.add(Gen);
}

for (String JERARQUIA : jerarquíasGenes)
{
    File myDir = new File(base+JERARQUIA);     // Crea un objeto
    myDir.mkdir();
}
```

3) Imagen de los genes como aparecen en la Consola del IDE, y de los cálculos que el programa realiza con ellos:

```
Jerarquía   : 12
Gene   : NFKB1
Dimensiones   : 115974
Hebra   : +
Entrez   : https://www.ncbi.nlm.nih.gov/gene/4790

Jerarquía   : 13
Gene   : IL10
Dimensiones   : 7006
Hebra   : -
Entrez   : https://www.ncbi.nlm.nih.gov/gene/3586

Jerarquía   : 14
Gene   : ACE
Dimensiones   : 21320
Hebra   : +
Entrez   : https://www.ncbi.nlm.nih.gov/gene/1636

Jerarquía   : 15
Gene   : BRCA1
Dimensiones   : 125951
Hebra   : -
Entrez   : https://www.ncbi.nlm.nih.gov/gene/672

La suma de la longitud de los genes de hebra negativa o minus (-) es: 274898.0
La suma de la longitud de los genes de hebra positiva o plus (+) es: 877105.0
La diferencia de longitud a favor de los genes plus comparada con los genes minus es: 602207.0
Se concluye que hay: ¡3.1906561706523875 veces más longitud en los genes plus que en los genes minus!
BUILD SUCCESSFUL (total time: 0 seconds)
```

4) Archivo CSV producido por la primera parte del código (en *Excell*), y los 15 directorios producidos por la segunda parte del programa (conteniendo cada uno su ficha informativa en texto, formato TXT):



Name	Date modified	Type
1	10/14/2016 9:57 PM	File folder
2	10/14/2016 9:57 PM	File folder
3	10/14/2016 9:57 PM	File folder
4	10/14/2016 9:57 PM	File folder
5	10/14/2016 9:57 PM	File folder
6	10/14/2016 9:57 PM	File folder
7	10/14/2016 9:57 PM	File folder
8	10/14/2016 9:57 PM	File folder
9	10/14/2016 9:57 PM	File folder
10	10/14/2016 9:57 PM	File folder
11	10/14/2016 9:57 PM	File folder
12	10/14/2016 9:57 PM	File folder
13	10/14/2016 9:57 PM	File folder
14	10/14/2016 9:57 PM	File folder
15	10/14/2016 9:57 PM	File folder
HumanidadGenesMasEstudiados.csv	10/14/2016 10:15 ...	Microsoft Excel Comma Separated Values File

5) Interior de un directorio individual, e interior del archivo con su información básica: