Polimorfismos de nucleótido único

Los **polimorfismos de nucleótido único** (SNPs) son el tipo de variación genética más habitual entre las personas. Un polimorfismo consiste en el cambio de un nucleótido (adenina, guanina, citosina, timina) en una posición concreta del ADN con respecto a un genoma de referencia, y que se encuentra al menos en el 1 por ciento de la población. La mayoría de estos polimorfismos no tienen efectos en el desarrollo y la salud, pero unos pocos sí pueden tener consecuencias en el fenotipo. El fenotipo es la expresión del genotipo que puede observarse (aunque está parcialmente determinado por el entorno). Estos polimorfismos afectarían a nuestra susceptibilidad a padecer determinadas enfermedades o a metabolizar ciertas sustancias de forma diferente.

Por ejemplo, en la población de origen europeo, la intolerancia genética a la lactosa (hipolactasia) está causada por un polimorfismo consistente en el cambio de un único nucleótido en la posición 13910 del gen MCM6, el cual regula la producción de lactasa (enzima que metaboliza la lactosa). En poblaciones de otro origen (asiáticos, africanos), la hipolactasia está causada por polimorfismos diferentes en este mismo gen. Existe además otro tipo de intolerancia a la lactosa (intolerancia secundaria), cuyo origen no es genético y que puede producirse por una gastroenteritis infecciosa, otras enfermedades intestinales, o incluso algunos medicamentos. La intolerancia secundaria desaparece cuando se resuelve la causa que la induce.

En otros casos (la mayoría), son necesarios varios polimorfismos (de un mismo gen o de genes distintos) para obtener un fenotipo concreto, es decir, el grado de expresión de dicho fenotipo está determinado por la combinación de sus polimorfismos relacionados. Por ejemplo, hay evidencia científica del papel de varios polimorfismos en genes del sistema dopaminérgico en la patogénesis de la migraña.

El objetivo de la práctica es la implementación de una aplicación que facilite el tratamiento de ciertos datos relacionados con algunos polimorfismos de nucleótido único.

De modo consensuado, cada polimorfismo se identifica con un código `rs`, que se emplea en las bases de datos de referencia (por ejemplo, *dbSNP*) y en investigación. En esta ocasión, toda la información de los polimorfismos se almacena en el fichero `dataSNP.txt`, cuyas líneas tienen la estructura siguiente:

- Código `rs`: cadena de caracteres de la forma $\texttt{rsd}^+$, donde $\texttt{d}^+$ es una cadena no vacía de dígitos.
- Nucleótidos: cadena de caracteres de la forma `N/N`, donde $\texttt{N} \in \{\texttt{A}, \texttt{G}, \texttt{C}, \texttt{T}\}$.
- Patologías: Las patologías relacionadas con el polimorfismo separadas por comas.

A continuación, mostramos las líneas del fichero `dataSNP.txt` correspondientes a los polimorfismos `rs4420638` y `rs6313`:

```
rs4420638 A/G: enfermedad de Alzheimer, hipercolesterolemia, cardiopatías
rs6313 C/C: artritis reumatoide
rs6313 T/T: depresión, trastorno de pánico
rs6313 C/T: artritis reumatoide
```

Estos datos muestran que el SNP `rs4420638` Adenina/Guanina está relacionado con la enfermedad de Alzheimer, la hipercolesterolemia y cardiopatías sin especificar. Análogamente, dos de las variaciones del polimorfismo `rs6313` (Citosina/Citosina, Citosina/Timina) participan en el desarrollo de la artitris reumatoide, mientras que el SNP `rs6313` Timina/Timina se liga tanto a la depresión como al trastorno de pánico. Por otra parte, se asume que el fichero puede contener líneas que comienzan con el símbolo `\#`, las cuales dan información provisional y/o adicional que no será procesada.

La información del fichero `dataSNP.txt` debe volcarse en un diccionario cuyas claves son los códigos `rs` de los polimorfismos. Cada código tiene asociado otro diccionario cuyas claves son los nucleótidos del SNP correspondiente y los valores son las las patologías asociadas al SNP. Por ejemplo, en el diccionario, el par correspondiente al polimorfismo `rs6313` debe ser equivalente al siguiente:

```
'rs6313':{'C/T':{'artritis reumatoide'},
          'C/C':{'artritis reumatoide'},
          'T/T':{'depresión', 'trastorno de pánico'}
```

1. En primer lugar vamos a realizar una función `get_data(line: str) -> (str, str, set[str])` que dada una línea del fichero de datos nos devuelva una terna `(rs_code, code_n, pats)` También podemos representar un SNP mediante una terna `(rs_code, code_n, pats)`, donde `rs_code` es un código `rs`, `code_n` es un par de nucleótidos `N/N`, y `pats` es un conjunto de patologías relacionadas. Por ejemplo, de la línea

   ```
   rs4420638 A/G': 'enfermedad de Alzheimer', 'hipercolesterolemia', 'cardiopatías'
   ```

   obtendermos la tupla

   ```
   ('rs4420638','A/G', {'enfermedad de Alzheimer', 'hipercolesterolemia', 'cardiopatías'})
   ```

   Las enfermedades se almacenarán en minúsculas. Usa correctamente los métodos `split`, `strip` y `lower` de las cadenas de caracteres de Python.

2. Implementa una función

   ```
   add_data(snp: dict, rs_code: str, code_n: str, pats: set[str]) -> None:
   ```

que dados un diccionario `snp` de polimorfismos, un código `rs_code`, un par de nucleótidos `code_n` y unas patologías `pats`, incluya la información del SNP en el diccionario. La función no solo debe incorporar nuevos polimorfismos sino también actualizar los SNPs existentes.

Por ejemplo, supongamos que `dict_snp` contiene el par siguiente:

```
'rs6313': {'C/T': {'artritis reumatoide'},
           'C/C': {'artritis reumatoide'},
           'T/T': {'depresión', 'trastorno de pánico'}}
```

entonces la ejecución de la llamada `add_data(dict_snp, 'rs6313','C/T', {'mesotelioma'})` lo modifica del modo siguiente:

```
'rs6313': {'C/T': {'artritis reumatoide', 'mesotelioma'},
           'C/C': {'artritis reumatoide'},
           'T/T': {'depresión','trastorno de pánico'}}
```

3. Diseña una función llamada `read_snp(filename: str) -> dict` que, dado el nombre de un fichero con la estructura indicada, lea su información y devuelva un diccionario que la contenga con de polimorfismos.

4. Diseña una función

   ```
   remove_snp(snp: dict, rs_code: str, code_n: str, pats: set[str])-> None
   ```

   que, dados un diccionario `snp` de polimorfismos y la tupla `tuple_snp` de un SNP, elimine los datos del SNP del diccionario, siguiendo el criterio indicado a continuación: sea `p` el SNP en `dict_snp` cuyos código y nucleótidos son `rs_code` y `code_n`. Si el conjunto `pats` es vacío, entonces borraremos de `p` todas sus patologías relacionadas, y en caso contrario, eliminaremos solo las patologías enumeradas en `pats`. Si tras esta modificación la lista de `p` esta vacía, entonces borraremos de `snp` el diccionario `code_n:{}`. Después de esto, si el polimorfismo no presenta variaciones en `snp` (esto es, el diccionario de `rs_code` está vacío), entonces eliminaremos de `snp` el diccionario `rs_code:{}`.

   Por ejemplo, supongamos que `snp` contiene el par siguiente:
   `'rs1800497':{'C/T':{'alcoholismo','tabaquismo','obesidad'}}` entonces la ejecución de la llamada
   `remove_snp(dict_snp,('rs1800497','C/T',{'tabaquismo'}))` lo modifica del modo siguiente:
   `'rs1800497':{'C/T':{'alcoholismo','obesidad'}}`
   y la ejecución de `remove_snp(dict_snp,('rs1800497','C/T',{'obesidad','alcoholismo'}))` lo elimina definitivamente del diccionario.

5. Implementa una función `get_snpinfo(snp: dict, pats: set[str]) -> str[str, str]` que, dados un diccionario `snp` de polimorfismos y un conjunto `pats` de patologías, devuelva un conjunto pares de la forma `(rs_code, code_n)` de modo que todas las patologías de `pats` estén relacionadas con todos los SNP del resultado.

   Por ejemplo, supongamos que `dict_snp` cocntiene el par siguiente:

   ```
   'rs4420638':{'A/G': {'enfermedad de alzheimer',
                        'hipercolesterolemia', 'cardiopatías'},
                'G/G': {'enfermedad de alzheimer',
                        'hipercolesterolemia'}
   ```

   entonces la lista que resulta de la llamada `get_snpinfo(dict_snp, {'hipercolesterolemia'})` incluye los pares `('rs4420638','A/G')` y `('rs4420638','G/G')`. Sin embargo, el último SNP no aparece entre los resultados de la llamada
   `get_snpinfo(dict_snp, {'cardiopatías', 'hipercolesterolemia'})`, ya que el el SNP `rs4420638` Guanina/Guanina no está relacionado con las cardiopatías.

6. A continuación vamos a realizar algunos análisis sobre los SNP.

   a) En primer implementa una función `get_snps(snp: list[dict]) -> dict` que devuelve un diccionario, las claves son códigos `rs` y el valor es un entero que indica las patologías asociadas al código `rs`.

   Por ejemplo, si el diccionario `snp` contiene el dato

   ```
   'rs4420638': {'A/G':{'enfermedad de alzheimer',
                        'hipercolesterolemia',
                        'cardiopatias'},
                 'G/G':['enfermedad de alzheimer',
                        'hipercolesterolemia']}}
   ```

   La lista resultante contendrá la clave `'rs4420638'` con valor 3.

b) Realiza una función `mean_ocurrences(snps: dict) -> float` Que devuelva la media de las apariciones en alguna patología en los SNPs, es decir las medias de los valores del duccionario.

c) Realiza una función `mode_ocurrences(snps: dict) -> int` que calcule la moda de las apariciones en alguna patología en los SNPs, es decir la moda de los valores.

d) Realiza una función `median_ocurrences(snps: dict) -> float` que calcule la mediana de las apariciones en alguna patología en los SNPs, es decir la mediana de los valores

7. Implementa una función llamada `write_snp(snp: dict, filename: str) -> None` que, dados un diccionario `snp` de polimorfismos y el nombre de un fichero nuevo, almacene en este último el contenido de `snp` respetando el formato especificado en el primer apartado.

   Por ejemplo, supongamos que `dict_snp` contiene el par siguiente:

```
'rs6313': { 'C/T': {'artritis reumatoide'},
            'C/C': {'artritis reumatoide'},
            'T/T': {'depresión','trastorno de pánico'}
```

   entonces la ejecución de la llamada `write_snp(dict_snp, 'dataSNPfinal.txt')` almacena en el fichero `dataSNPfinal.txt` las líneas siguientes:

```
rs6313 C/T: artritis reumatoide
rs6313 C/C: artritis reumatoide
rs6313 T/T: depresión, trastorno de pánico
```