El problema

El procesamiento de grandes volúmenes de datos es generalmente un reto cuando se trabaja en problemas de bioinformática.

La evolución de las técnicas ha conducido a que actualmente la disposición de datos biológicos sea masiva. El desafío ahora es evolucionar tanto en hardware como en software de una manera tal que dichos datos se puedan procesar y obtener así información con un sentido biológico.

En el caso de las proteínas el caso no es la excepción. Cada vez más se tiene acceso a servidores con gran cantidad de secuencias de proteínas listas para ser procesadas y analizadas. Algunos centros de investigación, que cuentan con gran capacidad de cómputo, colocan a disposición de la comunidad científica datos que han sido procesados en sus máquinas, los cuales pueden ser accesados libremente a través de una descarga o previa solicitud de los centros.

Tal es el caso del supercomputador Anton, una máquina especialmente diseñada que ha simulado cambios en la estructura tridimensional de una proteína en un periodo de un milisegundo, una escala bastante superior comparado con otros casos. Los resultados obtenidos se representan con la trayectoria de la proteína, es decir, cientos o miles de secuencias de proteínas que indican la posición de cada uno de sus átomos en un instante de tiempo.

Teniendo en cuenta que el acceso a supercomputadores como Anton es algo con lo que en general no se puede contar, es preciso idear técnicas que nos permitan trabajar con dichos datos a una escala más cercana a la capacidad computacional con la que normalmente contamos. Es acá donde métodos de análisis de datos tales como el clustering pueden ser útiles a la hora de hacer manejables la cantidad de datos que hay para su posterior procesamiento.
El clustering se puede definir como la agrupación de un conjunto particular de objetos basándose en sus características y agregandolos teniendo en cuenta sus similitudes. Hemos decidido usar esta técnica para, de la cantidad original de secuencias de proteínas, obtener una muestra reducida que sea representativa de las secuencias y que su tamaño haga que sea más manejable con los recursos computacionales con los que contamos.

Reducción de la escala

La solución propuesta permite obtener un porcentaje determinado de la población total de secuencias, aprovechando las ventajas de la programación paralela. Este porcentaje se obtiene de una manera tal que las secuencias resultantes representan las características estructurales del conjunto inicial. De esta manera, los análisis que posteriormente se realicen sobre este subconjunto, generarán resultados similares a los que se obtendrían si se analizara la totalidad de las secuencias.

Las secuencias de proteínas que proporciona Anton corresponden a los resultados de una simulación de dinámica molecular para una cantidad específica de tiempo. La dinámica molecular, así como otras simulaciones de sistemas físicos, genera “snapshots” del sistema simulado. En la simulación se tienen en cuenta “steps”, que corresponden al cálculo del siguiente estado del sistema, y “timesteps” que corresponden al intervalo de tiempo para pasar de un estado a otro.

La longitud de los timesteps determina por lo tanto la cantidad de snapshots que se obtienen en una simulación dada, suponiendo constante el tiempo de la simulación. Por ejemplo, los datos analizados en este artíiculo corresponden a una simulación de cerca de 200 microsegundos con timesteps de 200 picosegundos. Con esta configuración se tiene alrededor de 1.000.000 snapshots. La reducción que se realiza corresponde a un aumento en la longitud del timestep.

De esta manera, si se decide partir de un timestep de 2000 picosegundos en lugar de 200 picosegundos, el aumento de 10 veces en dicha longitud corresponde con una reducción de 10 veces en la cantidad de snapshots, es decir, se obtienen 10.000 snapshots. Esta cantidad es la que se define como el número de elementos de la población original que se tomarán como elementos “representativos” y permitirá que sea más manejable el procesamiento posterior.

El proceso

Existen varias librerías que permiten aplicar el proceso de clustering a una cantidad de secuencias de proteínas. El RMSD es uno de los parámetros que se puede utilizar como medida de la similitud entre una secuencia y otra. Encontramos, sin embargo, que aplicar el método de clustering con las más de 1.000.000 de conformaciones que se pueden obtener de la trayectoria de Anton para la proteína con la que estamos trabajando es demasiado costoso computacionalmente y el uso de memoria sobrepasa rápidamente la capacidad con la que contamos. Por esta razón, además de paralelizar el proceso de tal manera que se aprovechen todos los procesadores de la máquina en la que se ejecutará el proceso, se divide el conjunto inicial de secuencias en bloques de una cantidad específica, y se aplica clustering sobre cada uno de esos bloques. El número de elementos por bloque (1000 en este caso) fue determinado experimentalmente probando distintos valores hasta que encontramos que no se sobrepasaba la memoria y al mismo tiempo que permitía obtener un muestreo significativo en cada bloque teniendo en cuenta la cantidad global de secuencias.

A continuación la descripción del proceso:

- Adicción (si es necesario) de átomos a las secuencias

  Dependiendo de la fuente, las secuencias estarán ”completas” o no. Cuando no están completas, las secuencias contienen únicamente los carbonos alfa, resultando esto en archivos mucho más livianos y fáciles de almacenar y distribuir. Otros servidores pueden proveer las secuencias con todos los átomos, de una manera similar a la que se puede encontrar en un repositorio como el PDB. Este paso de adición de a´tomos, por lo tanto, se realiza para aquellas secuencias que contienen u´nicamente los carbonos alfa. En este caso, se hace uso de algún algoritmo que

complete la secuencia, esto es, que basándose en información de posiciones de átomos en estructuras completas, reconstruyen la secuencia asignando las coordenadas más probables para los átomos faltantes. Pulcra es un ejemplo de una herramienta que implementa la funcionalidad de reconstrucción.

- Paralelización del procesamiento
  Usando un parámetro que indica el número de núcleos que se pueden usar para el procesamiento, se puede dividir el total de proteínas entre dicho número y esa es la cantidad que cada uno de ellos procesa.

- Aplicación del método de clustering
  Cada núcleo debe aplicar el método de clustering al bloque de proteínas que le corresponde debido a la división del trabajo. Sin embargo, cada bloque puede contener un número arbitrariamente alto de proteínas a procesar, ya que el conjunto inicial de proteínas no tiene un límite en cuanto a cantidad. Por eso, se hace una nueva división, esta vez generando bloques de un número fijo de proteínas, de tal manera que se vayan procesando secuencialmente y no se tengan problemas de memoria. Estos bloques de tamaño fijo son los que finalmente se usan como entrada para el algoritmo de clustering, el cual analiza las secuencias y genera tantos grupos como se ha especificado a través de un parámetro dependiente del porcentaje que el usuario desea obtener de la población total. La salida de este paso es un archivo de texto con un representante de cada grupo generado. Cada representante se obtiene eligiendo el primer elemento de cada grupo.

- Unificación de resultados
  En este paso se leen todos los archivos generados en el paso anterior y se copian a un directorio los archivos pdb correspondientes a los representantes de cada grupo obtenido. Este directorio contiene la cantidad de secuencias que representa el porcentaje indicado por el usuario, del total de proteínas.