

El proceso se divide en los siguientes pasos:

1. Se descomprime el archivo que contiene la toda la trayectoria en formato pdb.
2. Se escanea el archivo para buscar las etiquetas de terminación de secuencia según el formato pdb y usando esta información se crean los archivos individuales para cada secuencia. N denotará el número de secuencias obtenidas de la trayectoria.
3. Dividir las N secuencias en P procesadores. Este número es definido por el usuario y depende del número de procesadores que tenga a su disposición para ejecutar el proceso. A cada procesador le corresponden entonces N/P secuencias. Este número lo definimos como k.
4. Siendo k un número que depende directamente del número de secuencias en la trayectoria, puede suceder que sea lo bastante alto como para generar problemas de memoria al momento de realizar el clustering. En este punto entonces, se divide k en bloques fijos de b elementos, garantizando así un procesamiento con un número máximo de secuencias con el que se controlan los problemas de memoria.
5. Para cada bloque de b secuencias se aplica el proceso de clustering. La librería utilizada permite definir el número de grupos o clústeres que se desean obtener. Definimos ese valor como G, que se calcula basándose en un parámetro definido por el usuario donde define el porcentaje de reducción que requiere usar en el proceso. Es decir, un porcentaje de reducción del 10% indica que de 1.000.000 de secuencias, por ejemplo, se desea obtener al finalizar un conjunto de 100.000 que actúen como muestra representativa de toda la población. Por cada cluster se selecciona la primera secuencia perteneciente al grupo, y esta será la representante de todo ese cluster, logrando de esta manera la reducción buscada.
6. Cada proceso de clustering en cada bloque genera un archivo con los resultados obtenidos. Al finalizar se leen todos estos archivos y se genera un directorio con las secuencias representativas.