Apéndice B

# Manual de uso

Los programas desarrollados son compatibles con el original *Dot2Dot* tanto en su forma de uso como en resultados que se obtienen. Conservan todas las opciones y parámetros de configuración, salvo por el número de hilos, que ahora no se aplica. Salvo algunos cambios, este manual de uso es una traducción del original.

## Compilación

Los programas se distribuyen mediante su código fuente, que es necesario compilar.

### Requisitos

Los requisitos para compilar los programas son:

- *compilador de C* con soporte para OpenMP (probado con GNU gcc)
- *biblioteca MPI* (probado con OpenMPI)
- utilidad `make`

### Cómo compilar

Para facilitar la compilación, se ha utilizado la herramienta de desarrollo `make`. Una vez satistechos los requisitos, para conseguir los ejecutables solo es necesario teclear `make` en el directorio correspondiente al código fuente del programa.

## Uso de los programas

El programa acepta parámetros a través de la línea de órdenes y mediante un fichero de configuración. En caso de que un parámetro sea definido en ambos lugares, el pasado por línea de comandos tiene preferencia.

### Parámetros generales

A continuación se listan los parámetros generales para controlar el comportamiento del programa desde línea de órdenes. Si un parámetro se puede especificar en el fichero de configuración, la opción correspondiente se muestra entre llaves ({}):

- **-s, –sequence {Sequence}**: nombre del fichero de entrada. Acepta formatos *fasta/multifasta/fastq*. El formato es automáticamente detectado, se ignora la extensión del fichero
- **-c, --config**: nombre del fichero de configuración (*obligatorio*)
- **-o, --output {Outfile}**: nombre del fichero de salida. Se usa el formato especificado por la extensión del fichero (*.dot* / *.bed*. Por defecto, .dot). Si no se especifica esta opción o no se puede abrir, se utiliza la salida estándar
- **-l, --minmotif {MinMotifLen}**: (*por defecto=2*) longitud mínima del motif en pb
- **-L, --maxmotif {MaxMotifLen}**: (*por defecto=30*) longitud máxima del motif en pb
- **-m, --minmatch {MinMatch}**: (*intervalo (0,1]. Por defecto=1*) Coeficiente de puntuación mínima de la suma ponderada de las coincidencias entre el motif y sus copias (véase B.2.4)
- **-G, --maxgaps {MaxGaps}**: (*por defecto=0*) número máximo de bases distintas entre el motif y sus copias
- **-I, --maxinsert {MaxInsert}**: (*por defecto=0*) número máximo de bases entre dos copias del motif. Si el valor es mayor que la longitud del motif, este valor se ajusta automáticamente durante la ejecución a $longitud - 1$
- **-v, --verbose**: salida verbosa
- **-V, --version**: muestra la versión del programa y finaliza
- **-h, --help**: muestra un mensaje con ayuda de uso y finaliza

Nota: las opciones `minmatch` y `maxgaps` se combinan en un único umbral que determina si una cadena puede considerarse copia del motif, de acuerdo a la siguiente fórmula:

$$umbral = max(|Motif| - MaxGaps, \lfloor |Motif| \cdot MinMatch \rfloor)$$

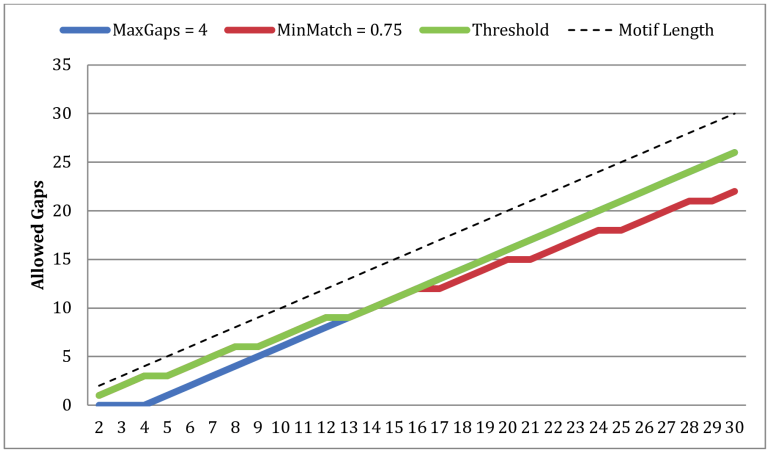


Figura B.1: Valor del umbral en relación a la longitud del *motif*
Fuente: Material complementario del artículo de *Dot2Dot* [22]

### Nuevos parámetros

Las versiones MPI e híbrida de los programas añaden dos nuevos parámetros:

- **-S, --schedule {Schedule}**: ( [BLOCK|BALANCED], por defecto: BALANCED) tipo de planificador a usar
- **-C, --corespernode {CoresPerNode}** (solo version hybrid-variable) indica el número total de núcleos que tiene cada nodo (*obligatorio*)

### Filtrado de los resultados

El filtrado adicional de los resultados se controla a través de una serie de variables especificadas en el fichero de configuración:

- **FilterType**: (*por defecto=NONE*) selecciona el tipo de filtrado, desde deshabilitado hasta el método más agresivo
  - **NONE**: desactiva el filtrado
  - **THRESHOLD**: descarta las TRs demasiado pequeñas (definido mediante el parámetro MinTRLen) o sin la pureza suficiente (parámetro MinPurity). El resto de los filtros aplican éste primero
  - **LIGHT**: en un conjunto de TRs que se solapan, se mantienen aquellas que cumplan al menos una de estas estadísticas: la más larga, la más pura o la que la longitud de su motif sea la más corta

  - **FAIR**: similar a la anterior. Se suma el número de las estadísticas anteriores que cumple cada TR de un conjunto que se solapan y se mantienen aquellas con la mayor puntuación
  - **HEAVY**: es el filtro más restrictivo. A cada TR de un conjunto que se solapan se le asigna una puntuación que es su pureza más su longitud normalizada (su longitud dividida por la máxima longitud del conjunto). Únicamente se mantienen aquellas TRs cuya puntuación esté por encima de un umbral. Este umbral se calcula restando un valor de tolerancia (parámetro Tolerance) a la puntuación más alta

- **MinTRLen**: (*por defecto=12*) longitud mínima de una TR para ser incluida en la salida
- **MinPurity**: (*intervalo (0,1), por defecto=0.1*) pureza mínima de una TR para ser incluida en la salida
- **Tolerance**: (*intervalo (0,1), por defecto=0*) (solo se aplica al filtro HEAVY) establece un grado de tolerancia. A menor valor, el filtro es más restrictivo
- **AllowOverlap**: (*valores (Y/N), por defecto=Y*) controla si pueden existir TRs que se solapen en la salida final. Es el último filtro que se aplica. Para cada par de TRs consecutivas que se solapen, se mantiene la más pura y se descarta la otra. En caso de empate, se escoge la más larga. Si el empate continúa, en ausencia de más discriminantes, se descarta la segunda TR. Aunque arbitraria, esta medida contribuye a reducir la probabilidad de solapamiento de una TR con la siguiente

### Ponderación de las comparaciones de bases

Se permite ponderar cualquier combinación de bases, incluso si no son las mismas, con un valor real entre 0 y 1. Los pesos se definen en el fichero de configuración mediante dos caracteres que representan a las bases, seguidas por un signo igual (=) y la puntuación correspondiente. Si una combinación no está definida, se considera que su valor es 0. Ejemplo:

```
AC = 0.5
TA = 0.7
```

Estos pesos se utilizan para calcular la puntuación de coincidencia usada en los filtros.

La matriz de pesos es simétrica. Estoy quiere decir que si definimos `AC=x`, implica que `CA=x`.

Se diferencia entre mayúsculas y minúsculas, por lo que si la secuencia contiene ambas, se han de especificar todas esas combinaciones (`Aa=1`, `Tt=1`, etc.).

Los caracteres permitidos son aquellos que representan a las bases, en mayúscula o minúscula, (AaCcGgTt), junto con la letra *ene* (Nn), que equivale a cualquier base. Se ignoran las combinaciones compuestas por caracteres distintos a estos.

Se desaconseja utilizar la combinación NN ya que los genomas contienen grandes regiones sin definir que podrían ser consideradas como TRs.