Práctica MPI

OPP 2013/14

Guillermo López Taboada, Sabela Ramos Garea

Computer Architecture Group
Department of Electronics and Systems
University of A Coruna (Spain)

sramos@udc.es

October, 2013

Motivación del proyecto

- Un Laboratorio Criminalístico se pone en contacto con la UDC con objeto de resolver un problema:
  - Problema: se necesita una alicación rápida para realizar tests de ADN con muestras degradadas.
  - Causa: Debido a que el test de ADN debe ser rápido, el algoritmo está implementado en hardware. Proporciona celeridad, pero a cambio de poca versatilidad.
  - Consecuencia: Numerosos crímenes no se están resolviendo.

Motivación del proyecto

- Tras consultar con una comisión de expertos, se ha dictaminado que:
  - Algoritmo: Los genetistas han recomendado la utilización de la técnica de alineamiento de cadenas de nucleótidos del ADN, utilizando el algoritmo Needleman-Wunsch.
  - Ingeniería Informática: Los ingenieros, a la vista de que se trata de un algoritmo de programación dinámica con alta demanda de recursos, tanto de memoria como de capacidad de cómputo, recomiendan su implementación paralela.
  - Implementación: Es necesario realizar una implementación paralela utilizando MPI que sea ejecutada en el clúster del CESGA.

Algoritmo Needleman-Wunsch

El Algoritmo de Needleman-Wunsch en un algoritmo de programación dinámica para resolver alineamientos locales de cadenas de nucleótidos.

- Cadenas de nucleótidos, que pueden ser A (Adenina), G (Guanina), C (Citosina) y T (Timina). Sin patrón de aparición aparente. Ejemplo: AGGGAAACTT
- Matriz de distancias entre nucleótidos (fuerzas de unión entre nucleótidos):

-	A	G	C	T
A	10	-1	-3	-4
G	-1	7	-5	-3
C	-3	-5	9	0
T	-4	-3	0	8

- Y la penalización por romper una cadena es de 5 unidades (fuerza de repulsión).

Se construye la matriz dinámica de fuerzas (F) sobre la que calcular el alineamiento óptimo.

- Las cadenas a comparar son: CGAGACGT y AGACTAGTTAC

```
          A,  G,  A,  C,  T,  A,  G,  T,  T,  A,  C,
      0, -5,-10,-15,-20,-25,-30,-35,-40,-45,-50,-55,
C  -5, -3, -8,-13, -6,-11,-16,-21,-26,-31,-36,-41,
G -10, -6,  4, -1, -6, -9,-12, -9,-14,-19,-24,-29,
A -15,  0, -1, 14,  9,  4,  1, -4, -9,-14, -9,-14,
G -20, -5,  7,  9,  9,  6,  3,  8,  3, -2, -7,-12,
A -25,-10,  2, 17, 12,  7, 16, 11,  6,  1,  8,  3,
C -30,-15, -3, 12, 26, 21, 16, 11, 11,  6,  3, 17,
G -35,-20, -8,  7, 21, 23, 20, 23, 18, 13,  8, 12,
T -40,-25,-13,  2, 16, 29, 24, 19, 31, 26, 21, 16,
```

- Pseudocódigo de construcción de la matriz F:

```
for i=0 to length(A)-1
  F(i,0) <- penalty*i
for j=0 to length(B)-1
  F(0,j) <- penalty*j
for i=1 to length(A)
  for j = 1 to length(B)
  {
    Choice1 <- F(i-1,j-1) + S(A(i), B(j))
    Choice2 <- F(i-1, j) + penalty
    Choice3 <- F(i, j-1) + penalty
    F(i,j) <- max(Choice1, Choice2, Choice3)
  }
```

Construcción de la matriz F (I).

```
          A,  G,  A,  C,  T,  A,  G,  T,  T,  A,  C,
    0, -5,-10,-15,-20,-25,-30,-35,-40,-45,-50,-55,
C  -5, -3, -8,
G -10, -6,
A -15,
G -20,
A -25,
C -30,
G -35,
T -40,
```

Construcción de la matriz F (II).

```
          A,  G,  A,  C,  T,  A,  G,  T,  T,  A,  C,
    0, -5,-10,-15,-20,-25,-30,-35,-40,-45,-50,-55,
C  -5, -3, -8,-13,
G -10, -6,  4,
A -15,  0,
G -20,
A -25,
C -30,
G -35,
T -40,
```

Construcción de la matriz F (III).

```
          A,  G,  A,  C,  T,  A,  G,  T,  T,  A,  C,
    0, -5,-10,-15,-20,-25,-30,-35,-40,-45,-50,-55,
C  -5, -3, -8,-13, -6,-11,
G -10, -6,  4, -1, -6,
A -15,  0, -1, 14,
G -20, -5,  7,
A -25,-10,
C -30,
G -35,
T -40,
```

Construcción de la matriz F (IV).

```
        A,  G,  A,  C,  T,  A,  G,  T,  T,  A,  C,
    0, -5,-10,-15,-20,-25,-30,-35,-40,-45,-50,-55,
C  -5, -3, -8,-13, -6,-11,-16,-21,-26,-31,-36,-41,
G -10, -6,  4, -1, -6, -9,-12, -9,-14,-19,-24,-29,
A -15,  0, -1, 14,  9,  4,  1, -4, -9,-14, -9,-14,
G -20, -5,  7,  9,  9,  6,  3,  8,  3, -2, -7,-12,
A -25,-10,  2, 17, 12,  7, 16, 11,  6,  1,  8,  3,
C -30,-15, -3, 12, 26, 21, 16, 11, 11,  6,  3,
G -35,-20, -8,  7, 21, 23, 20, 23, 18, 13,
T -40,-25,-13,  2, 16, 29, 24, 19, 31,
```

Construcción de la matriz F (y V).

```
        A,  G,  A,  C,  T,  A,  G,  T,  T,  A,  C,
    0, -5,-10,-15,-20,-25,-30,-35,-40,-45,-50,-55,
C  -5, -3, -8,-13, -6,-11,-16,-21,-26,-31,-36,-41,
G -10, -6,  4, -1, -6, -9,-12, -9,-14,-19,-24,-29,
A -15,  0, -1, 14,  9,  4,  1, -4, -9,-14, -9,-14,
G -20, -5,  7,  9,  9,  6,  3,  8,  3, -2, -7,-12,
A -25,-10,  2, 17, 12,  7, 16, 11,  6,  1,  8,  3,
C -30,-15, -3, 12, 26, 21, 16, 11, 11,  6,  3, 17,
G -35,-20, -8,  7, 21, 23, 20, 23, 18, 13,  8, 12,
T -40,-25,-13,  2, 16, 29, 24, 19, 31, 26, 21, 16,
```

Una vez finalizada la construcción de la matriz F, se recorre desde el extremo final hacia atrás. De tal modo que:

- Si el valor se obtuvo de la fila anterior, se avanza en las filas un nucleótido y se mantienen las columnas ('-').
- Si el valor se obtuvo de la columna anterior, se avanza en las columnas un nucleótido y se mantienen las filas ('-').
- Si el valor se obtuvo de la diagonal, se avanza tanto en filas como en columnas un nucleótido.

La cadena de alineamiento se construye de atrás a adelante.

Reconstrucción del camino (I):

```
          A,  G,  A,  C,  T,  A,  G,  T,  T,  A,  C,
     0, -5,-10,-15,-20,-25,-30,-35,-40,-45,-50,-55,
C  -5, -3, -8,-13, -6,-11,-16,-21,-26,-31,-36,-41,
G -10, -6,  4, -1, -6, -9,-12, -9,-14,-19,-24,-29,
A -15,  0, -1, 14,  9,  4,  1, -4, -9,-14, -9,-14,
G -20, -5,  7,  9,  9,  6,  3,  8,  3, -2, -7,-12,
A -25,-10,  2, 17, 12,  7, 16, 11,  6,  1,  8,  3,
C -30,-15, -3, 12, 26, 21, 16, 11, 11,  6,  3, 17,
G -35,-20, -8,  7, 21, 23, 20, 23,(18) 13,  8, 12,
T -40,-25,-13,  2, 16, 29, 24, 19, 31,(26)(21)(16)
```

```
Columnas: TAC
Filas:    T--
```

Reconstrucción del camino (II):

```
          A,  G,  A,  C,  T,  A,  G,  T,  T,  A,  C,
    0, -5,-10,-15,-20,-25,-30,-35,-40,-45,-50,-55,
C  -5, -3, -8,-13, -6,-11,-16,-21,-26,-31,-36,-41,
G -10, -6,  4, -1, -6, -9,-12, -9,-14,-19,-24,-29,
A -15,  0, -1, 14,  9,  4,  1, -4, -9,-14, -9,-14,
G -20, -5,  7,  9,  9,  6,  3,  8,  3, -2, -7,-12,
A -25,-10,  2, 17, 12,  7, 16, 11,  6,  1,  8,  3,
C -30,-15, -3, 12, 26, 21,(16) 11, 11,  6,  3, 17,
G -35,-20, -8,  7, 21, 23, 20,(23)(18) 13,  8, 12,
T -40,-25,-13,  2, 16, 29, 24, 19, 31,(26)(21)(16)

Columnas: GTTAC
Filas:    G-T--
```

Reconstrucción del camino (y III):

```
          A,  G,  A,  C,  T,  A,  G,  T,  T,  A,  C,
   (0) -5,-10,-15,-20,-25,-30,-35,-40,-45,-50,-55,
C (-5) -3, -8,-13, -6,-11,-16,-21,-26,-31,-36,-41,
G(-10) -6,  4, -1, -6, -9,-12, -9,-14,-19,-24,-29,
A -15, (0) -1, 14,  9,  4,  1, -4, -9,-14, -9,-14,
G -20, -5, (7)  9,  9,  6,  3,  8,  3, -2, -7,-12,
A -25,-10,  2, (17) 12,  7, 16, 11,  6,  1,  8,  3,
C -30,-15, -3, 12,(26)(21)(16) 11, 11,  6,  3, 17,
G -35,-20, -8,  7, 21, 23, 20,(23)(18) 13,  8, 12,
T -40,-25,-13,  2, 16, 29, 24, 19, 31,(26)(21)(16)
```

```
Columnas: --AGACTAGTTAC
Filas:    CGAGAC--G-T--
```

Estrategia de paralelización. Básico: división del dominio del problema entre procesadores:

```
                  P0          P1              P2
                              |               |
        A,  G,  A,  C,|  T,  A,  G,  T,| T,  A,  C,
    0, -5,-10,-15,-20,|-25,-30,-35,-40,|-45,-50,-55,
C  -5, -3, -8,-13, -6,|-11,-16,-21,-26,|-31,-36,-41,
G -10, -6,  4, -1, -6,|-9,-12, -9,-14,|-19,-24,-29,
A -15,  0, -1, 14,  9,| 4,  1, -4, -9,|-14, -9,-14,
G -20, -5,  7,  9,  9,| 6,  3,  8,  3,|-2, -7,-12,
A -25,-10,  2, 17, 12,| 7, 16, 11,  6,| 1,  8,  3,
C -30,-15, -3, 12, 26,|21, 16, 11, 11,| 6,  3, 17,
G -35,-20, -8,  7, 21,|23, 20, 23, 18,|13,  8, 12,
T -40,-25,-13,  2, 16,|29, 24, 19, 31,|26, 21, 16,
```

Enunciado de la práctica

Implementar el Algoritmo Needleman-Wunsch con MPI.

- Paradigma SPMD.
- El procesador 0 encargado de realizar Entrada/Salida. Lee de ficheros de texto según formato estándar en bioinformática. Reparto de datos con MPI_Scatter de forma óptima.
- Medición de tiempos de la construcción de la matriz F y generación de la secuencia de alineamiento (tras su obtención en el procesador 0). Utilización de MPI_Wtime para medir tiempos.

Enunciado de la práctica

Los archivos necesarios para la evaluación de la práctica están en Moodle y se trata de:

- Residuos hallados en la investigación de la escena del crimen
  - CSI1.dna
  - CSI2.dna
  - CSI3.dna
  - CSI4.dna
- Existen dos sospechosos. En sus prendas se ha hallado restos de material biológico. Se trata de la anterior pareja de la víctima (A) y del instalador de la televisión por cable (B). Hay tres muestras de A, y dos de B.
  - SUSPECTA1.dna
  - SUSPECTA2.dna
  - SUSPECTA3.dna
  - SUSPECTB1.dna
  - SUSPECTB2.dna

Enunciado de la práctica

Las pruebas que hay que realizar, pedidas por la CSI a cargo del caso, son las siete siguientes (en secuencial y paralelo):

- Test 1 "Needleman-Wunsch" CSI1.dna vs. SUSPECTA1.dna
- Test 2 "Needleman-Wunsch" CSI1.dna vs. SUSPECTA2.dna
- Test 3 "Needleman-Wunsch" CSI1.dna vs. SUSPECTA3.dna
- Test 4 "Needleman-Wunsch" CSI2.dna vs. SUSPECTB1.dna
- Test 5 "Needleman-Wunsch" CSI2.dna vs. SUSPECTB2.dna
- Test 6 "Needleman-Wunsch" CSI3.dna vs. SUSPECTB2.dna
- Test 7 "Needleman-Wunsch" CSI4.dna vs. SUSPECTB2.dna

Si el test de alineamiento de nucleótidos obtiene una "Similarity>90 %" se puede afirmar que los restos biológicos proceden de la misma fuente y por tanto incriminar al sospechoso que los porta.

Enunciado de la práctica

El proyecto se realizará y defenderá de forma individual.

- Se presentará una memoria explicativa del proyecto realizado.
- Dicha memoria ha de incluir una breve introducción, una explicación de las estrategias de paralelización abordadas y las aportaciones más interesantes.
- Además, ha de contener una exhaustiva evaluación del rendimiento de la aplicación con secuencias a alinear que se os proveerán, presentando de forma descriptiva (tablas y gráficas) los tiempos de ejecución utilizando 1, 2, 4, 8 y 16 procesadores. Se han de mostrar Speedups, Eficiencias y Tiempos de ejecución de forma clara.
- El código ha de presentarse en un anexo.
- Se valora especialmente el nivel de optimización obtenido. En particular, el solapamiento de computación con comunicación por medio de la utilización de comunicaciones no bloqueantes es especialísimamente valorado.
- Se ha de indicar el número de horas dedicadas al proyecto.

Boletines

Para orientaros en la realización del proyecto, se irán subiendo boletines con actividades al Moodle.

- Estos boletines son optativos pero se recomienda su seguimiento para mantener el trabajo al día.
- Los resultados podéis enviármelos al correo indicando en el asunto [OPP-Boletin X].
- La fecha de entrega es indicativa.