Universitat Autònoma de Barcelona

etse
Escola Tècnica Superior d'Enginyeria

ENGINYERIA INFORMATICA

MEMORIA PREVIA DEL PROYECTO:

2026 BIOINFORMÁTICA:

BÚSQUEDA DE ANCESTROS COMUNES ENTRE GENOMAS DE DIFERENTES ESPECIES

Signatura de l'estudiant Nom: Jonas Rodriguez Data: 15/1/2010	Signatura del director/a o directors/es Nom/s: Jordi Gonzalez / Mario Huerta Dpt: Bioinformatica Data: 15/1/2010

1. Objetivo del proyecto

El objetivo del proyecto es el desarrollo de unos algoritmos que permitan calcular los ancestros comunes entre diferentes especies a partir de la secuencia de sus genomas. Se ha dividido el trabajo a realizar en dos fases:

- Desarrollo de un algoritmo que a partir los *Maximal string matchings* entre las secuencias de los genomas a comparar devuelva datos que representan el factor de similitud aproximada entre ellos.

- A partir de los datos obtenidos en la fase anterior para todo par de especies se rastreara como una secuencia se conserva, aproximadamente, de una especie a otra.

Estos análisis formarán se utilizarán desde las aplicaciones web del servidor http://revolutionresearch.uab.es [6].
Para poder realizar estos algoritmos se deberán tener muy claros los diferentes aspectos teóricos sobre los que trabajar. En el siguiente apartado están explicados brevemente.

2. Estado del Arte

El genoma de todo ser vivo contiene su información genética. Esta está formada por una larga secuencia de 4 bases (A, C, T i G). La longitud de la secuencia cambia de una especia a otra, por ejemplo el genoma de bacteria Escherichia coli tiene 4,2 millones de pares de bases, la mosca del vinagre 140 millones de pares i la especie humana alrededor de 3300 millones. La longitud de las secuencias de las diferentes especies no para de augmentar a medida que se va secuenciando una mayor parte de su genoma.
Para comparar diferentes genomas se buscan similitudes entre sus secuencias, es decir, determinadas subsecuencias que compartan ambos. Una de las posibles estrategias para dicha comparación es el uso de *Maximal Unique Matchings* (MUMs). Las MUMs serían las subsecuencias comunes más largas y únicas, es decir no repetidas en el resto de las secuencias comparadas. Esta estrategia ya esta implementada en diversas aplicaciones, como el MUMmer [1] una aplicación para la alineación de genomas completos y el MUMs On-Line [3][4][5], una evolución del cálculo de MUMs a partir de una variación del algoritmo de *sufix-trees*.

Ejemplo de MUM:

…taggc**ATGCTAGA**aagta…

…agcacta**ATGCTAGA**cta…

También hay que tener en cuenta que en algunos casos durante el proceso evolutivo algunas partes de la secuencia se puede invertir completamente, estos casos son denominados MUMs inversos.

Ejemplo de MUM inverso:

…taggcA**TGCTAGA**aagta…

…agcacta**AGATCGTA**cta…

Existen centros en todo el mundo donde se realizan estudios sobre genoma. El mayor repositorio de genomas corresponde al *National Center for Biotechnology Information* (NCBI)[2] dedicado a diversos aspectos de investigación y enseñanza que permite además la descarga de los genomas de más de 800 especies.

3. Estudio de viabilidad del proyecto

Como ya se comentó en los objetivos, el proyecto esta diferenciado en dos fases:

- Calculo de superMUMs:
  - Buscar superMUMs teniendo en cuenta los diferentes casos que se pueden encontrar (MUMs invertidos, …)
  - Hacer el juego de pruebas para testear los datos de salida
- Búsqueda de ancestros comunes:
  - Desarrollar la función que determina la conservación aproximada de una subsecuencia de un genoma a otro (ancestros comunes).
  - Buscar para diferentes genomas los ancestros comunes todos con todos.
  - Buscar para diferentes genomas los ancestros comunes siguiendo el *minimum spanning tree* o árbol de mínima dispersión.

Para el desarrollo de los algoritmos se ha escogido el lenguaje de programación C. Los genomas para las comparaciones se obtienen del servidor del NCBI. Para obtener los MUMs entre cada par de genomas se utilizará una implementación del MUMs On-Line[3][4].

Con esto tenemos todos los elementos necesarios para el correcto desarrollo del proyecto es por este motivo que considero que es un proyecto viable.

4. Planificación temporal del trabajo

Enero 2010	Finalizar el programa de cálculo superMUMS.
Marzo 2010	Finalizar el programa de cálculo de ances comunes entre 2 secuencias.
Abril 2010	Finalizar el programa que calcula los ancestros comparando los genomas todos con todos.
Junio 2010	Finalizar el programa que obiene los ances siguiendo el *Minimum Spanning Tree* de genomas
Julio 2010 Agosto 2010	Documentación final del proyecto
Septiembre 2010	Presentación del Proyecto

5. Bibliografia

[1] Alignment of Whole Genomes. A.L. Delcher, S. Kasif, R.D. Fleischmann, J. Peterson, O. White, and S.L. Salzberg, *Nucleic Acids Research*, 27:11 (1999), 2369-2376.

[2] National Center for Biotechnology Information
http://www.ncbi.nlm.nih.gov/

[3] Efficient space and time multicomparison of genomes. Huerta,M. and Messeguer,X. Research Report LSI-02-64-R Dep. Llenguatge i Sistemes Informàtics, Universitat Politècnica de Catalunya.(2002).

[4] Suffix Tree Construction with slide nodes . Mario Huerta . technical report LSI-02-63-R Dep. Llenguatge i Sistemes Informàtics, Universitat Politècnica de Catalunya (2002).

[5] Identification of patterns in biological sequences at the ALGGEN server. PROMO and MALGEN. Domènec Farré, Romà Roset, Mario Huerta, José E. Adsuara,

Llorenç Roselló, M. Mar Albà, Xavier Messeguer.. Nucleic Acids Research 31(13): 3651-3653 (2003).

[6] http://platypus.uab.es : Web server for the all-known-genomes comparison by web. Server supported by the Institute of Biotechnology and Biomedicine of the Autonomous University of Barcelona (IBB-UAB).