Universidad de Chile
Facultad de Ciencias Físicas y Matemáticas
Departamento de Ciencias de la Computación

Semestre Primavera 2007
Fundamentos de Ciencias de la Computación
CC30B

Profesor: Gonzalo Navarro.
Auxiliares: Gonzalo Ríos, Esteban Allende
Fecha: 26 de Agosto de 2007

Tarea 1: PROSITE

1 Introducción

El método PROSITE consiste en una base de datos de patrones y perfiles biológicamente significativos establecidos de manera tal que con las herramientas computacionales adecuadas es posible de manera rápida y eficiente determinar a cuál familia conocida de proteínas pertenece la nueva secuencia, o qué dominios conocidos contiene.

La sintaxis de los patrones usados en la base de datos PROSITE es:

- Los aminoácidos son codificados con el estándard IUPAC de una letra:

Aminoácido	**Abreviatura**
Alanina	A
Cisteína	C
Ácido aspártico	D
Ácido glutámico	E
Fenilalanina	F
Glicina	G
Histidina	H
Isoleucina	I
Lisina	K
Leucina	L
Metionina	M

Aminoácido	**Abreviatura**
Asparagina	N
Pirrolisina	O
Prolina	P
Glutamina	Q
Arginina	R
Serina	S
Treonina	T
Selenocisteína	U
Valina	V
Triptófano	W
Tirosina	Y

- El símbolo 'x' es usado para una posición en donde cualquier aminoácido es aceptado.
- Los paréntisis cuadrados '[]' indican una lista de los posibles aminoácidos aceptados. Por ejemplo, [AD] indica A ó D.
- Las llaves '{}' indican una lista de los aminoácidos que no son aceptados. Por ejemplo, {NP} indica que se puede ser cualquier aminoácido diferente a N y P.
- Cada elemento en un patrón es separado de sus vecinos con el símbolo '-'. Por ejemplo: A-C-D.
- Repeticiones consecutivas de un elemento del patrón se puede representar como la letra seguido de los paréntesis '()' que contienen el número de repeticiones del elemento. Por ejemplo, A(3) es lo mismo que A-A-A.
- Los gap's ('x') permiten un rango entre los paréntesis '()'. Por ejemplo, x(3) corresponde a x-x-x , y x(2,3) corresponde a x-x ó x-x-x.

2 Tarea

En esta tarea lo que se pide es hacer una herramienta que analice proteínas. Para esto, su programa debe:

1. Recibir un patrón PROSITE
2. Convertir el patrón en expresión regular (usando | para convertir los "[]"y "{}" ; y usando x(x|ε)(x|ε) para convertir cosas como x(1,3), donde "x" será un símbolo especial que represente todos los caracteres). Imprimir la expresión regular.
3. Convertir la expresión regular en AFND. Modificar el AFND, para que reconozca todas las cadenas que terminan en la expresión regular. Imprimir el AFND.
4. Convertir el AFND en AFD, e imprimirlo.
5. Recibir una proteína.
6. Entregar todas las posiciones finales de ocurrencias del patrón en la proteína
7. Volver al punto 5. hasta que la proteína recibida sea la cadena vacía.

Se les adjunta un archivo patrones.txt que contiene patrones PROSITE y un archivo proteinas.txt que contiene porteinas del tipo Mycobacterium tuberculosis, que los pueden usar de pruebas ;-).

3 Entrega

El plazo de entrega vence el día Lunes 24 de Septiembre hasta las 23:59:59 hrs. Su implementación debe ser en C o Java. La entrada y la salida del programa deben ser a través de la salida estándar. Puede realizarse en grupos de máximo 3 personas y la tarea debe ser enviada por U-Cursos (código de fuente), además de entregar en secretaría un pequeño informe indicando:

- Breve descripción del programa.
- Instrucciones de compilación y ejecución.
- Ejemplos de uso.

Este lo pueden entregar hasta el día Martes 25 de Septiembre hasta las 16:30 hrs.