Área Temática: Aplicaciones de sistemas inteligentes a la Bioinformática

MODELOS LINGÜÍSTICOS BIOINFORMÁTICOS

Gamma Z. Galindo Pérez
IPN - UPIBI
Av. Milán 173 Col. Izcalli Pirámide Tlalnepantla Ciudad de México. MÉXICO 54140
Tel: (+52)(55) 53 91 64 92 e-mail: zaragato@hotmail.com

Palabras clave: *bioinformática, estructuras lingüísticas de cadenas proteicas*

Introducción

Existe una gran cantidad de información sobre secuenciación genética y proteica en la red, que constantemente esta aumentando, al decodificarse nuevas proteínas y genes en todo el mundo, esto abre grandes necesidades

.

Es necesario ir diseñando y desarrollando herramientas que permitan ir encontrando patrones en la información biológica, que a la larga sirvan para modelar desde un marco teórico secuencias de genes y/o proteínas de interés

Esto abriría infinidad de posibilidades que van desde las mas practicas como el diseño de mejores matrices para la purificación de proteínas hasta la posibilidad de poder diseñar genes y proteínas de nuestro interés programando los genes de los m.o.

Como ejemplo de lo anterior en este proyecto se pretende encontrar patrones de comportamiento en secuencias de genes y de proteínas mediante la aplicación de la Informática.

Metodología

Primero se recopila información en bancos de secuencias de proteínas y de secuencias de genes, esta información es clasificada e integrada para la búsqueda de patrones de comportamiento a nivel genético para lo cual se utilizan herramientas informáticas como Sistemas Evolutivos y Lingüística Matemática. En base a los patrones encontrados se proponen los modelos informáticos

Resultados y Discusión

En el presente trabajo se utilizaron como proteínas de estudio a las celulasas y amilasas debido a su importancia económica, por su parte el modelo de citocromo C que obtengamos al analizar las cadenas completas nos servirá de referencia para validar los otros modelos, ya que existen suficientes estudios en base al citocromo C para comparar con nuestros resultados.

Cabe señalar que los procedimientos aquí utilizados pueden ser aplicados para el análisis de proteínas y de ser el caso también para el estudio de genes, ya que a las herramientas que creamos para poder analizar nuestras secuencias les es indistinto, puesto que visualizan a las proteínas o a los genes como una secuencia de caracteres.

Como primer paso se recopila información en bancos de secuencias de proteínas y de genes, estos bancos se encuentran en internet[15] y la primera parte del proyecto fue localizar su dirección electrónica, después de buscar en internet se decidió utilizar el sitio del Instituto Europeo de Bioinformática (European Bioinformatic Institute EBI) que se encuentra entrando a internet con *www.ebi.ac.uk,*

En base a la secuencia de la proteína de interés se realiza un análisis de Blast-p.

La *información es clasificada e integrada con el fin de utilizarla para la búsqueda de patrones de comportamiento a nivel genético y proteico*, para lo cual se utilizan varias herramientas informáticas, como lingüística matemática y sistemas evolutivos.

En esta etapa lo primero que se realiza es el *establecimiento de la estructura lingüística para la proteína o gen analizado*. Para ilustrar lo anterior y como ejemplo de los resultados obtenidos partiremos de la tabla 2.1 que muestra un pequeño fragmento del Blast p realizado para citocromo C.

Tabla 2.1. Fragmento del blast-p realizado a citocromo C

Citocromo C (CYC)											
Organismo	Secuencia de a.a										
	1	2	3	4	5	6	7	8	9	10	11
EUGGR	G	D	A	E	R	G	K	K	L	F	E
EUGVI	G	D	A	E	R	G	K	K	L	F	E
MOUSE	G	D	A	E	A	G	K	K	I	F	V
EQUAS	G	D	V	E	K	G	K	K	I	F	V
HORSE	G	D	V	E	K	G	K	K	I	F	V
BOVIN	G	D	V	E	K	G	K	K	I	F	V

De la tabla 2.1 observamos como en todos los CYC el primer aminoácido es G, el segundo es D, mientras en el tercero observamos dos posibles a.a A y V en este lugar lo marcamos como x_1, donde la x indica que se tiene una variable y el 1 por ser el primer lugar donde se presenta variación, lo mismo sucede en el 5° a.a. aquí lo marcamos como x_2 por ser el 2° sitio donde existen diversos a.a. para la misma posición, de esta forma vamos obteniendo la estructura lingüística

$$G\ D\ x_1\ E\ x_2\ G\ K\ K\ x_3\ F\ x_4$$

Mediante este proceso *se ha desarrollado el análisis de cadenas de diferentes tipos de proteínas, encontrando su estructura lingüística***.** Para el caso del citocromo C pudimos establecer su estructura en base al análisis visual de su blast-p ya que se trata de una proteína relativamente pequeña, pero *cuando empezamos a tratar con los blast-p de las celulasas el análisis se complicó al tener algunas de las proteínas secuencias de más de 500 a.a. por lo cual nos vimos en la necesidad de crear varios programas*

En base al análisis lingüístico se determinó que al parecer los sitios activos, de las amilasas y el grupo hemo del citocromo C tienden a ser altamente conservativos.

Así el posible sitio activo consenso en el caso de las amilasas sería

x	**x+1**	**x+2**	**x+3**	**x+4**	**x+13**
D	**A**	**A**	**K**	**H**	**D**

Y la posible conformación del grupo hemo en los citocromos C sería

1	**17**	**18**	**79**	**85**
x	**x+16**	**x+17**	**x+78**	**x+84**
G	**C**	**H**	**M**	**K**

Donde x es el sitio del aminoácido donde comienza el sitio activo o el grupo hemo según sea el caso

Para un mejor entendimiento de las propiedades lingüísticas de las proteínas estudiadas es necesario comenzar un análisis a nivel de caracteres, donde *cada aminoácido se visualiza como un carácter de una oración*, en este sentido se crearon programas que nos permitieran visualizar el porcentaje de aparición de un aminoácido en un sitio determinado obteniendo resultados muy interesantes.

Para poder hacer esto se creó un programa, al cual si le suministramos por ejemplo los siguientes datos

G	D	A	E	R	G	K	K	L	F	E
G	D	A	E	A	G	K	K	I	F	V
G	D	V	E	K	G	K	K	I	F	V
G	D	V	E	K	G	K	K	I	F	V
G	D	V	E	K	G	K	K	I	F	V

Nos da como resultado la siguiente tabla:

A	0	0	**40**	0	**20**	0	0	0	0	0	0
B	0	0	0	0	0	0	0	0	0	0	0
C	0	0	0	0	0	0	0	0	0	0	0
D	0	**100**	0	0	0	0	0	0	0	0	0
E	0	0	0	**100**	0	0	0	0	0	0	**20**
F	0	0	0	0	0	0	0	0	0	**100**	0
G	**100**	0	0	0	0	**100**	0	0	0	0	0
H	0	0	0	0	0	0	0	0	0	0	0
I	0	0	0	0	0	0	0	0	**80**	0	0
K	0	0	0	0	**60**	0	**100**	**100**	0	0	0
L	0	0	0	0	0	0	0	0	**20**	0	0
M	0	0	0	0	0	0	0	0	0	0	0
N	0	0	0	0	0	0	0	0	0	0	0
P	0	0	0	0	0	0	0	0	0	0	0
Q	0	0	0	0	0	0	0	0	0	0	0
R	0	0	0	0	**20**	0	0	0	0	0	0
S	0	0	0	0	0	0	0	0	0	0	0
T	0	0	0	0	0	0	0	0	0	0	0
V	0	0	**60**	0	0	0	0	0	0	0	**80**
W	0	0	0	0	0	0	0	0	0	0	0
	G	**D**	**.**	**E**	**.**	**G**	**K**	**K**	**.**	**F**	**.**

En la cual, la computadora nos entrega el análisis de la siguiente forma, en la primera columna imprime una lista de todos los aminoácidos y al lado establece los porcentajes de aparición del aminoácido para cada sitio específico de la proteína, abajo se va escribiendo la ecuación lingüística

Estos resultados nos permiten apreciar entre otras cosas los patrones de variación en un sitio determinado de la proteína, *pudiéndose apreciar si los aminoácidos que aparecen en un sitio determinado son todos del mismo tipo* p. ej. 100% neutros *o si existe combinaciones* p ej. 70% neutros 20% ácidos 10% básicos *e incluso* 30% neutros-aromáticos 70% neutros no aromáticos

A continuación se muestran algunos de los resultados que obtuvimos de forma preliminar del análisis de las tablas resultantes,

Al observar la primera tabla de variaciones de citocromo C notamos que en el sitio 14 el 88% de los CYC analizados presentaban Cisteína, mientras solo el 12% Alanina, al buscar otros sitios donde existiera esa proporción en variación encontramos que tanto en el sitio 39 como en el 48 había una proporción 88-12% pero en estos casos la Timina estaba en el 88% de los casos y la serina en 12%, como se ve a continuación:

	A	**B**	**C**	**D**	**E**	**F**	**G**	**H**	**I**	**K**	**L**	**M**	**N**	**P**	**Q**	**R**	**S**	**T**	**V**	**W**	**Y**
14	**12**	**,**	**88**	**,**	**,**	**,**	**,**	**,**	**,**	**,**	**,**	**,**	**,**	**,**	**,**	**,**	**,**	**,**	**,**	**,**	**,**
39	**,**	**,**	**,**	**,**	**,**	**,**	**,**	**,**	**,**	**,**	**,**	**,**	**,**	**,**	**,**	**,**	**12**	**88**	**,**	**,**	**,**
48	**,**	**,**	**,**	**,**	**,**	**,**	**,**	**,**	**,**	**,**	**,**	**,**	**,**	**,**	**,**	**,**	**12**	**88**	**,**	**,**	**,**

al ir al blast-p nos encontramos lo siguiente:

Organismo	14	39	48
1 :CYC_EUGGR	A	S	S
2 :CYC_EUGVI	A	S	S
3 :CYC2_MOUSE	C	T	T
4 :CYC2_RAT	C	T	T
5 :CYC_EQUAS	C	T	T
6 :CYC_HORSE	C	T	T
7 :CYC_BOVIN	C	T	T
8 :CYC_CYPCA	C	T	T
9 :CYC_MACGI	C	T	T
10 :CYC_HIPAM	C	T	T
11 :CYC_THELA	C	T	T
12 :CYC_CRIFA	A	S	S
13 :G298836	C	T	T
14 :CYC_HUMAN	C	T	T
15 :CYC_CANFA	C	T	T
16 :CYC_MIRLE	C	T	T
17 :CYC_KATPE	C	T	T
18 :CYC_MACMU	C	T	T
19 :CYC_ESCGI	C	T	T
20 :CYC_NEUCR	C	T	T
21 :CYC_RANCA	C	T	T
22 :CYC_MINSC	C	T	T
23 :CYC_APTPA	C	T	T
24 :CYC_ENTTR	C	T	T
25 :CYC_MOUSE	C	T	T

Al observar el blast-p podemos dividir a los organismos en dos categorías, aquellos que tienen cisteína en el lugar 17 y en los lugares 39 y 49 Timina y aquellos que tienen Alanina en el lugar 17 y en los otros dos Serina, en base a estos resultados se puede decir que es posible que la secuencia de a.a. CTT sea substituíble por ASS o en otras palabras es posible que CTT y ASS actúen como sinónimos

Finalmente se desarrolló otra herramienta que conjugando lo anterior permite tanto el establecimiento de la ecuación lingüística como el comparar ésta con secuencias de aminoácidos de forma tal que selecciona aquellas que cumplen la ecuación.

Este programa se podría utilizar por ejemplo para ir creando un identificador de secuencias, de tal manera que *en vez de tener almacenadas las secuencias de múltiples proteínas solo se tenga la ecuación lingüística de cada tipo de proteína* y con ella se realice una identificación primaria.

CONCLUSIONES

Es posible encontrar estructuras lingüísticas en base al análisis de la información proteica, ya que, en diferentes enzimas que catalizan la misma reacción se encuentran zonas de a.a. que se repiten independientemente del organismo del cual son extraídas la enzimas

De lo observado en los resultados de los diversos Blast-p realizados para amilasas podemos decir que es posible que existan ciertas secuencias de aminoácidos que actúen como verbos y otra serie de aminoácidos que actúen como sujeto y que al conjugarse logran la alta especificidad en las enzimas Además en base a el análisis de las variaciones del Citocromo C es posible que existan secuencias de aminoácidos equivalentes que podrían actuar como sinónimos

Es posible el desarrollo de programas que facilitan ir estableciendo las diferentes características gramaticales de las secuencias de las proteínasy /o genes.

FUENTES DE INFORMACIÓN

1. Fernando **Galindo Soria,** Marina **Vicario Solorzano,** *Rumbo a la Fundamentación de la Informática Educativa,* en Memorias del XII Simposio Internacional de Computación en la Educación organizado por la SOMECE, Cd. de México, Octubre de 1996.
2. José Luis **Carrillo Aguado**, *Entrevista sobre Informática a Fernando Galindo Soria y Marina Vicario Solorzano,* en la Revista Investigación hoy #79, pag. 22 y 23, Cd. de México, Dic. de 1997.
3. *El Origen de las formas*, edición especial de Mundo Científico #188, Barcelona, Marzo de 1998.
4. Jagjit **Singh**, *Teoría de la Información, del lenguaje y de la cibernética*, Ed. Alianza Editorial AU-29, Madrid, 1982
5. Fernando **Galindo Soria,** *Algunas propiedades matemáticas de los sistemas lingüísticos* en: las Memorias sobre "Sistemas Evolutivos" del ler Congreso Internacional de Investigación en Ciencias Computacionales, Instituto Tecnológico de Toluca, Metepec Edo. de México, Septiembre de 1994.
6. Fernando **Galindo Soria,** *Sistemas Evolutivos de Reescritura,* en Memorias sobre "Sistemas Evolutivos" del ler. Congreso Internacional de Investigación en Ciencias Computacionales, Instituto Tecnológico de Toluca, Metepec Edo. de México, Septiembre de 1994.
7. Fernando **Galindo Soria,** *Sistemas Evolutivos de Lenguajes de Trayectoria,* En las Memorias de la VI Reunión Nacional de Inteligencia Artificial, Ed. Limusa, Querétaro, Qro., Junio de 1989.
8. Rémi **Jullien**, Robert **Botet** y Max **Kolb**, *Los Agregados*, en Mundo Científico vol. 6, #54, pag. 36, Ed. Fontalba, S.A., Barcelona, España.
9. Eliezer **Braun**, Caos, *Fractales y cosas raras,* Ed. FCE., México, 1996
10. Vicente **Talanquer**, *Fractus, fracta, fractal,* Ed. FCE., México, 1996

11. Gamma Z. **Galindo Pérez** y Patricia **Rodríguez Pascual**, *Modelos Bioinformáticos*, en las memorias del VIII Congreso Nacional de Biotecnología y Bioingeniería y IV Congreso Latinoamericano de Biotecnología y Bioingeniería, pag 599, Huatulco, Oaxaca, México, septiembre de 1999.
12. Lorenzo **Segovia**, *Bioinformática: Análisis de la familia estructural de las Beta-lactamasas*, en las memorias del VIII Congreso Nacional de Biotecnología y Bioingeniería y IV Congreso Latinoamericano de Biotecnología y Bioingeniería, pag 598, Huatulco, Oaxaca, México, septiembre de 1999.
13. Albert L. **Lehninger**, *Biochemistry, 2. Edición*, Nueva York, 1975
14. C.U.M: **Smith**, *Biología Molecular*, Ed, Alianza Editorial AU-7, Madrid 1971
15. *www.ebi.ac.uk*, Página del Instituto Europeo de Bioinformática