Revista de la Sociedad Química de México, Vol. **47***, Núm. 1 (2003) 6-21*

Investigación

Memorias conformacionales en la predicción de estructura terciaria de polipéptidos

Ramón Garduño Juárez,*[1] y Luis B. Morales[2]

[1] Centro de Ciencias Físicas, Universidad Nacional Autónoma de México, Apdo. Postal 48-3, 62250 Cuernavaca, Morelos, México. Tel. (777)3291-749; Fax (777)3291-775; E-mail: ramon@fis.unam.mx.
[2] Instituto de Investigación en Matemáticas Aplicadas y Sistemas, Universidad Nacional Autónoma de México, Apdo. Postal 70-221, 04510 México D.F.

Recibido el 23 de septiembre del 2002; aceptado el 23 de enero del 2003

Resumen. Para reducir el espacio conformacional presente en la predicción de la estructura terciaria de polipéptidos, el cual crece exponencialmente con el número de residuos, hemos diseñado un algoritmo de búsqueda heurística que es robusto y que puede proporcionar estructuras cercanas a la nativa con relativa facilidad. Específicamente, hemos desarrollado un algoritmo genético híbrido (AGH) que usa números reales en lugar de bits para describir a los genes de cualquier cromosoma, y que incluye un operador de perfeccionamiento cuya función es reorientar la búsqueda hacia los individuos mejor adaptados. Al final de cada ciclo del AGH, un análisis de la población resultante indica la presencia de cúmulos bien definidos en los valores de ángulos de torsión para cada cromosoma. Estos segmentos corresponden a diferentes conformaciones de baja energía, y son relativamente constantes cada vez que se practica un nuevo experimento de AGH sobre la misma molécula. A estos segmentos les hemos llamado "memorias conformacionales", y son usados como los límites de un espacio conformacional reducido dentro del cual se realiza la siguiente ronda del AGH. El empleo de las memorias conformacionales acelera y afina la localización de la estructura correspondiente al Mínimo Global de Energía (MGE). Este algoritmo ha sido empleado para localizar con éxito el MGE de la Met- y la Leu-encefalina.
Palabras Clave: Predicción de la estructura de proteínas, algoritmos genéticos, plegado de proteínas, búsqueda en el espacio conformacional, memorias conformacionales.

Abstract. To reduce the conformational space existing in the process of searching for the tertiary structure of polypeptides, which grows exponentially with the number of residues, we have designed a heuristic search algorithm that is robust and that can provide with structures near the native with ease. Specifically, we have developed a hybrid genetic algorithm (HGA) that uses real numbers instead of bits in order to describe the genes of any given chromosome, and that includes an improvement operator which function is to reorient the search towards the best fitted individuals. At the end of each HGA cycle, an analysis of the resulting population shows the presence of well defined dihedral angle clusters for each chromosome. These segments correspond to different low energy conformations, and are relatively constant each time a new HGA experiment is performed on the same molecule. We have called "conformational memories" to these segments, which are used as a reduced conformational space in which the new round of the HGA is performed. Use of conformational memories speeds up and refines the localization of the structure at the Global Energy Minimum (GEM). This algorithm has been used to predict successfully the GEM for Met- and Leu-enkephalin.
Key words: Protein structure prediction, heuristic algorithms, genetic algorithms, protein folding, conformational space search, conformational memories.

Introducción

Debido a su papel fisiológico central en la mayoría de los sistemas biológicos, la estructura y función de las proteínas son foco de un estudio intenso. La estructura 3D biológicamente activa de una proteína está codificada en la secuencia particular de sus aminoácidos; sin embargo, no se conoce con precisión cómo esta secuencia puede conducir a esa estructura. Constantemente se reporta en la literatura la secuencia de aminoácidos de muchas proteínas, y muchas más serán incorporadas en un futuro cercano debido a la explosión genómica. Aún así, el número de estructuras tridimensionales, biológicamente activas, que son depositadas en el Protein Data Bank (PDB) permanece aún muy pequeño. Por lo tanto, la predicción de estructura para las proteínas secuenciadas demandará un desarrollo continuo de nuevos métodos para el estudio de problemas cada vez más complejos donde habrá poca información experimental de rayos X o espectroscópicos. Elucidar la estructura tridimensional de estas macromoléculas biológicas solamente a partir de la información de la secuencia de sus aminoácidos es un problema fundamental para la biología molecular, para la biofísica, y para la química computacional. Contar con un algoritmo de predicción exitoso podría tener un efecto profundo en la biotecnología y en el tratamiento de enfermedades; sin embargo, una solución práctica todavía parece difícil de encontrar debido al problema de los múltiples mínimos.

El problema de los múltiples mínimos está relacionado con encontrar el Mínimo Global de Energía (MGE) entre un enorme número de mínimos locales presentes en la hípersuperficie de energía potencial de las moléculas flexibles como las proteínas. La hipótesis termodinámica de Anfisen [1] implica que la estructura nativa de una proteína es aquella que se encuentra en el MGE y que es una estructura única. El

grado de libertad rotacional en una proteína está ligado al número de ángulos de torsión de ésta. Si cada uno de estos ángulos tuviera igual probabilidad de movimiento, el número de posibles conformaciones que una proteína dada pudiera adoptar es astronómico. El encontrar el mínimo global en esta hípersuperficie en una cantidad finita de tiempo constituye lo que se conoce como el problema del plegado de proteínas, el cual se ha probado que pertenece a los problemas matemáticos del tipo *NP*-completos [2] La búsqueda sistemática en esta hípersuperficie es impráctica ya que aún para una molécula de mediano tamaño el número más pequeño de posibles conformeros es de 2^N, donde N es el número de variables (ángulos de torsión). Una forma de resolver este problema es por medio del diseño de métodos computacionales que puedan revelar las diversas conformaciones de baja energía para estas moléculas, lo cual en principio se puede obtener usando el muestreo conformacional.

Las técnicas de muestreo conformacional pueden agruparse en métodos determinísticos y estocásticos. Los primeros incluyen cualquier método por el cual la generación y evaluación de una conformación de una molécula está determinada por la conformación anterior. Los así llamados métodos estocásticos están basados en la técnica de la búsqueda de Monte Carlo, donde una conformación es generada cada vez que se toma de manera azarosa los valores para cualquier parámetro que define esta conformación. Sin embargo, todavía no existe un fundamento teórico para decidir cuándo el muestreo es suficiente, por lo que esta decisión es hecha empíricamente. Para muchos problemas estas técnicas han sido la única herramienta válida, sin embargo, ambos métodos consumen una gran cantidad de tiempo de cómputo. Con el afán de reducir los costos de cómputo, nosotros hemos propuesto el uso de meta heurísticos tales como el recosido simulado [3, 4, 5], el threshold accepting [6], y la búsqueda tabú [7].

Existen otros métodos estocásticos que copian las muchas manifestaciones de la evolución natural y se han diseñado para su uso en las computadoras, éstos forman parte de la así llamada computación evolutiva.

La computación evolutiva es esencialmente heurística, esto es, contiene un componente al azar; y no se puede garantizar que este tipo de algoritmos puedan encontrar una solución óptima o simplemente una solución. Los algoritmos evolutivos son usados preferentemente para las aplicaciones donde los métodos determinísticos o analíticos fallan, porque el modelo matemático no está bien definido o porque el espacio de la búsqueda es demasiado grande para una búsqueda sistemática completa. Los algoritmos genéticos (AG) son un ejemplo de los algoritmos evolutivos que se parecen al paradigma del proceso de la información desarrollado y exhibido por la naturaleza. La naturaleza usa los principios de herencia genética y evolución de manera impresionante. Muchas manifestaciones de la evolución natural pueden ser copiadas en la inteligencia artificial, como la supervivencia del más apto y el apareamiento con el más fuerte del grupo, ambas se caracterizan por dar lugar a organismos bien adaptados que pueden ajustarse a un ambiente potencialmente adverso. Otras manifestaciones como el comportamiento altruista, la cooperación y otros pueden imaginarse como el resultado de una evolución superior, pero son muy difíciles de evaluar.

El uso de los AG en la búsqueda conformacional no es una idea nueva [8,9]. Estos se emplearon inicialmente para optimizar ya sea una función objetivo o el esfuerzo de generar individuos bien adaptados en generaciones sucesivas. Las estrategias de evolución en su forma original fueron básicamente algoritmos estocásticos de escalamiento hacia arriba y fueron usados para la optimización de funciones objetivo multiparamétricas, que en la práctica no pueden ser tratadas analíticamente. Los AG en su forma original no fueron diseñados para la optimización de funciones sino para demostrar la eficiencia del cruzamiento genético en la creación de candidatos exitosos dentro de espacios de búsqueda complicados. Una de las áreas más prometedoras y rápidamente creciente en la biología molecular es la aplicación de los AG en el análisis de datos y la predicción de estructura. Los AG ya han sido usados para interpretar los datos de resonancia magnética nuclear en la determinación de la estructura del ADN [10], encontrar el orden correcto para un grupo desordenado de fragmentos de ADN [11], y la predicción de la estructura de proteínas [12, 13, 14] entre otras cosas.

Dado que los AG tienen una convergencia muy lenta, y que hasta la fecha no se ha diseñado algún algoritmo genético que provea con estructuras cercanas al mínimo global de péptidos cuando se toma un método *ab initio* [13,15], nos hemos dado a la tarea de construir un algoritmo genético híbrido (AGH) que nos permita distinguir rápidamente las regiones mas probables del espacio conformacional entre las cuales se pudiera encontrar la estructura del MGE.

Nuestro AGH incluye algunas modificaciones en la forma tradicional de programar los algoritmos genéticos. La primera modificación permite el uso de números reales en lugar de dígitos binarios para describir a cada uno de los genes, o ángulos de torsión dentro de un péptido. La segunda modificación tiene que ver con la aparición de lo que hemos llamado memorias conformacionales al final de cada ciclo del algoritmo genético. La tercera modificación es la inclusión de un operador de perfeccionamiento que nos permite refinar a los cromosomas, o individuos. A cada ciclo de AGH donde el nuevo espacio conformacional está delimitado por las memorias conformacionales obtenidas en el ciclo anterior le hemos llamado un cedazo. Un cedazo, es una técnica de programación combinatoria que toma un conjunto finito y elimina a aquellos de sus miembros que no son de interés. Los cedazos son muy útiles en el cálculo de la teoría de números, *e.g.*, el cedazo de Eratosthenes y los números de Fibonacci [16].

Varias moléculas fueron usadas para valorar nuestra implementación de AGH. Entre éstas, las moléculas prueba correspondientes al pentapéptido de la Met-encefalina (YGGFM) y su análogo la Leu-encefalina (YGGFL), dado que la estructura tridimensional de estos péptidos está muy documentada de manera experimental y teórica [17, 18].

Los experimentos llevados a cabo con las moléculas prueba se realizaron en la ausencia y en la presencia del operador

de refinamiento. Aquellos donde no se aplicó este operador el número total de evaluaciones de la función objetivo estuvieron más allá de cualquier esperanza razonable, los cúmulos de ángulos de torsión (diedros) donde se localizan las memorias conformacionales fueron poco definidos, y para arribar al MEG de las moléculas prueba se necesitaron cuatro pasos de cedazo. Aquellos donde se aplicó este operador, durante el primer ciclo del AGH se encontraron cúmulos de ángulos diedro más definidos que en los experimentos donde se omitió, y por lo tanto, fue más fácil asignar las memorias conformacionales correspondientes. Cuando estos segmentos se usaron para un segundo ciclo de cedazo, en la población resultante se encontró un buen número de individuos en el GEM o muy cerca de éste, y no fue necesario realizar más ciclos de cedazo. Este procedimiento donde se aplicó el operador de refinamiento necesitó de un número mucho menor de evaluaciones de la función objetivo, este número a su vez es menor a los reportados para otros métodos heurísticos.

Métodos

Generalidades de los algoritmos genéticos

Los algoritmos genéticos fueron propuestos por Holland [19] y son procedimientos de búsqueda probabilística basados en la simulación parcial de la evolución y selección natural observada en la naturaleza. Muchas variantes de los AG han sido descritas, sin embargo, todos ellos comparten las mismas características básicas tales como:

1. Los AG codifican el dominio del problema, no las variables.
2. Los AG generan una población inicial de tamaño N_{pob} de posibles soluciones. Estos no trabajan en un solo individuo.
3. Los AG evalúan una función objetivo para cada individuo en la muestra. Esa función no necesita propiedades especiales como ser continua o si ésta es derivable o no.
4. Los AG son azarosos en la selección de los individuos más apropiados.
5. Con una probabilidad *pc* los AG combinan (entrecruzan o aparean) los genes de los individuos seleccionados como padres dando lugar a otros individuos diferentes a los originales.
6. Con una probabilidad *pm* los AG cambian (mutan) el valor almacenado en cualquier lugar de la cadena genética.
7. Los individuos resultantes, apareados o mutados, constituyen la nueva generación de soluciones posibles.
8. Los AG proporcionan soluciones cerca del óptimo (mínimo) en tiempo finito.

Generalmente hablando un AG es un método heurístico que opera en pedazos de información, como lo hace la naturaleza en los genes durante el curso de la evolución. Los individuos están representados por una cadena lineal (cromosoma) de letras (genes) de un alfabeto (en la naturaleza son los nucleótidos; en los algoritmos genéticos son los bits, números, o cualquier otro dato estructural) y a éstos se les permite reproducirse, mutar, y morir. Los individuos son evaluados en cada generación por una función de aptitud, o función objetivo. Los individuos de la generación actual que tienen un mejor desempeño (o aptitud) tienen una probabilidad más alta de participar en la construcción de la siguiente generación. En la operación de cruzamiento normalmente se producen dos hijos por pareja de padres. En la operación de mutación, el valor de un gen que se selecciona al azar dentro de un cromosoma que también se selecciona al azar, será cambiado por otro valor azaroso. Dependiendo de la modalidad de reemplazamiento para cada generación, un subgrupo de padres y los hijos generados pueden o no entrar en el siguiente ciclo de reproducción. Después de un número de iteraciones la población consistirá de individuos que están bien adaptados en términos de la función de aptitud, sin embargo, no se puede probar que entre los individuos de la generación final estará la solución óptima.

Para el problema simplificado del plegado de proteínas Unger y Moult [20] han demostrado que el desempeño de un AG es mucho más eficiente que cualquiera de las estrategias de Monte Carlo, ya que los algoritmos genéticos proveen mejores valores de aptitud, y con mucho menor esfuerzo computacional.

En la aplicación tradicional de los AG existen parámetros inherentes tales como el número de individuos en la población, la probabilidad de mutación, y la probabilidad de apareamiento entre los padres; generalmente todos éstos se mantienen constantes a través de la ejecución del algoritmo, y hay que realizar muchos experimentos, costosos en tiempo, para encontrar los valores que proporcionen la mejor respuesta. Sin embargo, es mucho más deseable encontrar los mejores valores de cada uno estos parámetros para cada problema a ser resuelto, esta es la base de los algoritmos genéticos auto adaptativos [21], donde cada parámetro varia dentro de cierto intervalo, de tal forma que durante la ejecución del algoritmo los valores más adecuados para estos parámetros son afinados. En este reporte usamos la manera tradicional de un AG, en el cual por medio de un operador de perfeccionamiento proponemos que es posible reducir significativamente el costo de los algoritmos genéticos en el modelado molecular a pesar de no contar con los parámetros adecuados.

AG en el modelado molecular

El desarrollo de nuestro algoritmo genético (AG) se basó en la así llamada aproximación híbrida. Esto significa que un AG se puede configurar para operar con números reales, no con cadenas de bits como en el algoritmo original [22], y que puede o no incluir un paso determinístico en la forma de un optimizador local. Un algoritmo genético híbrido (AGH) es más fácil de implementar y también facilita el uso de operadores de dominio específicos en donde, por ejemplo, el proceso de codificación / decodificación es evitado. Para una representación del AGH para proteínas uno puede usar coordenadas cartesianas, ángulos de torsión, rotameros, o de cual-

quier otra descripción simplificada de sus residuos. Estrictamente hablando los fundamentos matemáticos de los algoritmos genéticos son válidos sólo para representaciones binarias, sin embargo, algunas de sus propiedades matemáticas también son válidas para una representación de punto flotante.

La estructura tridimensional de nuestras proteínas modelo se codificó con una secuencia de ángulos de torsión bajo la suposición de que sus distancias y ángulos de unión permanecieron constantes durante la búsqueda. Los ángulos de torsión codificados corresponden a los ángulos definidos como φ, ψ, ω, y χ. Esta suposición es ciertamente una simplificación de la situación real donde las longitudes y ángulos de unión cambian hasta cierto punto dependiendo del ambiente de cada átomo. Sin embargo, el subgrupo de ángulos de torsión provee suficientes grados en libertad para representar cualquier conformación permitida dentro de un margen pequeño de desviación estándar. Una característica de la representación geométrica en base a ángulos de torsión es el hecho de que pequeños cambios en los ángulos φ y ψ de un péptido pueden inducir cambios significativos en la conformación entera. Esta propiedad es útil para crear la diversidad necesaria dentro de una población al comienzo de la simulación.

En el funcionamiento de un AGH la búsqueda dentro del espacio conformacional es llevada a cabo en varios pasos. En el primer ciclo el espacio conformacional comprende los ángulos euclidianos de una circunferencia entre 0° y 360°, donde los ángulos de torsión (variables o genes) pueden tomar cualquier valor, es muestreado de acuerdo a la hipótesis de probabilidad igual *a priori*. En el primer paso se coleccionan resultados de unas cuantas corridas del AGH sobre una población fija de conformeros. En cada corrida la población inicial es generada al azar a la cual se le permite evolucionar dentro de unas cuantas generaciones. Cada conformero está sujeto a una minimización local después de cada reproducción y mutación. En la generación final se espera la presencia de varios conformeros con energía suficientemente baja. En principio, estos conformeros de baja energía deben proveer una distribución de valores de ángulos diedro estéricamente permitidos para cada uno de los ángulos de torsión que definen a la molécula prueba. Estos valores deben pertenecer a un espacio conformacional reducido donde estos conformeros de baja energía existen, valores fuera de esta distribución corresponderán a conformeros de alta energía. Al mismo tiempo, la distribución de ángulos esterícamente permitidos puede ser usada como entrada para el siguiente ciclo del algoritmo, el cual ahora trabajará en un espacio conformacional reducido. En los ciclos subsecuentes este proceso es repetido y se espera que una combinación adecuada de estos intervalos angulares podría converger al MGE y sus conformaciones vecinas de baja energía.

A cada ciclo de AGH donde el nuevo espacio conformacional está delimitado por las memorias conformacionales obtenidas en el ciclo anterior le hemos llamado un cedazo. Un cedazo, es una técnica de programación combinatoria que toma un conjunto finito y elimina a aquellos de sus miembros que no son de interés. Los cedazos son muy útiles en el cálculo de la teoría de números, *e.g.*, el cedazo de Eratosthenes y los números de Fibonacci [16].

Asignación de las memorias conformacionales

Para facilitar la designación de las memorias conformacionales hemos tomado en cuenta lo siguiente: 1) al final de cada ciclo del AGH elaborar un histograma de energía conformacional contra los valores de ángulo diedro para cada gen de los individuos presentes en la población final, 2) los cúmulos alrededor de {–180°, –170°} y {170°,180°} son en realidad partes del mismo conjunto de memoria conformacional ya que este es el intervalo del ángulo ω, o unión peptídica en posición *trans*, 3) los puntos que pertenecen a estructuras de alta energía pueden ser eliminados de cada conjunto ya que estos no presentan un peso estadístico importante cuando se les compara con otras regiones altamente pobladas, 4) después de esta acción, detectamos si el conjunto se rompe en otros subconjuntos, esto es, que puedan adoptar límites bien definidos (el criterio empleado para que se forme una nueva memoria conformacional es que los conjuntos deben estar separados por bandas vacías de 40° o más), 5) los conjuntos formados por tres o menos valores de ángulos diedro son eliminados, a menos que uno de ellos sea o esté muy cercano al mínimo energía más bajo encontrado en el experimento, 6) los conjuntos donde a pesar de estas consideraciones no se logró obtener una mejor definición, se mantuvo el conjunto original como una memoria conformacional muy ancha en espera de que ésta sea refinada en los cedazos subsecuentes. Este procedimiento fue codificado en un programa de post proceso que facilita la generación de la tabla de consulta para las memorias conformacionales usadas en el segundo cedazo.

Los operadores AGH

En nuestra implementación un individuo o conformero, es un cromosoma hecho de genes, donde cada gen representa el valor de un ángulo de torsión representado por números reales. Un individuo con una conformación dada define a un padre (diferentes padres deben tener la misma secuencia de aminoácidos, pero diferentes conformaciones). La aptitud de un conformero está determinada con un campo de fuerza empírico donde los conformeros con energía baja son considerados como sobrevivientes, mientras que los conformeros de energía alta son considerados no aptos para evolucionar. El operador de cruzamiento permite a los padres el aparearse con una probabilidad *pc*, los genes recombinantes crean N_{prole} individuos (la conformación de la prole debe ser diferente de cualquiera de sus padres) para restaurar el número original de individuos. Para asegurar la variación genética, las mutaciones se introducen al azar después del proceso de apareamiento con una probabilidad *pm*. Con el operador de perfeccionamiento relajamos los conformeros con contactos estéricos fuertes, y por lo tanto se reorienta la búsqueda a los individuos mejor adaptados. Esta operación se lleva cabo empleando el método

Tabla I. Parámetros establecidos en la operación del AGH.

Parámetro	Descripción	Valor
N_{pob}	número de individuos en la población	20-30
N_{gen}	número de generaciones	20-70
pc	probabilidad de realizar el cruzamiento uniforme	0.6-0.7
pm	probabilidad de realizar la mutación	0.2-0.3
cm	método para realizar el cruzamiento	HU@1[a] UP@2
m	número de mutaciones puntuales	1
om	método para realizar las mutaciones	gaussian@ 1[b] lineal@ 2
oo	probabilidad de optimización de la prole	1.0
mo	probabilidad de optimización de los mutantes	1.0
t	control de gemelos	T
N_{buenos}	número de individuos buenos que pasan a la siguiente generación	2-4
N_{malos}	numero de individuos malos a ser eliminados	16-18
N_{genes}	número de genes en los cromosomas	24-78
N_{cyc}	número de ciclos de optimización	10

[a] HU@1 = Cruzamiento Heurístico Uniforme en el cedazo número uno. UP@2 = Cruzamiento de Un Punto en el cedazo número dos.
[b] Gaussiana@1 = Mutación basada en una distribución gaussiana en el cedazo numero uno. Lineal@2 = Mutación basada en una ditribución lineal en el cedazo numero dos.

de minimización de Newton-Raphson. Para asegurar la sobrevivencia del más apto, N_{buenos} conformeros de baja energía (uno o más) de la generación actual pasarán sin perturbarse a la siguiente generación. La nueva población consiste de $N_{pob} = N_{buenos} + N_{prole}$ en la que no hay preferencias, esto significa que todos los individuos de la nueva población tienen la misma oportunidad para aparearse.

La función de aptitud

Para evaluar la aptitud de cada individuo se usó la función de potencial conformacional conocida como ECEPP / 2 [23] como función objetivo. La función de potencial ECEPP / 2 ha sido descrita en otros lados, pero hablando generalmente, ésta es una suma de expresiones para el potencial electrostático, el potencial de van der Waals por pares, puentes de hidrógeno, el potencial del ángulo de torsión, el potencial de formación de anillos, y los puentes disulfuro.

$$E_{conf} = E_{el} + E_{vdW} + E_{hb} + E_{tor} + E_{loop} + E_{S\text{-}S}$$

En este potencial se supone que las longitudes y los ángulos de unión son constantes. En todas las corridas el plegado fue simulado en el vacío.

Codificación de parámetros, precisión y límites

En nuestro AGH el algoritmo está limitado a explorar el espacio conformacional definido por todo el conjunto de ángulos euclidianos en una circunferencia, particularmente está limitado a la búsqueda de los ángulos diedro dentro de los límites $-180\ ^\circ < \theta \leq +180^\circ$. Para la aplicación de la función de aptitud, es necesario convertir la geometría de la proteína representada por ángulos de torsión a una representación en coordenadas cartesianas. En este formato, los parámetros son continuos y pueden tomar cualquier valor, así que la precisión de estos parámetros depende solamente de la precisión y error de redondeo del procesador de la computadora usada. El código correspondiente está en lenguaje FORTRAN, está compilado en una PC con doble precisión (32 bits) con el sistema operativo Linux, y fue empleado como una extensión del programa FANTOM v4.2 [24,25].

Población inicial

La población inicial se construyó de acuerdo con:

```
do i = 1, N_pob
    do j = 1, N_genes
        if (memoria.eq.false) then
            N_pob{i, j} = (al − ba) × azar {0.0, 1.0}
            + ba
        else
            N_pob{i, j} = azar {liminf, limsup}_j
        endif
    enddo
enddo
donde
```

al = número más alto en el intervalo = +180.0
ba = número más bajo en el intervalo = –180.0

Cuando la población inicial se genera usando las memorias conformacionales, los individuos surgen dentro de los respectivos límites inferior (*liminf*) y superior (*limsup*) definidos por la memoria conformacional correspondiente a cada gen.

Selección natural

Dado que los cromosomas no son creados iguales, cada uno de estos fue evaluado por la función de aptitud y catalogado desde el costo más bajo al costo más alto. Los N_{buenos} miembros de la generación *i*-ésima son parte de los N_{pob} miembros de la generación $(i + 1)$ésima donde la nueva prole ha reemplazado a la parte inferior de N_{malos} miembros de la generación *i* ($N_{malos} = N_{pob} - N_{buenos}$). Este proceso ocurrió en cada iteración del algoritmo para permitir que la población evolucione a través de las generaciones hacia los miembros mejor adaptados. El tamaño de la población en cada generación se mantuvo constante a N_{pob} individuos.

Tabla 2 Memorias conformacionales puras para la Met- y Leu-encefalina.

Residuo	Ángulo	Intervalo A(*ba*, *al*)	Intervalo B (*ba*, *al*)	Intervalo C (*ba*, *ali*)	Intervalo D (*ba*, *al*)
Tyr_1	ϕ	(–180 ,–150)	(–120°,–60°)	(50°,70°)‡	(170 ,180)
	ψ	(–60 ,–30)	(70°,100°)†	(130°,160°)	
	ω	(–180°,–170°)	(170°,180°)		
	χ_1	(–180°,–167°)	(50°,70°)‡	(50°,70°)‡	(170°,180°)
	χ_2	(–130°,–80°)	(70 ,100)		
	χ_6	(–180°,–170°)	(–15°,–5°)‡	(0°,20°)	(155°,180°)
Gly_2	ϕ	(–166°,–140°)	(–92°,–69°)	(69°,80°)	(146°,166°)
	ψ	(–160 ,–140)‡	(–100°,–70°)	(50°,100°)	(150°,170°)‡
	ω	(–180°,–170°)	(170°,180°)		
Gly_3	ϕ	(–160°,–140°)	(–90°,–70°)	(70°,90°)	(150°,160°)‡
	ψ	(–90°,–30°)	(–20°,0°)†	(0°,30°)	(70°,100°)‡
	ω	(–180°,–170°)	(170°,180°)		
Phe_4	ϕ	(–160°,–140°)	(–80 ,–60)		
	ψ	(–60°,–30°)	(10°,40°)‡	(135°,175°)	
	ω	(–180°,–170°)	(170°,180°)		
	χ_1	(–180 ,–170)‡	(–60 ,–50)	(50°,70°)	(170 ,180)
	χ_2	(–110°,–80°)	(70°,100°)		
Met_5 / Leu_5	ϕ	(–175°,–140°)	(–85°,–60°)	(50 ,65)†	
	ψ	(–80 ,–60)‡	(–50 ,–30)	(60°,170°)	
	ω	(–180°,–170°)	(170°,180°)		
	χ_1	(–180°,–170°)	(–75 ,–50)	(50 ,75)‡	(170 ,180)†
	χ_2	(–180°,–170°)‡	(–75°,–65°)‡	(60°,75°)	(174°,179°)
	χ_3	(–180°,–170°)‡	(–90°,–80°)	(80°,90°)	(175°,180°)
	χ_4	(–180°,–170°)	(–60°,–60°)	(61°,61°)	(175°,180°)

†) Intervalo que no está presente en la memoria conformacional para la Met-encefalina.
‡) Intervalo que no está presente en la memoria conformacional para la Leu-encefalina.

Mejoramiento genético

El operador de mejoramiento consistió en aplicar 10 ciclos del minimizador de Newton-Raphson a cada individuo nuevo que aparece en la población y a cada individuo que ha sido mutado antes de regresarlo al grueso de la población. La acción neta del minimizador local fue la de relajar una posible conformación de baja energía, pero que por un contacto estérico ésta no pueda ser considerada como candidato a sobrevivir; por lo tanto se recupera la tendencia de la evolución hacia los individuos mejor adaptados.

Apareamiento

El operador de cruzamiento fue diseñado para trabajar en uno de los esquemas siguientes: a) cruzamiento en un punto, b) cruzamiento en dos puntos, y c) el cruzamiento uniforme. Este último se diseño para trabajar en una de dos esquemas, el cruzamiento lineal uniforme y el cruzamiento heurístico uniforme. Los puntos de cruzamiento se seleccionan al azar y es el mismo para los cromosomas de cada uno de los dos padres.

En el cruzamiento en un punto y en dos puntos cada gen se propaga entre la prole en combinaciones diferentes, ya que solo se intercambian los segmentos de igual longitud entre los padres. Estos mecanismos están bien documentados en la literatura [26] y generalmente no introducen nueva información genética.

El cruzamiento lineal uniforme entre dos padres que se aparean consiste de recorrer la secuencia de los genes para cada cromosoma, y en cada posición intercambiar esta información con una probabilidad *pc*. En este caso tampoco se introduce nueva información.

El cruzamiento heurístico uniforme fue presentado por Michalewicz [27]. Este es un tipo de método de mezclado que encuentra formas de combinar los valores en los genes de cada uno de los padres para dar valores nuevos en los genes de la prole. El cruzamiento heurístico comienza al seleccionar al azar una pareja de padres entre los N_{pob} miembros de la población donde cada padre consiste de:

$$\text{padre}_1 = [p_{m1}, p_{m2}, \ldots \boldsymbol{p_{mi}}, \ldots p_{mN\text{genes-1}}, p_{mN\text{genes}}]$$
$$\text{padre}_2 = [p_{p1}, p_{p2}, \ldots \boldsymbol{p_{pi}}, \ldots p_{pN\text{genes-1}}, p_{pN\text{genes}}]$$

donde los subscriptos *m* y *p* indican al individuo *mamá* y al individuo *papá,* respectivamente. La nueva prole se genera al combinar cada uno de los genes de acuerdo con la siguiente regla:

$$p_{\text{nuevo1}} = p_{mi} - \beta[p_{mi} - p_{pi}]$$
$$p_{\text{nuevo2}} = p_{pi} + \beta[p_{mi} - p_{pi}]$$

donde β es un número azaroso entre 0.0 y 1.0.

Para el entrecruzamiento dos padres son seleccionados al azar de entre los N_{pob} individuos de la población, el aparea-

miento se lleva a cabo en uno de los esquemas mencionados, y la cantidad de prole que se genera es equivalente al número de individuos N_{malos} de forma tal que se pueda regenerar el número de N_{pob} requerido para una nueva generación.

Mutación

Con una probabilidad *pm*, se selecciona al azar un individuo para su mutación. Para este individuo uno de sus genes se escoge al azar y su valor real es remplazado por otro valor real generado de manera azarosa en el intervalo (–180.0°, + 180.0°]. Cuando se usan las memorias conformacionales el nuevo ángulo se genera dentro de los respectivos límites inferior y superior definidos por su memoria conformacional.

Ya que la mutación se lleva cabo con números reales y no con cadenas de bits, la generación del nuevo valor fue programado para hacerse en una de las tres siguientes maneras. La primera de éstas fue con el uso de un generador de números aleatorios con una distribución lineal uniforme, la segunda fue con el uso de un generador de números aleatorios con una distribución gaussiana, y la tercera con el uso de un generador de números aleatorios con una distribución de Poisson. De éstos, los números aleatorios en base a una distribución lineal y gaussiana dieron el mejor desempeño para nuestro algoritmo. La distribución de Poisson dio resultados muy pobres.

Tratamiento de gemelos

Correr un AG con una población de pocos individuos genera demasiados gemelos. Dado que esta sociedad de cromosomas no trabaja bajo una democracia, para evitar que el algoritmo se atrape en una población que no evoluciona, uno de los gemelos es eliminado y es remplazado por otro hijo que se genera al cambiar a uno de los padres por otro diferente entre los N_{pob} individuos de la generación.

Implementación

El algoritmo descrito en este reporte fue desarrollado como una extensión del programa FANTOM v4.2 [24,25]. Las energías son calculadas usando la función de potencial ECEPP/2 [23] el cual es una función de los ángulos de torsión. FANTOM trabaja con base en a coordenadas cartesianas, que son calculadas a partir de una matriz Z donde las longitudes y los ángulos de unión permanecen constantes durante toda la simulación. Los ángulos de torsión ϕ, ψ, ω y χ de los péptidos son generados de acuerdo con las reglas de nuestro algoritmo genético, donde el ángulo w se mantuvo constante a 180° excepto cuando el cromosoma se sometió a una minimización local.

Las subrutina que contiene el desempeño de nuestro algoritmo fue escrito de manera separada del programa principal FANTOM, y fue incorporada posteriormente como otro de sus comandos. El diagrama de flujo se muestra en la Fig. 1.

Los parámetros de control para nuestro algoritmo genético fueron: el tamaño de la población (N_{pob}), el número de generaciones (N_{gen}), la tasa de apareamiento (*pc*), la tasa de mutación (*pm*), el número de mutaciones puntuales (*m*), la optimización de la prole (*oo*), la optimización de los mutantes (*mo*), el control de los gemelos (*t*), el número de genes en el cromosoma (N_{genes}), el número de ciclos en los pasos de minimización (N_{cyc}), y los números de N_{buenos} y N_{malos} miembros de la población.

Todos los cálculos se realizaron en una PC con un procesador dual Pentium III a 550 MHz bajo el sistema operativo Linux Red Hat v6.2. Los algoritmos fueron escritos en FORTRAN ya que el programa FANTOM estaba escrito en este lenguaje. El programa fue compilado como un ejecutable de doble decisión a 32 bits.

Las moléculas prueba

El pequeño neuropéptido de la Met-encefalina (YGGFM) es una molécula comúnmente usada como prueba en la evaluación de varios programas de predicción de estructura y de técnicas de búsqueda conformacional (véase la referencia 16 para una revisión). Dado que la mayoría de los trabajos reportados han usado la función de potencial empírica conocida como ECEPP /

PSEUDO CÓDIGO PARA EL ALGORITMO GENÉTICO HYBRIDO

Entrada		ga.seq (datos del péptido), ga.inp (parámetros del AGH)
Paso 1		Inicializar
	(A)	Poner el contador de las iteraciones a $k = 1$
	(B)	Establecer el espacio conformacional para cada ángulo dentro de intervalos fijos
Paso 2		Correr *N* veces el AGH con una población inicial azarosa de tamaño N_{pop} dentro de estos intervalos
		(A) Aplicar el operador de mejora
		(B) Con una probabilidad ***pc*** escoger a los cónyuges para producir tanta prole como individuos N_{malos}
		(C) Con una probabilidad *pm* mutar a la población
		(D) Sustitución y remoción de gemelos
		(E) N_{buenos} individuos más la prole constituyen la nueva población
		(F) Si el número de generaciones no se satisface vaya al paso 2(A) si el operador de mejora está activo, en caso contrario vaya al paso 2(B)
Paso 3		Obtención de las memorias conformacionales
	(A)	Al final de la iteración *k* realizar la reducción del espacio conformacional empleando los resultados de las *N* corridas del AGH
Paso 4		$k = k + 1$
Paso 5		Si la condición de término no se satisface, vaya al Paso 1(B)
Salida		ga.out (población final), ga.ang (información de los ángulos diedros), ga.pdb (estructuras en formato PDB)

Fig. 1. Diagrama de flujo del algoritmo genético híbrido.

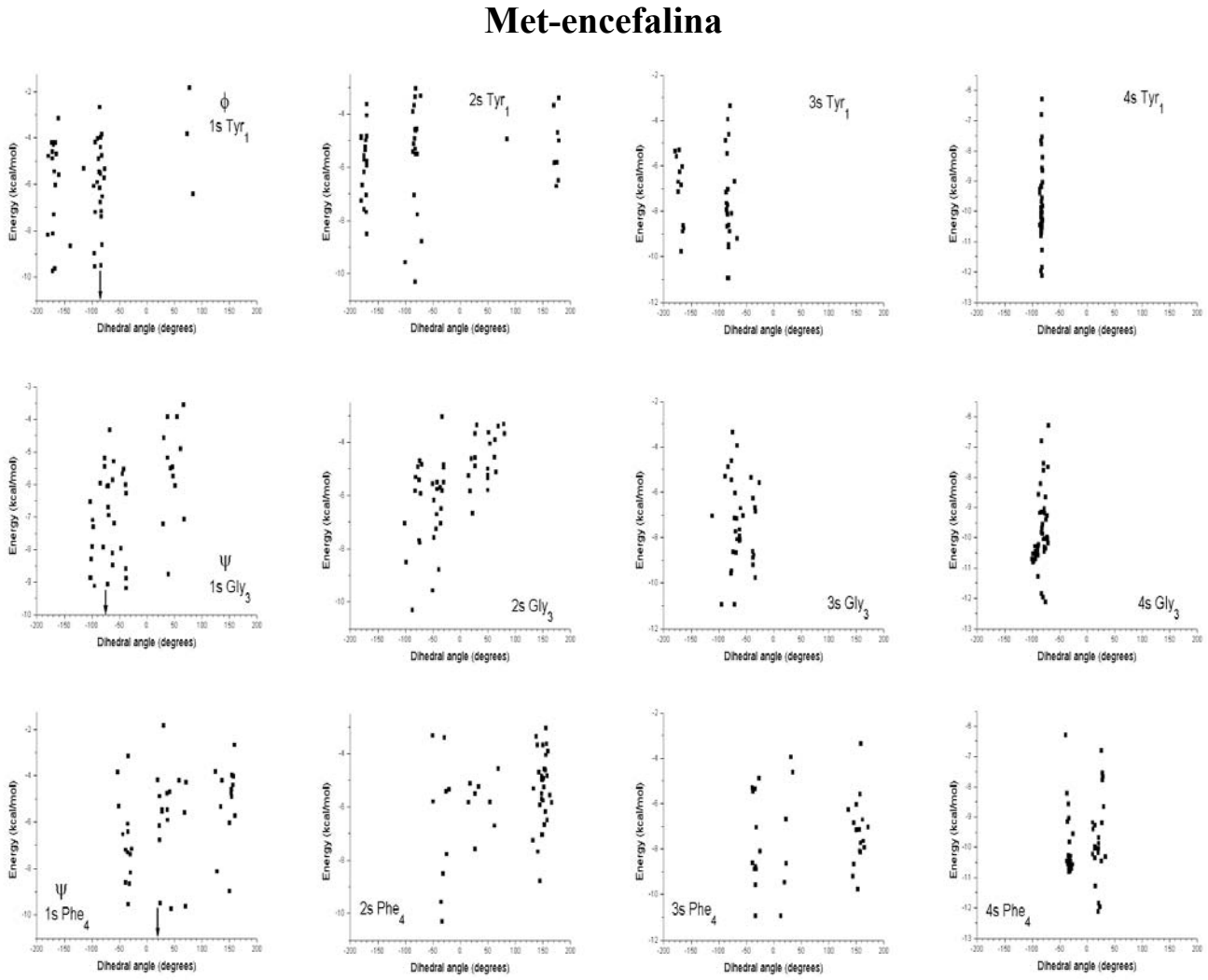

Fig. 2. Histogramas representativos de los cuatro cedazos realizados en la fase preliminar del AGH para encontrar la estructura del MGE de la Met-encefalina. Cada uno de los cedazos están numerados como 1s, 2s, 3s y 4s. La flecha en negrita señala el valor que adopta cada ángulo en la estructura del MGE. El proceso de mejoramiento hacia la obtención del MGE es evidente. Al final del cuarto cedazo se observó que el cúmulo donde se encontró el individuo con más baja energía corresponde a la región indicada por la flecha en el primer cedazo.

2, y que reportan la misma estructura tridimensional para el conformero correspondiente a la energía más baja, esto sugiere fuertemente que esta estructura pertenece al MEG cuando se usa este campo de fuerza. Ninguna otra conformación de la Met-encefalina ha sido reportada con una energía potencial más baja.

La Met-encefalina tiene un análogo de mayor dificultad computacional, la Leu-encefalina (YGGFL) y por esta razón también incluimos a este neuropéptido en este estudio, ya que representa un desafío interesante para probar la fuerza predictiva de nuestro algoritmo.

Cedazo sin mejoramiento en las moléculas prueba

Primer cedazo. Esta fase consistió en llevar a cabo 50 corridas independientes del AGH con poblaciones iniciales generadas al azar para estas moléculas. La búsqueda conformacional para todos los genes se realizó en el espacio de ángulos diedros definido por el conjunto de los números reales dentro de una distribución lineal uniforme en el intervalo (–180.0, + 180.0]. Las estructuras de la Met- y la Leu-encefalina se definieron con $N_{genes} = 24$, y los parámetros empleados para el AGH fueron $N_{pob} = 20$, $N_{gen} = 20$, $pc = 0.6$ para el algoritmo de cruzamiento heurístico uniforme, y $pm = 0.2$ usando el algoritmo de una distribución gaussiana para la mutación.

Al final del primer cedazo se coleccionó el mejor individuo encontrado en cada uno de los 50 experimentos. A partir de este conjunto se construyó un histograma de energía contra la distribución angular para cada uno de los 24 ángulos de torsión. Cada uno de estos histogramas fue tratado como se describió arriba para obtener las memorias conformacionales correspondientes. A los conjuntos de ángulos diedros formados por cada individuo en la población final los hemos llamado memorias conformacionales crudas.

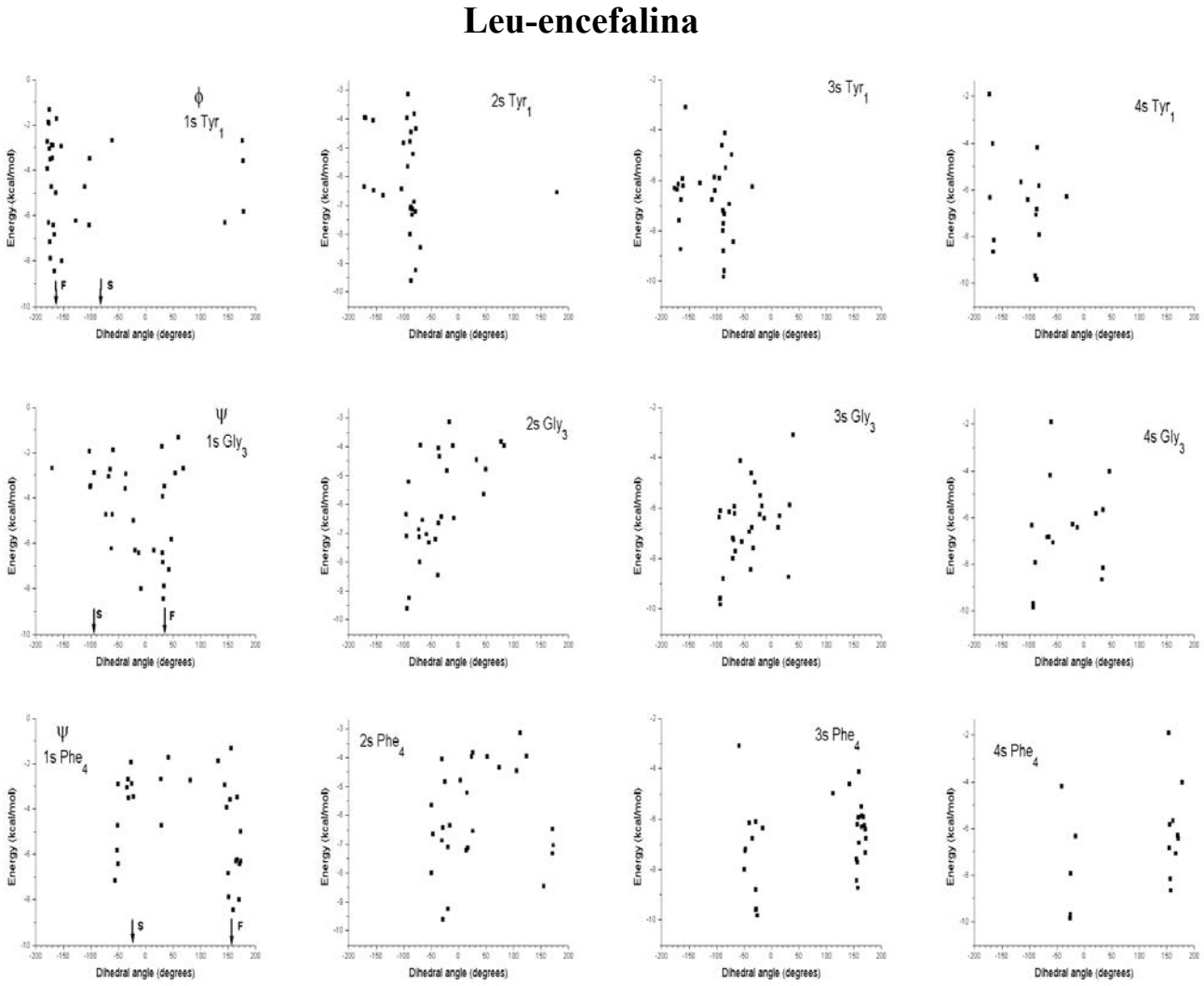

Fig. 3. Histogramas representativos de los cuatro cedazos realizados en la fase preliminar del AGH para encontrar la estructura del MGE de la Leu-encefalina. Cada uno de los cedazos están numerados como 1s, 2s, 3s y 4s. Las flechas en negrita señalan el valor que adopta cada ángulo en las dos estructuras reportadas como el MGE de este péptido. El proceso de mejoramiento hacia la obtención del MGE es evidente. Al final del cuarto cedazo se observó que los cúmulos donde se encontraron los individuos con más baja energía corresponden a las regiones indicadas por las flechas en el primer cedazo.

Cedazos posteriores. El segundo cedazo consistió en llevar a cabo 50 corridas independientes del AGH donde la búsqueda se llevó a cabo en el espacio conformacional reducido definido por las memorias conformacionales de los ángulos diedros del cedazo anterior. Al final de este proceso se realizó nuevamente una reducción del espacio conformacional de manera similar al cedazo número uno. Este proceso se repitió una tercera y una cuarta vez.

Cedazo con mejoramiento en las moléculas prueba

Primer cedazo. Esta fase consistió en llevar a cabo 10 corridas independientes del AGH con poblaciones iniciales generadas al azar para cada una de las moléculas estudiadas. La búsqueda conformacional para todos los genes se realizó en el espacio de ángulos diedros definido por el conjunto de números reales dentro de una distribución lineal uniforme en el intervalo (–180.0, + 180]. Los parámetros empleados para el AGH fueron N_{pob} = 20, N_{gen} = 30, pc = 0.6 para el algoritmo de cruzamiento heurístico uniforme, y pm = 0.2 usando el algoritmo de una distribución gaussiana para la mutación. De la población final en cada corrida se seleccionó aquellos individuos con energía por debajo o igual a 0.0 kcal / mole para construir los histogramas de cada gen, tal como se describió anteriormente. A los conjuntos de ángulos diedros que aparecen en cada uno de los histogramas los hemos llamado memorias conformacionales puras.

Segundo cedazo. En esta parte del experimento se realizaron 10 corridas independientes del AGH, donde cada una de las poblaciones iniciales fueron generadas de manera azarosa dentro de los límites obtenidos para las memorias conformacionales del primer cedazo. Los parámetros empleados para el

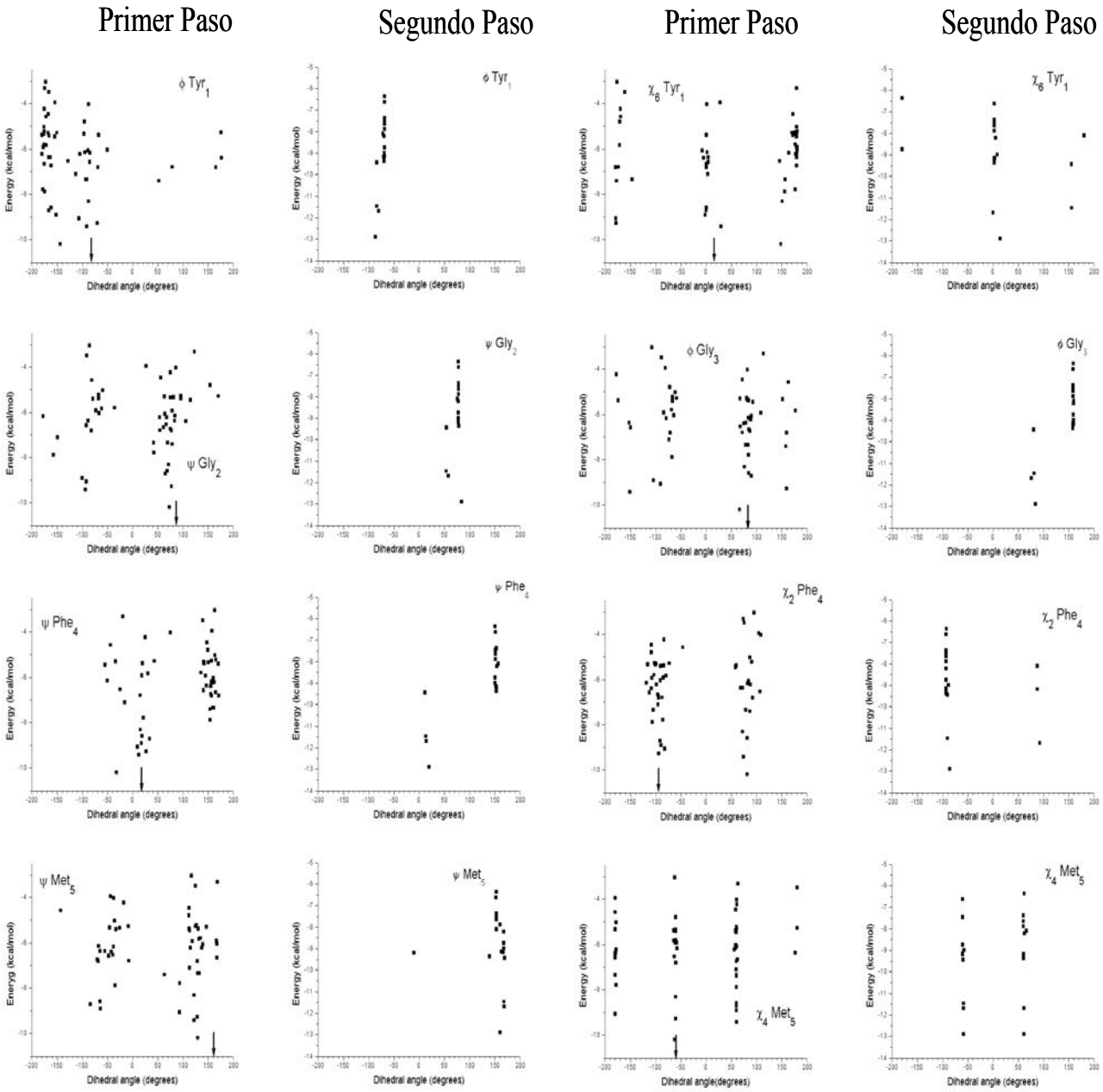

Fig. 4. Histogramas representativos para el AGH donde los *N* individuos de la Met-encefalina fueron sometidos al operador de mejoramiento. La flecha en negrita indica el valor que adopta cada ángulo en la estructura del MGE para este péptido. En las graficas señaladas como primer cedazo se observa una mejor definición de las memorias conformacionales debido al proceso mejoramiento, y por lo tanto en las graficas señaladas como segundo cedazo se observa que los cúmulos de ángulo diedro son mucho más delgados y que definen con claridad la posición del MGE.

AGH fueron N_{pop} = 20, N_{gen} = 30, pc = 0.6 usando el algoritmo de cruzamiento en un solo punto, y pm = 0.2 usando el algoritmo de una distribución lineal. En la mayoría de los casos la población total final de este cedazo proporcionó un 25 % de individuos con una conformación cercana al MGE conocido para las moléculas prueba. Es claro que un tercer cedazo no fue necesario para detectar el MGE.

Resultados

Cedazos sin mejoramiento

Al final del primer cedazo coleccionamos los datos estructurales del mejor individuo encontrado en cada uno de los 50 experimentos propuestos. Para cada uno de los 24 ángulos de torsión se construyó un histograma de energía contra su distribución angular (Figs. 2 y 3). En la mayoría de los casos estas

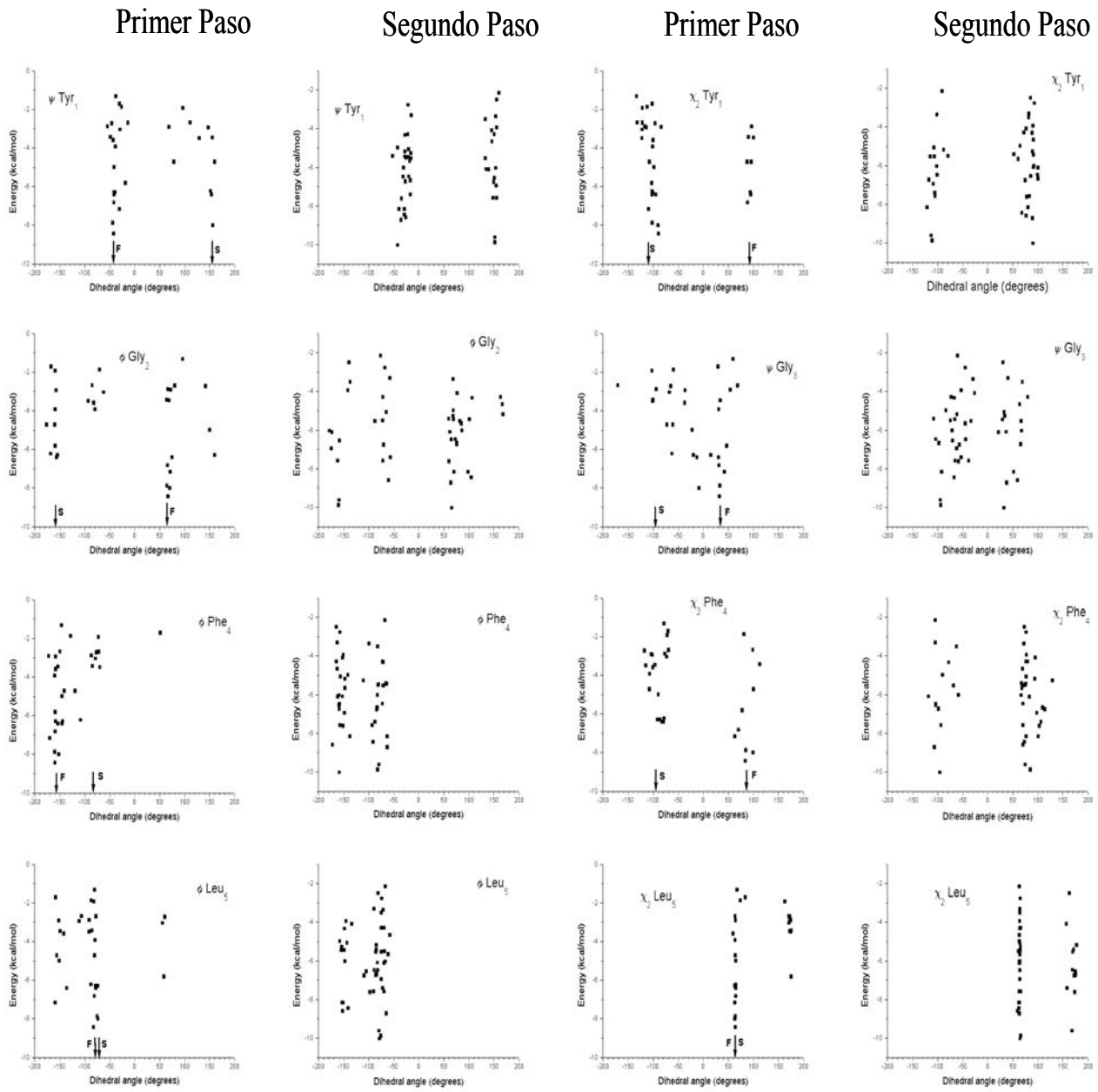

Fig. 5. Histogramas representativos para el AGH donde los *N* individuos de la Leu-encefalina fueron sometidos al operador de mejoramiento. Las flechas en negrita indican el valor que adopta cada ángulo en las dos estructuras reportadas como el MGE para este péptido. La letra ***S*** es para el MGE reportado por el grupo de Scheraga y la letra ***F*** es para el MGE reportado por el grupo de Floudas. Debido al proceso de mejoramiento en las graficas señaladas como segundo cedazo se observa que los cúmulos de ángulo diedro son mucho más delgados y que definen con claridad las dos posiciónes del MGE.

gráficas fueron fáciles de interpretar en términos de las memorias conformacionales, ya que los intervalos para estructuras de baja energía formaron conjuntos bien definidos de ángulos diedros. Sin embargo, hubo otras gráficas donde los intervalos fueron difusos y difíciles de agrupar en un conjunto dado. El grupo completo de estos histogramas se puede encontrar en http:// www.fis.unam.mx/CCF/areas_invest /acad/ academicos/ ramong.html. Cada vez que se inicia un nuevo experimento sobre la misma molécula, los conjuntos de ángulo diedro que se obtienen son semejantes entre sí, y dado que estos conjuntos de ángulos se comportan de manera similar a las memorias conformacionales propuestas por Guarnieri *et al*. [28, 29], hemos decidimos adoptar el mismo nombre para estos conjuntos.

Al final del segundo cedazo, donde la búsqueda del AGH se llevó a cabo en el espacio reducido de ángulos diedro definido por las memorias conformacionales del cedazo anterior, también se coleccionaron los datos estructurales del mejor individuo encontrado en cada una de las 50 corridas

Tabla 3. Tiempos de CPU y número de evaluaciones de la función objetivo reportados por varios métodos de búsqueda del mínimo global para la Met-encefalina.

Método	N_{var}[a]	CPU (hr)	Núm. de evaluaciones (10^5)[b]	Computadora	Mflops[c]
Monte Carlo[37a, 37b]	19	2-3	1.0	IBM 3090	7.5
	24	10	—		
Recocido Simulado[5]	24	2.5	2.5	Apollo DN10000	5.8
Threshold Accepting[6]	24	1.5	2.0	Apollo DN10000	5.8
Monte Carlo con Minimización[30]	24	1.5-4	—	IBM 3090	7.5
Algoritmo Multicanonico[38]	19	6	1.5	IBM RS/6000 320H	12
Recocido del Espacio Conformacional[39]	24	0.75	1.7	SG Indigo 2	32
Ecuación de Difusión[40a, 40b]	19	0.33	—	IBM 3090	7.5
Teoría del Campo Medio[41]	10[b]	1.6	—	IBM 3090	7.5
αBB[32]	24	1.3	3.9	HP/9000 730	24
Búsqueda Tabu[7]	24	0.07	1.7	SG Origin 2000	114
Este trabajo	20	0.014	0.30	Pentium III, 550MHz	80

[a] Los métodos mencionados emplearon este número de variables en la búsqueda del mínimo global aparente.
[b] Este es el número promedio de evaluaciones basado en 100 corridas.
[c] Jack Dongarra, http://performance.netlib.org/performance/html/linpack.data.col0.html.

independientes y se construyeron histogramas de energía contra distribución angular para cada uno de los 24 ángulos de torsión. De estos histogramas se asignaron nuevas memorias conformacionales para un tercer cedazo. Al final del tercer cedazo se repitió el proceso para obtener nuevas memorias conformacionales para realizar un cuarto cedazo.

La Fig. 2 contiene una serie de histogramas representativos para la búsqueda conformacional de la Met-encefalina. Como ejemplo seguiremos el comportamiento del ángulo ϕ del residuo T_{yr1} durante cuatro cedazos. Cada una de las gráficas está marcada de manera progresiva del cedazo uno al cedazo cuatro (1s al 4s). La primera gráfica muestra tres conjuntos, uno alrededor de –180°, otro alrededor de –90°, y un conjunto de pocos elementos alrededor de +80°. La flecha en negrita indica el valor que este ángulo toma en la estructura de MGE para este péptido [30]. Se puede observar que estos conjuntos se fueron definiendo conforme avanzó el número de cedazos. La figura que corresponde al cedazo número 4 muestra un solo conjunto de ángulos diedros en la región que se indicó en la flecha del primer histograma. Este conjunto muestra una distribución de puntos sobre un intervalo muy amplio de energía conformacional que es debido a la variabilidad genética desarrollada por el algoritmo.

La Fig. 3 contiene una serie de histogramas representativos para la búsqueda conformacional de la Leu-encefalina. Como ejemplo seguiremos el comportamiento del ángulo ψ y del residuo Phe_4 durante cuatro cedazos. La primera gráfica muestra tres conjuntos, uno alrededor de –40°, un conjunto de pocos elementos alrededor de +50°, y un conjunto alrededor de +160°. Las flechas en negrita muestran los valores reportados para dos estructuras consideradas como el MGE para esté péptido. La flecha marcada con una *S* indica el valor reportado por el grupo de H. Scheraga [31], y la flecha marcada con una *F* indica el valor reportado por el grupo de C. Floudas [32, 33]. En esta ocasión se puede observar que a pesar de haber rasurado el contenido de estos cúmulos para formar las memorias conformacionales correspondientes, las dos regiones marcadas aunque difusas siempre estuvieron altamente pobladas y en ocasiones se fueron definiendo conforme avanzó el número de cedazos. La figura que corresponde al cedazo número 4 muestra solamente dos cúmulos de ángulos diedro y que corresponden a las regiones marcadas por las flechas del primer histograma.

En resumen, la serie de cuatro cedazos a los que se sometieron las estructuras para la Met- y para la Leu-encefalina necesitaron aproximadamente de 1×10^6 evaluaciones de la función de aptitud para completar el primer cedazo, y de 7.5×10^5 evaluaciones para completar cada uno de los cedazos restantes. Hasta este punto, este proceso no es mejor que el descrito por Jin *et al.* [14], ya que el número total de evaluaciones está más allá de cualquier expectativa razonable para ser de utilidad práctica en la predicción de estructura terciaria de proteínas.

Cedazos con mejoramiento

Primer cedazo. Al final de cada una de las 10 corridas independientes se seleccionaron aquellos individuos con energía conformacional por debajo o igual a 0.0 kcal / mole. Con los datos estructurales de estos individuos se construyeron 20 histogramas, uno para cada gen, tal como se describió en la sección anterior. Para cada uno de estos histogramas se asignaron las memorias conformacionales correspondientes. Ejemplos de éstas se ilustran en la Fig. 4 para la Met-encefalina y en la Fig. 5 para la Leu-encefalina. La Fig. 4 contiene ocho histogramas representativos de los cinco residuos de aminoácido presentes en la Met-encefalina, la primera y tercera columnas corresponden a los resultados obtenidos después del primer cedazo. En cada uno de éstos se presentan las memorias conformacionales sin depurar. La flecha en negrita indica el valor de ángulo diedro observado [30] en el MGE para este péptido.

Es notorio que existe un cúmulo de ángulos diedros bien definido en la región marcada con la flecha. La Fig. 5 contiene ocho histogramas representativos de los cinco residuos de aminoácido presentes en la Leu-encefalina, la primera y tercera columnas corresponden a los resultados obtenidos después del primer cedazo. En estos histogramas se presentan dos flechas en negrita, una marcada con una *S* y la otra marcada con una *F*. La *S* indica el valor de ángulo diedro observado para el MGE reportado por el grupo de H. Scheraga [31], y la *F* indica el valor de ángulo diedro observado para el MGE reportado por el grupo de C. Floudas [32, 33]. Es notorio que para cada región marcada por las flechas existe un cúmulo de ángulos diedro bien definido.

Cada uno de los histogramas anteriores fue sometido al proceso de depuración descrito en la sección de métodos, y su resultado fue empleado como el espacio conformacional restringido para la ejecución del segundo cedazo. El número promedio de evaluaciones de la función de aptitud durante el primer cedazo fue de 20,000, las cuales incluyen las evaluaciones requeridas por el operador de mejoramiento.

Segundo cedazo. En esta parte del experimento las poblaciones iniciales fueron generadas de manera azarosa dentro de los límites obtenidos para las memorias conformacionales del primer cedazo, y se realizaron 10 corridas independientes del AGH para cada uno de los péptidos prueba. Al final de cada experimento se seleccionaron aquellos individuos con energía menor o igual a 0.0 kcal / mole. Con estos datos estructurales se construyeron los histogramas correspondientes, y de éstos se asignaron las memorias conformacionales para cada uno de los genes. Ejemplos de estas se muestran en la segunda y cuarta columnas de la Fig. 4 para la Met-encefalina y en la Fig. 5 para la Leu-encefalina.

De los 200 individuos de la población final para la Met-encefalina 25 % de éstos tuvieron una energía por debajo de 0.0 kcal / mole, entre los cuales alrededor de doce individuos presentaron la estructura del MGE conocido para estas molécula, véase Fig. 6. Cada vez que repetimos este proceso siempre encontramos una conformación única en el mínimo de energía más bajo. Esta conformación corresponde a la estructura sugerida [30] para el MGE de este péptido bajo el campo de fuerza ECEPP / 2 con una energía de –12.938 kcal / mole. Otras conformaciones similares a aquella del MGE aparecen con un DEs entre 0.5 a 2.0 kcal / mole arriba del MGE. Todas estas estructuras exhiben una vuelta de horquilla β tipo II' centrada en la unión peptídica entre Gly_3 y Phe_4 (Fig. 7A). Conformaciones significativamente diferentes siempre se encuentran a un valor más alto de energía potencial. La segunda y cuarta columnas de la Fig. 4 claramente muestran que el mínimo de energía más baja que se obtuvo con nuestro AGH corresponde a las regiones marcadas con la flecha en negrita del primer cedazo. Dadas estas condiciones, para nuestros propósitos realizar un tercer cedazo no fue necesario. El número promedio de evaluaciones de la función objetivo fue de 10,000 las cuales también incluyen las evaluaciones requeridas por el operador de mejoramiento.

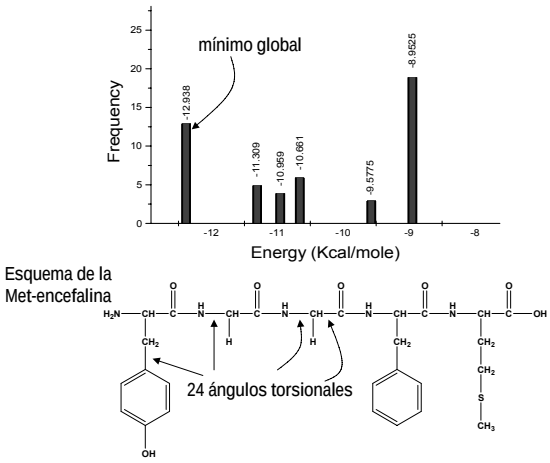

Fig. 6. Histograma de la distribución de energía en la población final de una corrida típica de nuestro AGH, donde aproximadamente 25 % de la población adopta la estructura del MGE para la Met-encefalina. En promedio cada corrida necesitó de 15,000 evaluaciones de la función objetivo. El esbozo de la molécula de Met-encefalina muestra los 24 ángulos torsionales que definen al cromosoma a ser mejorado.

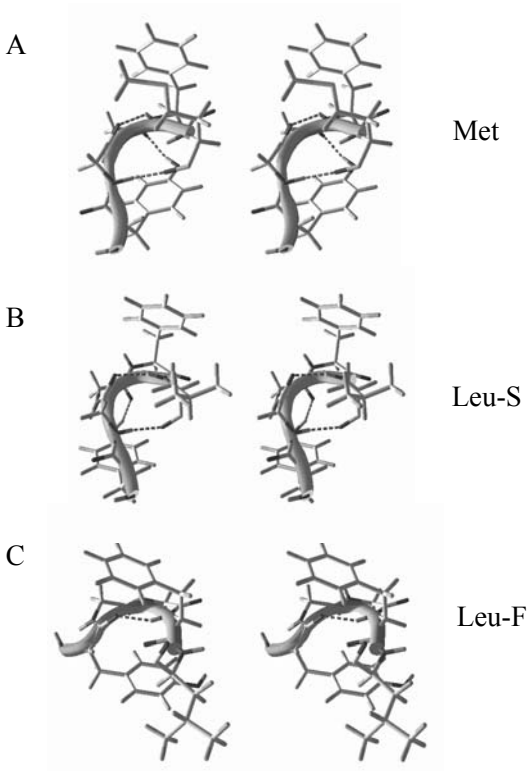

Fig. 7. A) Dibujo estereoscópico de la estructura correspondiente al MGE obtenida con nuestro AGH para la Met-encefalina. El valor de energía ECEPP/2 es –12.938 kcal / mole. B) Dibujo estereoscópico de la estructura correspondiente al segundo mínimo de energía más bajo obtenido con nuestro AGH para la Leu-encefalina. Esta estructura coincide con la reportada por Glasser y Scheraga [30] como el MGE. Su energía ECEPP / 2 es –9.8951 kcal / mole. C) Dibujo estereoscópico de la estructura correspondiente al mínimo de energía más bajo obtenido con nuestro AGH para la Leu-encefalina. Esta estructura coincide con la reportada por Androulakis *et al.* [31] como el MGE. Su energía ECEPP / 2 es de –10.03 kcal / mole. Todas las figuras fueron generadas con el Swiss PDB Viewer [36]. La vuelta de horquilla β tipo II' está representada por una varilla gris.

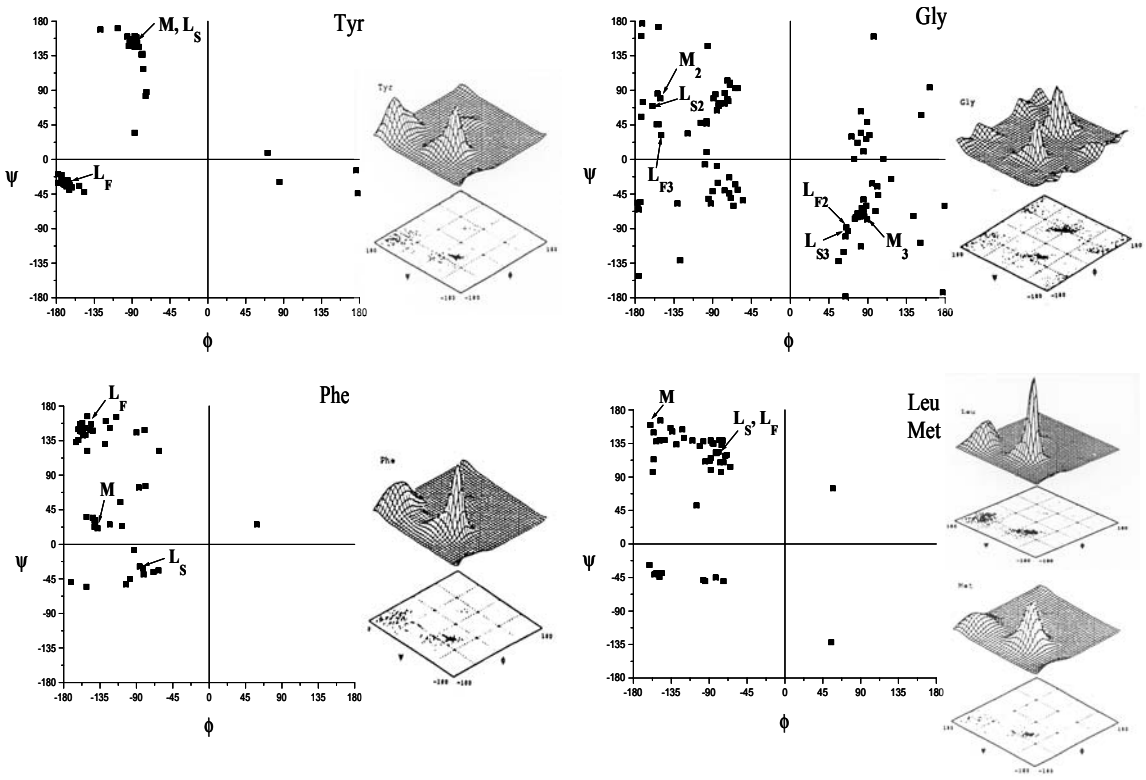

Fig. 8. Mapas de Ramachandran para los aminoácidos que componen a la Met y a la Leu-encefalina. Cada mapa a la izquierda contiene a los cincuenta mejores conformeros al final de nuestro AGH. Estos mapas se comparan con los mapas experimentales correspondientes. Las flechas señalan las coordenadas (ϕ, ψ) correspondientes a la estructura del MGE putativo para estos péptidos. Las etiquetas M, L_S y L_F indican el MGE de Met-encefalina, el MGE de Scheraga y el MGE de Floudas para Leu-encefalina, respectivamente. El mapa de Ramachandran para Gly contiene las (ϕ, ψ) coordenadas para los dos residuos de glicina (Gly_2, Gly_3) presentes en Met y Leu-encefalina, así los sufijos 2 y 3 hacen referencia a estos residuos. Los mapas a la derecha de cada gráfica son los mapas de Ramachandran experimentales correspondientes (tomados con permiso de la Ref. 34).

Para la Leu-encefalina entre el conjunto de estructuras de más baja energía que se obtuvieron al final de nuestro algoritmo, se encontraron dos conformeros estructuralmente diferentes, uno a –9.8951 y el otro a –10.03 kcal / mole (Figs. 7B y 7C). Estas estructuras difieren con un DE de 0.135 kcal / mole y con un RMS de 3.24 Å para el mejor ajuste entre ellas. Ambas estructuras han sido reportadas previamente por otros grupos de investigación como aquellas que corresponden al MGE. La estructura con –9.8951 kcal / mole fue reportada por Glasser y Scheraga [31] al usar el método de la ecuación de difusión. La otra estructura fue reportada por Androulakis *et al.* [32] y por Klepeis and Floudas [33] quienes usaron el método determinístico de Branch and Bound. En ambos trabajos se reportó el uso del campo de fuerza ECCEP/2 como función objetivo, pero no se reportó la presencia de otro conformero con energía similar pero estructuralmente diferente. La estructura reportada por Glasser y Scheraga posee una vuelta de horquilla β tipo II' centrada en la unión peptídica entre Gly_3 y Phe_4, la cual es similar a la estructura del MGE para la Met-encefalina (Fig. 7B). En contraste la estructura de Klepeis y Floudas la vuelta de horquilla β tipo II' está centrada en la unión peptídica entre Gly2 y Gly3 (Fig. 7C). La segunda y cuarta columnas de la Fig. 5 claramente muestran que los mínimos de energía más baja que se obtuvieron con nuestro AGH corresponden a las regiones marcadas con la flecha en negrita del primer cedazo. Dadas estas condiciones, realizar un tercer cedazo no fue necesario para localizar el MGE de estas moléculas.

Mapas de Ramachandran

La mayoría de los grados de libertad de las proteínas son muy rígidos, y la desviación de su valor de equilibrio está limitada a un intervalo muy corto. Un ejemplo es el ángulo ω, el cual siempre está cercano a los 180° y tiene una importancia relativa en la búsqueda conformacional. Dos excepciones son los ángulos de torsión ϕ y ψ, los que pueden variar con relativa facilidad. Estos ángulos proveen la flexibilidad necesaria en la molécula de proteína y le permiten plegarse sin mucho esfuerzo. Por lo tanto, estos ángulos son las variables más importantes en la conformación de una proteína. Cualquier conformación local puede ser vista como un punto en el plano (ϕ, ψ) o mapa de Ramachandran. Los mapas de Ramachandran han sido compilados para cada uno de los 20 aminoácidos (gráficas para 403 entradas no homólogas al PDB, donde 30 % de las conformaciones más lejanas han sido omitidas, pueden ser

consultadas en http://alpha2.bmc.uu.se/gerard/supmat/ramarev.html) y pueden ser usadas para valorar el desempeño de nuestro AGH. Para este fin, empleando los datos estructurales de 50 individuos con energía menor o igual a 0.0 kcal / mole se construyeron los mapas de Ramachandran correspondientes para cada uno de los residuos constitutivos de la Met- y la Leu-encefalina, y comparamos éstos con aquellos mapas reportados en la literatura.

En la Fig. 8 presentamos dos tipos de mapas de Ramachandran. Aquellos de la mano izquierda son los mapas de Ramachandran para cada uno de los cinco residuos de la Met- y la Leu-encefalina al final del segundo cedazo. Cada punto representa a uno de los 50 mejores conformeros y la flecha indica las coordenadas (ϕ, ψ) donde se encuentra la estructura del MGE para estos péptidos, y están marcados como M, LS y LF para la Met-encefalina, la Leu-encefalina de Glaser y Scheraga, y la Leu-enkephalin de Klepeis y Floudas. El mapa de Ramachandran para la Gly contiene de manera conjunta los resultados para los dos residuos de glicina presentes en ambos neuropéptidos, por lo que he hemos añadido los sufijos 2 y 3 para diferenciar las coordenadas (ϕ, ψ) correspondientes a las Gly_2 y Gly_3. Aquellas gráficas de la mano derecha son las correspondientes a los mapas de Ramachandran que resultan de un análisis de 67 entradas no homólogas al PDB reportadas por Kamimura and Takahashi [34].

Los mapas de Ramachandran muestran que los puntos obtenidos a partir de los resultados computacionales se encuentran distribuidos dentro de las regiones conformacionalmente permitidas para cada uno de los aminoácidos constitutivos de las moléculas prueba, a pesar de que nuestro AGH fue aplicado a una población relativamente pequeña de conformeros. Asímismo, se observa que las coordenadas (ϕ, ψ) del MGE para estas moléculas están dentro de cúmulos bien definidos que son favorecidos por el uso de las memorias conformacionales correspondientes.

Discusión

El objetivo de este estudio fue el de encarar la tarea de reducir la dificultad y el tiempo de cómputo para encontrar la estructura del MGE para moléculas de oligopéptidos con 24 o más grados de libertad torsional. Aunque nuestro laboratorio ha sido capaz de encontrar estas estructuras al emplear otros métodos heurísticos [4-7], nuestra principal preocupación ha sido reducir la enorme cantidad de evaluaciones de la función objetivo que se han necesitado para encontrarlas. En los primeros pasos del desarrollo de nuestro AGH pudimos observar que alrededor del 60 % del espacio conformacional no estaba poblado cuando la población final se aproximó a regiones de baja energía. Más aún, los ángulos diedros ω permanecieron exclusivamente en la conformación *trans* durante toda la búsqueda conformacional, por lo tanto decidimos dejar a este ángulo diedro con un valor constante de 180° durante todos los experimentos subsecuentes, excepto donde se aplicaba el operador de mejoramiento. A través de los histogramas de energía contra distribución angular encontramos que las poblaciónes finales forman cúmulos en regiones específicas del espacio conformacional, a las que hemos llamado memorias conformacionales. Estas memorias conformacionales fueron usadas posteriormente para reducir en varios órdenes de magnitud el volumen del espacio conformacional que debía ser muestreado en la segunda fase de este algoritmo. Aún así, el número de evaluaciones de la función objetivo resultaba ser poco práctico. Este número se redujo significativamente cuando incluimos el operador de mejoramiento en el desarrollo de nuestro AGH.

El hecho que dentro de la población final para la Leu-encefalina nuestro algoritmo encontró los dos conformeros reportados como el MGE, es una prueba que nuestro procedimiento es capaz de encontrar individuos con energía semejante pero estructuralmente diferentes entre ellos. Esta cualidad es de mucha importancia ya que es sabido que el fondo del pozo de energía de los potenciales empíricos para proteínas es rugoso, y que por lo tanto permite la coexistencia de muchos conformeros. La pregunta principal es cuáles de estos conformeros tienen parámetros de estructura secundaria similares, y cuál de ellos corresponde al MGE. Para ayudar a resolver esta encrucijada nuestro grupo ha reportado [35] una propuesta para detectar al verdadero MGE.

Los mapas de Ramachandran resultaron ser una herramienta excelente para demostrar que nuestro AGH puede muestrear de manera eficiente el espacio de ángulos diedros, ya que la población final presenta una correspondencia excelente entre los datos predichos con los datos experimentales, aún si el AG es corrido con una población de pocos individuos.

La Tabla 3 compara la demanda en recursos de cómputo para otros métodos que han sido aplicados al problema de plegar la Met-encefalina, esto es, que han considerado todos los 24 ángulos diedros de esta molécula como variables y que además han usado también la función objetivo ECEPP/2. Esta claro que nuestro algoritmo representa una ventaja significativa sobre los métodos más populares al reducir en aproximadamente un 70 % el número de evaluaciones de la función objetivo.

El AGH aquí reportado abre la posibilidad de predecir la estructura terciaria cercana al MGE para moléculas mucho más grandes que simples pentapéptidos. Estamos conscientes que una población de $N_{pob} \leq 30$ podría no contener suficientes individuos para asegurar que todas las posibles combinaciones entre las memorias conformacionales han sido muestreadas. El considerar moléculas con más de 24 ángulos diedros y probar el efecto que una población mayor a 30 individuos pueda tener sobre el desempeño de nuestro algoritmo son acciones para investigación futura.

Agradecimientos

Este trabajo fue parcialmente apoyado por los proyectos 25245-E del CONACyT a R.G.J., IN109999 e IN107701 por parte del PAPIIT-DGSCA-UNAM a R.G.J. y a L.B.M.

Referencias

1. Anfisen, C.B. *Science* **1973**, *181*, 223-230.
2. Hart, W. E.; Istrail, S. *J. Comput. Biol.*, **1997**, *4(1)*,1-22.
3. Garduño-Juárez, R.; Morales, L. B.; Pérez-Neri, F. *J. Mol. Struct. (THEOCHEM),* **1990**, *208*, 279-300.
4. Garduño-Juárez, R.; Pérez-Neri, F. *J. Biomol. Struct. Dyn.* **1991**, *8(4)*, 737-758.
5. Morales, L. B.; Garduño-Juárez, R.; Romero, D. *J. Biomol. Struct. Dyn.*, **1991**, *8*, 721-735.
6. Morales, L. B.; Garduño-Juárez, R.; Romero, D. *J. Biomol. Struct. Dyn.*, **1992**, *9*, 951-957.
7. Morales, L. B.; Garduño-Juárez, R.; Aguilar-Alvarado, J. M.; Riveros-Castro, F. J. *J. Comp. Chem.*, **2000**, *21*, 147-156.
8. Judson, R. S.; Jaeger, E. P.; Treasurywala, A. M.; Peterson, M. L. *J. Comp. Chem.*, **1993**, *14*, 1407-1414.
9. Brodmeier, T.; Pretsch, E. *J. Comp. Chem.*, **1994**, *15(6)*, 588-595.
10. Lucasius, C. B.; Blomerss, M. J. J.; Werten, S.; van Aeert, A. H. J. M.; Kateman, G. In *Proceedings of the First Workshop on Parallel Problem Solving from Nature*, (H. P. Schwefel & R. Männer, Editors), Soringer, Berlin, **1991**, pp. 90-97.
11. Forrest, S. *Science*, **1993**, *261*, 872-878.
12. Schulze-Kremer, S. In *Proccedings of the Second Workshop on Parallel Problem Solving from Nature*, (B. Männer & B. Manderick, Editors), Elsevier, Amsterdam, **1992**, pp. 391-400.
13. Jin, A. Y.; Leung, F. Y.; Weaver, D. F. *J. Comp. Chem.*, **1997**, *18*, 1971-1984.
14. Jin, A. Y.; Leung, F. Y.; Weaver, D. F. *J. Comp. Chem.*, **1999**, *20*, 1329-1342.
15. Le Grand, S. M.; Merz, K. M. *J. Global Opt.* **1993**, *3*, 49-66.
16. Reinhold, E. M.; Niebergelt, J.; Deo, N. *Combinatorial Algorithms: Theory and Practice*, Prentice Hall, New Jersey, **1977**.
17. Carlacci, L. *J. Comp. Aid. Mol. Des.*, **1998**, *12*,195-213.
18. Nishimura, L.; Naito, A.; Tuzi, S.; Saito, H.; Hashimoto, C.; Aida, M. *J. Phys. Chem. B*, **1998**, *102*, 7476-7483.
19. Holland, J. H. *Adaptation in Natural and Artificial Systems*, Ann-Arbor, University of Michigan Press, **1975**.
20. Unger, R.; Moult, J. *J. Mol. Biol.*, **1993**, *231*, 75-81.
21. Bäck, T. "*Self Adaptation in Genetic Algorithms"*, In *Proceedings of the First European Conference on Artificial Life*, Paris, France, MIT Press, Cambridge, MA, **1991.**
22. Goldberg, D. E. *Genetic Algorithms in Search, Optimization and Machine Learning*, Addison-Wesley Pub. Co., Inc., **1989.**
23. Némethy, G.; Pottle, M. S.; Scheraga, H. A. *J. Phys. Chem.*, **1983**, *87*, 1883-1887.
24. Schaumann, T.; Braun, W.; Wüthrich, K. Biopolymers, **1990**, *29*, 679-694.
25. Soman, K.; Franczkiewicz, R.; Mumenthaler, C.; von Freyberg, B.; Schaumann, T.; Braun, W. (1998), FANTOM v4.2, University of Texas Medical Branch, Galveston, Texas, U.S.A. and Institut für Molekularbiologie und Biophysik, Eidgenossische Technische Hochschule, Zürich, Switzerland.
26. Haupt, R.L.; Haupt, S.E. *Practical Genetic Algorithms*, Wiley Interscience, **1998.**
27. Michalewicz, Z. *Genetic Algorithms + Data Structures = Evolution Programs*. Artificial Intelligence, Springer Verlag, New York, **1992.**.
28. Guarnieri, F.; Wilson, S. R. *J. Comp. Chem.*, **1995**, *16*, 648-643.
29. Guarnieri, F.; Weinstein, H. *J. Am. Chem. Soc.*, **1996**, *118*, 5580-5589.
30. Nayeem, A.; Vila, J.; Scheraga, H. A. *J. Comp. Chem.*, **1991**, *12*, 594-605.
31. Glasser, L.; Scheraga, H. A. *J. Mol. Biol.*, **1988**, *199*, 513-524.
32. Adroulakis, I. P.; Maranas, C. D.; Floudas, C. A. *J. Global Opt.*, **1997**, *11*, 1-34.
33. Klepeis, J. L.; Floudas, C.A. *J. Comp. Chem.*, **1999**, *20*, 636-654.
34. Kamimura, M.; Takahashi, Y. *CABIOS*, **1994**, *10*, 163-169.
35. Garduño-Juárez, R.; Morales, L. B.; Flores-Pérez, P. *J. Mol. Struct. (Theochem),* **2001**, *543*, 277-284.
36. Guex, N.; Peitsch, M.C. *Electrophoresis*, **1997**, *18*, 2714-2723
37. (a) Li, Z.; Scheraga, H. A. *Proc. Natl. Acad. Sci. USA*, **1987**, *84*, 6611-6615; (b) Li, Z.; Scheraga, H. A. *J. Mol. Struct. (Theochem)*, **1988**, *179*, 333-338.
38. Hansmann, U. H.; Okamoto, Y. *J. Comp. Chem.*, **1993**, *14*, 1333-1340.
39. Lee, J.; Scheraga, H. A.; Rackovsky, S. *J. Comp. Chem.*, **1997**, *18*, 1222-1232.
40. (a) Kostrowicki, J.; Piela, L.;, Cherayil, B. J.; Scheraga, H. A. *J. Phys. Chem*., **1991**, *95*, 4113-4119; (b) Kostrowicki, J.; Scheraga, H. A. *J. Phys. Chem.*, **1992**, *96*, 7442-7447.
41. Olszewski, K. A.; Piela, L.; Scheraga, H. A. *J. Phys. Chem.*, **1992**, *96*, 4672-4680.