matemática: Una publicación del ICM – ESPOL
2012, Vol. 10, No. 1

APLICACIÓN DE ALGORITMOS EVOLUTIVOS A LA BÚSQUEDA DE MOTIVOS BIOLÓGICOS EN REGIONES PROMOTORAS DEL GENOMA

C. I. Jordán [1], C. J. Jordán[2]

Resumen. *El control en la producción de proteínas en las células es un problema importante de la biología molecular, del cual dependen un sinnúmero de aplicaciones en los campos de la medicina, la agricultura y ganadería. Esta producción está regulada al interior de la célula por un interruptor biológico basado en la fijación de una proteína (el factor de transcripción) sobre sitios determinados al interior de los genes portadores de la información genética. Este sitio de fijación se conoce como el motivo de una proteína. Existen métodos en la biología para identificar estos sitios de fijación, pero son procedimientos muy costosos y toman mucho tiempo. Los métodos basados en la computación evolutiva han demostrado ser algoritmos más eficientes y eficaces a la hora de identificar las posiciones del motivo que todo otro método desarrollado hasta la fecha. En este trabajo se implementaron dos métodos propios de búsqueda de motivos, uno basado en los algoritmos genéticos (MBMAG) y otro en la estimación de distribuciones (MBMEDA), los cuales evalúan el contenido de información de los individuos de la población para discriminar las mejores soluciones en cada generación. Los resultados obtenidos sobre bases de ADN fueron evaluados utilizando métricas estándar para medir el desempeño de métodos computacionales de búsquedas. Estos resultados muestran que los métodos evolutivos son superiores respecto a otros métodos conocidos, en cuanto a encontrar un mayor número de secuencias correctas que constituyen el motivo.*

Palabras Clave: *Bioinformática, TFBS, Dogma Central Biología Molecular, Computación Evolutiva, Algoritmo Genético, Algoritmo por Estimación de Distribuciones.*

Abstract. *The control in the production of proteins in cells is an important problem in molecular biology, which depend on a number of applications in the fields of medicine ,agriculture and livestock. This production is regulated within the cell by a biological switch based on binding of a protein on binding sites within certain genes carrying genetic information. These binding sites are known as the motif of a protein. There are methods in biology to identify these binding sites, but are very expensive procedures and time consuming. The methods based on evolutionary computation algorithms have proven to be more efficient and effective in identifying the positions of the sequences that every other method developed so far. In this work we implemented two search methods , one based on genetic algorithms (MBMAG) and the other in the estimation of distributions (MBMEDA), which assess the information content of the individuals in the population to discriminate the best solutions in each generation. The results of these methods in the search for motifs on DNA bases were evaluated using metrics to measure the performance of different search methods. These results demonstrate that evolutionary methods are superior in precision and recall to other methods in the task of finding the correct sequences of a motif.*

Keywords: *Bioinformatics, TFBS, Central Dogma of Molecular Biology, Evolutionary Computation, Genetic Algorithm, Estimation of Distribution Algorithm.*

Recibido: Agosto 2012
Aceptado: Septiembre 2012

1. INTRODUCCIÓN

Todo organismo depende de un número muy grande de proteínas para cumplir sus funciones vitales; se clasifican en dos tipos: proteínas esenciales y no esenciales; a las primeras se las llama así porque el organismo no las produce y requiere ingerirlas mediante la alimentación; las proteínas no esenciales, en cambio, son aquellas que se producen al interior de las células mediante un proceso de biosíntesis conocido como el dogma central de la biología molecular [7].

No obstante su denominación, las proteínas no esenciales son vitales a los organismos; por ejemplo: el colágeno, la insulina, gran parte de las hormonas y un sinnúmero de enzimas, son proteínas no esenciales, de las cuales depende la vida de manera significativa.

El dogma central de la biología molecular explica como la información genética se transcribe y traduce en cadenas de aminoácidos que son las proteínas. Un componente importante de este mecanismo es la transcripción, proceso por el cual la información contenida en los genes, originalmente en forma de molécula de ADN, pasa a una moléculas de ARN mensajero; este proceso depende de una clase de interruptor biológico conocido como el TFBS, por sus siglas en inglés (*transcription factor binding site*), cuyo funcionamiento requiere que una proteína conocida como factor de transcripción (TF) se fije en un

[1] *Jordán Carlos I., Facultad de Ingeniería en Electricidad y Computación. Escuela Superior Politécnica del Litoral (ESPOL); (e_mail: cjjordan@espol.edu.ec).*
[2] *Jordán Carlos J., Facultad de Ingeniería en Electricidad y Computación. Escuela Superior Politécnica del Litoral (ESPOL); (e_mail: cjordan@espol.edu.ec).*

conjunto de patrones de nucleótidos (BS, por binding site) ubicados en la región promotora del gen, en su cabecera. Por lo tanto, estos sitios de fijación deben poder diferenciarse de otras secuencias de nucleótidos, y permitir la fijación del factor de transcripción asociado a la producción de una cierta proteína. Este sitio de fijación recibe también el nombre de motivo de dicha proteína. En el campo de la biología molecular, la identificación de motivos en un genoma constituye uno de los problemas más importantes en la actualidad, debido a los enormes beneficios potenciales que esto tendría; por ejemplo: curar enfermedades de manera más natural, estimulando la producción de ciertas proteínas que un organismo hubiera dejado de producir por motivos hasta ahora desconocidos, y cuyo déficit sea responsable de la patología. La identificación o reconocimiento de motivos biológicos es, sin duda, un verdadero reto; esto debido a que se desconoce a priori cual es el patrón de similitud subyacente a las secuencias de nucleótidos que lo constituyen. Tomando en cuenta que el alfabeto genético está compuesto únicamente por los símbolos A, C, G y T, no es difícil reconocer que hallar diferencias y similitudes entre varias secuencias de nucleótidos constituya un desafío. Por otro lado, se desconoce también cual es la ubicación exacta de los motivos en las regiones promotoras, pues no siempre se encuentran en las mismas posiciones: pudieran ocurrir al inicio, al final o en el centro de la zona de regulación o promotora. En las ciencias biológicas existen métodos confiables y precisos para identificar los TFBS, por ejemplo: el análisis ADN footprint [5] y la electroforesis en gel [6]; sin embargo, estos métodos requieren mucho tiempo y su implementación es muy costosa. Por esta razón, en la actualidad los métodos computacionales han surgido como una alternativa viable para la búsqueda de motivos; los métodos informáticos clásicos pueden clasificarse en dos grupos: aquellos con base en secuencias de nucleótidos y los que se basan en modelos probabilísticos [2]. Los métodos del primer grupo garantizan que encuentran el motivo óptimo; no obstante, sus tiempos de ejecución exponenciales determinan que sólo sean útiles para motivos de tamaño pequeño. Un ejemplo de este grupo es el algoritmo MITRA [16]. Por otro lado, los métodos que utilizan en la búsqueda modelos probabilísticos no siempre encuentran la solución óptima, pero en cambio son más eficientes en cuanto a tiempos de ejecución, y los resultados son generalmente aproximadamente correctos. Ejemplos de estos métodos son los algoritmos MEME [17] y Gibbs Sampler [18]. Entre los métodos computacionales de búsqueda de motivos, aquellos que aplican computación evolutiva han ganado recientemente importancia debido a sus buenos resultados. Los métodos con base en los algoritmos genéticos [1] [13] y por estimación de distribuciones [9] a pesar de no ser muy conocidos estos últimos- presentan los mejores resultados. El principal objetivo de este trabajo es presentar dos métodos de búsqueda de motivos con base en la computación evolutiva. Cada uno se ha implementado utilizando un motor de búsqueda diferente: algoritmos genéticos y algoritmos por estimación de distribuciones. Estos métodos se llamarán en adelante según sus siglas: MBMAG (método de búsqueda de motivos con base en algoritmos genéticos) y MBMEDA (método de búsqueda de motivos basado en estimación de distribuciones), respectivamente. Una vez implementados, se probaron utilizando primero bases de datos sintéticas y luego cierto número de bases reales de ADN correspondientes a diferentes organismos; cada una de estas bases reales consiste de un conjunto de secuencias promotoras del genoma de un organismo donde se sabe que existe por lo menos una instancia del motivo que fija cierto factor de transcripción común. La calidad de los resultados se midió mediante métricas que fueron tomadas del campo de la recuperación de información (IR, por Information Retrieval), a saber: *precisión* y *exhaustividad* [9]. Lo que sigue de este documento tiene la siguiente estructura: en la Sección 2 se hace una hace una breve introducción a la computación evolutiva y a los paradigmas utilizados en el desarrollo de los métodos antes indicados; luego, en la Sección 3, se explican los detalles de la aplicación de estos métodos a la solución del problema concreto de la búsqueda de motivos biológicos; en la sección 4 se indicará que datos fueron utilizados para probar los métodos y que métricas fueron usadas para medir su desempeño; en la sección 5 se presentan los resultados obtenidos y se comparan con los de otros métodos que aparecen en la literatura; en la sección 6 se dan algunas conclusiones; y, finalmente, en la sección 7, se indican maneras en que este trabajo podría extenderse en el futuro.

2. MÉTODOS EVOLUTIVOS

Los métodos evolutivos son métodos metaheurísticos [3] que resuelven problemas haciendo búsquedas globales en un espacio de soluciones potenciales. La búsqueda se hace en base a poblaciones, que son subconjuntos del universo de soluciones potenciales del problema. Los

métodos evolutivos evalúan grupos de soluciones de forma paralela, lo que reduce el tiempo de ejecución del método al descartar simultáneamente grupos de soluciones que no son idóneas al problema, reduciendo el espacio sobre el cual realizar la búsqueda. Los métodos evolutivos son excelentes para resolver problemas con espacios de soluciones de gran cardinalidad. Generalmente, los problemas de optimización presentan este tipo de características, por lo que los métodos evolutivos se utilizan ampliamente en problemas de optimización en una variedad de campos. Un algoritmo evolutivo tiene tres componentes principales:

1. Una población de individuos.
2. Una función para evaluar la calidad de los de los individuos como soluciones.
3. Operadores de variación sobre los individuos.

Cada individuo de una población se representa por una estructura de datos que almacena las características que hacen única a cada solución. Se dice que la representación de cada individuo constituye un "cromosoma", que a su vez agrupa sus "genes".

La función de evaluación o función de fitness es una función que asigna a cada individuo de la población un valor numérico para representar el grado de aptitud como solución correcta del problema. La función de fitness es el criterio principal que conduce el proceso evolutivo hasta que el método converge a una solución. La convergencia del algoritmo genético ocurre cuando el valor de fitness del mejor individuo no mejora después de un cierto número de generaciones. La elección de una función de fitness apropiada depende de las características del problema. La función de fitness podría ser igual a la función objetivo en un problema de optimización, o para problemas complejos que no cuentan con una ecuación definida, existen funciones alternativas que quedan a criterio del investigador para su uso. Los operadores de variación modifican el contenido genético de los individuos de una población generando nuevos individuos para constituir una nueva población. Los individuos sobre los que se ejecutan los operadores de variación se denominan padres, mientras que los individuos resultantes se conocen como hijos. Se espera que individuos de la nueva población tengan mejores valores de función de fitness que de la generación anterior. La mecánica tras un método evolutivo es bastante simple: 1) se genera de forma aleatoria una población inicial P, 2) se evalúan los individuos de la población, 3) se ejecutan los operadores de variación sobre los individuos padres seleccionados de la población y 4) se genera una nueva población en base a operaciones de selección entre los individuos padres y los hijos. Los pasos 2) al 4) se ejecutan de manera iterativa hasta que el método converge, o hasta que se satisface una condición de terminación conveniente. Los algoritmos genéticos son métodos adaptivos de búsqueda sobre el espacio de soluciones del problema. El proceso de adaptación consiste en utilizar operadores de exploración y explotación para modificar los individuos de una población de tal manera que el método converge a la mejor solución del problema. La operación de exploración busca nuevos individuos dentro del espacio de búsqueda; generalmente recibe el nombre de mutación y es un operador unario, es decir, que tiene un solo operando: un individuo de la población actual, al que modifica de manera aleatoria en uno o varios genes. De esta manera se obtiene un individuo nuevo que normalmente no se encuentra en la población actual. La operación de mutación brinda diversidad a la población, evitando la convergencia pronta del algoritmo genético a soluciones óptimas locales. La operación de explotación combina las características genéticas de dos individuos con el propósito de producir nuevos individuos con características mejores que las sus padres tenían. La operación de explotación se conoce como de cruce o recombinación; en general, este operador selecciona de forma aleatoria uno o varios genes de dos individuos padres e intercambia su contenido genético entre ellos, generando otros dos nuevos individuos. Los operadores de mutación y cruce se ejecutan sobre una población en función de tasas de mutación y cruce que toman valores entre en el intervalo [0,1]. Estas tasas determinan la frecuencia con que se ejecutan estos operadores de variación. Los procesos de evaluación y las operaciones de cruce y mutación se ejecutan iterativamente hasta que se cumpla un criterio de parada apropiado, que puede depender del número de generaciones o de la convergencia del algoritmo a una solución. En la siguiente figura se describe mediante un pseudocódigo el funcionamiento básico de un algoritmo genético:

Pseudocódigo 1.
Código del funcionamiento de un algoritmo genético

```
P <- P ∪ {individuo generado al azar}
Repetir
para Pi e P hacer
FuncionFitness(Pi)
para childsize hacer
Padres Pa, Pb <- Seleccion(P)
Hijos Ha, Hb <- Cruce(Pa, Pb)
```

Q <- Q U {Mutar(H_a), Mutar(H_b)}
P <- Q
Best <- Mejor (P)
Hasta (Criterio de Parada)
Retornar Best

El algoritmo por estimación de distribución es una variación de un algoritmo genético [3]. A diferencia de este, donde los individuos se generan aplicando los operadores de cruce y mutación, el algoritmo por estimación de distribuciones (EDA) estima un modelo probabilístico a partir de los mejores individuos de la población, y genera los nuevos individuos de la siguiente generación tomando una muestra del modelo. Los métodos basados en la estimación de una distribución (ED) permiten mediante el modelo probabilístico expresar de manera explícita la relación entre las variables del problema, lo que permite encontrar soluciones más eficientes al problema que la mayoría de métodos evolutivos [8]. El siguiente pseudocódigo describe el funcionamiento básico de un algoritmo por estimación de distribución:

Pseudocódigo 2.
Código del funcionamiento de un algoritmo de estimación de distribuciones

P <- P ∪{individuo generado al azar}
Repetir
para cada individuo Pi e P **hacer**
Fitness(P_i)
Q <- Seleccionar(P)
M <- Generar_ModeloP(Q)
para childsize **hacer**
H <- Muestrear_Ind(M)
Q <- Q U {H}
P <- Q
Best <-Mejor(P) Hasta (Criterio de parada)
Retornar Best

El algoritmo por estimación de distribuciones genera al azar la primera población de individuos. Luego se evalúa el valor de la función de fitness para cada uno de los individuos de la población, y se eligen los mejores según cierta función de selección. El modelo probabilístico M se estima entonces a partir de los individuos seleccionados de P. La operación de muestreo genera nuevos individuos a partir de M. La nueva población estará conformada por la unión de los individuos padres y los nuevos individuos obtenidos muestreando el modelo. El criterio de parada suele ser similar al presente en los algoritmos genéticos. Finalmente, el algoritmo retorna la mejor solución de la última generación.

3. APLICACIÓN DE LOS MÉTODOS EVOLUTIVOS EN LA BÚSQUEDA DE MOTIVOS

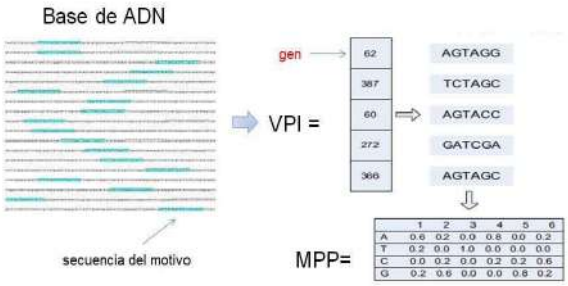
En este trabajo se implementaron dos métodos evolutivos de búsqueda de motivos: MBMAG y MBMEDA, basados respectivamente en los algoritmos genéticos y los algoritmos por estimación de distribución. Ambos métodos tienen elementos comunes, a saber:

1. La representación de los individuos
2. La función de fitness
3. El operador de selección de los mejores individuos

Como se mostró en la sección anterior, la diferencia esencial entre estos métodos estriba en la forma en la que generan los individuos de la siguiente generación. A continuación se muestra los componentes utilizados por ambos métodos evolutivos:

Representación de los individuos
Cada individuo de la población P se representa por un vector posiciones iniciales VPI. Cada celda en el VPI representa un gen, es decir, la posición donde inicia una secuencia del motivo en la fila correspondiente en la base de ADN. El valor que toma cada celda del vector se encuentra dentro del intervalo [0, n - w +1], donde w es el tamaño de las palabras del motivo y n el tamaño de la región promotora en la base de ADN. Esta representación se basa en el modelo de motivos 1-S que supone la presencia de una sola instancia del motivo por cada fila de la base de ADN. El vector de posiciones iniciales (VPI) determina a su vez un vector S cuyos elementos son las instancias del motivo correspondientes a dichas posiciones iniciales. A partir de S se calcula la matriz de pesos posicionales (MPP) $M_{b\ x\ l}$, donde l representa el tamaño de las instancias del motivo y b el número de símbolos del alfabeto genético que tiene cardinalidad 4, por lo tanto b = 4. Los elementos de la matriz representan la frecuencia con que cada símbolo de dicho alfabeto aparece en la posición correspondiente del motivo; de esta manera, S representa un modelo probabilístico de como se distribuyen los nucleótidos en el motivo. La matriz de pesos posicionales es una representación más completa del patrón detrás de las palabras del motivo candidato. La siguiente figura ilustra el proceso de creación de un nuevo individuo a partir de una base de ADN:

FIGURA 1
Aplicación de algoritmos evolutivos a la búsqueda de motivos biológicos en regiones promotoras del genoma
Construcción de un VPI a partir de la información seleccionada de la Base de ADN



3.1 FUNCIÓN DE FITNESS

La función de fitness utilizada para ambos métodos es la función de contenido de información [12]. Esta función mide el grado de similitud en la distribución de los nucleótidos del motivo con respecto a la distribución de los nucleótidos en la base de ADN. La utilidad de esta función como función de fitness estriba en el supuesto que los nucleótidos en las secuencias del motivo presentan una distribución diferente a la distribución de los nucleótidos en la Base. Por ello, mientras mayor sea la diferencia entre la distribución de los nucleótidos en el motivo con la base de ADN, mayor será la probabilidad que ese sea el motivo buscado. El contenido de información está definido como:

$$IC = \sum_{i=1}^{L} \sum_{b} \frac{f_b(i) \log(f_b(i))}{p_b} \qquad [1]$$

Donde f_b representa la frecuencia del símbolo b en la matriz de pesos posicionales MPP y p_b es la frecuencia del símbolo b en la base de ADN.

3.2 OPERADOR DE SELECCIÓN

El operador de selección utilizado para ambos métodos es la ***selección por torneo***, que consiste en realizar una competencia entre dos o más soluciones escogidas aleatoriamente; producto de esta competencia se escoge el individuo que tenga el mejor valor de fitness.

3.3 OPERADORES DE VARIACIÓN

El método con base en los algoritmos genéticos (MBMAG) utiliza como operadores de variación la mutación y el cruce en un punto, con tasas de mutación y cruce de 0.1 y 0.9, respectivamente. Asignar valores óptimos a estas tasas de mutación y de cruce consiste en sí mismo un problema de optimización; por esta razón los valores escogidos en este trabajo se fueron obtenidos en base a varios experimentos realizados, a partir de los cuales se eligieron las tasas donde los resultados fueron mejores.

El método basado en la estimación de distribuciones (MBMEDA) utiliza cuatro modelos probabilísticos basados en la distribución normal univariada para cada símbolo del alfabeto genético. Para estimar los modelos probabilísticos se utilizaron los estimadores estándares de la media y la varianza. Una vez estimados los parámetros de los modelos probabilísticos, el método MBMEDA genera una nueva población con el operador de muestreo sobre los cuatro modelos probabilísticos. Además de los operadores de variación clásicos, los métodos evolutivos que son objeto de este trabajo requieren de la aplicación posterior de otros operadores para mejorar los resultados obtenidos. Estos operadores son utilizados también por otros métodos descritos en la literatura, y reciben el nombre de operadores de desplazamiento [19] y filtrado local [19]. El operador desplazamiento modifica el contenido de los genes del mejor individuo de la población considerando la posibilidad de que el motivo buscado se encuentre desplazando unas cuantas posiciones del mejor individuo de la población. Este operador se aplica cada 10 generaciones, buscando siempre mejorar el valor de fitness del individuo al encontrar probablemente el individuo más cercano a la solución óptima. La operación de filtrado local modifica el vector de posiciones iniciales de un individuo en base al criterio de similitud entre las instancias del motivo definidas por VPI. Si una palabra es poco similar al resto, quiere decir que se necesita buscar una palabra con mayor similitud dentro de la fila correspondiente en la base de ADN. La operación de filtrado local se ejecuta sobre toda la población cada 10 generaciones.

Las tablas siguientes muestran valores de los parámetros utilizados en cada uno de los métodos de búsqueda implementados.

TABLA I

Aplicación de algoritmos evolutivos a la búsqueda de motivos biológicos en regiones promotoras del genoma

Parámetros del método MBMAG

Representación:		Vector Posición Inicial
Tamaño Población:		500
Número de Hijos:		250
Función Fitness:		Contenido de Información
Operador Selección:		Torneo
Operador Reproducción:	**Modelo P.**	Normal Univariado
	Muestreo	Func. Asociada al Modelo
Métodos Adicionales:		Transformación
		Filtrado Local
		Desplazamiento

TABLA II

Aplicación de algoritmos evolutivos a la búsqueda de motivos biológicos en regiones promotoras del genoma

Parámetros del método MBMEDA

Representación:		Vector Posición Inicial
Tamaño Población:		500
Número de Hijos:		250
Función Fitness:		Contenido de Información
Operador Selección:		Torneo
Operador Reproducción:	**Mutación**	Un punto(0.1)
	Cruce	1 Punto (0.9)
Métodos Adicionales:		Desplazamiento
		Filtrado Local

El pseudocódigo siguiente describe el funcionamiento del método con base en un algoritmo genético (MBMAG):

Pseudocódigo 3. Código del MBMGA

```
Repetir{
P<- {}
para popsize hacer
P <- P U {individuo generado al azar}
Best < valor inicial repeat
para cada individuo Pi e P hacer
Information_Content(Pi)
si(Best = valor incial O Fitness(Pi) > Fitness(Best))
Best <- Pi
para childsize hacer
Padre Pa <- Seleccionar_Tournament(P)
Padre Pb <- Seleccionar_Tournament(P)
Hijos Ha,Hb <- Cruce_1Punto(Pa,Pb)
Q <- Q U {Mutar_1Pto(Ha),Mutar_1Pto(Hb)}
para (%10)
Filtrado Local(Q)
Desplazamiento(Mejor(Q))
P <- Q
Hasta(condición de terminación) Retornar Best
}Hasta(const)
```

El siguiente pseudocódigo en cambio describe el funcionamiento del método con base en EDA (MBMEDA):

Pseudocódigo 4. Código del MBMEDA

```
Repetir{
P <- {} Pad <-{}
para popsize times hacer
    P <- P U {individuo generado al azar}
Filtrado_Local(P) Best
<- valor frontera repeat
      Para cada individuo Pi e P hacer
          Information_Content(Pi)
          si(Best = valor frontera o Fitness(Pi) >
          Fitness(Best)
          Best <- Pi
       Q <- {}
       Padres Pad <- Seleccionar_Tournament(P)
       M[4] <- Modelo_Normal_Univariado(Pad)
       para childsize hacer
          H <- Muestrear_MNU(M)
          I <- Mapeo(H)
          Q <- Q U {I}
       para (%10) Filtrado_Local(Q)
          Desplazamiento(Mejor(Q))
       P <- Q
  Hasta (condición de terminación)
  Retornar Best
}Hasta(const)
```

La variable const presente en ambos métodos representa el número de ejecuciones del algoritmo evolutivo en la búsqueda del motivo. En cada iteración se elige el mejor individuo de toda la población, representado por la variable Best, y una vez terminada la ejecución del algoritmo evolutivo, se elige al mejor de cada iteración como la solución final al problema. Este es un procedimiento estándar en varios métodos evolutivos de búsqueda de motivos [1][9][15], que tiene el fin de mejorar las posibilidades de encontrar un buen resultado.

4. DATOS Y MÉTRICAS

Los métodos evolutivos desarrollados en este trabajo se probaron utilizando dos tipos de bases de ADN: sintéticas y reales. Las bases sintéticas se utilizaron como casos de prueba elementales para medir el rendimiento de los métodos aun antes de aplicarlos sobre las bases de ADN reales. Las bases sintéticas fueron construidas según el método propuesto en [15], y consisten de un conjunto de t líneas de n nucleótidos generadas aleatoriamente;

luego, en cada línea se sobrescribe una secuencia de *l* nucleótidos generada al azar, que empieza en una posición aleatoria; finalmente, los nucleótidos de cada instancia del patrón sobrescrito sufren una mutación. El grado al que se modifican las instancias del patrón así "sembrado" depende de factores tales como: el tamaño del motivo, la conservación de los nucleótidos y la presencia de más de una instancia por fila en la base de ADN. Los métodos que aquí se presentan fueron probados con doce bases sintéticas generadas con este procedimiento.

Las bases de ADN reales son un conjunto de regiones promotoras de uno o varios genes regulados por el mismo factor de transcripción, para los que se ha determinado experimentalmente la ubicación de las instancias del motivo. La denominación de cada una de estas bases corresponde al factor de transcripción que se fija en los TFBS de estas regiones promotoras. En la siguiente tabla se listan las bases de ADN reales que fueron utilizadas en este trabajo, donde w representa el tamaño del motivo y N_t el número total de instancias que se encuentran en la base de ADN.

TABLA III

Aplicación de algoritmos evolutivos a la búsqueda de motivos biológicos en regiones promotoras del genoma

Bases reales de ADN

Base	Número de secuencias (T)	Tamaño de cada Secuencia(bp)	*w*	N_t
CRP	18	105	22	23
ERE	25	200	13	25
E2F	25	200	11	27
MYOD	17	200	6	17
ME2F	17	199	7	21

Con el propósito de medir el rendimiento de los métodos evolutivos se utilizaron dos métricas tomadas del campo de la recuperación de información (IR) y conocidas como: precisión y exhaustividad. La precisión mide la exactitud del método para encontrar las palabras correctas del motivo; por otro lado, la exhaustividad mide la capacidad del método para encontrar el mayor número posible de instancias correctas del motivo. Las métricas de precisión y exhaustividad contestan las siguientes preguntas a partir de la búsqueda en una base de ADN de los motivos: ¿están todos los que son?, para el caso de la precisión, y ¿son todos los que están? Para el caso de la exhaustividad. Estas métricas han sido utilizadas extensamente por otros investigadores [11], lo que facilitará comparar sus resultados con los aquí obtenidos.

Dichas métricas se definen de la manera siguiente:

$$\text{precisión} = N_c/N_p \qquad [2]$$

$$\text{exhaustividad} = N_c/N_t \qquad [3]$$

Donde N_c representa el número de motivos correctos encontrados por el método y N_p representa el número supuesto de palabras del motivo presentes en la base de ADN.

5. RESULTADOS

5.1 RESULTADOS DEL MÉTODO CON BASE EN ALGORITMOS GENÉTICOS (MBMAG)

Para probar el desempeño del MBMAG sobre las bases sintéticas se generaron 12 bases aplicando el método propuesto por [15]. Los resultados obtenidos para las métricas precisión y exhaustividad se encuentran en la siguiente tabla:

TABLA IV

Aplicación de algoritmos evolutivos a la búsqueda de motivos biológicos en regiones promotoras del genoma

Resultados del método MBMAG sobre las bases sintéticas

Número de Secuencias	Tamaño motivo	Conservación	Ruido			
			Sin Ruido		Con Ruido	
			Precisión	Exh.	Precisión	Exh.
100	16	1	1	1	0.99	0.97
20	16	1	0.99	0.99	0.98	0.97
100	8	1	0.99	0.99	0.97	0.93
20	8	1	0.98	0.98	0.91	0.89
100	16	0	0.95	0.95	0.88	0.80
20	16	0	0.94	0.94	0.89	0.85

En el caso de las bases sintéticas, el parámetro conservación es una medida de la resistencia a la mutación en las palabras del motivo en las diferentes filas de base. Una conservación de 1 representa que los nucleótidos en las palabras del motivo se conservan con una tasa del 90%, mientras que una conservación de 0 significa una tasa de menos del 50% de cohesión en las palabras del motivo.

Los resultados obtenidos demuestran que el método MBMAG encuentra el mayor número correcto de instancias del motivo. En base a estos resultados, se aplica la búsqueda de motivos sobre 5 bases reales de regiones promotoras de ADN. Los resultados se muestran a continuación:

TABLA V

Aplicación de algoritmos evolutivos a la búsqueda de motivos biológicos en regiones promotoras del genoma

Resultados del MBMAG sobre las bases reales

Base	Precision	Exhaustividad
CRP	0.88	0.69
ERE	0.76	0.76
E2F	0.76	0.70
MYOD	0.94	0.76
ME2F	0.94	0.94

TABLA VI

Aplicación de algoritmos evolutivos a la búsqueda de motivos biológicos en regiones promotoras del genoma

Comparación del desempeño entre el método **MBMAG con GAME Y GALF**

				MBMAG		GAME		GALF	
Base	**l**	**T**	**N_t**	**Pr.**	**Ex.**	**Pr.**	**Ex.**	**Pr.**	**Ex.**
CRP	22	18	23	0.88	0.69	0.94	0.70	**0.94**	**0.74**
ERE	9	25	25	0.76	0.76	0.73	0.76	**0.84**	**0.84**
E2F	11	25	27	0.76	0.70	**0.96**	**0.85**	0.80	0.74
MYOD	6	17	21	**0.94**	**0.76**	0.48	0.48	0.88	0.71
ME2F	9	17	17	0.94	0.94	0.88	0.88	**1.00**	**1.00**

En la Tabla VI se puede observar que de las cinco bases reales de ADN, en cuatro los métodos GAME y GALF son más precisos y exhaustivos que MBMAG, lo que se debe principalmente a que ambos métodos realizan una vez terminado el proceso evolutivo un procesamiento posterior sobre las instancias del motivo obtenido, lo que les permite reconocer un mayor número de palabras del patrón buscado. A pesar de que MBMAG no cuenta con este post-procesamiento, en promedio la diferencia entre sus valores de precisión y exhaustividad con los de los otros métodos es 0.05; lo que significa que los resultados obtenidos por el método MBMAG son suficientemente buenos para ser tomados en consideración.

5.2 RESULTADOS DEL MÉTODO BASADO EN LA ESTIMACIÓN DE DISTRIBUCIONES (MBMEDA)

Luego se procedió a aplicar el método MBMEDA a las bases de datos sintéticas como reales de ADN, midiendo en cada caso los valores de precisión y exhaustividad; los resultados obtenidos para las primeras se tabulan en la Tabla 7 siguiente:

TABLA VII

Aplicación de algoritmos evolutivos a la búsqueda de motivos biológicos en regiones promotoras del genoma

Resultados del método MBMEDA sobre las bases sintéticas

Número de Secuencias	Tamaño motivo	Conservación	Ruido			
			Sin Ruido		Con Ruido	
			Pr. (%)	Ex. (%)	Pr. (%)	Ex. (%)
100	16	1	1	1	0.99	0.97
20	16	1	0.99	0.99	0.98	0.95
100	8	1	1	1	0.99	0.93
20	8	1	0.99	0.99	0.98	0.92
100	16	0	0.97	0.97	0.84	0.79
20	16	0	0.95	0.95	0.93	0.86

En la Tabla VII se observa que los resultados obtenidos con este método sobre las bases sintéticas son superiores a los del método MBMAG, lo cual constituyó un estimulo para aplicar dicho método a las bases de ADN reales. De hecho, en la siguiente tabla se muestran los resultados obtenidos con las cinco bases de ADN anteriores:

TABLA VIII

Aplicación de algoritmos evolutivos a la búsqueda de motivos biológicos en regiones promotoras del genoma

Resultados del MBMEDA sobre las bases reales

Base	Precisión	Exhaustividad
CRP	0.83	0.65
ERE	0.80	0.80
E2F	0.80	0.74
MYOD	1.00	0.80
ME2F	1.00	1.00

Al comparar los valores con los de la Tabla V se observa que el método MBMEDA obtiene mejores resultados que el MBMAG cuando se buscan motivos reales con longitudes menores a 10 bps; prueba de ello son los resultados obtenidos para las bases MYOD y ME2F, en donde la precisión del método MBMEDA es 1 y su exhaustividad mayor a

0.8. Al comparar el desempeño del MBMEDA con el otro método de búsqueda de motivos con base en la estimación de distribuciones conocido como EDAMD [9] se obtienen los siguientes resultados:

TABLA IX

Aplicación de algoritmos evolutivos a la búsqueda de motivos biológicos en regiones promotoras del genoma

Comparación del desempeño del método MBMEDA con EDAMD

	MBMEDAM		EDAMD	
Base	**Pr.**	**Ex.**	**Pr.**	**Ex.**
CRP	0.83	0.65	**0.94**	**0.74**
ERE	**0.80**	**0.80**	0.76	0.76
E2F	**0.80**	0.74	0.71	**0.80**
MYOD	**1.00**	**0.80**	0.86	0.9
ME2F	1.00	1.00	1.00	1.00

En la Tabla IX se observa que el método MBMEDA presenta mayor precisión en los resultados que EDAMD; sin embargo, este último método presenta mejor exhaustividad, es decir, encuentra un mayor número de patrones correctos que el método implementado en este trabajo. Esto se debe a que el método EDAMD utiliza un modelo multivariado para estimar la distribución probabilística de los individuos de la población, lo que permite tomar en cuenta las interrelaciones entre las variables del problema.

5.3 COMPARACIÓN DEL DESEMPEÑO ENTRE LOS MÉTODOS EVOLUTIVOS Y OTROS MÉTODOS DE BÚSQUEDA

La Tabla X permite comparar los resultados obtenidos al aplicar los dos métodos de computación evolutiva desarrollados en este trabajo a las bases de ADN reales:

TABLA X

Aplicación de algoritmos evolutivos a la búsqueda de motivos ***biológicos en regiones promotoras del genoma***

Comparación entre MBMEDA y MBMA

	MBMEDA		MBMAG	
Base	**Pr.**	**Ex.**	**Pr.**	**Ex.**
CRP	0.83	0.65	**0.88**	**0.69**
E2F	**0.80**	**0.74**	0.76	0.70
ERE	**0.8**	**0.8**	0.76	0.76
ME2F	**1.00**	**1.00**	0.94	0.94
MYOD	**1.00**	**0.80**	0.94	0.76

Como se puede observar en la tabla anterior, en cuatro de las cinco bases reales de ADN, el método MBMEDA presenta mejores valores para la precisión y la exhaustividad en la búsqueda de motivos que el método MBMAG; esto se debe a que el desempeño de este último depende de que se utilicen valores apropiados para las tasas de mutación y de cruce; dependencia que no presenta el MBMEDA, que trabaja sobre un modelo más natural a la distribución de las variables del problema, lo que permite que las soluciones obtenidas por este método sean más próximas a las correctas que el MBMAG.

Finalmente, la Tabla XI permite comparar el desempeño de los métodos aquí desarrollados con los de dos métodos estadísticos que resuelven el mismo problema: MEME [17] y BioProspector; los métodos estadísticos están entre aquellos que mejores resultados obtienen en la búsqueda de motivos.

TABLA XI

Aplicación de algoritmos evolutivos a la búsqueda de motivos biológicos en regiones promotoras del genoma

Comparación del desempeño de los métodos desarrollados con los algoritmos MEME y Bioprospector

	MBMEDA		**MBMAG**		**MEME**		**BioProspector**	
Base	**Pr.**	**Ex.**	**Pr.**	**Ex.**	**Pr.**	**Ex.**	**Pr.**	**Ex.**
CRP	0.83	0.65	0.88	**0.69**	0.92	0.52	**1.00**	0.35
E2F	**0.80**	**0.74**	0.76	0.70	0.80	0.70	0.52	0.41
ERE	0.80	**0.80**	0.76	0.76	**0.88**	0.60	0.46	0.56
ME2F	**1.00**	**1.00**	0.94	0.94	0.93	0.82	0.71	0.71
MYOD	**1**	**0.80**	0.94	0.76	0.00	0.00	0.00	0.00

Con las bases CRP y ERE, tanto MEME como Bioprospector obtienen valores para Np y NT menores que los de métodos evolutivos. Esta es la razón por la que presentan mejores resultados en precisión y exhaustividad que los métodos evolutivos desarrollados. Para las demás bases reales, donde ambas variables tienen los mismos valores, los métodos evolutivos son más precisos y exhaustivos que los métodos estadísticos.

Por último, se muestra en las siguientes figuras representaciones gráficas de los motivos mediante los logos de secuencias [13]; en la Figura 2 se observa el logo para el motivo que corresponde al factor de transcripción CRP obtenido de manera experimental, mientras que en las figuras 3 y 4 se muestran los logos para los logos encontrados por los métodos aquí desarrollados: MBMAG y MBMEDA, respectivamente.

FIGURA 2

Aplicación de algoritmos evolutivos a la búsqueda de motivos biológicos en regiones promotoras del genoma

Logo de las secuencias reales donde se fija el factor de transcripción CRP

FIGURA 3

Aplicación de algoritmos evolutivos a la búsqueda de motivos biológicos en regiones promotoras del genoma

Logo de las secuencias encontradas por el método basado en AG

FIGURA 4

Aplicación de algoritmos evolutivos a la búsqueda de motivos biológicos en regiones promotoras del genoma

Logo de secuencias encontradas por el método basado en ED

Una comparación visual de estos logos permite concluir que el logo del método basado en AG es más similar al logo del método basado en ED para el caso de la base CRP. Este criterio concuerda con los valores de precisión y exhaustividad de ambos algoritmos para esta base real de ADN.

6. CONCLUSIONES

En este trabajo se han presentado dos métodos evolutivos para resolver el problema de encontrar motivos biológicos en una base de ADN. Se ha denominado a tales métodos MBMAG y MBMEDA, según el algoritmo que utilizan como motor de búsqueda: algoritmos genéticos y algoritmo por estimación de distribuciones, respectivamente.

Según las métricas utilizadas: precisión y exhaustividad, estos métodos son en general tan precisos y completos como otros métodos computacionales que resuelven el mismo problema, y en algunos casos mejores. Entre los dos, el que obtienen mejores resultados es aquel que utiliza estimación de distribuciones para generar la próxima población de soluciones potenciales, especialmente cuando la longitud del motivo buscado es pequeña, que es cuando son mas difíciles de reconocer. A fin de evitar la pronta convergencia de estos métodos evolutivos a mínimos locales, fue necesario introducir al proceso de búsqueda corrección y procesamiento posterior mediante los operadores desplazamiento y filtrado local, en adición a los ya clásicos de variación.

Los métodos aquí presentados utilizaron un modelo 1-S que limita a 1 el número de instancias del motivo por secuencia promotora, restricción que no está presente en las bases de ADN reales. No obstante, los resultados obtenidos fueron buenos si se los compara con los de otros métodos que no aplican tal restricción, pues en promedio la diferencia entre métricas correspondientes es inferior al 5%. Esta pequeña perdida en la calidad de los resultados se ve recompensada con el aumento en eficiencia en cuanto a tiempos de ejecución, debido a la mayor rapidez de los métodos aquí propuestos.

Finalmente, no existe una relación directa entre la calidad de los resultados obtenidos al probar un método con bases sintéticas o artificiales con aquella que se obtiene al utilizar bases reales de ADN; esto, debido a que el procedimiento aquí utilizado [15] para generar tales bases es completamente aleatorio, en tanto que se desconoce el método utilizado por la naturaleza en el diseño de motivos biológicos. Una muestra de ello ocurrió con motivos de gran tamaño; a pesar de que el método MBMEDA obtuvo mejores resultados que el MBMAG sobre bases de datos sintéticas, sobre la base CRP que tiene un motivo de 23 bps -considerado de gran tamaño- se obtuvieron mejores resultados con el método basado en los algoritmos genéticos.

7. TRABAJOS FUTUROS

El diseño de algoritmos de búsqueda de motivos biológicos con base en la computación evolutiva abre muchos frentes de investigación y desarrollo. Para el caso de los algoritmos aquí presentados debe estudiarse la posibilidad de trabajar con modelo M-S, que incluyen la posibilidad de que ocurra mas de 1 motivo por región promotora, con lo que podría mejorar significativamente los resultados para la métrica exhaustividad.

Este trabajo podría extenderse en el caso del método MBMEDA utilizando otros modelos probabilísticos para generar próximas poblaciones; por ejemplo, que tomen en cuenta de manera explícita la dependencia entre variables del

problema, que aquí fueron consideradas como independientes.

Finalmente, también existe la posibilidad de mejorar el desempeño de los métodos aquí propuestos, incorporando a la búsqueda nuevos operadores de variación y procesamiento posterior, para evitar la tendencia casi natural de estos procedimientos a converger a mínimos locales, lo que da lugar a encontrar muchos falsos positivos.

REFERENCIAS BIBLIOGRÁFICAS Y ELECTRÓNICAS

[1]. **CHAN, TALK MING, KWONG SAK LEUNG, AND KIN HONG LEE.** (2008). "*TFBS Identification Basen on Genetic Algorithm with Combined Representations and Adaptive Post-processing*". *Bioinformatics* 24.3 (2008): 341-49.

[2]. **DAS MK, DAI HK.** (2007). "*A survey of DNA motif finding algorithms*". *BMC Bioinformatics*. Nov 1;8 Suppl. 7:S21.

[3]. **EIBEN, AGOSTON E., AND J. E. SMITH.** (2003). "*¿What is an Evolutionary Algorithm?*" Introduction to Evolutionary Computing. New York: Springer, 2003. 15-35

[4]. **EIBEN, AGOSTON E., AND J. E. SMITH.** (2003). "*Genetic Algorithms*". Introduction to Evolutionary Computing. New York: Springer, 2003. 37-69

[5]. **GALAS, DAVID J., AND ALBERT SCHMITZ.** (1978). "*DNAase Footprinting a Simple Method for the Detection of Protein-DNA Binding Specificity*". Nucleic Acids Research 5.9 (1978): 3157-170.

[6]. **GARNER, MARK M., AND ARNOLD REVZIN.** (1981). "*A Gel Electrophoresis Method for Quantifying the Binding of Proteins to Specific DNA Regions: Application to Components of the Escherichia Coli Lactose Operon Regulatory System*" Nucleic Acids Research 9.13: 3047-060.

[7]. **JONES, NEIL C., AND PAVEL PEVZNER.** (2004). "*Molecular Biology Primer*" An Introduction to Bioinformatics Algorithms. Cambridge, MA: MIT, 2004. 57-68.

[8]. **LARRAÑAGA, PEDRO, AND JOSÉ A. LOZANO.** (2002). "*Estimation of Distribution Algorithms: a New Tool for Evolutionary Computation*". Boston: Kluwer Academic.

[9]. **LI, GANG, TAK MING CHAN, KWONG SAK LEUNG, AND KIN HONG.** (2008). "*An Estimation of Distribution Algorithm for Motif Discovery*". *Evolutionary Computation* (2008): 2411-418. .

[10]. **LUKE, SEAN.** (2009). "*Essencials of Metaheuristics*". 1st ed. Washington: Lulu.

[11]. **MANNING, CHRISTOPHER D., PRABHAKAR RAGHAVAN, AND HINRICH SCHUTZE.** (2008). "*Introduction to Information Retrieval*". New York: Cambridge UP. 151-158.

[12]. **SCHNEIDER, T., G. STORMO, L. GOLD, AND A. EHRENFEUCHT.** "Information Content of Binding Sites on Nucleotide Sequences" *Journal of Molecular Biology* 188.3 (1986): 415-31.

[13]. **SCHNEIDER, THOMAS D., AND R.MICHAEL STEPHENS.** (1990). "*Sequence Logos: a New Way to Display Consensus Sequences*". *Nucleic Acids Research* 18.20: 6097-100.

[14]. **USSERY, DAVID W., TRUDY M. WASSENAAR, AND STEFANO BORINI.** (2009). "*Sequences as Biological Information: Cells Obey the Laws of Chemistry and Physics*" *Computing for Comparative Microbial Genomics Bioinformatics for Microbiologists*. London: Springer. 3-17.

[15]. **WEI, Z.** (2006). "*GAME: Detecting Cis-regulatory Elements Using a Genetic Algorithm*". *Bioinformatics* 22.13: 1577-584.

[16]. **ESKIN, ELEAZAR, AND PAVEL A. PEVZNER.** (2002). "*Finding Composite Regulatory Patterns in DNA Sequences*". *Bioinformatics* 18 (2002): 354-63.

[17]. **BAILEY, TIMOTHY L., AND CHARLES ELKAN.** (1993). "*Unsupervised Learning of Multiple Motifs in Biopolymers Using Expectation Maximization*". La Jolla, CA: Dept. of Computer Science and Engineering, University of California, San Diego.

[18]. LAWRENCE, C., S. ALTSCHUL, M. BOGUSKI, J. (1993). "*Detecting Subtle Sequence Signals: a Gibbs Sampling Strategy for Multiple Alignment*". *Science*262.5131: 208-213

[19]. MANNING, CHRISTOPHER D., PRABHAKAR RAGHAVAN, AND HINRICH SCHUTZE. (2000). "*Introduction to Information Retrieval*" New York: Cambridge UP.