IMPORTANCIA DE LA MATEMÁTICA DISCRETA EN EL DESARROLLO DE LA BIOLOGÍA Y LA BIOINFORMÁTICA

Gregorio Martín Quetglás
Instituto de Robótica
Departamento de Informática
Universidad de Valencia

Bernardo Cuenca Grau
Instituto de Robótica
Departamento de Informática
Universidad de Valencia

1. INTRODUCCIÓN

La Matemática Discreta es la disciplina científica que incluye la Combinatoria, la Teoría de Grafos, la Lógica y la Teoría de Cuerpos Finitos.

La Matemática Discreta se ha convertido en una disciplina clave para la Bioinformática y la Biología Computacional por dos razones: la primera, porque éstas se basan en la Informática, y la Matemática Discreta está íntimamente ligada a todos los campos de las Ciencias de la Computación; la segunda, porque el lenguaje y los conceptos que se emplean en Biología permiten que muchos problemas puedan formalizarse utilizando la Matemática Discreta. Así, los conceptos de jerarquía, interacción, combinación o secuencia de bases nucleicas son directamente trasladables a la Combinatoria y a la Teoría de Grafos.

El desarrollo de la Bioinformática y, en general, el de la Biología moderna, irá unido al de la Matemática Discreta. La definición y desarrollo de nuevas estructuras que describan problemas biológicos y el diseño de algoritmos eficientes para resolución de determinados problemas serán instrumentos decisivos en la investigación biológica.

De este modo, la Matemática Discreta permitirá el desarrollo de la Biología y, a su vez, la Biología estimulará el desarrollo de la Matemática Discreta planteando nuevos problemas.

Sin embargo, ésta no es la única manera en que la Biología puede contribuir a la resolución de problemas en Matemática Discreta. El desarrollo de las técnicas de bio-computación basadas en DNA ha permitido resolver de forma eficiente varios problemas clásicos de la Teoría de Grafos, como son el problema de la búsqueda de un camino hamiltoniano en un grafo dirigido y el problema de la supercadena común más corta.

Finalmente, la Matemática Discreta se utilizará en el desarrollo de estructuras más eficientes (grafos conceptuales) para el almacenamiento de información biológica. Los actuales métodos de búsqueda y análisis de información almacenada en grandes bases de datos y en Internet son insuficientes para clasificar y dar sentido a la ingente cantidad de datos biológicos que se generan en los laboratorios.

2. APLICACIONES DE LA MATEMÁTICA DISCRETA A LA RESOLUCIÓN DE PROBLEMAS EN BIOLOGÍA Y BIOINFORMÁTICA

2.1 Ensamblaje de secuencias

El proceso de secuenciación del código genético de un organismo comienza “rompiendo” físicamente el ADN en millones de fragmentos aleatorios. La información contenida en cada uno de estos fragmentos se analiza experimentalmente. El problema surge a la hora de ensamblar de nuevo los fragmentos para formar una única secuencia completa. La técnica que se utiliza se basa en la hipótesis de que dos fragmentos son adyacentes si se solapan de alguna manera, es decir, si la parte final de una secuencia coincide con la inicial de la siguiente. Esta hipótesis no es del todo válida debido a la aparición de secuencias repetitivas y también debido a que la superposición parcial de dos fragmentos puede deberse simplemente al azar, y no a que dichos fragmentos sean realmente adyacentes en la secuencia original.

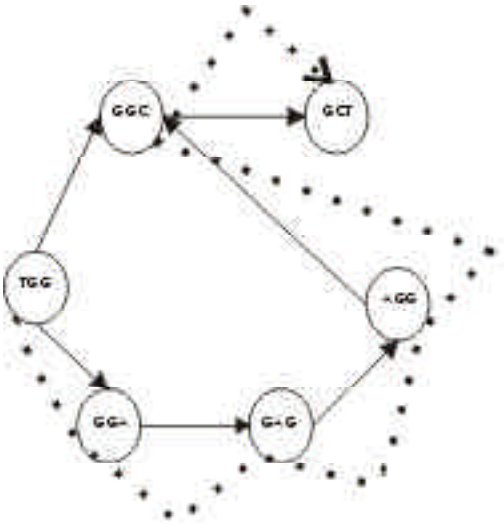
Además, la secuencia original no aparece cubierta en su totalidad por el conjunto de fragmentos y, por otra parte, se producen errores experimentales en la lectura de cada fragmento. Estos errores hacen que dar con la secuencia original correcta sea una tarea muy compleja.

Todas las aproximaciones a este problema se basan en reducirlo a problemas clásicos de Combinatoria y de Teoría de Grafos.

Existen varias soluciones al problema:

- *Algoritmos de "superposición-trazado-consenso"*

  Se basan en el concepto de grafo de superposición, o grafo de intersección. Cada secuencia obtenida experimentalmente de un fragmento es un vértice del grafo, de forma que dos vértices están conectados por una arista si las secuencias correspondientes se traslapan. El problema de reconstrucción de los fragmentos se reduce entonces a encontrar un camino en el grafo de superposición tal que se visite cada nodo solamente una vez [2], es decir, se trata de encontrar un camino hamiltoniano en el grafo (figura 1).



La secuencia TGGAGGCT, formada por ocho nucleótidos, se rompe físicamente en seis fragmentos, con tres bases nucleicas cada uno.

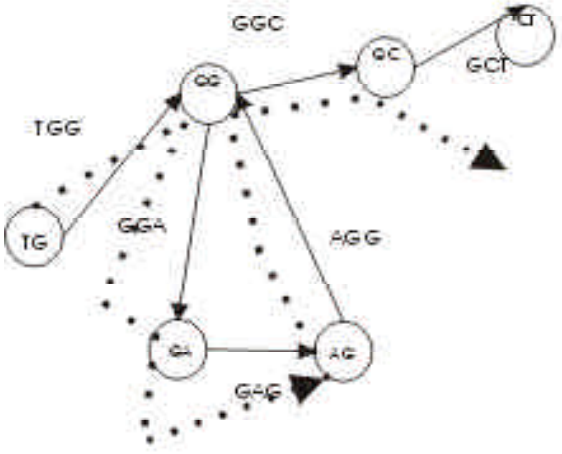
Con la información obtenida experimentalmente se construye un grafo dirigido, denominado grafo de intersección, en el que cada vértice representa un fragmento y cada arista indica si existe superposición entre los fragmentos que conecta. La solución del problema se reduce a encontrar un camino que pase por cada nodo una sola vez (camino hamiltoniano), está representado por la línea punteada en el dibujo.

Se conocen las condiciones necesarias y suficientes para que exista tal camino en el grafo considerado, pero no un algoritmo eficiente que lo halle.

Figura 1: Grafo de intersección

- *Algoritmo EULER.*

  Basado en el concepto de "grafo de De Brujin". Cada secuencia leída experimentalmente se corresponde con una arista en el grafo. El grafo de De Brujin es un grafo dirigido que se puede construir solamente a partir de las secuencias determinadas en los experimentos. El problema de encontrar la secuencia original se reduce al de hallar un camino euleriano en el grafo [4], es decir, un camino que pase por cada arista solamente una vez (figura 2).



En esta figura se considera el mismo ejemplo que el tratado en la figura 1. Se construye un grafo de De Brujin, en el que las aristas "A" son los fragmentos.

A = { TGG, GGA, GAG, AGG, GGC, GCT }

Los nodos están formados por todas las cadenas de dos bases nucleicas que se pueden construir con los fragmentos. Por ejemplo, los dos primeros fragmentos darían lugar a los siguientes nodos

Nodos : { TG, GG, GG, GA } De entre todos los nodos repetidos tomamos solo uno y nos queda : { TG, GG, GA }
La secuencia original se obtiene hallando un camino que atraviese cada arista una sola vez (camino euleriano).
Dado un grafo se conocen las condiciones necesarias y suficientes para que exista el camino, así como un algoritmo eficiente para calcularlo.

Figura 2. Grafo de De Brujin

2.2 Comparación de secuencias

Cuando se descubre una secuencia nueva de ADN y se dispone de una base de datos de secuencias ya estudiadas, se plantea el problema de hallar la secuencia almacenada en la base de datos que más se parezca a la nueva secuencia.

Para medir el grado de similitud entre dos secuencias es necesario alinearlas, añadiendo adecuadamente espacios en ambas para que coincidan el mayor número de caracteres en la misma columna. Después se suman el número de caracteres coincidentes y se penalizan los emparejamientos espacio-carácter.

El problema se formaliza utilizando grafos con dos ejes, que representan posiciones en las dos cadenas. Las aristas del grafo representan posibles columnas en el alineamiento. Se definen un punto fuente y un sumidero, de tal modo que todos los posibles alineamientos vienen dados por caminos que unen ambos puntos en el grafo [7].

Este grafo, denominado "de rejilla" (grid graph) es un grafo dirigido, acíclico y en el que hay definidos dos nodos especiales: la fuente y el sumidero. El mejor alineamiento posible de dos secuencias vendrá dado por el camino más largo posible entre el punto fuente y el sumidero. Una vez hallado este camino, se calcula su puntuación según el grado de similitud entre ambas secuencias y se repite el proceso para todas las secuencias almacenadas en la base de datos. Al final se obtienen las secuencias que más se parecen a la nueva secuencia (figura 3).

Existe una variación del problema, que consiste en hallar patrones localmente similares contenidos en dos secuencias, en lugar de analizar el parecido global entre ambas. En este caso, el problema consiste en hallar la subcadenas de las dos secuencias que más se parezcan. En el lenguaje de Teoría de Grafos, esto significa buscar el camino más largo entre dos puntos cualesquiera del grafo dirigido, y no buscar el camino más largo entre dos puntos previamente definidos.

La mejora del rendimiento en la comparación de cadenas depende del desarrollo de algoritmos eficientes para encontrar el camino más largo entre puntos del grafo.

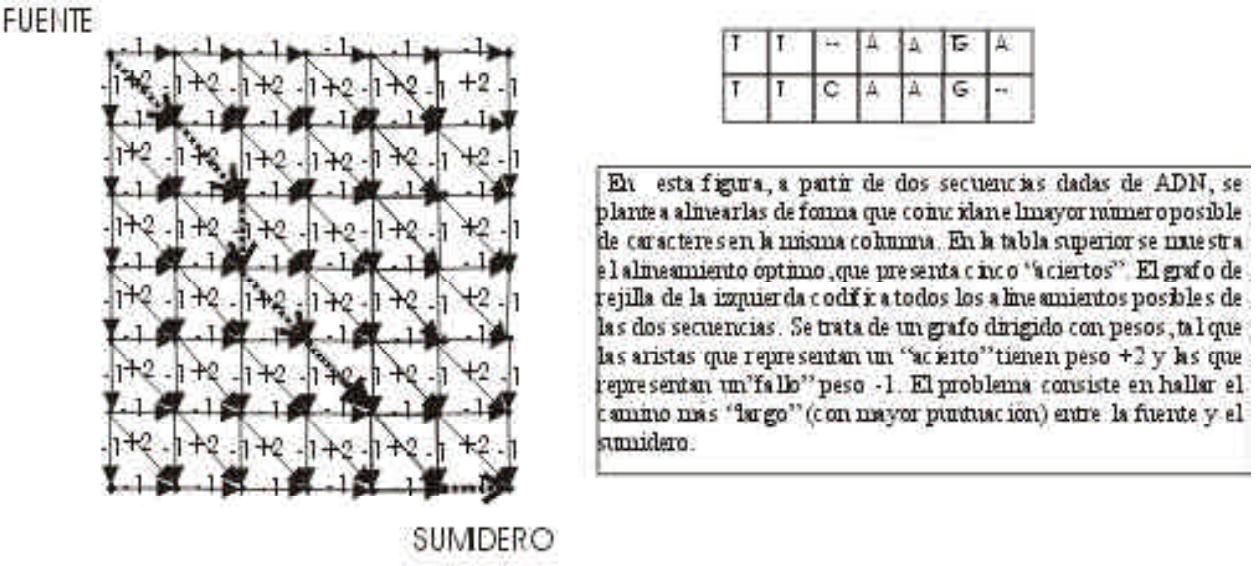


Figura 3. Grafo de rejilla

2.3 Árboles filogenéticos

Dado un conjunto de especies y cierta información sobre las similitudes ("distancia evolutiva estimada") entre ellas, se plantea la reconstrucción del árbol filogenético, o historia evolutiva de dicho conjunto de organismos. El análisis filogenético comienza con la adquisición de datos biomoleculares, como por ejemplo secuencias de ADN. Después, se construye un árbol que representa la hipotética evolución de las secuencias, o de los organismos estudiados.

Generalmente, las hojas del árbol representan a los organismos sobre los que disponemos de información, mientras que los nodos interiores representan las especies de las que supuestamente evolucionaron.

El proceso evolutivo se modeliza mediante un tipo especial de árboles con pesos en sus ramas, denominados árboles filogenéticos. El objetivo consiste en, dada una matriz de distancias evolutivas entre especies, obtener el árbol y los pesos de sus ramas que se adecuen lo más posible a la matriz de distancias. En caso de que el árbol hallado reproduzca exactamente los valores de la matriz, se dice que la reconstrucción ha sido perfecta.

Recientemente se han propuesto nuevas estructuras matemáticas para la reconstrucción de árboles filogenéticos: los "grafos filogenéticos" [5][6].

El concepto de grafo filogenético deriva del de grafo de competencia (competition graph), de utilidad en problemas de ecología.

El diseño de mejores algoritmos para la reconstrucción de árboles filogenéticos dependerá del estudio intensivo de las propiedades de los grafos filogenéticos y de competencia.

2.4 Reordenación de genomas

Dados dos genomas y un conjunto de operaciones permitidas a nivel de gen sobre ellos, se plantea cuál es el conjunto más pequeño de transformaciones que han podido convertir una de las secuencias en la otra. El problema es de importancia, por ejemplo, en evolución.

Se supone en todo momento que los genomas que se comparan tienen los mismos genes, pero colocados en orden u orientaciones diferentes. Las operaciones permitidas pueden ser inversiones, transposiciones, translocaciones, divisiones o fusiones de genes.

El problema matemático consiste en transformar una permutación de pares número-orientación en otra concreta utilizando para ello el menor número posible de operaciones permitidas. Para resolver el problema de nuevo se emplean estructuras basadas en grafos

2.5 Nuevas estructuras semánticas para el almacenamiento de la información biológica. Grafos conceptuales

Uno de los problemas importantes de la Biología moderna es la búsqueda de información concreta en grandes bases de datos, o en Internet. Los algoritmos que se utilizan para recuperar esta información no parecen ser suficientes ante la ingente cantidad de datos almacenados y el ritmo con que se acumulan.

Hoy en día, la búsqueda en Internet es puramente *sintáctica*, es decir, se basa en hallar documentos Web que incluyan la palabra o palabras buscadas. Esto da lugar a una desbordante cantidad de enlaces, imposibles de analizar en su totalidad y que, en su mayoría, no guardan relación con lo que realmente se está buscando.

Las investigaciones actuales se centran en orientar las búsquedas, no hacia la sintaxis de las palabras clave, sino hacia su significado real, es decir, hacia la *semántica*. El objetivo es transformar la búsqueda de cadenas en búsqueda de conceptos.

Para ello, se ha definido una nueva estructura para almacenar conceptos y relaciones entre conceptos, denominada *grafo conceptual* [8].

Los grafos conceptuales expresan conceptos de forma legible, precisa desde el punto de vista de la lógica matemática, y tratable computacionalmente. Los grafos conceptuales representan un lenguaje intermedio entre la lógica formal, que emplean los ordenadores, y el lenguaje natural, que utilizamos los seres humanos.

Un grafo conceptual G es un grafo bipartido que contiene dos tipos de nodos: conceptos y relaciones conceptuales. Cada arista de G es un par ordenado (r,c) formado por una relación conceptual "r" y un concepto "c", pertenecientes ambos a G, de forma que puede haber conceptos que no estén unidos a ninguna relación conceptual (figura 4).

Un aspecto importante de los grafos conceptuales es que son trasladables a predicados lógicos, es decir, todo grafo conceptual lleva asociado un conjunto de predicados lógicos de primer orden o de orden superior. Ello permite trasladar las reglas de inferencia del lenguaje lógico a operaciones sencillas sobre los grafos conceptuales y viceversa.

El lenguaje de la lógica se puede representar de formas muy diferentes: los grafos conceptuales son una forma gráfica de hacerlo. Cada una de estas formas está orientada a un tipo diferente de aplicaciones. Así, para favorecer la lectura, es posible representar los operadores lógicos en un lenguaje natural *controlado*, que utiliza la sintaxis y el vocabulario de los lenguajes naturales. Aunque la tarea de trasladar el lenguaje natural (no restringido) a una notación formal es todavía un problema abierto, es mucho más fácil traducir los grafos conceptuales y otras notaciones formales a lenguaje natural.

Además de la notación, la lógica formal posee sus propias reglas de definición y de inferencia, que permiten que una determinada representación se pueda traducir a otra equivalente. Así, la lógica posee la capacidad de generar expresiones semánticamente equivalentes.

Sin embargo, la lógica por sí misma no contiene significado semántico. Para representar el conocimiento en un determinado campo es necesario añadir a la lógica una ontología que defina cómo se clasifican los conceptos en dicho dominio.

Los grafos conceptuales y las estructuras semánticas de almacenamiento de la información son un intento de formalizar el lenguaje natural humano. Esta tarea es difícil, quizás incluso imposible, pero estas nuevas estructuras para codificar la información, aunque limitadas de por sí, representan un importante paso adelante para mejorar las búsquedas en la red y en bases de datos de información específica, especialmente de información biológica.

El éxito de estos sistemas depende del diseño de algoritmos eficientes para identificar, recorrer e interpretar grafos conceptuales. Este es un trabajo todavía pendiente.

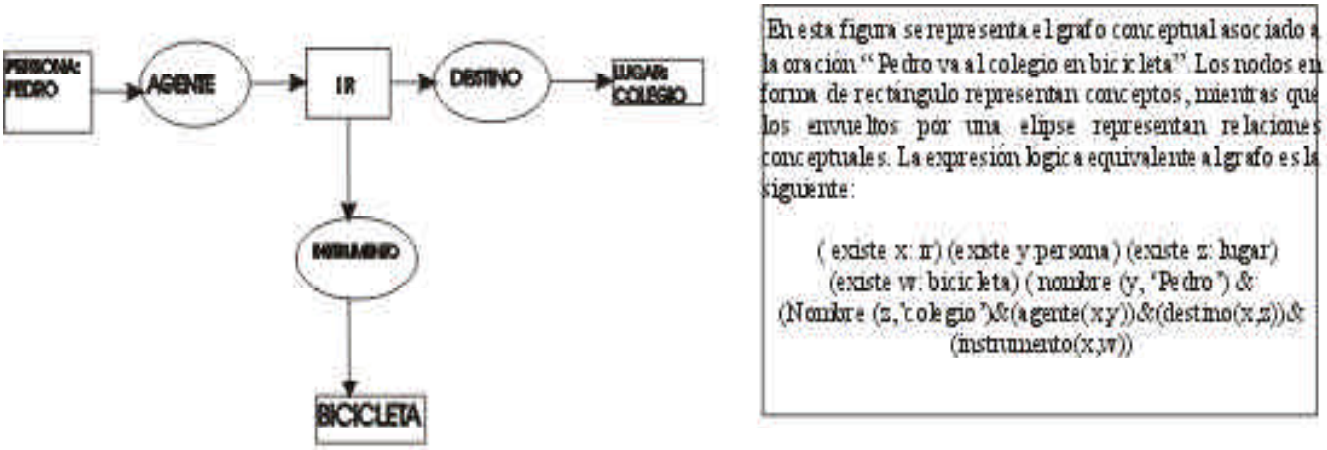


Figura 4: Grafo Conceptual

1. APLICACIONES DE LA BIOLOGÍA A LA RESOLUCIÓN DE PROBLEMAS EN MATEMÁTICA DISCRETA.

En 1994, Leonard Adleman utilizó ADN para resolver el problema de la búsqueda de un camino hamiltoniano en un grafo dirigido [1].

El ADN posee algunas características que permiten acelerar la resolución de algunos problemas intratables, por las siguientes razones [3]:

- La densidad de información almacenada en el DNA es muy elevada, muy superior a la de cualquier disco duro actual.
- El ADN posee una estructura en doble hélice, de forma que toda secuencia genética tiene una secuencia complementaria, lo cual puede utilizarse para corregir errores. Así, si ocurre algún error en una de las hélices del ADN, se puede reconstruir la secuencia original tomando como referencia la otra hebra.

- El ADN permite realizar operaciones en paralelo. En una célula, determinados enzimas son capaces de modificar el ADN. Por ejemplo, hay enzimas que permiten reparar ciertas zonas, cortar una secuencia o unir dos secuencias. La posibilidad de utilizar diferentes copias de un mismo enzima actuando al mismo tiempo sobre diversas moléculas de DNA es lo que permite realizar cálculos de forma paralela.

Sin embargo, la computación mediante ADN también posee importantes inconvenientes. La complejidad del problema se manifiesta en que la cantidad de ADN necesaria para hallar la solución aumenta exponencialmente con el tamaño del problema. En consecuencia, si el tamaño del grafo aumenta hasta varios cientos de nodos, se necesitaría una masa de ADN superior a la masa de la Tierra para resolver el problema.

Aunque las limitaciones de esta técnica son muy importantes, la evolución tecnológica podría llegar a solventarlas en un futuro.

4. ENSEÑANZA DE LA MATEMÁTICA DISCRETA EN ESPAÑA

La Matemática Discreta se imparte como asignatura obligatoria en las facultades de Matemáticas y de Ingeniería Informática españolas. Sin embargo, después de examinar los planes de estudios en España de Biología y otros estudios afines, hemos comprobado que dichas titulaciones no ofertan la Matemática Discreta, ni siquiera como asignatura optativa. Como hemos visto, la Matemática Discreta se ha convertido en un elemento fundamental para entender la formalización matemática de problemas biológicos y ello debe quedar reflejado en los planes de estudio de las titulaciones afines a la Biología.

Además, los profesores encargados de enseñar Matemática Discreta en las titulaciones de Matemáticas e Ingeniería Informática deben ser conscientes de sus aplicaciones en Biología e incluirlas en los temarios de las asignaturas que imparten, por dos razones: la primera, porque la Bioinformática se convertirá en un futuro cercano en una salida profesional importante para los ingenieros informáticos; la segunda, porque las aplicaciones de la Matemática Discreta a la Biología son muy atractivas conceptualmente y pueden servir para motivar al estudiante en el estudio de la asignatura.

5. CONCLUSIONES

- En los últimos 15 años, en todos los avances relevantes de Bioinformática y Biología Computacional se han utilizado técnicas propias de la Matemática Discreta.
- La Matemática Discreta, que ha sido un campo de "segunda fila" dentro de las Matemáticas, se ha convertido en un instrumento básico para el desarrollo de la Ciencia de la Computación y de la Biología moderna.
- Si bien se imparten uno o varios cursos de Matemática Discreta en las facultades españolas de Ingeniería Informática, no se ha establecido todavía como asignatura fundamental en la formación matemática de los biólogos, en cuyo currículo académico figuran solamente el Cálculo Funcional, el Álgebra y la Estadística.
- Surge pues la necesidad de establecer la Matemática Discreta como asignatura importante en los planes de estudio de Biología y Bioinformática.

6. BIBLIOGRAFÍA

[1]Adleman, L.M. (1994) *Molecular computation of solutions to combinatorial problems,* Science, 226, 1021-1024

[2] Kececioglu J.D. and Myers E.W. (1995), *Combinatorial algorithms for DNA sequence assembly,* Algorithmica, Vol.13.

[3] Paun,G. (1998) *Computing with bio-molecules,* Springer Verlag

[4] Pevzer P.A., Tang H and Waterman M.S. (2001), *An eulerian path approach to DNA fragment assembly,* Proc. Nat. Academy of Science USA, vol.98 nº17

[5] Roberts F.S. (1999) *Competion graphs and phylogeny graphs*. In L.Lovasz, editor, Graph Theory and Computational Biology, Bolyai Studies. J.Bolyai Mathematical Society, Budapest.

[6] Roberts F.S. and Sheng L (1998) *Phylogeny Numbers*. Discrete Applied Mathematics, 87: 213-228

[7] Setubal, J and Meidanis (1997), *Introduction to Computational Molecular Biology.* PWS publishing company.

[8] Sowa, J.F.(2000) *Ontology, metadata and semiotics.* Conceptual structures: logical, linguistic and computational issues, Lecture Notes in AI#1867 Springer Verlag, Berlin