Memoria del trabajo a realizar

Sergi Elizalde

1. Interés de la investigación

El trabajo que pretendo realizar es sobre la aplicación de métodos combinatorios a problemas de biología computacional.

Los modelos estadísticos se usan para resolver algunos problemas en bioinformática, como determinar qué partes del genoma se traducen a proteínas, o cómo una secuencia de ADN se transformó en otra durante la evolución, a través de una serie de mutaciones, inserciones y delecciones. Cada posible respuesta tiene una cierta probabilidad que depende de los parámetros del modelo. Cuando éstos se conocen, la respuesta más probable, llamada *explicación*, se obtiene resolviendo un problema de optimización combinatoria. La aplicación que envía cada observación a su explicación correspondiente se llama *función de inferencia* ("*inference function*").

Parte importante de mi investigación se centra en determinar el número de funciones de inferencia. Este número es importante porque indica cómo los parámetros de modelo pueden afectar la solución. Un caso particular muy importante en biología es el de alineación óptima de secuencias. En este caso, las funciones de inferencia indican, dado un par de secuencias de ADN, cuál es el proceso de evolución más probable que siguió una de ellas para transformarse en la otra.

Otro caso particular de funciones de inferencia son las llamadas *funciones de búsqueda de genes* ("*gene finding functions*"). Estas funciones determinan qué partes de una secuencia dada de ADN son *exones* y qué partes son *intrones*. Esta distinción es importante porque los exones son las partes que codifican proteínas, mientras que los intrones, que forman la mayor parte del genoma, son segmentos intermedios de los que no se conoce su función, pues nunca son traducidos a proteínas.

Aunque la mayoría de estos problemas provienen de situaciones reales en biología y en genética, pueden ser fácilmente traducidos a cuestiones puramente matemáticas, lo cual permite estudiarlos usando herramientas combinatorias. Esta colaboración entre el álgebra, la combinatoria, la estadística, la biología y la computación es relativamente reciente, y está resultando muy fructífera.

2. Experiencia previa del candidato en el tema

En junio de 2004 terminé mi tesis doctoral en Matemáticas en el Massachusetts Institute of Technology (MIT). La tesis, en el área de combinatoria, se titula "Statistics on pattern-avoiding permutations", y la realicé bajo la dirección del profesor Richard Stanley, que es un experto en combinatoria reconocido a nivel mundial.

Al acabar el doctorado recibí una beca postdoctoral del Mathematical Sciences Research Institute (MSRI), en Berkeley, donde he estado haciendo investigación durante el pasado año. Allí he participado en los programas "Hyperplane arrangements and applications" y "Probability, algorithms and statistical physics". En Berkeley entré en contacto con los profesores Bernd Sturmfels y Peter Winkler. Ellos me introdujeron en el apasionante tema de la biología matemática, y sugirieron que los conocimientos que había adquirido durante mi doctorado en MIT podían ser aplicados a la investigación que ellos llevaban a cabo en biología computacional.

Rápidamente empecé a trabajar con ellos en este proyecto, y el primer fruto de esta colaboración son los dos capítulos que he escrito del libro *Algebraic Statistics for Computational Biology*, editado por L. Pachter and B. Sturmfels, y que va a ser publicado en octubre de

2005 por *Cambridge University Press*. Mis capítulos se titulan *Inference functions* (del que soy el único autor) y *Bounds for optimal sequence alignment* (del que soy coautor junto a Fumei Lam).

3. Plan de trabajo

Uno de los resultados principales que aparece en el primero de estos capítulos es que el número de funciones de inferencia de cualquier modelo gráfico es polinómico en el tamaño del modelo. La demostración de este hecho usa la teoría sobre politopos integrales, un tema muy actual en combinatoria.

Este resultado, lejos de dejar el tema zanjado, abre una nueva línea de investigación muy prometedora. Sabiendo ahora que el número de funciones de inferencia de un modelo gráfico es polinómico en el tamaño del modelo, el siguiente paso es encontrar un modo eficiente de precomputar todas las funciones de inferencia para modelos dados. Esto nos permitiría dar la explicación correspondiente a cada observación muy eficientemente.

Aunque de entrada pueda parecer un trabajo muy teórico, existen una gran cantidad de aplicaciones de este hecho a problemas concretos. Por ejemplo, en el caso de alineación de secuencias ("*sequence alignment*"), sería posible encontrar la alineación óptima de cualquier par de secuencias de ADN en tiempo polinómico en la longitud de las secuencias (es decir, eficientemente desde el punto de vista computacional). El paso siguiente en este trabajo consistirá en implementar estos algoritmos y usarlos con datos reales.

El caso de detección de genes, la aplicación de este proyecto permitirá determinar eficientemente qué partes de una secuencia de ADN son exones y qué partes son intrones.

Otro problema abierto que pretendo estudiar en mi investigación es el siguiente. Dadas dos secuencias de ADN, y fijando un modelo estadístico de alineación de secuencias, se desconoce cuántas de las posibles maneras de alinearlas (el número total de maneras es exponencial en la longitud de las secuencias) pueden ser óptimas para algun valor de los parámetros del modelo. Se sabe que si n es la longitud de dichas secuencias, el número de alineaciones que pueden ser óptimas es como mucho $O(n^{2/3})$. Sin embargo, nosotros conjeturamos que nunca se alcanza esta cota, sino que el número máximo de alineaciones que pueden ser óptimas nunca supera $O(n^{1/2})$. Dentro de mi plan de trabajo intentaré demostrar este resultado. Encontrar el valor exacto de esta cota es importante porque nos permite conocer la complejidad del problema de alineación de secuencias de ADN, y desarrollar algoritmos mejores para encontrar alineaciones óptimas.

4. Razones de mi elección del centro de destino

El centro de destino que he elegido para realizar mi investigación es Dartmouth College, en New Hampshire. La razón principal de esta elección es que en el departamento de matemáticas de esta universidad se encuentra el profesor Peter Winkler. El profesor Winkler fue uno de los organizadores principales del programa en "Probability, algorithms and statistical physics" en MSRI el semestre pasado, donde le conocí y donde empecé a trabajar con él en la aplicación de herramientas combinatorias y probabilísticas en problemas de biología computacional. El profesor Winkler es sin duda uno de los expertos mundiales en probabilidad y algoritmos, y su trabajo en procesos aleatorios y cadenas de Markov tiene aplicaciones muy interesantes a la bioinformática.

Considero que P. Winkler, junto con los otros miembros de su grupo de investigación en Dartmouth College, es la persona más indicada para dirigirme en este proyecto de investigación.

Finalmente, Dartmouth College es una de las universidades más prestigiosas de Estados Unidos. Me siento muy afortunado por haber sido aceptado en el Departamento de Matemáticas

de dicha universidad y poder realizar mi investigación post-doctoral dentro de un equipo tan competente.

5. Repercusión del trabajo a mi vuelta a España

En verano de 2007, una vez concluída esta estancia post-doctoral en Dartmouth College, mi intención es volver a España para incorporarme como profesor en alguna universidad del país. Desde que acabé mi carrera de Matemática Aplicada en la Universidad Politécnica de Cataluña he mantenido contacto y he seguido colaborando con varios profesores de esta universidad, en particular con los profesores Marc Noy, Sebastià Xambó, Josep Grané y Oriol Serra. Con el profesor Noy, director del Departamento de Matemática Aplicada II, tenemos varios proyectos conjuntos. Mis planes consisten en encontrar una plaza de investigador en España a mi regreso de Estados Unidos.

La bioinformática, y en concreto la aplicación de herramientas matemáticas a la biología computacional, es un tema que en España esta empezando a desarrollarse. El nivel no es actualmente comparable al de algunos centros norteamericanos. Mi intención a la vuelta a España es contribuir al desarrollo de este tema, aportando las ideas y las técnicas untilizadas en mi trabajo en los Estados Unidos.