ejemplo los órganos de los animales huésped. Por tanto, los *pili* son un determinante de virulencia de algunas bacterias patógenas. Un sistema bien estudiado es el que constituyen algunas cepas patógenas de *E. coli*. Que una de estas bacterias exhiba o no *pili* en su superficie depende del estado de activación de un operón llamado *Pap* (Pyelonephritis-associated pilus). Se ha comprobado que la activación/inactivación de este operón está estrictamente correlacionada con el estado de metilación de una región reguladora, que posee dos sitios de metilación, GATC1 y GATC2. La metilación de GATC2 (GmATC2) es necesaria para que se transcriba el operón (Fig. 2A). Hay una proteína reguladora, cuyo número por célula es bajo, que tiene una alta afinidad por los sitios GATC no metilados, de manera que se une a GATC1 o a GATC2 de forma aleatoria. Esta unión previene la metilación de dicho sitio. Si por azar la proteína reguladora se une a GATC2 protegiéndola de la metilación, se inhibe la transcripción del operón (Fig. 2B).

En una población genéticamente homogénea y en condiciones ambientales idénticas, encontraremos bacterias que posean *pili* y otras que no, ya que cada individuo echa a cara o cruz la decisión de encender o apagar el operón responsable de la síntesis del *pilus*. No pretendemos ocultar que en ocasiones, la moneda utilizada puede estar trucada y que bajo determinadas condiciones ambientales las caras (digamos GmATC1) sean más frecuentes que las cruces (GmATC2); pero en cualquier caso, la expresión de *pili* es un proceso estocástico ya que aunque conozcamos el genotipo, el ambiente y la historia, nunca podremos precisar con grado de determinismo cuál será el fenotipo para un individuo dado. ¿Qué ventaja representa dejar al azar la expresión de *pili*? Para el individuo puede que ninguna, pero para la especie supone un enriquecimiento muy ventajoso. Los *pili* son estructuras altamente inmunogénicas, contar en todo momento con una subpoblación de bacterias que carezcan de estas estructuras, puede permitir evadir las defensas del huésped, para que más tarde, por azar, algunas de ellas vuelvan a desarrollar *pili* e infectar el tracto urinario.

Para finalizar nos gustaría hacer nuestras las palabras de Henrik Kacser, diciendo que cuando tratamos con seres vivos, el todo es mucho más que la suma de sus partes. Obstinarse en ignorarlo es condenarse al fracaso.

PREDICCIÓN DE ESTRUCTURA DE PROTEÍNAS EMPLEANDO *SOFTWARE* LIBRE

Aurelio A. Moya García

Las técnicas experimentales de caracterización estructural, principalmente cristalografía de rayos X y resonancia magnética nuclear, proporcionan estructuras de alta resolución, aunque desafortunadamente sólo una pequeña parte de las proteínas se pueden caracterizar así. Para una gran parte de la fracción de secuencias cuya estructura no se puede determinar experimentalmente, los métodos computacionales de predicción de estructura nos ofrecen información valiosa, útil para explicar gran parte de los aspectos funcionales que se pueden derivar del conocimiento estructural.

Por otro lado, es obvio que la implementación de estas técnicas pasa por el desarrollo de *software* apropiado y que para resolver problemas biológicos mediante esta aproximación necesitamos usar ordenadores y programas específicos. Existe a nuestra disposición gran variedad de servidores y programas asequibles a coste ínfimo. En este artículo explicaré las bases de la predicción estructural de proteínas y las ventajas de realizar esta tarea empleando herramientas de código abierto o con más propiedad *software* libre.

Métodos de predicción de estructura

El primer grupo de métodos de predicción de estructura que consideraré se denomina genéricamente métodos *de novo* o *ab initio*.

Estos métodos parten de la asunción de que la información necesaria para conocer la estructura tridimensional de una proteína está en su secuencia de aminoácidos. Buscan la estructura nativa como la conformación que corresponde al mínimo global de una función potencial determinada, que «representa» a la proteína y que se construye desde su secuencia. Para optimizar la función potencial emplean distintos métodos de búsqueda en el espacio de las conformaciones, los cuales suelen implementar algoritmos de mecánica molecular combinados con dinámica molecular [van Gunsteren W.F. y Beredsen H.J.C. **1977**. *Molecular Physics* **34**:1311], simulaciones Monte Carlo (Combination, Rosseta [Simons K. et al. (2000) *J. Mol.Biol.* **306**:1191]) o el empleo de bases de datos de elementos de estructura secundaria estándar

(FRAGFOLD [Jones D.T. 2001. *Proteins* ;Suppl **5**:127]).

Los métodos *ab initio* son computacionalmente costosos y su fiabilidad disminuye con el tamaño de la proteína, generalmente funcionan bien con péptidos menores a 150 aminoácidos. La principal ventaja que tienen es que sólo se necesita la secuencia como información de partida, de modo que en principio es posible modelar proteínas que corresponden a plegamientos no conocidos.

En la práctica no se emplean para deducir la estructura de una proteína completa, sino como apoyo a otras técnicas más potentes y que consiguen más éxitos. Este conjunto de técnicas constituyen el segundo grupo de métodos de predicción, el modelado por homología (*Homology, Comparative Modelling*).

La idea básica de la que surge esta aproximación descansa en el hecho de que todas las parejas de proteínas que presentan una identidad de secuencia mayor al 30% tienen estructura tridimensional similar [Sander C. y Schneider R. 1991. *Proteins* **9**:56]. De este modo se puede construir el modelo tridimensional de una proteína de estructura desconocida, partiendo de la semejanza de secuencia con proteínas de estructura conocida [Blundell T.L. et al. 1987. *Nature* **326**:347].

Las estapas necesarias para el modelado por homología son básicamente cuatro:

- Identificación de estructuras conocidas (moldes) relacionadas con la secuencia diana. Se emplean métodos de comparación de secuencias como FASTA [http://fasta.bioch.virginia.edu/], BLAST o PSI-BLAST [http://www.ncbi.nlm.nih.gov/BLAST/].
- Alineamiento de la diana con los moldes. Es la etapa más importante y la más delicada ya que la construcción del modelo se realizará conforme a este alineamiento. En este paso se emplean programas típicos de alineamiento de secuencias como CLUSTAL [http://www.ebi.ac.uk/clustalw/].
- Construcción del modelo. Existen varias aproximaciones para construir las coordenadas espaciales de la secuencia diana desde el alineamiento realizado. Como ejemplo tenemos ProModII en el servidor SWISS-MODEL [http://www.expasy.org/swissmod/SWISS-MODEL.html].
- Evaluación del modelo. La información que se puede obtener del modelo depende de su calidad, de modo que es importante poder evaluarla. Existen muchas pruebas que se pueden realizar sobre un modelo que incluyen comprobaciones estéricas, químicas, representaciones de Ramachandran, etc.

Para una revisión más detallada de estas etapas ver Martí-Renom et al. 2000. *Annu. Rev. Biophys. Biomol. Struct.* **29**:291.

Se estima que el modelado por homología es aplicable a un tercio de todas las secuencias proteicas conocidas [Rost B. y Schneider R. Pedestrian Guide to Analysing Sequence Databases. http://www.cbi.cnptia.embrapa.br/SMS/bbnet/pedestrian/Springer96.html] y esta cifra aumenta un 4% cada año, a medida que se determinan más estructuras correspondientes a nuevos plegamientos por vías experimentales. Aún así ¿qué pasa con el resto?

Cuando la similitud entre la secuencia diana y el molde es demasiado baja no es posible realizar un buen alineamiento y no se puede aplicar con éxito el modelado por homología. En estos casos aún podemos obtener información estructural de la proteína empleando técnicas de *threading*. En resumen consiste en colocar la secuencia problema en diferentes plegamientos conocidos y evaluar cómo se «encuentra de bien» o cómo encaja en cada uno de ellos. Por «encajar» se entienden cosas diferentes según el programa de *threading*: coincidencia de estructura secundaria, residuos en ambientes parecidos a como se encuentran en la base de datos, etc.

Pero para hacer todo esto hay que usar programas, ¿de qué tipo?.

Programas libres

Existe una gran variedad de programas para realizar predicciones de estructura de proteínas. Gran parte de ellos son desarrollados por investigadores que construyen sus propias herramientas para ser empleadas en la investigación, y que se ponen a disposición de la comunidad científica con el código fuente y permiso de realizar las copias necesarias, es decir, como *software* libre (no confundir con *shareware*. En la dirección http://www.fsf.org/ hay abundante información sobre este tipo de software).

Estos programas además de herramientas son resultado de investigación científica, de modo que es de esperar que ese conocimiento sea puesto al alcance de toda la comunidad. Este intercambio de conocimiento, entre otras cosas, es lo que favorece el desarrollo de programas de ordenador libres.

Para una discusión más extensa sobre la difusión libre del conocimiento ver El futuro de la información: ¿vamos hacia donde queremos? de Jesús M. González Barahona, disponible en http://sindominio.net/biblioweb/telematica/futuroinfo.pdf

En definitiva, el uso de programas de ordenador libres en el ámbito científico ofrece muchas ventajas y pocos inconvenientes, los cuales, por otro lado, son fácilmente resueltos si se consigue vencer la inercia al cambio.