Cristina Marino Buslje y Gustavo Parisi
Fundación Instituto Leloir

Aprender para predecir
La bioinformática como herramienta de estudio en biología

Aprender para predecir

Desde que adquirió conciencia sobre el tiempo, el ser humano sintió la necesidad de predecir situaciones y eventos: tormentas, abundancia de animales para cazar, períodos de lluvias para sus cosechas o cosas más abstractas como el resultado de alguna batalla o el amor de algún pretendiente. Estas predicciones tomaron distintas formas: desde oráculos que respondían inquietudes humanas recurriendo a la comunicación con un dios todopoderoso hasta, cuando no se contaba con el apoyo divino, relaciones entre características físicas como las formas de las entrañas –humanas o animales–, las líneas de las manos, la ocurrencia de fenómenos climáticos y astronómicos, con posibles evidencias de un futuro. La falta de correlación entre esas evidencias y los hechos a predecir hacían que dichas predicciones fallaran enormemente.

La ciencia actual propone aprender primero como método para mejorar la capacidad predictiva acerca de determinados hechos. En particular, en bioinformática es fundamental procesar la información contenida en sistemas biológicos conocidos con el fin de utilizarla en el desarrollo de herramientas capaces de realizar predicciones. Esta disciplina científica se dedica a responder preguntas de índole biológica mediante el procesamiento

¿DE QUÉ SE TRATA?

El uso de la informática permite el almacenamiento y manejo de grandes cantidades de datos biológicos.

computacional de los datos, además de almacenar, extraer, organizar, analizar, interpretar y utilizar distintos tipos de información biológica. Así, cumple una función esencial para comprender aspectos biológicos derivados del procesamiento de una gran cantidad de datos, predecir fenómenos y propiedades donde la investigación experimental no ha llegado, optimizar el diseño de experimentos, acelerar los tiempos de investigación y abaratar sus costos.

Transformar los datos en información relevante: bases de datos

Las bases de datos biológicas son repositorios de información de muy distintos tipos; en la mayoría de los estudios bioinformáticos son la materia prima de la cual extraemos información para aprender. Las hay de secuencias de proteínas, de genes, de estructuras proteicas, de rutas metabólicas, de mutaciones asociadas a enfermedades, bibliográficas, para mencionar algunos pocos ejemplos de la variedad que existe en la actualidad. Los investigadores en bioinformática crean, *curan*, mantienen y analizan estas bases de datos. Y si bien es importante que estén actualizadas y fuertemente relacionadas entre sí, es esencial que la información que contienen sea confiable. Muchas pasan por un proceso de verificación o *curado* a mano, lo que quiere decir que un investigador experimentado (y no un programa o método automatizado) contrasta las distintas evidencias biológicas contenidas con publicaciones científicas para evaluar su validez. Como el curado de las bases de datos implica un alto costo de tiempo, algunas han comenzado a promover la participación activa de la comunidad científica para este propósito. Por ejemplo, la base de datos DisProt (*http://www.disprot.org/*), de *proteínas desordenadas* (proteínas que no tienen una estructura tridimensional definida) es curada a mano y en cada proteína figura el nombre del curador correspondiente. La base Uniprot está anotada con información biológica obtenida en forma automática (en su versión actual contiene unas 150 millones de proteínas). Sin embargo, una subdivisión de UniProt, denominada Swiss-Prot, está curada a mano y contiene solo 500.000 proteínas, lo que resalta el esfuerzo que requiere la revisión de tal cantidad de información.

El ritmo extraordinario al que crecen las publicaciones científicas, la facilidad relativa de acceso a las nuevas técnicas de secuenciación de ADN y las enormes cantidades de información genética que generan estas y otras técnicas de alto rendimiento plantean un desafío computacional extraordinario para extraer información biológica en tiempos razonables.

Nombre	Tipo de información que contiene	Sitio web	Cantidad de datos almacenados
DisProt	Proteínas desordenadas	*http://www.disprot.org/*	803
PDB	Estructuras tridimensionales de proteínas	*http://www.rcsb.org/*	150.145
UniProt	Secuencias de proteínas y todo tipo de información disponible sobre ellas	*www.uniprot.org*	146.106.279
GenBank	Secuencias de ácidos nucleicos	*https://www.ncbi.nlm.nih.gov/nuccore/*	212.260.377
PubMed	Publicaciones científicas	*https://www.ncbi.nlm.nih.gov/pubmed/*	~29.000.000
Kegg	Rutas metabólicas	*https://www.kegg.jp/kegg/pathway.html*	530
IntAct	Interacciones proteína-proteína	*https://www.ebi.ac.uk/intact/*	882.962

Algunos ejemplos de bases de datos comúnmente usadas en bioinformática y los datos contenidos en ellas en marzo de 2019.

Predicción de la estructura de proteínas

La estructura de una proteína es en general esencial para llevar a cabo su función biológica. Por esta razón los científicos dedican mucho esfuerzo para determinarla en forma experimental. Sin embargo, este proceso es largo, costoso y a veces tedioso. Evidencia de esto es el número de estructuras de proteínas conocidas depositadas en la base de datos PDB, que alcanza a unas 150.000, mientras que el número de secuencias proteicas conocidas es cien veces mayor. Para achicar esta brecha es posible utilizar herramientas bioinformáticas que crean modelos computacionales de una proteína de la cual aún no se conoce la estructura, y lo hacen con bastante fidelidad. Se evitan así los gastos de recursos y de tiempo necesarios para su resolución y el modelo permite avanzar con la investigación, por ejemplo, para el diseño de una droga que inhiba a dicha proteína.

En bioinformática, como en todas las disciplinas científicas, primero debemos aprender para desarrollar métodos predictivos. Si quisiéramos utilizar la información de la secuencia de aminoácidos para predecir la estructura de una proteína, primero tendríamos que comprender cómo varían las estructuras en función de los cambios en dichas secuencias. Estos estudios fueron realizados por el inglés Cyrus Chothia y el norteamericano Arthur Lesk en la década de 1980. Ellos encontraron que dicha capacidad predictiva crece en forma exponencial con el aumento de similitud entre secuencia y estructura. Básicamente, cuanto más se asemejen las secuencias de dos proteínas, más parecidas serán sus estructuras. Esto permitió desarrollar métodos denominados de modelado molecular para predecir la estructura de las proteínas. Tales modelos utilizan la estructura conocida de una proteína (proteína molde) para predecir el modelo estructural de una proteína buscada (proteína blanco). Según Chothia y Lesk, cuanto más se parezcan entre sí las secuencias de las proteínas blanco y molde, mejor será el modelo predicho.

A modo de ejemplo, en la figura 1 se muestra en violeta la estructura de la proteína denominada factor de crecimiento del fibroblasto número 8 (FGF8) y, en otros colores, varios modelos estructurales obtenidos utilizando como molde proteínas cada vez más alejadas, o sea que comparten cada vez menos aminoácidos idénticos. Puede observarse entonces que, al disminuir de izquierda a derecha el porcentaje de identidad, menos parecidos son los modelos obtenidos a la estructura real en violeta. Dicha proteína FGF8 desempeña un papel importante en la regulación del desarrollo embrionario y como molécula de señalización en la inducción y el patrón del cerebro embrionario; de ahí que resulte una proteína esencial para el desarrollo normal del cerebro. Por lo tanto, es de suma utilidad conocer su estructura para descubrir su función.

Cambios genéticos y enfermedad

Si la conservación de la estructura de las proteínas es esencial para que estas puedan llevar a cabo su función biológica, sucede que durante su evolución, las proteínas pueden presentar cambios en sus secuencias. La enorme mayoría de estas sustituciones o mutaciones no tiene casi ningún efecto sobre la estructura de la proteína y por ende tampoco en su función biológica. Sin embargo, una mínima fracción de estos cambios puede ciertamente afectarlas. Varias enfermedades que ocurren en humanos pueden originarse por el mal funcionamiento,

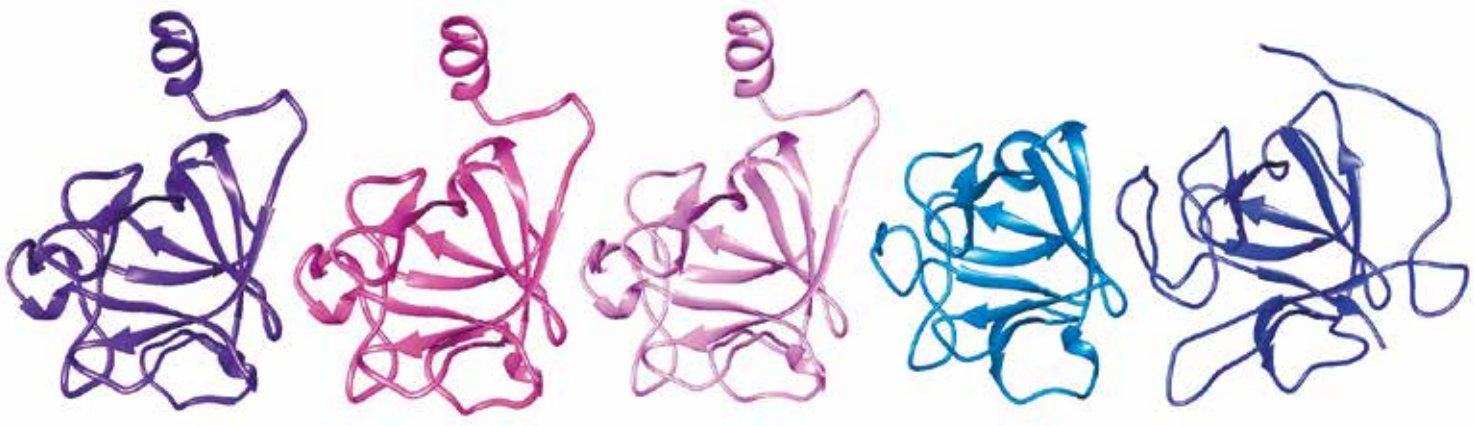

Figura 1. Representación en cintas de varios modelos estructurales para la proteína FGF8. A la izquierda, su estructura conocida experimentalmente (código PDB: 2FDB). Hacia la derecha, distintos modelos de ella ideados usando diferentes proteínas como moldes, con distintos porcentajes de identidad. Los modelos fueron realizados y representados con el programa Chimera (Universidad de California en San Francisco, 2004). A medida que la identidad de secuencia entre el molde y la proteína blanco decrece, disminuye la calidad del modelo, es decir, se parece cada vez menos a la estructura real en violeta.

la ausencia o la exacerbación de la actividad biológica de una determinada proteína. Existe una rama de la bioinformática que se dedica a desarrollar métodos para predecir el efecto de los cambios secuenciales en una proteína, y es actualmente un área en extremo activa debido a sus potenciales aplicaciones en el ámbito de la salud. Nuevamente, existen bases de datos que han acumulado millones de estas variaciones genéticas en distintas proteínas que podrían estar relacionadas con distintas enfermedades humanas. Por ejemplo, la base de datos COSMIC (*Catalogue Of Somatic Mutations in Cancer*, Welcome Sanger Institute, Universidad de Cambridge, 2018) contiene cerca de 2.800.000 mutaciones genéticas provenientes de, aproximadamente, 1.200.000 muestras correspondientes a 47 tipos de cáncer, incluyendo los de mama, páncreas, piel, colon, pulmón y ovarios. Detectar patrones entre las mutaciones en determinadas proteínas y poder relacionarlos con un algún tipo de cáncer es un área de gran interés ya que podrían ser utilizados para realizar pronósticos de la evolución de la enfermedad, su diagnóstico, predecir cambios de la respuesta a medicamentos, y ser útiles para la prescripción de las terapias adecuadas. Poder anticipar cuándo los cambios en el nivel molecular, en el nivel de los ácidos nucleicos o en el nivel de las proteínas, se correlacionan con la ocurrencia de enfermedades, implica un enorme desafío para las predicciones computacionales. La complejidad del sistema es aún mayor ya que se ha observado que los cambios en un determinado gen pueden afectar a muchos otros genes. La denominada medicina personalizada, que utiliza información genética del paciente para lograr el diagnóstico o el mejoramiento del tratamiento adecuándolo a las características particulares del individuo, constituye una de las principales áreas emergentes de la bioinformática y muestra una enorme potencialidad en el ámbito de la salud.

Los límites de la predicción en bioinformática

Si bien las predicciones computacionales en bioinformática pueden llegar a ser muy precisas, en ciertos casos distan mucho de tener una calidad tal como para reemplazar la medición y caracterización experimental. Es interesante mencionar que existen distintos 'torneos' en los cuales decenas de grupos de bioinformática en el mundo contrastan sus metodologías y predicciones computacionales con datos obtenidos en forma experimental. El denominado CASP (*Critical Assessment of Techniques for Protein Structure Prediction*) está dedicado a la evaluación de métodos de predicción de estructuras de proteínas. La organización provee a los grupos participantes de todo el mundo de una o varias secuencias de proteínas de las cuales cada uno estima computacionalmente su estructura, y obtiene distintos posibles modelos. Una vez finalizado el período de predicción, cada modelo es comparado con la estructura establecida en forma experimental para evaluar cuál es el mejor modelo y el mejor método de predicción (figura 2). De esta manera se pone a la luz cuáles son los métodos que funcionan mejor en ese momento.

Existen distintos trabajos comunitarios de este tipo para evaluar la confianza de las predicciones en distintos aspectos biológicos. Así tenemos que el de predicción de la función de la proteína se denomina CAFA (*Critical Assessment of Functional Annotation*); el de predicción del efecto de mutaciones relacionadas con patologías, CAGI (*Critical Assessment of Genome Interpretation*), solo por citar unos ejemplos. Estos torneos son de suma utilidad para diagnosticar el estado del arte en la predicción en una determinada área, detectar casos difíciles en los cuales las predicciones fallan sistemáticamente y ver los avances relativos con el pasar de los años.

¿Llegará el día en el que por aumento de nuestro conocimiento y de la capacidad de cálculo de nuestras computadoras las predicciones reemplacen a los experimentos en biología? Quizá esto se cumpla para muchos sistemas en un futuro cercano, pero esta afirmación nos tienta a una reflexión final. Las predicciones

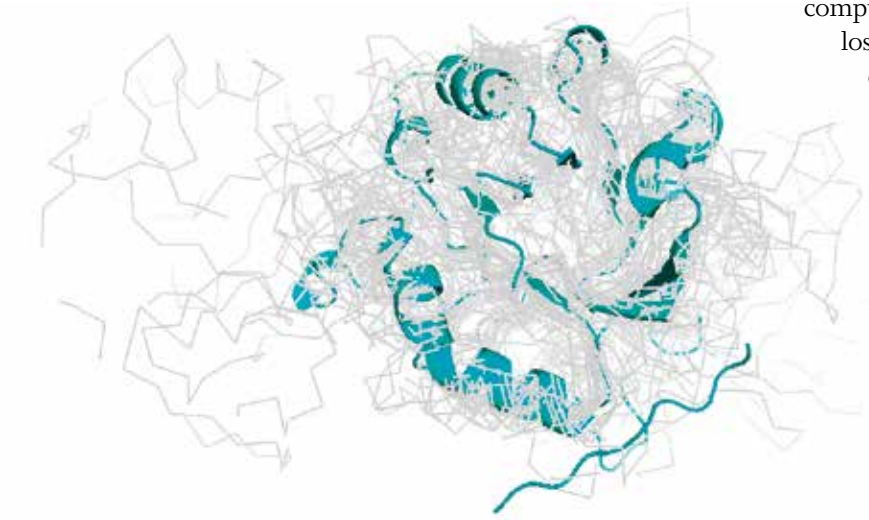

Figura 2. Resultados del CASP 2011. En formato cinta y en celeste se indica la estructura de la proteína obtenida en forma experimental por cristalografía de rayos X. Las líneas grises indican las distintas predicciones computacionales hechas por distintos grupos de investigadores alrededor del mundo (mostrando con una línea solo el esqueleto proteico). Observar que algunos modelos son más semejantes a la estructura experimental que otros.

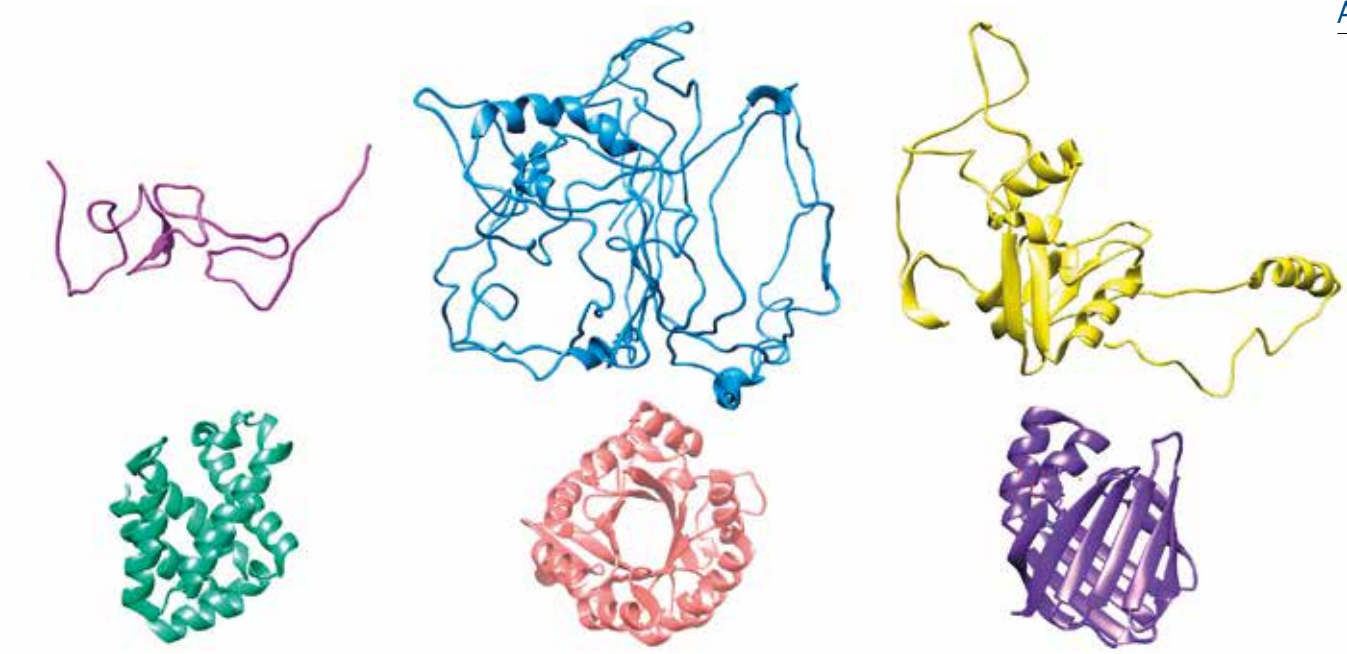

Figura 3. En la fila superior, ejemplos de proteínas desordenadas, y en la inferior, estructuras de proteínas ordenadas. Las proteínas ordenadas adoptan determinada estructura que les permite cumplir su función, la cual no sería imprescindible para que las desordenadas cumplan con la suya.

bioinformáticas se basan en extraer información de los sistemas biológicos que conocemos. ¿Qué pasaría si descubrimos en la Tierra sistemas biológicos con tales particularidades que hagan inaplicables nuestro conocimiento acumulado? Algo así ocurrió hacia principios de 2000 cuando se descubrieron las primeras proteínas desordenadas (figura 3), proteínas que por carecer de estructuras definidas y que por poseer una composición de aminoácidos muy particular hicieron ineficientes los métodos de predicción computacionales descriptos anteriormente. A pesar de numerosos avances, nuestro conocimiento sobre las proteínas desordenadas parece aún incompleto, lo cual hace ineficaces los métodos de predicción para estudiar la relación entre este desorden y su función biológica.

Avanzando más aún con esta idea, ¿qué pasaría si se descubriera vida en Marte o en algún otro planeta y estos sistemas vivos tuviesen, por ejemplo, proteínas compuestas por 30 aminoácidos en vez de los 20 comúnmente encontrados en las proteínas terrestres conocidas hasta ahora? Otra vez, nuestros métodos fallarían enormemente ya que todo lo que hemos aprendido es sobre la vida que conocemos en la Tierra. Como dice Isaiah Berlin en *Two Concepts of Liberty*: 'Knowledge liberates not by offering us more open possibilities amongst which we can make our choice, but by preserving us from the frustration of attempting the impossible' (El conocimiento libera no por ofrecernos más posibilidades entre las cuales podamos elegir, sino por preservarnos de la frustración de intentar lo imposible).

A diferencia de otras disciplinas científicas como la física y la química, que en principio son universales, la biología como la conocemos en la actualidad, y por ende de la bioinformática, solo es aplicable en nuestro planeta, la Tierra. Si recordamos los sistemáticos fracasos de las predicciones en la antigüedad, debemos concluir que, en bioinformática, primero necesitamos aprender para mejorar nuestras predicciones, reconocer sus límites y delinear su campo de aplicabilidad. CH

LECTURAS SUGERIDAS

CHOTHIA C & LESK AM, 1986, 'The relation between the divergence of sequence and structure in proteins', *The EMBO Journal*, 5 (4): 823-826.

GU, J. & P. BOURNE, 2009, *Structural Bioinformatics*, Nueva Jersey, Wiley-Blackwell.

TOMPA P, 2012, 'Intrinsically disordered proteins: A 10-year recap.', *Trends in Biochemical Sciences*, 37, 12: 509-516.

Cristina Marino Buslje
Doctora en ciencias biológicas, Universidad Autónoma de Barcelona.
Investigadora independiente del Conicet.
Presidenta de la Asociación Argentina de Bioinformática y Biología Computacional.
cmb@leloir.org.ar

Gustavo Parisi
Doctor en ciencias biológicas, UNLP.
Profesor titular, UNQ.
Investigador principal del Conicet.
Vicepresidente de la Asociación Argentina de Bioinformática y Biología Computacional.
gusparisi@gmail.com