11 Bioinformática e Inteligencia Artificial

MATÍAS GERARD

Introducción

Bioinformática e Inteligencia Artificial son disciplinas que han estado acompañadas de grandes avances, principalmente en la última década. Sin embargo, es poco sabido que sus orígenes se remontan a mediados del siglo XX. Por este motivo, se propone hacer un breve recorrido por la historia, para conocer cómo se gestaron estas disciplinas y cómo evolucionaron hasta la actualidad. Repasaremos algunos de los hechos más relevantes, y se describirán algunos desarrollos en los que Inteligencia Artificial y Bioinformática han convergido para dar respuestas a problemas reales.

Bioinformática

Las computadoras y el software especializado se han vuelto una pieza esencial del trabajo diario en biología. Esto es evidente dado que todos los proyectos de investigación requieren, en alguna medida, del uso de computadoras. Por esto es que resulta fácil pensar que la bioinformática surgió recientemente, para ayudar al análisis de los grandes volúmenes de datos generados en la actualidad. Sin embargo, sus inicios se ubican a mediados del siglo XX, cuando las computadoras personales (PCs, por sus siglas en inglés) eran todavía una hipótesis, y el ADN aún no había sido secuenciado. Para ubicarnos en esta historia, podemos mencionar que Hershey y Chase (Hershey y Chase, 1952) probaron que la información genética está codificada en el ADN en 1952, y Watson, Crick y Franklin (Watson y Crick, 1953) descubrieron su estructura de doble hélice un año después. Además, debieron pasar 13 años para descifrar el código genético (Nirenberg y Leder, 1964), y otros 25 años para que el primer método de secuenciación estuviera disponible (Sanger y Nicklen, 1977). Todo esto podemos verlo reflejado en la línea temporal presentada en figura 11.1, donde se resaltan algunos hechos importantes para la historia de la bioinformática, junto a hechos que marcaron la evolución de la inteligencia artificial (que analizaremos más adelante). Como puede verse, dado que son muchos los acontecimientos involucrados en el nacimiento y evolución de la bioinformática, dividiremos su historia en cuatro grandes bloques.

sinc(*i*) Research Institute for Signals, Systems and Computational Intelligence (sinc.unl.edu.ar)
M. Gerard; "Bioinformática e Inteligencia Artificial, Introducción a la Biotecnología. Sus aplicaciones y alcances"
, pp. 127-141, 2022.

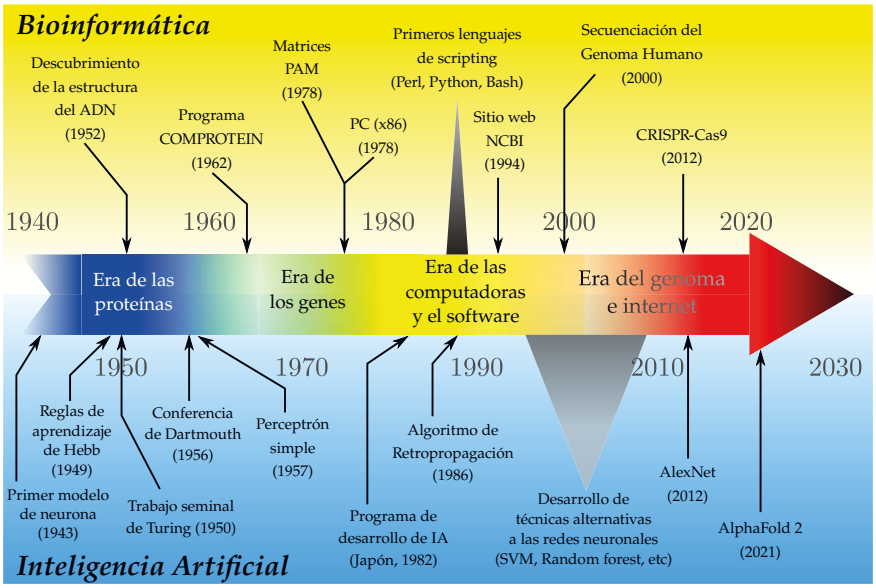

Figura 11.1. Línea temporal con algunos eventos relevantes para bioinformática e inteligencia artificial

La era de las proteínas

No es sencillo definir cuando nació la bioinformática, pero podemos considerar su inicio a finales de los 50, cuando se publicó la primera secuencia de una proteína: la insulina (Sanger y Thompson, 1953). Esto despertó el debate acerca de la relación entre la secuencia y la estructura de las proteínas, acelerando la carrera para desarrollar métodos eficientes para obtener secuencias. Si bien la automatización del método de Edman (Li y Chen, 2014) permitió obtener las secuencias en forma relativamente rápida, su eficiencia se limitaba a secuencias con hasta 50–60 aminoácidos. Por lo tanto, las proteínas de mayor tamaño debían ser fragmentadas para poder emplear esta técnica. Esto creó la necesidad de empalmar las secuencias parciales para conocer la secuencia original. Aunque el empalme manual era posible, la tarea era compleja.

En respuesta a este desafío, Margaret Dayhoff y Robert Ledley desarrollaron el programa COMPROTEIN en 1962, uno de los primeros desarrollos bioinformáticos (Dayhoff y Ledley, 1962). Este ensamblador de secuencias, escrito usando tarjetas perforadas, permitía reconstruir la secuencia completa de forma automática. Originalmente empleaba tres letras para la codificación de los aminoácidos, pero Dayhoff lo simplificó proponiendo la codificación de una letra que se usa en la actualidad. Este avance hizo posible que se creara en 1966 la primera base de datos de secuencias de proteínas (Hersh,

1967), y abrió las puertas para el estudio comparativo de secuencias y el análisis de ancestros comunes. Sin embargo, dos grandes avances fueron claves para lograr esto. Por un lado, el desarrollo de algoritmos eficientes para alineamiento de secuencias, que permitió identificar variaciones en la composición de aminoácidos para proteínas similares (Needleman y Wunsch, 1970). Por el otro, la elaboración del primer modelo probabilístico de substitución de aminoácidos (Dayhoff y Schwartz, 1978) hizo posible cuantificar la probabilidad de que un aminoácido sea reemplazado por otro en un intervalo de tiempo dado. Estos modelos se conocen hoy en día como matrices PAM, y se emplean frecuentemente para la identificación de ancestros.

En paralelo, descubrimientos como la estructura del ADN y su identificación como fuente del transporte de información biológica comenzaron a darle mayor relevancia al papel que cumple esta molécula. Además, cuando Francis Crick propone en 1958 lo que conocemos como el *Dogma Central de la Biología* (ADN→ARN→Proteína) (Lodish *et al.*, 2005), el foco de las investigaciones giró drásticamente hacia el estudio del ADN. Esto llevó a pensar que el simple conocimiento de la secuencia del ADN, de la estructura de los 64 codones, y cómo traducirlos a aminoácidos sería suficiente para desentrañar los misterios de la biología.

La era de los genes

El método de secuenciación de ADN desarrollado por Frederick Sanger (Sanger y Nicklen, 1977) en 1977 fue esencial para el avance en el estudio de los genes y el genoma. Mientras que las proteínas deben purificarse individualmente antes de secuenciarlas, este avance permitía obtener la información completa del organismo de una forma simple. Sin embargo, el análisis de la información presentó mayores desafíos dado que además del volumen de datos producidos, también era necesaria la traducción desde el lenguaje del ADN al de las proteínas. Claramente, antes de estudiar y comparar secuencias de proteínas para diferentes genomas debía realizarse este proceso. Sin embargo, el trabajo con proteínas demostró que el análisis de la información podía realizarse fácilmente empleando computadoras. Así, en 1979 se desarrolló el primer conjunto de herramientas bioinformáticas para manipular secuencias de ADN y buscar solapamientos.

Pese a las posibilidades que brinda la secuenciación, los genes presentan naturalmente una abundancia que es varios órdenes de magnitud menor que las proteínas, por lo que es necesario contar con grandes cantidades de muestra para cualquier tipo de análisis. Fue gracias al desarrollo de la Reacción en Cadena de la Polimerasa (PCR) en 1983 (Mullis y Faloona, 1987) que esta limitación fue superada, ya que permitía incrementar la cantidad de material genético disponible incluso en pequeñas cantidades de muestra. Esto no solo permitió abordar estudios más complejos, sino que también generó una explosión en el volumen de datos de secuencias disponibles.

sinc(*i*) Research Institute for Signals, Systems and Computational Intelligence (sinc.unl.edu.ar)
M. Gerard; "Bioinformática e Inteligencia Artificial, Introducción a la Biotecnología. Sus aplicaciones y alcances"
, pp. 127-141, 2022.

La era de las computadoras y el software

Antes de 1970, una *minicomputadora* tenía las dimensiones y el peso de una heladera. Su tamaño y costo hacían que su adquisición fuera engorrosa para pequeños grupos de trabajo. La primera oleada de *microcomputadoras* llegó al mercado de consumo en 1977. Eran pequeñas, económicas y relativamente fáciles de usar, lo que impulsó el desarrollo de software bioinformático para estos equipos. Incluso, algunos desarrolladores comenzaron a ofrecer copias gratuitas del código bajo demanda, dando relevancia a un movimiento que promovía el intercambio de software en el mundo de la programación. Luego, en 1985 surgió la filosofía del software libre de la mano de Richard Stallman (fundador del movimiento GNU), y fue el núcleo de varias iniciativas para el desarrollo de alternativas libres a herramientas bioinformáticas existentes.

Durante la década de 1980, diferentes laboratorios reconocidos internacionalmente (EMBL, GenBank, DDBJ) se unieron para estandarizar el formato y la información mínima de las secuencias que se debe proporcionar. Fue también durante esta época que la bioinformática se hizo lo suficientemente presente en la ciencia moderna como para tener una revista dedicada (Gauthier *et al.*, 2019). Dada la mayor facilidad de acceso a las computadoras y el enorme potencial para realizar análisis que brindaban, en 1985 se creó una revista especializada en bioinformática: *Computer Applications in the Biosciences*. Esta revista actualmente continúa existiendo bajo el nombre *Bioinformatics*.

Una nueva clase de computadora personal surgió a principios de los '80, con la llegada de los microprocesadores x86. Estas brindaban la posibilidad de disponer de computadoras de propósito múltiple, tanto para aplicaciones técnicas y científicas como para el hogar. Por otro lado, hasta ese momento la escritura de programas estaba dominada por lenguajes compilados como el C y el Fortran. Sin embargo, a mediados de esta década también surgieron varios lenguajes de *scripting* que continúan siendo populares en bioinformática. Estos permitían escribir programas elaborados empleando simples secuencias de comandos, de manera similar a como si se estuviera escribiendo un texto. Dos ejemplos muy conocidos en el ámbito de la bioinformática son Perl y Python. Perl fue creado en 1987 con el objetivo de facilitar el manejo de texto, pero debido a su flexibilidad y a la incorporación de funcionalidades para trabajar en bionformática (librería *BioPerl*) en 1996, fue muy utilizado hasta finales de la década de 2000. Python fue un lenguaje creado en 1989 con el propósito de tener un vocabulario y una sintaxis que facilitaran la lectura y el mantenimiento de los programas. A partir de la incorporación de funcionalidades para el trabajo en bioinformática en el año 2000, este se ha vuelto uno de los lenguajes más importantes en el análisis bioinformático. Como podría esperarse, estas herramientas facilitaron el análisis de secuencias y permitieron realizar estudios cada vez más complejos, tales como el análisis de genomas completos desde principios de los '90.

sinc(*i*) Research Institute for Signals, Systems and Computational Intelligence (sinc.unl.edu.ar)
M. Gerard; "Bioinformática e Inteligencia Artificial, Introducción a la Biotecnología. Sus aplicaciones y alcances"
, pp. 127-141, 2022.

La era del genoma e internet

A pesar de algunos resultados previos, se considera que la era del genoma da inicio oficialmente con la secuenciación del Genoma Humano a principios del siglo XXI. Dos actores fueron parte de este hito. El Instituto Nacional de Salud de Estados Unidos (NIH, por sus siglas en inglés) y la empresa biotecnológica Celera Genomics. Mientras que el NIH requirió 13 años para completar la tarea, Celera Genomics lo hizo en sólo tres años empleando otra estrategia de secuenciación. Pese a esta diferencia, en ambos casos fue necesario el desarrollo de nuevos procedimientos de laboratorio y el diseño de herramientas bioinformáticas especializadas. Estas tecnologías han continuado su evolución, permitiendo que la secuenciación de un genoma humano en 2018 costara alrededor de U$D 1000 y llevara menos de una semana (Gauthier *et al.*, 2019). Este proyecto es uno de los muchos que dio origen a una nueva forma de hacer investigación, que hoy constituyen lo que se conoce como ciencias-*ómicas*. Estas tienen por objetivo estudiar un gran número de moléculas implicadas en el funcionamiento de un organismo. Así, mientras que para la genómica el objeto de estudio es el genoma, la proteómica busca estudiar el conjunto completo de proteínas en un organismo y sus interacciones.

Otro actor importante en esta etapa fue internet, que comenzó a tomar relevancia a principios de los '90. Fue a través de esta red de redes que el proyecto de secuenciación del genoma humano financiado por el NIH puso sus datos a disposición del público (Gauthier *et al.*, 2019). Pronto, esta red se volvió en omnipresente en el mundo científico, facilitando la forma en que se comparte información y software. Esta tecnología permitió que desde 1993 esté disponible la primera base de datos de secuencias de nucleótidos del mundo, la *Biblioteca de Datos de Secuencias de Nucleótidos del EMBL*, y el sitio web del NCBI desde 1994 (incluyendo la herramienta BLAST, que permite realizar alineaciones de secuencias de forma eficiente) (Gauthier *et al.*, 2019). El auge de los recursos web también amplió y simplificó el acceso a las herramientas bioinformáticas, principalmente a través de servidores web mediante interfaces gráficas simples. Esta tendencia ha continuado siendo tan importante, que la revista *Nucleic Acids Research* publica cada año un número especial sobre estas herramientas.

Un aspecto que ha acompañado a estos desarrollos es el aumento exponencial de las secuencias en las bases de datos públicas. De hecho, la comunidad científica ha generado datos que ya superan el nivel de los exabytes (para tener una idea, un exabyte equivale a 1,5 billones de CDs) (Li y Chen, 2014). Eso ha evidenciado la necesidad de disponer de importantes recursos computacionales para almacenar y manipular estos grandes volúmenes de datos y facilitar su acceso. Así es como han surgido nuevas infraestructuras para organizar la información en base a «organismos modelo», como Drosophila,

Saccharomyces y Humanos, que proporcionan recursos unificados y especializados para la comunidad científica que trabaja en estos organismos.

Claramente, los avances en bioinformática han sido siempre en respuesta a los desarrollos logrados en el campo de la biología y de la computación. Tecnologías como CRISPR–Cas9 y la secuenciación de células individuales (*single cell sequencing,* en inglés) prometen ser claves para avanzar en el entendimiento de los sistemas biológicos y sus componentes. Desde el punto de vista computacional, el aumento en la capacidad de cómputo y la accesibilidad creciente a dispositivos con mayor potencia y a menores costos será clave para el modelado de interacciones complejas. Además, la Inteligencia Artificial, un campo floreciente en los últimos tiempos, ha demostrado ser un actor fundamental para los avances actuales y futuros en bioinformática.

Inteligencia artificial

La Inteligencia Artificial (IA) se encuentra en casi cada aspecto de nuestras vidas, y su objetivo es desarrollar programas con la habilidad de aprender y razonar cómo los seres humanos (Poole, Mackworth y Goebel, 1998). Sin embargo, esta disciplina no es nueva y lleva años siendo desarrollada. Aunque no es fácil precisar su origen, el trabajo publicado por Alan Turing en 1950 (Turing, 1950) podría considerarse como el punto de partida. Allí estableció las bases para la creación de máquinas inteligentes y la evaluación de su inteligencia a través del *test de Turing* (Turing, 1950). Este propone que una máquina puede ser considerada inteligente, cuando un humano es incapaz de identificar si está interactuando con otro humano o con dicha máquina. También introdujo otras ideas como el aprendizaje maquinal, los algoritmos genéticos y el aprendizaje por refuerzo (Russell y Norvig, 2010). Estas ideas dieron lugar al desarrollo de la disciplina a través de dos enfoques. El «modelado del contenido» buscaría construir representaciones que permitieran organizar y manipular el conocimiento para resolver diversas tareas. En cambio, el «modelado del continente» buscaría modelar directamente las estructuras que manipulan el conocimiento, como es el caso de las neuronas biológicas.

Modelado del contenido

Ciertamente, el término IA tuvo su aparición por primera vez en la conferencia de Dartmouth en 1956. De ella participaron matemáticos y científicos informáticos destacados, como Marvin Minsky y Claude Shanon. Muchos de

sinc(*i*) Research Institute for Signals, Systems and Computational Intelligence (sinc.unl.edu.ar)
M. Gerard; "Bioinformática e Inteligencia Artificial, Introducción a la Biotecnología. Sus aplicaciones y alcances"
, pp. 127-141, 2022.

ellos estaban interesados en la demostración de teoremas y algoritmos que pudieran ser comprobados mediante computadoras. Las ideas discutidas en este encuentro dieron como resultado casi dos décadas de desarrollos prometedores en IA. Un ejemplo es el *Programa General de Resolución de Problemas* (Haenlein y Kaplan, 2019), capaz de resolver automáticamente el problema de las *Torres de Hanoi*. Este tipo de enfoques suponía que cualquier problema que pudiera describirse mediante un código de programación tenía solución, sin importar si se trataba de la demostración de un teorema matemático o una partida de ajedrez. Su resolución implicaría primero representar el conocimiento mediante algún formato informático legible, y luego se buscaría la solución explorando todos los posibles escenarios. Por ejemplo, en una partida de ajedrez existiría una representación del tablero, las piezas y los posibles movimientos, y la mejor movida se obtendría luego de explorar todas las posibilidades.

Sin embargo, el optimismo que acompañó a los primeros éxitos condujo a predicciones acerca de la IA que resultaron exageradas. Una de las razones que limitaron el éxito de estos «sistemas expertos» es que pretendían simular la inteligencia humana mediante conjuntos de reglas, ya que la resolución de problemas reales no siempre es tan simple. La otra razón es la dificultad para manejar el número de escenarios que deben explorarse en situaciones reales. La famosa superorcomputadora Deep Blue era capaz de jugar al ajedrez construyendo un árbol con los posibles movimientos propios y de su oponente. Luego elegía la respuesta más conveniente a partir de la exploración de las posibles combinaciones. Este proceso debía ser realizado de forma muy rápida y era necesario disponer de grandes cantidades de memoria para almacenar el árbol construido. Si bien el juego de ajedrez puede parecer muy complejo, sólo se trata de un tablero con 64 casilleros y 32 piezas con movimientos predefinidos. En los problemas que se encuentran en la vida real, la cantidad de opciones es mucho mayor, por lo que resulta poco práctica esta aproximación.

Luego de un período de retroceso sufrido por la falta de financiamiento a causa de estos resultados, en 1982 Japón lanzó un plan para estimular el desarrollo integral de la IA, que contemplaba tanto software como hardware (Haenlein y Kaplan, 2019). Esto sirvió como catalizador e hizo que en EE. UU., Europa y el Reino Unido se gestara un movimiento centrado en la construcción de *Sistemas Basados en Conocimiento Inteligente*. Este nuevo enfoque buscaba modelar el conocimiento en un dominio de aplicación específico. Por ejemplo, con estos *nuevos sistemas expertos*, se esperaba que fuera posible realizar el diagnóstico médico de una enfermedad infecciosa incorporando al sistema conocimiento específico. Este podría adquirirse a partir de expertos humanos, ya que a menudo este tipo de conocimiento puede describirse en forma de reglas. Luego, un software de inferencia podía utilizar este conocimiento para extraer conclusiones y realizar un diagnóstico. Sin embargo,

a pesar de estos prometedores desarrollos, los sistemas expertos todavía carecen del sentido común que los seres humanos adquirimos desde el día en que nacemos. Por tal motivo, resulta necesario continuar investigando nuevos mecanismos para modelar adecuadamente estas características humanas.

Modelado del continente

El campo de la IA no sólo ha buscado construir programas capaces de manipular representaciones simbólicas para resolver una tarea. La inteligencia computacional es una disciplina que ha emergido desde la IA trayendo nuevos modelos capaces de adaptarse automáticamente a partir de los datos, utilizando estrategias más relacionadas con los métodos numéricos que con el procesamiento simbólico. Dicha capacidad le permite a estos sistemas aprender directamente a partir de los datos, construir reglas de inferencia y generalizar el conocimiento. Así, esta disciplina busca modelar explícitamente los elementos que contienen al conocimiento mismo.

De entre las primeras contribuciones en esta rama se pueden destacar los trabajos de McCulloch y Pitts, quienes propusieron en 1943 el primer modelo de una neurona artificial. También el de Donal Hebb, quien demostró en 1949 la existencia de un mecanismo sencillo responsable de actualizar la fuerza de conexión entre neuronas biológicas (Russell y Norvig, 2010). Este proceso, hoy conocido como *aprendizaje hebbiano*, establece que la sinapsis entre dos neuronas se ve reforzada cuando ambas interactúan en respuesta a un estímulo. Tomando estas ideas, fue Frank Rosenblatt quien luego propuso en 1957 el modelo del perceptrón, una neurona artificial capaz de aprender automáticamente a partir de datos (Rosenblatt, 1957). Este modelo constituye el bloque de construcción para la mayor parte de las redes neuronales de la actualidad.

A pesar de estos avances, las neuronas artificiales y las investigaciones asociadas fueron dejadas de lado por mucho tiempo debido al trabajo publicado en 1969 por Minsky y Papert (Minsky y Papert, 1969). Este demostraba que aunque estas neuronas eran capaces de aprender, sólo podían resolver problemas simples y sin aplicación práctica.

Un desarrollo clave realizado en 1986 reavivó el interés por los modelos neuronales. Aunque se habían hecho avances para construir redes de neuronas interconectadas, todavía no existía un mecanismo sencillo para entrenarlas. Así, fue la propuesta del algoritmo de retropropagación realizada por Rumelhart, Hinton y Williams lo que permitió el desarrollo de las redes neuronales (Rumelhart, Hinton y Williams, 1986). Este nuevo verano experimentado por los modelos neuronales llegó a su fin a finales del siglo XX, debido a diversas dificultades encontradas para entrenar modelos de gran tamaño y complejidad. Sin embargo, esto permitió el desarrollo de otras técnicas que

sinc(*i*) Research Institute for Signals, Systems and Computational Intelligence (sinc.unl.edu.ar)
M. Gerard; "Bioinformática e Inteligencia Artificial, Introducción a la Biotecnología. Sus aplicaciones y alcances"
, pp. 127-141, 2022.

brindaron mejores desempeños (Engelbrecht, 2007), tales como las Máquinas de Soporte Vectorial (SVM, por sus siglas en inglés) y los métodos de ensambles de clasificadores, donde Random Forest es el más conocido.

En 2012 el equipo de Geoffrey Hinton desarrolló AlexNet (Krizhevsky, Sutskever y Hinton, 2017), un modelo con ocho capas de neuronas capaz de resolver tareas de clasificación de imágenes. Con este modelo participaron de ImageNet, una competencia internacional de clasificación de imágenes, consiguiendo un desempeño sorprendente. AlexNet logró una precisión de aproximadamente 85 %, superando en casi 10 % a los demás participantes. Esto también marcó un punto de quiebre respecto de los desempeños alcanzados durante la década previa. Otro aspecto novedoso fue el entrenamiento del modelo, el cual se realizó utilizando placas gráficas (comúnmente conocidas como GPU, similares a las usadas en las PC para jugar juegos en alta definición). Esto permitió reducir drásticamente los tiempos de entrenamiento de *AlexNet*. En una comparación realizada por Ciresan (Cireşan *et al.*, 2011), donde entrenaron un modelo similar a *AlexNet* empleando una computadora personal y una GPU, mostraron que el entrenamiento era 60 veces más rápido con el uso de estas placas. Este hito despertó una vez más el interés en los modelos neuronales, el cual continúa hasta la actualidad.

El desarrollo que ha experimentado la tecnología de cálculo detrás de las GPU y la aparición de herramientas como TensorFlow y PyTorch, que permiten programar estos modelos usando el lenguaje Python, han hecho que la construcción y entrenamiento de modelos neuronales sea cada vez más simple. Esto sumado a los logros alcanzados con estas «redes profundas» (nombre que reciben debido a que poseen más de tres capas de neuronas) ha llevado a que sean cada vez más aplicadas en todos los campos de la ciencia, e incluso en la vida cotidiana. Un claro ejemplo es la capacidad que poseen actualmente las máquinas para entender comandos de voz, traducir entre idiomas en tiempo real, o la habilidad que han desarrollado algunos programas para jugar juegos igual o mejor que un ser humano. Incluso se ven demostraciones de cómo las redes profundas son capaces de reemplazar en una fotografía o video el rostro de una persona por el de otra (famosos *deepfakes*).

Claramente, el éxito de las redes profundas está llevado a que la investigación en IA se vuelque principalmente hacia este tipo de técnicas de aprendizaje automático. Sin embargo, aunque hasta ahora se ha hecho énfasis en las redes neuronales, debe recordarse que la IA es mucho más amplia e involucra un gran número de áreas que no se han abordado en estas páginas. Del mismo modo que ha ocurrido antes, es posible que los próximos avances vengan de la mano de otras áreas de la IA, por lo que no deben ser descuidadas. Importantes técnicas como la lógica difusa, los algoritmos evolutivos y los algoritmos de inteligencia colectiva también forman parte de lo que conocemos como IA y han contribuido en gran medida al desarrollo de la disciplina y a la resolución de problemas. De este modo, dejo al lector la tarea de explorar este fascinante mundo de la inteligencia computacional.

sinc(*i*) Research Institute for Signals, Systems and Computational Intelligence (sinc.unl.edu.ar)
M. Gerard; "Bioinformática e Inteligencia Artificial, Introducción a la Biotecnología. Sus aplicaciones y alcances"
, pp. 127-141, 2022.

Aprendizaje automático

El aprendizaje automático es una de las áreas de investigación dentro de la IA que ha tomado gran relevancia en las últimas décadas. Tiene como objetivo el desarrollo de programas capaces de *aprender* sin la necesidad de ser explícitamente programados para la tarea. En muchos casos, estos programas suelen estar inspirados en la biología y buscan replicar su funcionamiento. Por ejemplo, algunas redes neuronales artificiales emulan el funcionamiento de las redes biológicas aprendiendo a realizar tareas de manera similar a como aprende un niño.

Los desarrollos llevados a cabo en el contexto del aprendizaje automático suelen basarse en tres enfoques: aprendizaje supervisado, aprendizaje no supervisado y aprendizaje por refuerzo. Estas pueden ser consideradas como las más generales, pero existen otras estrategias que combinan diferentes características. Además, si bien estas estrategias de aprendizaje suelen asociarse principalmente a las redes neuronales, las mismas son generales y pueden ser aplicadas también con otro tipo de modelos. A modo de resumen, en la figura 11.2 podemos ver un esquema con las tres estrategias y algunos ejemplos de uso, así como un rápido pantallazo mostrando cómo el aprendizaje automático está inserto dentro del campo de la Inteligencia Artificial.

El aprendizaje supervisado es una estrategia que emula la forma en que un niño aprende a partir de ejemplos. Al realizar una tarea, el niño compara su resultado con el intenta imitar, y aprende de los errores que comete al intentar completar la tarea. Siguiendo esta idea, una red neuronal puede ser entrenada para que aprenda a establecer relaciones (encontrar una función matemática) a partir de los datos disponibles. En este caso los datos provistos son pares (dato de entrada, resultado deseado) entre los que se desea aprender una relación, y la función corresponde a un valor numérico (problemas de regresión) o una etiqueta de clase (problemas de clasificación). El objetivo perseguido con esto es lograr que la red neuronal generalice el conocimiento adquirido durante su entrenamiento, y sea capaz de generar una respuesta correcta frente a datos no vistos previamente.

El aprendizaje no supervisado implica un enfoque diferente, en el que el objetivo es descubrir relaciones en los datos, pero sin disponer de un conocimiento previo y sin información acerca de la correctitud de las soluciones. Entre los métodos que emplean este mecanismo de aprendizaje se encuentran los algoritmos de agrupamiento (*clustering*, en inglés), capaces de agrupar los datos de acuerdo con algún criterio y poner en evidencia

relaciones entre ellos, y la reducción de dimensiones, que puede emplearse para hacer compresión de datos.

El aprendizaje por refuerzo es otra área del aprendizaje automático cuyo objetivo consiste en determinar qué acciones debe escoger un agente de software en un entorno dado con el fin de maximizar algún tipo de recompensa. Podemos pensar en un agente como un programa capaz de percibir su entorno mediante sensores y realizar acciones en base a los estímulos recibidos. Así, mediante esta forma de aprendizaje, el agente puede aprender directamente de la experiencia adquirida durante la interacción con el entorno.

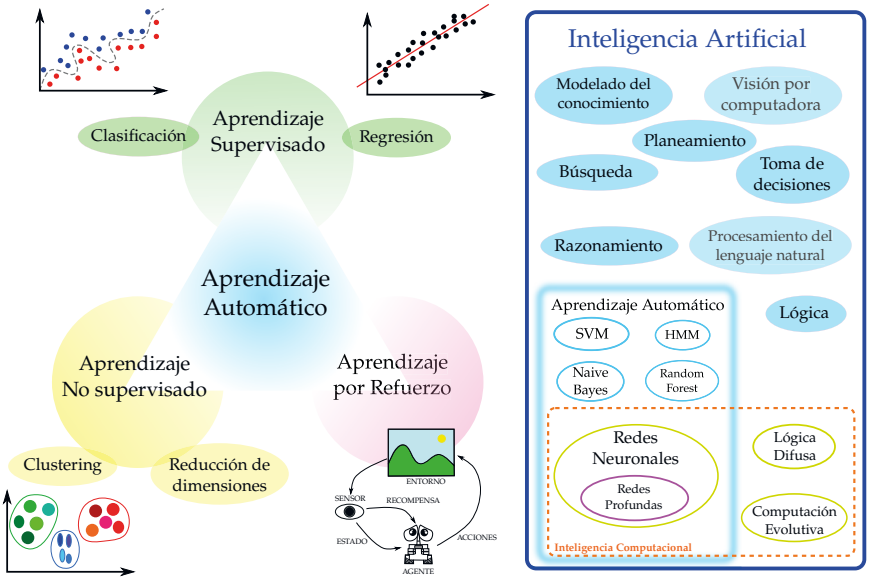

Figura 11.2. Aprendizaje automático: (izquierda) estrategias de aprendizaje; (derecha) algoritmos representativos

La posibilidad de aprender de manera automática, y la capacidad para descubrir patrones ocultos justifican los sorprendentes resultados proporcionados en los últimos años por estos métodos, y en particular por los métodos que emplean redes profundas (aprendizaje profundo). A pesar de esto, los sistemas de aprendizaje automático siguen estando todavía limitados a la resolución de tareas en dominios específicos. Afortunadamente, cada día se avanza más en la integración de estos desarrollos hacia una Inteligencia Artificial General, capaz de aprender y razonar para resolver tareas complejas en múltiples escenarios, e interactuar con los seres humanos como iguales.

Bioinformática y aprendizaje automático

El aprendizaje automático se ha vuelto un aliado de suma utilidad en tareas de análisis, interpretación y descubrimiento de conocimiento en grandes volúmenes de datos, permitiendo la resolución de una amplia variedad de problemas bioinformáticos. Un ejemplo reciente es *AlphaFold 2* (Jumper *et al.*, 2021), un sistema desarrollado en 2021 por la división de IA de Google. Este se basa en un modelo de aprendizaje profundo, y busca predecir cómo se plegará una proteína partiendo de la secuencia de aminoácidos. El modelo fue entrenado utilizando el equivalente a 200 GPUs durante algunas semanas, y alrededor de 170 000 estructuras de proteínas conocidas. Como resultado este modelo es capaz de predecir con una elevada precisión (cercana a 90 %) las estructuras a partir de sus secuencias.

La Universidad Nacional del Litoral (UNL) también cuenta con un grupo de investigación en bioinformática. Desde hace tiempo que el Instituto de Investigación en Señales, Sistemas e Inteligencia Computacional (sinc(i)), dependiente de UNL y CONICET, realiza investigaciones de vanguardia en los campos de la inteligencia computacional y la bioinformática. Si bien el listado completo de los algoritmos, trabajos científicos y herramientas se encuentran disponibles en la página web, se describirán brevemente algunos de estos desarrollos.

omeSOM (Milone; Stegmayer;...; Carrari, 2010), por ejemplo, es una herramienta basada en redes neuronales que utiliza aprendizaje no supervisado para encontrar relaciones entre genes y metabolitos (por ejemplo, glucosa). Esto permite agrupar genes y metabolitos que presentan un comportamiento similar y coordinado, facilitando la identificación de aquellos que participan de un mismo proceso biológico. Otro ejemplo es DL4Papers (Bugnon; Yones;...; y Stegmayer, 2020), una herramienta basada en aprendizaje profundo, capaz de analizar documentos científicos e identificar relaciones entre palabras clave. Esta herramienta puede ser de utilidad en el campo de la salud, dado que es capaz de analizar de forma automática un gran volumen de artículos

sinc(*i*) Research Institute for Signals, Systems and Computational Intelligence (sinc.unl.edu.ar)
M. Gerard; "Bioinformática e Inteligencia Artificial, Introducción a la Biotecnología. Sus aplicaciones y alcances"
, pp. 127-141, 2022.

de la literatura médica. A partir de este análisis se puede generar un ranking indicando, por ejemplo, aquellos trabajos en los que se describen tratamientos que muestran resultados prometedores para algún tipo específico de cáncer.

Además de los modelos basados en redes neuronales, también se desarrollan herramientas basadas en otras ramas de la inteligencia computacional. EvoMS (Gerard; Stegmayer y Milone, 2016), por ejemplo, está basado en computación evolutiva. Esta herramienta emplea una población de soluciones y el principio de supervivencia del más apto para buscar vías metabólicas que relacionen simultáneamente múltiples metabolitos. PhDseeker (Gerard; Stegmayer y Milone, 2018) es otra herramienta bioinspirada que se basa en el comportamiento de las hormigas para diseñar vías metabólicas lineales y ramificadas.

Comentarios finales

La aparición de las computadoras en el siglo XX, sumado a la mejora continua de las tecnologías de laboratorio han permitido llevar adelante investigaciones cada vez más desafiantes en el campo de la biología. Por otra parte, la IA ha tomado un papel preponderante en los desarrollos bioinformáticos de los últimos años, contribuyendo a la integración de diversas fuentes de información y al entendimiento de relaciones de gran complejidad. Claramente, esta conjunción ha permitido que la biología adopte un enfoque más holístico hacia la revolución de las ciencias «-ómicas», aunque todavía con escasa interrelación entre ellas. En base a esto es posible anticipar el siguiente paso: en lugar de investigar de forma independiente genomas, transcriptomas o metabolomas completos, el desafío será considerar las interacciones buscando así modelar computacionalmente organismos vivos enteros y sus entornos, teniendo en cuenta todas las partes de forma simultánea. Es ahí donde la IA y los nuevos modelos que se desarrollen jugarán un papel clave para la bioinformática.

sinc(*i*) Research Institute for Signals, Systems and Computational Intelligence (sinc.unl.edu.ar)
M. Gerard; "Bioinformática e Inteligencia Artificial, Introducción a la Biotecnología. Sus aplicaciones y alcances"
, pp. 127-141, 2022.

Referencias bibliográficas

Bugnon, L.; Yones, C.; ...; Stegmayer, G. (2020). DL4papers: a deep learning approach for the automatic interpretation of scientific articles. *Bioinformatics*, 36(11), 3499–3506.

Cireşan, D.; Meier, U.;...; Schmidhuber, J. (2011). Flexible, High Performance Convolutional Neural Networks for Image Classification. En *Proceedings of the Twenty–Second international joint conference on Artificial Intelligence* (pp. 1237–1242). Volume Two (IJCAI'11). AAAI Press.

Dayhoff, M.; Ledley, R. (1962). COMPROTEIN, a Computer Program to Aid Primary Protein Structure Determination. International Workshop on Managing Requirements Knowledge 1 262.

Dayhoff, M.; Schwartz, R. (1978). Chapter 22: A model of evolutionary change in proteins. En *Atlas of Protein Sequence and Structure* (pp. 345–352). National Biomedical Research Foundation.

Engelbrecht, A. (2007). *Computational Intelligence. An Introduction*. Wiley (2nd Ed).

Gauthier, J.; Vincent, A.;...; Derome, N. (2019). A brief history of bioinformatics. *Briefings in Bioinformatics*, 20(6), 1981–1996.

Gerard, M.; Stegmayer, G.; Milone, D.H. (2016). Evolutionary algorithm for metabolic pathways synthesis. *Biosystems*, 144, 55–67.

Gerard, M.; Stegmayer, G.; Milone, D.H. (2018). Metabolic pathways synthesis based on ant colony optimization. *Scientific Reports*, 8(1), 16398.

Haenlein, M.; Kaplan, A. (2019). A brief history of artificial intelligence: On the past, present, and future of artificial intelligence. *California Management Review*, 61(4), 5–14.

Hersh, R. (1967). Atlas of Protein Sequence and Structure, 1966. *Systematic Biology*, 16(3), 262–263.

Hershey, A.; Chase, M. (1952). Independent functions of viral protein and nucleic acid in growth of bacteriophage. *Journal of General Physiology*, 36(1), 39–56.

Jumper, J.; Evans, R.;...; Hassabis, D. (2021). Highly accurate protein structure prediction with AlphaFold. Nature 596(7873) págs 583–589.

Krizhevsky, A.; Sutskever, I.; Hinton, G. (2017). Imagenet classification with deep convolutional neural networks. *Communications of the ACM*, 60(6), 84–90.

Li, Y.; Chen, L. (2014). Big Biological Data: Challenges and opportunities. *Genomics, Proteomics & Bioinformatics*, 12(5), 187–189.

sinc(*i*) Research Institute for Signals, Systems and Computational Intelligence (sinc.unl.edu.ar)
M. Gerard; "Bioinformática e Inteligencia Artificial, Introducción a la Biotecnología. Sus aplicaciones y alcances"
, pp. 127-141, 2022.

LODISH, H.; BERK, A.;...; DARNELL, J. (2005). *Biología Celular y Molecular*. Editorial Médica Panamericana (5 Ed.)

MILONE, D.H.; STEGMAYER, G.;...CARRARI, F. (2010). *omeSOM: a software for clustering and visualization of transcriptional and metabolite data mined from interspecific crosses of crop plants. *BMC Bioinformatics*, 11(1), 438, 1–10.

MINSKY, M.; PAPERT, S. (1969). *Perceptrons: An Introduction to Computational Geometry*. MIT Press.

MULLIS, K.; FALOONA, F. (1987) [21] Specific synthesis of DNA in vitro via a polymerase–catalyzed chain reaction. En *Recombinant DNA* (pp. 335–350). Part F. Academic Press.

NEEDLEMAN, S.; WUNSCH, C. (1970). A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of Molecular Biology*, 48(3), 443–453.

NIRENBERG, M.; LEDER, P. (1964). RNA codewords and protein synthesis. The effect of trinucleotides upon the binding of sRNA to ribosomes. *Science*, 145(3639), 1399–1407.

POOLE, D.; MACKWORTH, A.; GOEBEL, R. (1998). *Computational Intelligence: A Logical Approach*. Oxford University Press.

ROSENBLATT, F. (1957). The Perceptron – A Perceiving and Recognizing Automaton. Report 85–460–1, project PARA, Cornell Aeronautical Laboratory.

RUMELHART, D.; HINTON, G.; WILLIAMS, R. (1986). Learning representations by back–propagating errors. *Nature*, 323(6088), 533–536.

RUSSELL, S.; NORVIG, P. (2010). *Artificial Intelligence: A Modern Approach*. Prentice Hall (2nd Ed.)

SANGER, F.; THOMPSON, E. (1953). The amino–acid sequence in the glycyl chain of insulin. 1. The identification of lower peptides from partial hydrolysates. *Biochemical Journal*, 53(3), 353–366.

SANGER, F.; NICKLEN, S.; COULSON, A. (1977). DNA sequencing with chain–terminating inhibitors. *PNAS*, 74(12), 5463–5467.

TURING, A. (1950). I.–Computing Machinery And Intelligence. *Mind*, LIX(236), 433–460.

WATSON, J.; CRICK, F. (1953). Molecular Structure of Nucleic Acids: A Structure for Deoxyribose. Nucleic Acid. *Nature*, 171(4356), 737–738.

sinc(*i*) Research Institute for Signals, Systems and Computational Intelligence (sinc.unl.edu.ar)
M. Gerard; "Bioinformática e Inteligencia Artificial, Introducción a la Biotecnología. Sus aplicaciones y alcances"
, pp. 127-141, 2022.