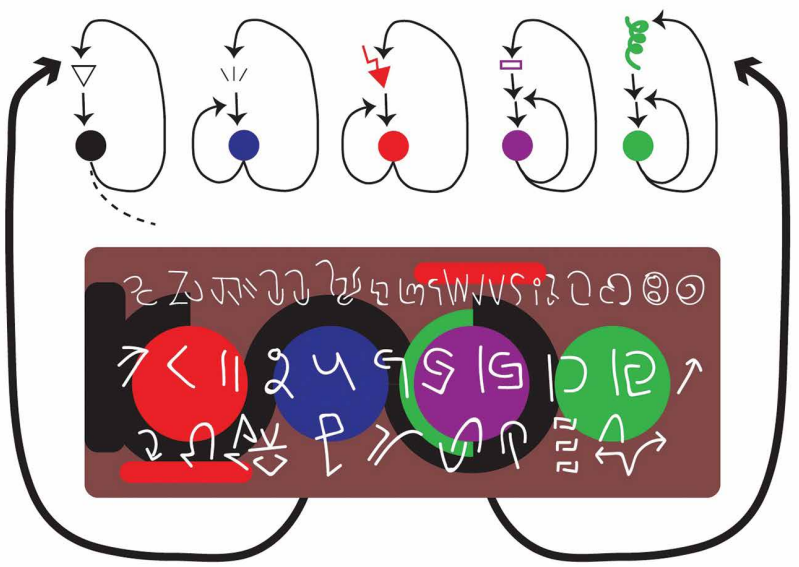
La Vida en Contexto

Doctor Julio Collado Vides

Scientia

LA VIDA EN CONTEXTO

¿Cómo podemos determinar conexiones científicas importantes cuando hay una avalancha de publicaciones nuevas cada día? El Dr. Julio Collado Vides y su equipo de la Universidad Nacional Autónoma de México parecen tener la respuesta.



Ya sean letras o ideogramas, señales viales estandarizadas o caritas sonrientes en mensajes instantáneos, la comunicación constituye una de las herramientas más importantes que tenemos los seres humanos. La capacidad de compartir información, gracias al lenguaje, nos ha llevado desde coordinar las primeras actividades de caza en la sabana hasta difundir los chismes de la familia real en la actualidad.

Con el lenguaje podemos transmitir información, pero el simple uso de las palabras no es suficiente. Con tan solo decir 'pollo', no se indica al interlocutor si hay un pollo viéndolo, escondido tras los arbustos, cocinado y listo para comerse, o detrás de él y robándose su comida. Las palabras deben enlazarse entre sí para proveer un contexto, empleando una serie de reglas que en conjunto constituyen nuestra gramática. Aunque la gramática nos hace sufrir en la escuela, los hablantes nativos de un idioma por lo general la entienden tan bien que incluso pueden identificar errores en oraciones que carecen de sentido (como el clásico ejemplo de Chomsky: 'Las ideas verdes incoloras duermen furiosamente', una oración que no tiene sentido, pero cuya gramática es aceptable).

Las reglas comunes y el contexto desempeñan una función esencial para comunicar la información. Esto no se limita al habla humana, sino que ocurre en una multitud de otros procesos que dependen del intercambio de información. El más fundamental de ellos, en los mismos cimientos de la vida, es el del ADN. Las hebras de ADN codifican toda la información necesaria para la vida, por medio de un alfabeto de cuatro letras (C, G, T y A) que forma alrededor de veinte sílabas (los aminoácidos y los codones de paro); a pesar de que la selección es limitada, las combinaciones de estos elementos componen todas las palabras (proteínas) que se requieren para sustentar la vida.

No obstante, como mencionamos anteriormente, las palabras no son nada si se desconoce el contexto. Los científicos que secuenciaron el genoma humano completo se sorprendieron al descubrir que éste contenía una cantidad de genes mucho menor a la esperada. De este modo, la diferencia entre, digamos, una célula de la piel y una neurona se debe a variaciones en el contexto: unos genes se activan en ciertas circunstancias, otros se activan en otras, mientras que algunos se recortan y se rearreglan durante la producción de proteínas para crear nuevas variantes empalmadas. Todos estos cambios surgen como resultado de las instrucciones genéticas presentes en la hebra de ADN, cada una de las cuales contribuye con parte del contexto necesario para entender correctamente la 'palabra' genética. Entonces, la pregunta es: ¿podemos tratar al ADN como si fuera un lenguaje, dado que tiene palabras y contexto?

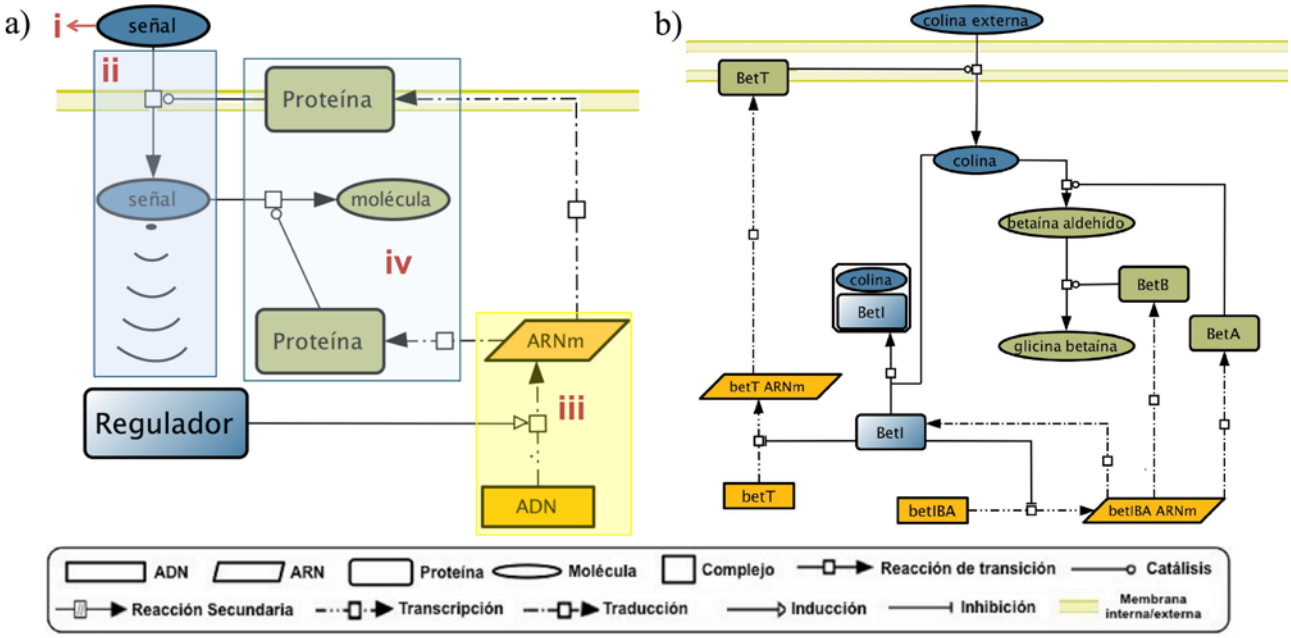
El Idioma del ADN

Intentar responder esa pregunta ha sido durante mucho tiempo una de las metas del Dr. Julio Collado Vides, de la Universidad Nacional Autónoma de México. Sus investigaciones iniciales giraron en torno a la función de los factores de transcripción en el control genético. Los factores de transcripción controlan el primer paso en la producción de proteínas: la transcripción de un gen de ADN en una copia efímera de ARN. Al unirse al ADN, ya sea en un gen o en sus inmediaciones, tienen la capacidad de controlar si la maquinaria de transcripción del ARN puede interactuar con el gen, y a qué grado. Aunque esto plantea un increíble grado de complejidad en la práctica, la mayoría de los factores de transcripción pueden clasificarse como *represores* (reducen la probabilidad de que se transcriba un gen) o *activadores* (aumentan la probabilidad).

La labor temprana del Dr. Collado Vides también involucró modelar la organización de los factores de transcripción y la regulación de los genes como si se tratara de una gramática genética. Al aprovechar la considerable cantidad de investigaciones sobre la gramática lingüística, pudo desarrollar un modelo de 'gramática de factores de transcripción', una serie de reglas que permiten identificar nuevos sitios de unión de factores de transcripción con una especificidad mucho mayor que la lograda hasta entonces.

Ese fue el punto de partida para la investigación que ha desarrollado por más de veinte años en el creciente campo de la bioinformática (el uso de computadoras y modelos matemáticos para entender los

‘El sueño es facilitar la exploración de grandes colecciones de datos, información y conocimientos’



(a) El concepto genérico de Unidad GENSOR incluye 4 componentes: (i) la señal, (ii) la transducción de la señal mostrada en azul, (iii) el switch genético en amarillo y (iv) la respuesta mostrada en verde. (b) Unidad GENSOR de BetI. Ejemplo de un GENSOR basado en un regulador transcripcional microbiano.

sistemas biológicos). Este enfoque sintáctico ayudó a predecir e identificar sitios de unión de factores de transcripción en los primeros días de la genómica, con la secuenciación del genoma completo de la bacteria *E. coli*. El Dr. Collado Vides sigue muy involucrado en el área de la bioinformática, y recientemente se ha centrado en ampliar la gramática genética inicial a un conjunto de datos más complejo: las unidades GENSOR.

La Esquematización de las Unidades GENSOR

Los factores de transcripción representan una parte de la maquinaria molecular que se ha estudiado a profundidad, desde el descubrimiento del factor de transcripción bacterial LacI en 1959. LacI es un represor: se une a una hebra de ADN y evita que se transcriban los genes cercanos. El represor se separa de la hebra de ADN sólo si se presentan las condiciones ambientales correctas, y de este modo actúa como enlace entre el ambiente y la respuesta genética.

En el caso de LacI, una señal ambiental (lactosa) se transduce en una señal que puede afectar directamente al regulador genético (en este ejemplo, al transformar la lactosa en alolactosa), que a su vez es responsable de efectuar un cambio en el ambiente genético (la expresión de los genes que metabolizan lactosa), lo cual genera una respuesta (la célula ahora puede metabolizar eficazmente la lactosa).

Los investigadores pueden enlazar esta combinación de elementos (señal, transducción de la señal, cambio genético y respuesta) y considerarla una sola unidad de mayor nivel. El Dr. Collado Vides y su equipo la bautizaron con el término ‘unidad de respuesta sensorial genética’, o unidad GENSOR, para abreviar. Las unidades GENSOR ayudan a registrar con rapidez la increíble complejidad de la señalización celular, que puede involucrar a decenas o cientos de componentes individuales, y permiten crear redes de flujo de información comparativamente sencillas. El Dr. Collado Vides explica: ‘Considero que las unidades GENSOR constituyen un nivel de integración superior por encima de conceptos como operón y regulón, una herramienta para comprender el análisis de experimentos genómicos globales y un primer paso para describir de forma sistemática los flujos de información regulada’.

Él y sus colaboradores intentaron elaborar un esquema integral de las unidades GENOSR de la bacteria *Escherichia coli*, famosa en todo el mundo por ser el caballo de batalla en los laboratorios de microbiología. En sus esfuerzos previos lograron recopilar 189 factores de transcripción, cada uno con blancos distintos, que responden a diferentes señales y generan efectos diversos. Ahora, el desafío consiste en agrupar estos factores de transcripción en una red clara y práctica.

El grupo utilizó un método que aprovechaba una serie de bases de datos distintas, a partir de su experiencia en enfoques bioinformáticos. Usaron proteínas de factores de transcripción como punto de partida, así como búsquedas automatizadas cada vez más amplias en bases de datos para encontrar proteínas con interacciones, genes conocidos controlados por el proceso y el efecto celular de esos genes. Posteriormente, filtraron la colección de los datos que obtuvieron para identificar los lazos más importantes. Se crearon 189 elementos GENSOR en total. Con base en nuestro conocimiento actual, sólo 89 constituyen sistemas genéticos completos de sensor y respuesta, mientras que en otros casos, este enfoque ayuda a predecir algunos de los componentes faltantes.

Se encontró que la mayoría de estas unidades GENSOR eran controladas por alguna forma de ciclo de retroalimentación; es decir, la actividad de la unidad GENSOR se ve afectada por su propia producción. Estos ciclos son predominantemente

simples, aunque se observó que algunos factores de transcripción eran controlados por medio de cambios metabólicos de varios pasos. Junto a estos ciclos de retroalimentación había una serie de factores metabólicos diferentes que controlaban directamente a la unidad GENSOR o eran controlados directamente por ésta. En otras palabras, una sola unidad GENSOR puede controlar la producción de varias moléculas diferentes, pero solamente se ve afectada por una de ellas.

No obstante, la verdadera ventaja de las unidades GENSOR consiste en la capacidad de fusionar varias unidades individuales en una red mayor. Por ejemplo, *E. coli* tiene una preferencia determinada respecto a la elección del carbohidrato que usará como fuente de alimento. Primero utiliza la glucosa, luego la lactosa, y sólo después de eso emplea otras moléculas, como la arabinosa y la xilosa. Varios factores de transcripción (y por ende, unidades GENSOR) están involucrados en el metabolismo de los carbohidratos, y las interacciones de estos pueden usarse para modelar el comportamiento de las bacterias ante cualquier combinación de deliciosos azúcares comestibles.

Antes de la era genómica, los microbiólogos se dedicaban a estudiar capacidades definidas de las células, como la degradación del carbono, la asimilación del nitrógeno, la división celular o las respuestas ante diferentes tipos de estrés ambiental. Los factores de transcripción se nombraron de acuerdo con esas capacidades celulares. Con el surgimiento de la genómica, ahora sabemos que hay una gran integración celular. Como señala el Dr. Collado, 'no hay una función sensorial elemental', porque la mayoría de las moléculas provocan múltiples repercusiones en la célula. La unidad GENSOR es un concepto adecuado para describir en diagramas explícitos dentro de bases de datos las interconexiones que dan pie a esta fisiología integrativa.

El Bibliotecario Artificial

Lo que funciona para la gramática genética también puede ser de utilidad en el ámbito del lenguaje normal, de modo que tal vez resulte natural que el equipo del Dr. Collado Vides también aplique sus conocimientos a la comunicación y organización del conocimiento científico. Durante mucho tiempo, el grupo ha recopilado y curado manualmente artículos sobre la regulación genética en bacterias, y su trabajo es el fundamento de varias bases de datos abiertas (por ejemplo RegulonDB) que los científicos utilizan con frecuencia para determinar el estado actual del conocimiento (ver http://regulondb.ccg.unam.mx/). Sin embargo, la curación manual tiene varias limitaciones: leer artículos consume mucho tiempo, se requiere el esfuerzo de expertos para registrar los hallazgos de forma adecuada y la información que puede buscarse se encuentra limitada por el formato de la base de datos.

Los bioinformáticos, por supuesto, son expertos en extraer información de fuentes de datos grandes y complejas. El grupo se preguntó si podría usar el aprendizaje automático y la curación basada en computadoras para crear una nueva base de datos de regulación genética con un alto grado de entrecruzamiento. 'El sueño', menciona el Dr. Collado Vides, 'es generar métodos que ejerzan un impacto en todo el dominio de quienes se dediquen a recopilar, organizar, integrar y facilitar la exploración de grandes corpus de datos, información y conocimientos de las ciencias biomédicas'.

El primer paso para cristalizar ese sueño es integrar la minería de datos y el análisis textual en varias de las bases de datos en las que participa el equipo. Se han desarrollado varias herramientas que se especializan en extraer información de publicaciones de las ciencias de la vida empleando el conocimiento de cómo se estructuran las oraciones en inglés para obtener un resumen de los resultados y los vínculos que se han demostrado. Esas herramientas se diseñaron con algoritmos de similitud que determinan palabras 'similares'; no es necesario que una búsqueda arroje el término exacto para ser relevante, sino que basta con una coincidencia cercana para que sea correcta. Posteriormente, se integraron a un sistema con interfaz que permite a los usuarios ver y decidir cuál es el conocimiento correcto con base en las propuestas que sugiere la computadora: resalta las palabras y las oraciones que considere que están involucradas en la regulación genética y muestra cómo se enlazan unas con otras.

El resultado final de toda esta labor es una base de datos inteligente que los curadores pueden utilizar. Conforme leen un artículo, el sistema les proporciona de manera automática enlaces a otras publicaciones que aborden el mismo tema, los cuales son definidos en su totalidad por la inteligencia artificial detrás de la base de datos. Esto implica que los curadores ya no tienen que dedicar su valioso tiempo a buscar publicaciones o referencias correlacionadas. Aunque el proceso final sigue siendo de naturaleza manual, el equipo de investigación ha determinado que el sistema puede acelerar el proceso de curación, y más importante aún, mejorará la detección de piezas de conocimiento al referirlas a la publicación original y así enriquecerá las bases de datos de manera novedosa.

Divulgación Científica fuera del Ámbito Universitario

Con el propósito de aplicar estas ideas de curación en otra área, el Dr. Collado Vides se embarcó en una aventura sin fines de lucro con la que espera crear una enciclopedia interactiva, en la cual el conocimiento se organice tanto dentro de una ontología (similar a la organización del conocimiento en la Enciclopedia británica) como en diferentes niveles de entendimiento, enlazando los textos para principiantes con aquellos más avanzados. Conogasi (http://conogasi.org/) iniciará actividades en el otoño de 2017.

El Contexto de la Avalancha

En la era de la genómica moderna, con la secuenciación de alto rendimiento y la expansión de la investigación a los laboratorios de todo el mundo, se genera una avalancha de conocimientos genéticos como nunca antes. Sin embargo, la simple cantidad de información evita que la entendamos o que incluso la leamos por completo, pues la mente humana simplemente no está a la altura de la tarea. En lugar de eso, los sistemas de aprendizaje automático nos permiten simplificar el flujo y extraer los enlaces más útiles, ya sea de publicaciones científicas o de bases de datos de regulación genética. Ambas fuentes requieren que entendamos el *contexto* de la información, es decir, los detalles, genéticos o escritos, que nos permiten dar sentido a lo que observamos.

Es justo en la solución de este problema que sobresale el trabajo del Dr. Collado Vides y su grupo. Al ayudar a automatizar el reconocimiento del contexto, nos brindan un medio para controlar y entender la avalancha de información. Esto significa, a su vez, que otros científicos pueden trabajar con eficacia y dedicar tiempo a su verdadera vocación: ampliar el conocimiento humano.

Conoce al investigador

Doctor Julio Collado Vides
Centro de Ciencias Genómicas
Universidad Nacional Autónoma de México
Cuernavaca
México

El Dr. Julio Collado Vides egresó de la maestría en ciencias en Físico Química en 1985 y obtuvo un doctorado en Investigación Biomédica en 1989 de la internacionalmente reconocida Universidad Nacional Autónoma de México. Luego del doctorado, estuvo tres años en una estancia de investigación posdoctoral en el MIT. Actualmente, es profesor de genómica computacional en el Centro de Ciencias Genómicas de la Universidad Nacional Autónoma de México. Con una carrera de investigación en bioinformática y genética que abarca más de dos décadas, ha publicado más de 100 artículos y lo han citado más de 21,000 veces, gracias a lo cual se le reconoce como uno de los científicos más citados del mundo. Ha supervisado a casi 20 estudiantes, es miembro de varios consejos directivos y ha recibido un gran número de reconocimientos que sigue en aumento. Su investigación ha sido un esfuerzo de equipo de los miembros antiguos y actuales de su laboratorio (http://www.ccg.unam.mx/es/ComputationalGenomics).

CONTACTO

Correo: collado@ccg.unam.mx
Teléfono: (+52) 777 313 2063
Web: http://www.ccg.unam.mx/es/personal/julio-collado-vides

COLABORADORES CLAVE

Daniela Ledezma-Tejeida (estudiante de doctorado que trabaja con unidades GENSOR)
David Rosenblueth del IIMAS, UNAM (científico computacional que implementó los programas del modelo gramatical; http://turing.iimas.unam.mx/~drosenbl/)
Jacques van Helden, Université d'Aix-Marseille, Francia (http://jacques.van-helden.perso.luminy.univ-amu.fr/)
Fabio Rinaldi, Instituto Suizo de Bioinformática (http://www.sib.swiss/rinaldi-fabio/rinaldi-fabio-sub)

UNAM

FINANCIAMIENTO

UNAM
CONACyT, México
NIH (NIGMS)

REFERENCIAS

D Ledezma-Tejeida, C Ishida, J Collado-Vides. Genome-wide mapping of transcriptional regulation and metabolism describes information-processing units in Escherichia coli, Frontiers in Microbiology, 8, 1466. DOI: 10.3389/fmicb.2017.01466

S Gama-Castro, H Salgado, A Santos-Zavaleta, D Ledezma-Tejeida, L Muñiz-Rascado, JS García-Sotelo, K Alquicira-Hernández, I Martínez-Flores, L Pannier, JA Castro-Mondragón, A Medina-Rivera, H Solano-Lira, C Bonavides-Martínez, E Pérez-Rueda, S Alquicira-Hernández, L Porrón-Sotelo, A López-Fuentes, A Hernández-Koutoucheva, VD Moral-Chávez, F Rinaldi, J Collado-Vides. RegulonDB version 9.0: high-level integration of gene regulation, coexpression, motif clustering and beyond, Nucleic Acids Research, 2016, 44, D133–43.

S Gama-Castro, F Rinaldi, A López-Fuentes, YI Balderas-Martínez, S Clematide, TR Ellendorff, A Santos-Zavaleta, H Marques-Madeira, J Collado-Vides. Assisted curation of regulatory interactions and growth conditions of OxyR in E. coli K-12, Database (Oxford), 2014, pii: bau049.

J Collado-Vides, Grammatical model of the regulation of gene expression, Proceedings of the National Academy of Sciences of the USA, 1992, 89, 9405–9409.

Este texto es la versión traducida al español de: Julio Collado-Vides (2017) Professor Julio Collado-Vides – Putting Life in Context. Scientia. doi: 10.26320/SCIENTIA26
http://www.scientia.global/professor-julio-collado-vides-putting-life-context/ gracias a nuestra colaboración conjunta.