UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO
DOCTORADO EN CIENCIAS BIOMÉDICAS
CENTRO DE CIENCIAS GENÓMICAS

DESARROLLO DE UNA ONTOLOGÍA PARA LA REPRESENTACIÓN DE CONOCIMIENTO SOBRE LA REGULACIÓN GENÉTICA

TESIS QUE PARA OPTAR POR EL GRADO DE :

DOCTORA EN CIENCIAS

PRESENTA:

CITLALLI MEJÍA ALMONTE

DIRECTOR DE TESIS
DR. JULIO COLLADO VIDES
Centro de Ciencias Genómicas

COMITÉ TUTOR
DR. FABIO RINALDI
Dalle Molle Institute for Artificial Intelligence

DR. ALEXANDER GELBUKH
Centro de Investigación en Computación, IPN

CUERNAVACA, MORELOS, ENERO DE 2021

Pour saisir le monde aujourd'hui, nous usons d'un langage qui fut établi pour le monde d'hier. Et la vie du passé nous semble mieux répondre à notre nature, pour la seule raison qu'elle répond mieux à notre langage.

Para comprender el mundo hoy, usamos un lenguaje creado para el mundo de ayer. Y la vida del pasado parece responder mejor a nuestra naturaleza por la sola razón de que responde mejor a nuestro lenguaje.

Antoine de Saint-Exupéry, Tierra de hombres

DEDICATORIA

Esta tesis está dedicada a mis padres y hermano.

AGRADECIMIENTOS

Agradezco:

A Julio Collado por aceptar ser mi tutor.

Al grupo de expertos que participó en la evaluación de las definiciones propuestas en este trabajo: Steve Busby, Joe Wade, Jacques van Helden, Adam Arkin, Karen Eilbeck, Bernhard Palsson, James Galagan y Julio Collado. Sin su colaboración, este trabajo no habría llegado a un final feliz.

A los miembros de mi comité tutor, Julio Collado, Fabio Rinaldi y Alexander Gelbukh, por el tiempo que semestre a semestre dedicaron a escuchar y criticar mis propuestas. A Fabio Rinaldi en particular por el curso personalizado de bioontologías.

A Chris Mungall, Becky Jackson, James Overton, Jim Balhoff, Nico Matentzoglu y todos en la OBO Foundry, así como a Rafael Gonçalves, Csongor Nyulas, Samson Tu y todos en la mailing list the Protégé que respondieron mis dudas sobre owl, ontologías formales y aspectos técnicos de la integración de ontologías biológicas. A Jean-Baptiste Lamy por resolver mis dudas sobre owlready2, su módulo de Python para crear ontologías programáticamente.

A los miembros del jurado de grado Alicia González, Osbaldo Resendis y de manera especial a Julio Collado, David Romero y Rosa María Gutiérrez por sus sugerencias para mejorar esta tesis.

A Jonatan Soffer por encaminarme en el entendimiento de la lógica y los menesteres de la integración continua y versionamiento con git.

El mero hecho de pertenecer al Laboratorio de Genómica Computacional y estar presente en los seminarios y reuniones es una experiencia de aprendizaje. Es la oportunidad de «subirte en los hombros de [los] gigantes» del laboratorio. En ese sentido, la iniciativa y el esfuerzo del Conocimiento U es fundamental. Cuanto mejor nos demos a entender, mejor podemos ayudar y ser ayudados. Por la oportunidad de participar en el grupo, agradezco a Julio Collado y a todo el laboratorio por compartir el conocimiento de manera U.

Especialmente a Yalbi Balderas, Daniela Ledezma, Alejandra López, Socorro Gama y Heladia Salgado por las pláticas biológicas, ontológicas e informáticas que ayudaron a esta analfabeta informática a comprender de qué se trataba el trabajo del laboratorio. Son mis ejemplos a seguir.

Al equipo de Microbial Conditions Ontology: Manuel Camacho, Víctor Tierrafría, Socorro Gama, Kevin Alquicira y Heladia Salgado, porque el desarrollo de la MCO fue una de las experiencias más educativas que tuve en el laboratorio.

A Victor del Moral y Concepción Hernández Levaro por siempre preocuparse y asegurarse de que los miembros del laboratorio tengamos todo lo que necesitamos y mantener el laboratorio andando, incluso cuando no es su deber. A Romualdo Zayas por el curso de Unix y por subir el volumen del micrófono cuando me tocaba exponer en el auditorio. A Luis José Muñiz por ayudarme a consultar RegulonDB. A César Bonavides por su apoyo técnico y administrativo, y porque su chistes sí me hacen reír.

A Denny Peralta, Gladys Avilés y la Dra. Susana Brom por resolver mis dudas sobre los trámites del posgrado y realizar los mismos de manera oportuna y eficaz.

A Liz González por tramitar amablemente las becas CONACyT del posgrado. A Angélica Téllez e Irasema Rodriguez por hacer sencillos los trámites del Apoyo a los Estudios de Posgrado (PAEP). A Ivonne Torres por sus avisos oportunos sobre temas de interés y trámites del Programa de Doctorado en Ciencias Biomédicas.

A Concepción Hernández Levaro y su hija Claudia por su generosidad y confianza al aceptar ser fiadoras en uno de los tantos contratos de arrendamiento que firmé en mi estancia en Cuernavaca. A Yalbi Balderas por la confianza para abrirme las puertas de su casa y facilitarme inmensamente la llegada y aclimatación a Cuernavaca. A ellas, larga vida y prosperidad.

Finalmente, a los arquitectos Fernando Lezama, Francisco Tlakuilo, Jorge Tamez y Batta, y los *Urban Sketchers Morelos*, por la oportunidad de conocer juntos el estado de Morelos «con un dibujo a la vez».

Tabla de contenidos

ABSTRACT

Recently, ontologies have become the fundamental informatic tool for the interoperability of biological data. One of the most important features of ontological representation of knowledge is the possibility of creating formal definitions that allow automatic reasoning. Reasoning in ontologies is based on symbolic logic representation. This requires that ontological definitions state either necessary conditions or necessary and sufficient conditions. Here we address the difficulties of finding the necessary and sufficient conditions to define biological entities universally and propose a manual approach to evaluate the necessity and sufficiency of biological entities. We follow this approach to define the fundamental concepts of bacterial transcriptional regulation. As a result, we propose precise, updated definitions that will support a logical, consistent ontological model of bacterial transcriptional regulation.

RESUMEN

Recientemente, las ontologías se han convertido en una herramienta informática fundamental para posibilitar la interoperabilidad de los datos biológicos. Una de las características más importantes de la representación ontológica de conocimiento es la posibilidad de crear definiciones formales que permiten el razonamiento automático. El razonamiento en las ontologías se basa en la representación de conocimiento en un lenguaje lógico. Esto requiere que las definiciones establezcan las condiciones necesarias y suficientes de un concepto. En este trabajo revisamos las dificultades de encontrar las condiciones necesarias y suficientes que definen a las entidades biológicas y proponemos un enfoque humano de revisión de la necesidad y suficiencia de las definiciones de estas entidades. Aplicamos este enfoque a los conceptos fundamentales de la

regulación transcripcional bacteriana y, como resultado, proponemos definiciones precisas y actualizadas que servirán de base para un modelo ontológico de la regulación bacteriana lógicamente consistente.

INTRODUCCIÓN

La meta última de este trabajo de investigación es desarrollar una ontología formal de la regulación genética bacteriana que pueda proponerse como un estándar de representación formal del conocimiento de la regulación genética bacteriana. Un prerrequisito fundamental para esto es contar con definiciones que especifiquen las condiciones necesarias y suficientes de pertenencia a las clases de entidades a representar. Sin embargo, la biología molecular es una ciencia experimental y no una ciencia formal, por lo que las entidades biológicas moleculares no se definen formalmente en el curso de la investigación experimental, aunque los resultados experimentales contribuyen a una definición formal. Por lo tanto, el primer paso para desarrollar una ontología es proponer definiciones ontológicas con alto potencial de ser adoptadas como estándares.

Esta tesis propone una aproximación para hacer definiciones ontológicas de las entidades fundamentales de la regulación transcripcional en bacterias y se bosqueja una posible representación ontológica del dominio.

En el *Marco conceptual* se explica lo que es una ontología formal. Primero se expondrá muy brevemente la definición de Ontología como disciplina filosófica. Gran parte del capítulo se dedica a exponer el conocimiento mínimo necesario para entender lo que es una ontología formal y cómo

deben ser las definiciones ontológicas. Para esto, explicaremos lo que son las condiciones necesarias y suficientes para la definición de una clase y veremos que la formalidad de las ontologías es conferida por el lenguaje lógico subyacente al artefacto de representación. Revisaremos de manera bastante breve las Lógicas de Descripción, los lenguajes lógicos detrás del lenguaje ontológico estándar de la *World Wide Web*, llamado *Web Ontology Language* cuyo acrónimo es OWL y usaremos en lo que resta de este documento. Destacaremos el carácter universal de las definiciones ontológicas y el problema epistemológico implicado por este carácter universal en la confección de definiciones ontológicas en las ciencias experimentales. Dado que el enfoque que proponemos para la construcción de definiciones ontológicas se basa en el método deductivo de prueba propuesto por Karl Popper, revisaremos brevemente este método. Las tres últimas secciones del capítulo se dedican a exponer la aplicación de las ontologías en la interoperabilidad de los modelos de representación de información y en la integración del conocimiento. Estas aplicaciones requieren una conceptualización común para todos los usuarios, lo que a su vez requiere una entidad de coordinación del desarrollo de ontologías biológicas: la *OBO Foundry*.

En el *Planteamiento del problema* se aborda cómo se han propuesto históricamente definiciones ontológicas en Biología y los problemas a los que nos enfrentamos al proponer definiciones ontológicas en Biología Molecular particularmente. En la sección *Método* se desglosa el flujo de trabajo para aplicar el método deductivo de prueba en la construcción de definiciones ontológicas. En *Resultados* se presentan los artículos publicados, incluidos el artículo corto de congreso en el que se explica el método propuesto; las definiciones recuperadas de la literatura para aplicar el método propuesto, mismas que se incluyeron como material suplementario en el artículo de las

definiciones propuestas; el artículo que incluye las definiciones derivadas mediante el método propuesto y un artículo corto de congreso en el que se esboza una ontología de aplicación de la regulación genética bacteriana. En *Discusión* analizaremos la idoneidad del método propuesto para generar las definiciones; las alternativas y dificultades que se presentan en la especificación formal de estas definiciones usando las lógicas de descripción y las consecuencias prácticas que las definiciones propuestas tienen en los esquemas de bases de datos como RegulonDB y EcoCyc.

Para exponer el proceso de formalización de las definiciones, se usarán varios ejemplos de conversión de enunciados expresados en lenguaje natural a lenguajes formales. Los ejemplos que se usan se toman tal cual de las referencias usadas para explicar el tema en cuestión, por lo que se verá una variedad de temas representados en los mismos.

MARCO TEÓRICO

Ontología

La palabra «ontología», al igual que muchas palabras, denota distintos conceptos dependiendo del contexto en el que se usa. Originalmente la ontología es una disciplina filosófica. La definición de la disciplina filosófica aún es debatida. Por un lado, la Ontología es el estudio de lo que existe. Algunos problemas ontológicos tratan de dilucidar si ciertas entidades existen o no, por ejemplo, ¿existen los números? Pero generalmente también se considera que la Ontología trata de encontrar las características y relaciones más generales entre las entidades que sí existen[1].

Si nuestras creencias implican la existencia de ciertas entidades, podemos decir que estamos comprometidos ontológicamente con la existencia de estas entidades. Descubrir los compromisos

ontológicos de un conjunto de creencias o de la aceptación de una teoría del mundo en particular, es parte de la disciplina de la ontología. Si un conjunto de enunciados que expresan mis creencias implica o no que existen entidades de cierto tipo podría no ser obvio y podría incluso ser controvertido. Los métodos formales pueden usarse para determinar qué implica qué. Las herramientas formales hacen explícitas las ambigüedades y las diferentes lecturas, y modelan sus comportamientos inferenciales respectivos[1].

Ontología formal

El uso cotidiano del lenguaje es vago y nuestro nivel de pensamiento cotidiano frecuentemente es confuso. La lógica proporciona una vía para usar el lenguaje y las ideas de manera precisa. Es particularmente útil el estudio de la teoría del razonamiento correcto, también conocida como la teoría de la inferencia lógica o teoría de la deducción[2].

La teoría de la inferencia lógica establece una serie de reglas para derivar conclusiones a partir de un conjunto de premisas. Estas reglas deben cumplir, entre otros, dos criterios importantes[2]:

Criterio 1: dado un conjunto de premisas, las reglas de derivación lógica deben permitirnos inferir SÓLO las conclusiones que siguen lógicamente a las premisas[2].

Criterio 2: dado un conjunto de premisas, las reglas de derivación lógica deben permitirnos inferir TODAS las conclusiones que siguen lógicamente a las premisas[2].

Los principios de la inferencia lógica se aplican universalmente en cada rama del conocimiento sistemático. Se dice frecuentemente que la prueba crucial más importante de cualquier teoría científica es su utilidad y precisión para predecir fenómenos antes de que los fenómenos sean observados. Cualquier predicción debe aplicar los principios de la inferencia lógica[2].

Para traer precisión lógica a nuestro análisis de ideas, ordinariamente no es suficiente construir inferencias válidas, también es esencial tener cierta maestría en los métodos para definir de manera exacta un concepto en términos de otros conceptos. En cualquier rama de la ciencia o las matemáticas, uno de los métodos más poderosos para eliminar la vaguedad conceptual es aislar un pequeño número de conceptos básicos del dominio en cuestión y luego definir los demás conceptos de la disciplina en términos de un conjunto básico[2].

Por otro lado, cualquier teoría científica puede axiomatizarse dentro de la teoría de conjuntos. La axiomatización dentro de la teoría de conjuntos es un paso inicial importante para hacer su estructura inicial exacta y explícita[2]. Así, un concepto, designado por un término, es una clase o conjunto de objetos o entidades que cumplen las propiedades de su definición. Las propiedades que definen un conjunto se conocen como condiciones necesarias y suficientes ya que expresan TODAS y SOLO las propiedades que definen al conjunto.

Para explicar esquemáticamente lo que son las condiciones necesarias y suficientes, analicemos el universo finito de juguete de la imagen 1. En este universo, cada figura es una entidad y el término A se usa para designar una clase o conjunto de entidades. ¿Cuál sería una definición precisa de la clase A en términos de sus propiedades?

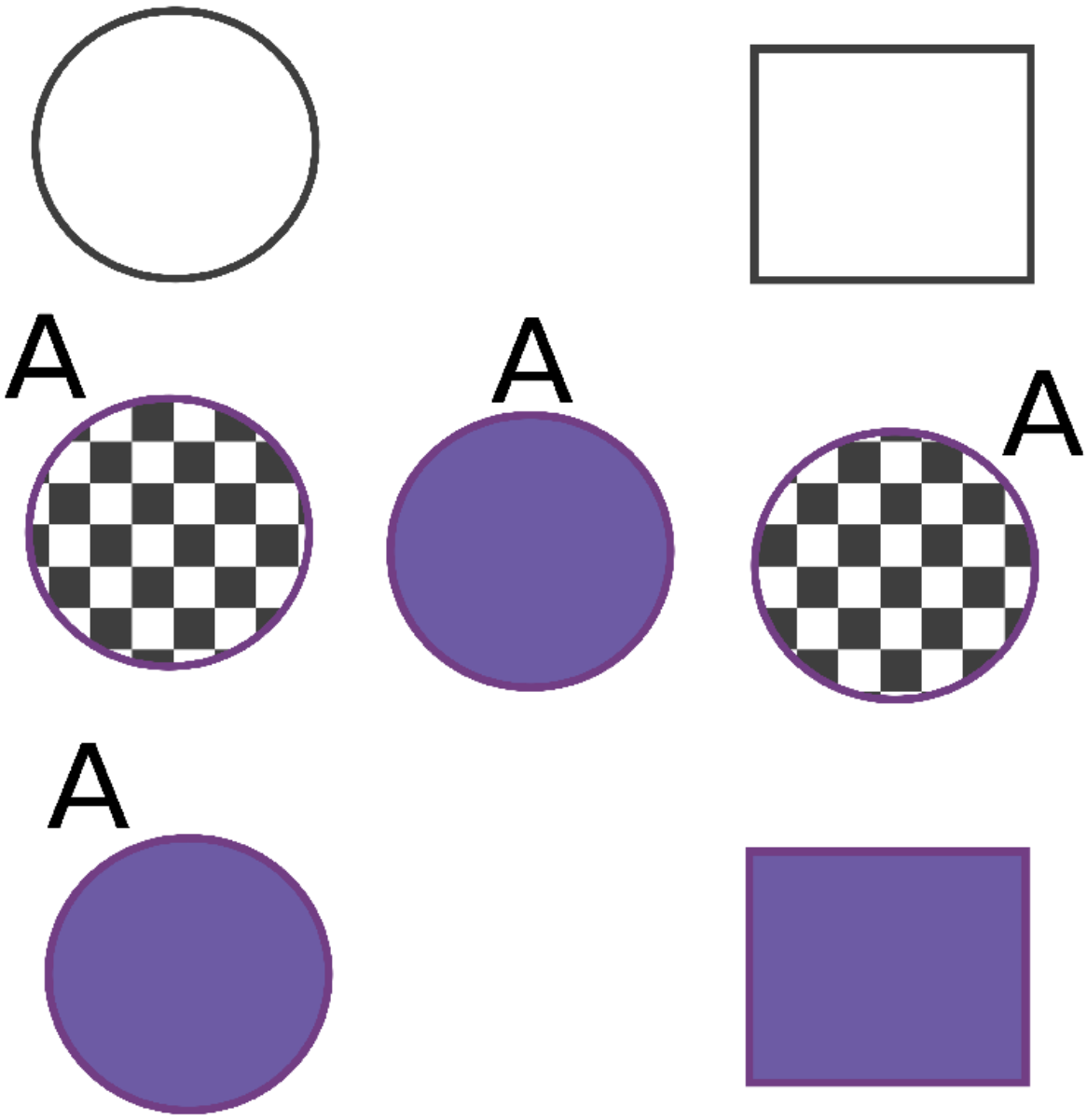

Imagen 1. Universo finito de juguete. El término A se ha utilizado para designar a un conjunto de objetos.

Si definimos la clase A como los objetos que tienen relleno sólido color morado, incluiremos el cuadrado con relleno color morado que no pertenece a la clase A, por lo que la propiedad no es

suficiente; además, excluiremos a los círculos que tienen el relleno cuadriculado, por lo que la propiedad no es necesaria.

Si definimos la clase A como los objetos que tienen relleno de patrón cuadriculado, excluiremos a todos los elementos que no pertenecen a la clase, por lo que la propiedad es suficiente, pero excluiremos también a los círculos con relleno morado que sí pertenecen a la clase, por lo que esta propiedad no es necesaria.

Si definimos la clase A como los objetos que tienen forma circular, incluiremos a todos los elementos de la clase, por lo que la propiedad es necesaria, pero incluiremos también al círculo sin relleno, por lo que esta propiedad no es suficiente.

Si definimos la clase A como los objetos que tienen forma circular y tienen cualquier relleno, incluiremos a TODOS los elementos de la clase y SÓLO a los elementos de la clase, por lo que las dos propiedades en conjunto representan las condiciones necesarias y suficientes para definir la clase A.

Si especificamos las condiciones necesarias de una clase lo que estamos diciendo es «si algo es miembro de esta clase, entonces es necesario que cumpla con estas condiciones». Con solo las condiciones necesarias no podemos decir que «si algo cumple con estas condiciones, entonces debe ser miembro de esta clase». Para hacer esto posible, tenemos que especificar las condiciones necesarias y suficientes. Esto significa que las condiciones no sólo son necesarias para que la cosa

sea miembro de la clase, sino que también son suficientes para determinar que cualquier individuo (aleatorio) que las satisface debe ser miembro de la clase[3].

Si especificamos las propiedades que definen los conceptos fundamentales de un dominio de conocimiento en un lenguaje lógico formal, el conjunto de propiedades define una teoría lógica del dominio y podemos aplicar las reglas de la inferencia lógica sobre las mismas. Especificar las definiciones como conjuntos de condiciones necesarias y suficientes nos permite obtener TODAS las conclusiones lógicas y SÓLO las conclusiones lógicas de nuestro modelo teórico del dominio. De esta forma, las herramientas formales permiten inferir los compromisos ontológicos de un conjunto de postulados.

Una ontología formal es una teoría lógica diseñada para capturar los modelos deseados que corresponden a una conceptualización determinada y para excluir los no deseados[4]. Es decir, las ontologías formales computacionales son especificaciones de las condiciones necesarias y suficientes en un lenguaje formal de los objetos de un dominio de interés. Las condiciones necesarias y suficientes se especifican representando los objetos y las relaciones que tienen con otros objetos. En nuestro ejemplo de juguete, podríamos representar los objetos usando tres clases: «forma geométrica», «relleno» y «A»; las primeras dos tendrían dos subclases, «círculo» y «cuadrado», y «patrón cuadriculado» y «sólido morado» respectivamente; así como los símbolos de propiedad «tiene la forma de» y «tiene relleno». Luego decimos que para ser miembro de A es necesario y suficiente también ser miembro de la clase círculo y estar relacionado mediante «tiene relleno» con al menos un miembro de la clase « relleno».

La ontología ideal es aquella cuyos modelos coinciden exactamente con los previstos. Sin embargo, lograr esto no es simple. Incluso una ontología así podría fallar en especificar exactamente la conceptualización deseada si el vocabulario y el dominio del discurso no se eligen adecuadamente. Esto se debe a la diferencia entre la noción lógica de modelo y la ontológica de un mundo posible. La primera es básicamente una combinación de asignaciones de estructuras relacionales abstractas (construidas sobre el dominio de discurso) a los elementos del vocabulario; la última es una combinación de estados de cosas reales (observados) de un determinado sistema[4].

El número de modelos posibles depende del tamaño del vocabulario y de la extensión del dominio de discurso, los cuales son elegidos de manera más o menos arbitraria sobre la base de lo que parece ser relevante representar. Por el contrario, el número de estados del mundo depende de las variables observadas, incluso aquellas que, a primera vista, se consideran irrelevantes[4].

Las lógicas de descripción detrás de los lenguajes ontológicos formales

Las lógicas de descripción (DL por Description Logics) son una familia de lenguajes formales lógicos que pueden usarse para representar el conocimiento de un dominio de manera estructurada y formalmente bien entendida. La idoneidad de las DL como lenguaje ontológico ha sido destacada por su rol como cimiento para varios lenguajes ontológicos de la web, incluido OWL, el estándar del consorcio de la World Wide Web[5].

El nombre «lógicas de descripción» está motivado por el hecho de que, por un lado, las nociones importantes del dominio se representan a través de descripciones de conceptos, i.e., expresiones que se construyen a partir de conceptos atómicos (predicados unarios) y roles atómicos (predicados

binarios) usando los constructores de conceptos y roles proporcionados por la DL particular[5]. Por otro lado, las DL difieren de sus predecesoras, como las redes semánticas y los *frames*, en que están equipadas con una semántica formal basada en la lógica. Los sistemas basados en alguna lógica de descripciones proporcionan a los usuarios varias posibilidades de deducir conocimiento implícito a partir del conocimiento representado explícitamente[5].

Para asegurar el comportamiento razonable y predecible de un sistema basado en DL, los problemas de la inferencia lógica deben ser al menos computacionalmente decidibles y, de preferencia, de baja complejidad. En consecuencia, el poder expresivo de las DL debe estar restringido de manera apropiada. Por el contrario, si las restricciones impuestas son muy severas, nociones importantes del dominio de aplicación no pueden ser expresadas. Las DL son más expresivas que la lógica proposicional, pero menos expresivas que la lógica de primer orden[5].

No trataré exhaustivamente los diferentes «sabores» ni los constructores de las DL. Sólo tomaré el ejemplo ilustrativo de los constructores típicos incluido en la referencia 5.

Supongamos que queremos definir el concepto de «un hombre que está casado con una médico y tiene por lo menos cinco hijos, todos los cuales son profesores». Este concepto puede describirse así:

Humano ⊓ ¬Mujer ⊓ ∃casadoCon.Médico ⊓ (≥5 tieneHijo) ⊓ ∀tieneHijo.Profesor

Este descripción usa cuatro conceptos: Humano, Mujer, Médico y Profesor, y dos roles: casadoCon y tieneHijo. Esta descripción también emplea los constructores booleanos de la conjunción (⊓, **and** o **intersectionOf** en OWL), que se interpreta como una intersección de conjuntos, así como el constructor de la negación (¬R.C, **not** o **complementOf** en OWL), el constructor de la restricción existencial (∃R.C, **some** o **someValuesFrom** en OWL), el constructor de restricción de valor (∀R.C, **only** o **allValuesFrom** en OWL) y el constructor de restricción numérica (≥nR, **min** o **minCardinality** en OWL). En todas las fórmulas R es un rol, C un concepto y n un número.

Una persona, digamos Beto, pertenece a la clase ∃casadoCon.Médico si y solo si existe al menos una persona que está casada con Beto (i.e., se relaciona con Beto mediante el rol casadoCon) y esa persona es médico (i.e., pertenece a la clase Médico). De manera similar, Beto pertenece a la clase (≥5 tieneHijo) si y solo si tiene al menos cinco hijos, y pertenece a la clase ∀tieneHijo.Professor si y solo si todos sus hijos (i.e., todos los individuos que se relacionan con Beto mediante el rol tieneHijo) son profesores.

Además de este formalismo de descripción, las DL normalmente están equipadas con un formalismo terminológico y uno asertivo. En su forma más simple, los axiomas terminológicos pueden usarse para introducir nombres o términos (abreviaciones) de descripciones complejas. Por ejemplo, podríamos introducir el nombre «Hombre feliz» para la descripción del concepto anterior. Para esto se usa el constructor de equivalencia (≡, equivalentClass en OWL). Por ejemplo:

Hombre feliz ≡ Humano ⊓ ¬Mujer ⊓ ∃casadoCon.Médico ⊓ (≥5 tieneHijo) ⊓ ∀tieneHijo.Profesor

Lo cual significa que la clase llamada «Hombre feliz» contiene exactamente los mismos individuos que los de la clase formada por los individuos que cumplen con la descripción. Es decir, la descripción especifica las condiciones necesarias y suficientes de pertenencia a la clase «Hombre feliz», por lo que decimos que «Hombre feliz» es una clase definida.

Formalismos terminológicos más expresivos permiten enunciados de restricciones como:

$\exists$tieneHijo.Humano $\sqsubseteq$ Humano,

Que expresa que solo los humanos pueden tener hijos humanos. $\sqsubseteq$ es el constructor «subclase de» o **subClassOf** en OWL. Un conjunto de axiomas terminológicos se llama *TBox*. El formalismo asertivo se usa para establecer las propiedades de los individuos. Por ejemplo, las aserciones:

HombreFeliz(BETO), tieneHijo(BETO, MARÍA)

establecen que Beto es una instancia del concepto «hombre feliz» y que María es uno de sus hijos. Un conjunto de aserciones así se llama ABox y los individuos nombrados que aparecen en las aserciones de la ABox se llaman individuos de la ABox.

He revisado brevemente la expresividad de las lógicas de descripción ya que una ontología formal es un conjunto de definiciones especificadas en un lenguaje formal. En el caso de las bioontologías, la comunidad ha adoptado OWL como lenguaje de representación, mismo que está soportado por las lógicas de descripción. Una vez que se tienen definiciones consensuadas en lenguaje natural de las clases a representar en la ontología, el siguiente paso es traducirlas a este lenguaje formal

tomando en cuenta sus limitaciones de expresividad. Aún no tengo resultados de este proceso de traducción, pero en la *Discusión* analizo algunas posibilidades y retos que dicho proceso implicará.

Las ontologías representan enunciados universales

El hecho de que las definiciones ontológicas establecen condiciones necesarias de pertenencia a una clase implica que los enunciados ontológicos están restringidos estrictamente a enunciados universales, es decir, afirmaciones que son verdaderas para TODAS las instancias de un determinado tipo. Todos los enunciados que expresan que algo es normalmente, pero no universalmente, verdadero no son ontológicos. Por ejemplo, podría ser tentador representar el enunciado «la elefantiasis normalmente es causada por la filariasis linfática» de la siguiente manera[6]:

*Elefantiasis **subClassOf** 'normalmente es causada por ' **some** 'filariasis linfática'

En la semántica de las DL del cuantificador existencial (**some**), esta afirmación significa que para cada miembro de la clase Elefantiasis (sin excepción), existe algún miembro de la clase filariasis linfática. La palabra «normalmente» que es parte del nombre del rol puede ser interpretada por humanos, pero desde el punto de vista de las DL, no tiene ninguna función lógica en absoluto. Esto da lugar a inferencias inesperadas. Por ejemplo, dada la siguiente afirmación[6]:

'Elefantiasis no infecciosa' **subClassOf** 'Elefantiasis' **and not** ('causada por' **some** 'filariasis linfática')

se puede inferir lógicamente el siguiente enunciado sin sentido[6]:

'Elefantiasis no infecciosa' **subClassOf** 'normalmente es causada por **some** 'filariasis linfática' **and not** ('causada por' **some** 'filariasis linfática')

El mensaje más importante es que los axiomas ontológicos solo pueden expresar lo que es verdadero para todos los miembros de la clase. Esto excluye de las ontologías a los enunciados contingentes o probabilísticos[6].

Enunciados universales en la ciencia experimental

Un científico, ya sea teórico o experimental, propone enunciados, o sistemas de enunciados, y los prueba paso a paso. Más particularmente en el campo de las ciencias empíricas, construye hipótesis, o sistemas de teorías, y las prueba confrontándolas con la experiencia por observación y experimentación[7].

Es usual decir que una inferencia es 'inductiva' si pasa de enunciados singulares o particulares a enunciados universales. Ejemplos de los primeros son los enunciados que dan cuenta de los resultados de las observaciones o experimentos, y de los segundos son las hipótesis y teorías. Pero, desde el punto de vista lógico, no se justifica la inferencia de enunciados universales a partir de los singulares, sin importar qué tan numerosos sean estos últimos; toda conclusión derivada de esa manera siempre puede resultar falsa: no importa cuántas instancias de cisnes blancos podamos haber observado, esto no justifica la conclusión de que todos los cisnes son blancos (imagen 2)[7]. Por otro lado, está claro que toda descripción de una experiencia —una observación o los

resultados de un experimento— sólo puede ser en primer lugar un enunciado singular y no universal[7].

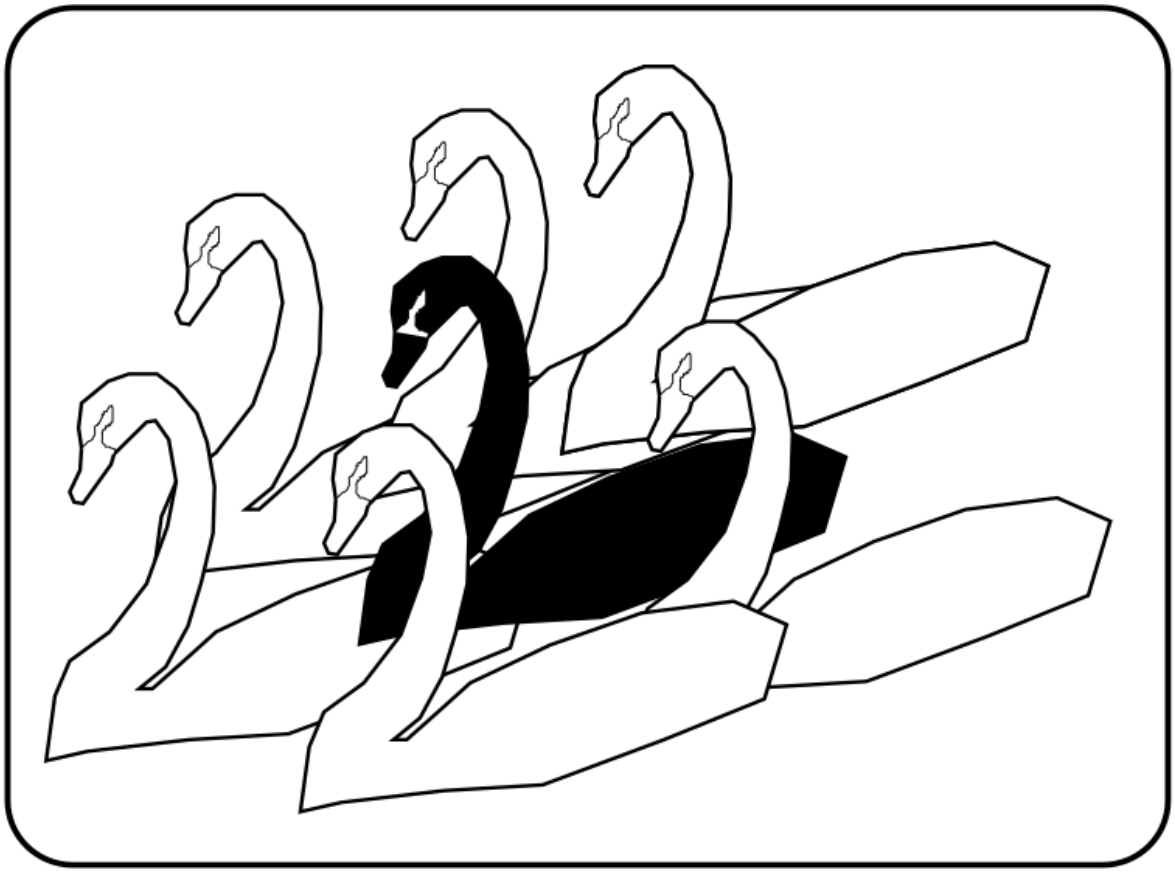

Imagen 2 El cisne negro. No importa cuántos cisnes blancos hayamos observado, eso no implica que TODOS los cisnes son blancos. Pero la observación de un solo cisne negro falsifica la afirmación de que TODOS los cisnes son blancos.

Método deductivo de prueba de las teorías

El propósito del método deductivo de prueba de las teorías es descubrir qué tanto las nuevas consecuencias de una teoría pueden resistir las demandas de la práctica. El procedimiento de prueba resulta ser deductivo. Con ayuda de otros enunciados previamente aceptados, de la teoría

se deducen algunos enunciados singulares, que podríamos llamar predicciones; especialmente predicciones que se pueden probar o aplicar fácilmente. Luego buscamos una decisión respecto a estos (y otros) enunciados derivados comparándolos con los resultados de las aplicaciones prácticas y los experimentos. Si esta decisión es positiva; es decir, si las conclusiones singulares resultan ser aceptables o verificadas, entonces la teoría por el momento ha pasado la prueba: no hemos encontrado una razón para descartarla. Pero si la decisión es negativa, o en otras palabras, si las conclusiones han sido falsificadas, entonces su falsificación también falsifica la teoría de la cual se dedujeron lógicamente. Los enunciados universales nunca pueden derivarse de los enunciados singulares, pero se pueden contradecir con enunciados singulares. Consecuentemente, es posible argumentar que, por medio de inferencias puramente deductivas (con ayuda del *modus tollens* de la lógica clásica), la verdad de los enunciados singulares falsifica los enunciados universales[7].

El *modus tollendo tollens* es una forma de argumento válida que se basa en el hecho de que el siguiente enunciado es una tautología: $[\neg Q \sqcap (P \rightarrow Q)] \rightarrow \neg P$, i.e., no Q y si P entonces Q implica no P. En otras palabras, si P implica Q y si la negación de Q es verdadera se puede concluir que la negación de P es verdadera. Una tautología es una fórmula bien formada que resulta verdadera para cualquier interpretación; es decir, para cualquier asignación de valores de verdad que se haga a sus fórmulas atómicas.

La importancia de tener una conceptualización común para varios usuarios

El valor de cualquier tipo de dato aumenta cuando puede integrarse con otros datos[8]. Dos personas comparten el mismo significado de un concepto si, en presencia de los mismos estados del mundo,

eligen los mismos objetos como instancias del concepto. Si no, tendrán diferentes conceptualizaciones, i.e., diferentes formas de interpretar sus datos sensoriales[4].

Para el uso práctico de las ontologías, resultó evidente rápidamente que sin al menos un mínimo compromiso ontológico compartido por las partes interesadas en la ontología, los beneficios de tener una ontología son limitados. La razón es que la ontología define formalmente la estructura de un dominio bajo la limitación de que el partícipe entiende los términos primitivos de la manera apropiada. En otras palabras, la ontología podría resultar inútil si se usa de manera contraria al compromiso ontológico compartido. Cualquier ontología siempre será menos completa y formal de lo deseable en teoría. Por eso es importante que las ontologías diseñadas para soportar la interoperabilidad a gran escala estén bien fundamentadas, en el sentido de que las primitivas sobre las que se basa se elijan y se axiomaticen de tal manera que sean entendidas generalmente[4].

Las ontologías locales son conceptualizaciones que pueden ser parciales, que cubren sólo una parte del dominio, o tienen una perspectiva específica que depende del objetivo. Las ontologías de referencia deben proporcionar conocimiento de dominio que pueda ser usado como marco común para la integración semántica de la información de diferentes fuentes[9]. La idea de estas ontologías es que funcionen como un artefacto de referencia independiente de cualquier propósito.

Tales ontologías independientes del propósito, en un segundo paso, se enlazan a artefactos computacionales centrados en un propósito como las bases de conocimiento y los sistemas de soporte de decisiones, proporcionando interoperabilidad con otros sistemas similares[6]. O bien, a partir de las ontologías de referencia se pueden derivar ontologías de aplicación para tareas

específicas[9]. Una propuesta para generar ontologías de dominio o de aplicación a partir de ontologías de referencia es trasladando el concepto de vista de las bases de datos relacionales a las ontologías. Una vista es una consulta que calcula una tabla nueva a partir de tablas preexistentes. En el caso de las ontologías, la consulta se expresa en un lenguaje de consulta para ontologías formales y se genera un módulo. Por ejemplo, un neurocientífico podría necesitar una ontología de aplicación que solo incluya la porción de la ontología de referencia de anatomía humana llamada *Foundational Model of Anatomy* relacionada con las partes del cerebro y que simplifique enormemente el conjunto de relaciones o roles entre estos conceptos neuroanatómicos, por ejemplo una partonomía simple (la jerarquía de relaciones «parte de»)[10].

El dominio de la Biología Molecular está marcado por la disponibilidad de grandes cantidades de datos que pueden usarse sin restricciones para el procesamiento algorítmico. El uso de esquemas de codificación local implica que estos datos se acumulen de una manera que dificulta la integración y la investigación. Por esta razón se ha reconocido la necesidad de estrategias de estandarización prospectiva para mejorar progresivamente y alinear recíprocamente los esquemas empleados para el manejo, descripción y publicación de los datos biomédicos[8].

Las ciencias de la vida se caracterizan por la variación que existe en la realidad. Una propiedad significativa de una ontología de referencia biomédica, por lo tanto, debe ser su capacidad de representar las variaciones que existen en la realidad. Las ontologías biomédicas no deben limitarse a los modelos canónicos y deben extenderse para modelar la variación en las entidades del dominio[9]. La representación de entidades canónicas versus no canónicas se ha discutido en el contexto de las ontologías de anatomía[11,12]. Por ejemplo, se propone una ontología de anatomía

canónica en la que las manos tienen cinco dedos y una ontología de anatomía no canónica o patológica donde puede haber manos con polidactilia, y mapeos entre estas[11,12].

En este trabajo, tratamos de abordar el problema de la variación en biología molecular que hay en un sistema «normal» o canónico. Al ser entidades moleculares cuya observación es complicada, su caracterización es parcial. Es un problema de investigación encontrar las características invariantes o definitivas de tales entidades a lo largo de un nivel taxonómico, por ejemplo en el dominio taxonómico de las bacterias. Proponemos un enfoque para generar definiciones ontológicas usando el conocimiento del dominio disponible a la fecha.

Coordinación del desarrollo de las ontologías biomédicas a través de la OBO foundry

La OBO Foundry (http://obofoundry.org) (OBO son las siglas en inglés para *Open Biomedical Ontologies*) trata de abordar el problema de la proliferación de ontologías como esquemas locales de representación y de coordinar el desarrollo de ontologías para que los compromisos ontológicos sean generalmente compartidos. Es un colectivo de ontólogos que aceptan voluntariamente un conjunto de principios en evolución para el desarrollo de ontologías. Estos principios incluyen y extienden a los que influyeron en el éxito de *Gene Ontology*. Estos principios establecen que las ontologías deben ser[8]:

- Abiertas: las ontologías mismas y los cuerpos de datos descritos por sus términos deben estar disponibles para su uso sin ninguna restricción o licencia, por lo que pueden aplicarse en todo propósito nuevo sin ninguna restricción.

- Ortogonales: cada concepto debe estar definido en una sola ontología de **referencia.** Esta característica ayuda a reducir la necesidad de decisiones arbitrarias entre términos aparentemente equivalentes de diferentes ontologías.
- Realizadas en una sintaxis bien definida y común: actualmente se aceptan el formato OBO y OWL.
- Compartir un espacio común de identificadores.
- Receptivas a modificaciones como resultado del debate de la comunidad.
- Desarrollarse mediante un esfuerzo colaborativo.
- Usar relaciones (roles) comunes definidas sin ambigüedades.
- Proporcionar procedimientos para recibir comentarios de los usuarios y para identificar versiones sucesivas.
- Tener un ámbito bien delimitado y, por ortogonalidad, ningún término debería aparecer en más de una ontología. Los desarrolladores de ontologías pequeñas pueden contribuir constructivamente al crecimiento de recursos compartidos al mismo tiempo de beneficiar a los usuarios de sus propias ontologías.
- Cada clase de la ontología debe estar definida con definiciones aristotélicas de la forma: A es un tipo de B que tiene la propiedad distintiva C. Por ejemplo, célula = estructura anatómica que está delimitada por una membrana plasmática.

El gran grafo del conocimiento biológico

Cada ontología de la OBO Foundry forma una estructura teórica en forma de grafo con términos o nodos conectados mediante aristas que representan relaciones o roles[8]. Para maximizar la coordinación entre ontologías, deben construirse términos compuestos usando términos y

relaciones de otras ontologías de la OBO Foundry siempre que se pueda. Pero este enfoque funcionará sólo si los términos resultantes no son ambiguos y la OBO Foundry ayuda a proporcionar el rigor necesario. Las ontologías de la OBO Foundry son creadas y mantenidas por biólogos con conocimiento profundo de la ciencia en cuestión. La meta a largo plazo es que los datos generados por la investigación biomédica formen un todo único, consistente, que se expanda por acumulación y que sea tratable algorítmicamente[8].

Si tenemos un modelo computacional en el que todas las propiedades (o al menos las relevantes o las que se conocen actualmente) de los objetos de un dominio están especificadas y si definimos precisamente una clase podemos escribir un programa que seleccione los miembros de la clase de manera automática. En un modelo suficientemente exhaustivo de un dominio esto podría ayudar a descubrir hechos que por la cantidad de información procesada no son obvios para la mente humana, a hacer evidente y explícito el conocimiento implícito en los compromisos ontológicos del dominio. La definición precisa de las entidades también permite a los humanos determinar si un objeto recientemente descubierto es una instancia del concepto definido.

PLANTEAMIENTO DE PROBLEMA

Definiciones ontológicas en Biología

El área de la Biología que históricamente se ha dedicado a definir las condiciones necesarias y suficientes de entidades biológicas es la Taxonomía. Describen las condiciones necesarias y suficientes para definir clases de organismos. Esta descripción y clasificación jerárquica de organismos sirve para determinar qué tipo de organismo es un ejemplar recuperado de la

naturaleza; en otras palabras, esta descripción ayuda a responder la pregunta ¿el ejemplar pertenece a una especie previamente descrita o se ha descubierto una nueva especie?

Los primeros taxónomos salían al campo, colectaban especímenes y armaban colecciones biológicas, mediante las cuales podían observar directamente las características morfológicas de los organismos. Estas colecciones han sido la base para el desarrollo de la clasificación de organismos, que aunque está sujeta a actualizaciones, ha sido bastante útil, y su lógica subyacente es robusta y persistente. Las definiciones taxonómicas pueden actualizarse al considerar información fisiológica, ecológica y molecular.

Análogamente, en Biología Molecular, las bases de datos pueden considerarse colecciones biológicas. Las colecciones biológicas contienen el ejemplar biológico en sí. Algunas bases de datos moleculares contienen una representación «cruda» de la estructura molecular de la entidad biológica en sí, por ejemplo la secuencia de nucleótidos o de aminoácidos. Sin embargo, la clasificación que podría derivarse del análisis «puro» de estas entidades se limitaría a clasificaciones por similitud de secuencia, lo cual arroja muy poca luz para la comprensión de los determinantes moleculares de los fenómenos biológicos. Por lo que existen otras bases de datos moleculares que intentan poner en un contexto biológico más amplio los datos moleculares estructurales. Sin embargo, estas bases de datos contienen información que ha sido seleccionada de manera más o menos arbitraria. El depósito más completo de conocimiento molecular situado en su contexto biológico continúa siendo la literatura científica.

Y aún sumando todas las publicaciones, la ignorancia sigue siendo mayor que el conocimiento. Por ejemplo, para definir las condiciones necesarias y suficientes de un promotor bacteriano, idealmente deberíamos tener la colección de todos los promotores de todas las bacterias que existen. Sin embargo, a la fecha sólo se conoce una fracción de genomas bacterianos completos. GenBank contiene las secuencias de 100305 genomas bacterianos completos (consulta hecha el 21 de agosto de 2020), mientras que se estima que existen 10^{12} especies bacterianas, aunque más recientemente se ha propuesto que una mejor estimación se quedaría solamente en el orden de 10^{6} (REF 13).

Por otro lado, cuando un biólogo experimental prueba sus hipótesis, busca refutar las explicaciones alternativas o las consecuencias lógicas del fenómeno estudiado. Por lo que el enunciado resultante de un estudio molecular ya es una hipótesis que ha resistido la prueba de la falsificación (hasta ahora). Sin embargo, en el modelado ontológico, es necesario proponer enunciados universales de orden superior, enunciados que tratan de hacer afirmaciones basadas en la integración de resultados de diversos estudios experimentales similares. Estos enunciados universales requieren someterse a la misma prueba de falsificación a medida que se va recopilando información nueva. En otras palabras, podemos ver las definiciones ontológicas propuestas como hipótesis que pueden ser falsificadas por resultados experimentales nuevos.

En este trabajo consideraremos las características usadas en algunas definiciones parciales, extraídas de referencias literarias individuales, de las entidades moleculares como hipótesis que someteremos a un proceso de falsificación. Usaremos enunciados contenidos en la literatura para

falsificar las hipótesis. La suposición subyacente es que la definición propuesta en un artículo puede estar falsificada por la información contenida en otro.

Mediante este enfoque tratamos de aumentar la objetividad en las definiciones propuestas de tal manera que puedan ser más fácilmente aceptadas por la comunidad, dado que la aceptación y comprensión intersubjetiva de las definiciones es un requisito para alcanzar el objetivo de la unificación formal de la Biología.

OBJETIVO

Proponer definiciones ontológicas de promotor, factor de transcripción, sitio de unión de factor de transcripción, operón, unidad de transcripción, regulón, efector y señal, mediante un enfoque de falsificación de definiciones propuestas en la literatura, usando los datos publicados en la literatura más reciente.

MÉTODO

Hay dos enfoques a considerar para elaborar definiciones:

- De abajo hacia arriba: derivar una definición intensional a partir de la definición extensional. Una definición extensional es una lista exhaustiva de las instancias incluidas en el concepto, mientras que una definición intensional establece las características que deben cumplir las instancias del concepto. En este enfoque, a partir de la lista completa de elementos que constituye una clase, observamos las propiedades distintivas que tienen en

común. En el ejemplo del universo finito discutido en el *Marco teórico* en la sección de *Ontología formal* se usó este enfoque.

- De arriba hacia abajo: proponer una definición intensional que abarcará sólo los elementos que cumplen con las propiedades y usarla para saber si un objeto es una instancia de la clase o no.

El enfoque de arriba hacia abajo es un enfoque prescriptivo que tiene el riesgo de excluir instancias que históricamente se han considerado miembros de una clase, por lo que lo ideal sería usar el enfoque de abajo hacia arriba. Sin embargo, no contamos con una lista exhaustiva de instancias bien caracterizadas de cada concepto, por lo que un enfoque de abajo hacia arriba no es posible.

Proponemos adaptar el método de arriba hacia abajo con el método deductivo de prueba para evitar la exclusión de entidades bien caracterizadas que no se ajustan a las definiciones que ya se han propuesto en la literatura. Como se discutió previamente, la verificación de enunciados universales de la necesidad y suficiencia en las ciencias empíricas es imposible. Dado que no podemos verificar las definiciones, para asegurar que están actualizadas, proponemos un enfoque de falsificación[14]:

• Podemos refutar la suficiencia encontrando algún objeto que tiene la propiedad P y no pertenece a la clase que se está definiendo.

• Podemos refutar la necesidad encontrando una instancia de la clase que se está definiendo y no tiene la propiedad P.

Las hipótesis que serán falsificadas se obtendrán de las siguiente manera: recuperamos algunas definiciones de la literatura y de ontologías existentes, y a partir de estas hacemos una lista de características que se han usado para definir las entidades. Estas características son las hipótesis que se someterán al proceso de falsificación. Se consideran hipótesis porque normalmente las definiciones propuestas en un artículo son válidas dentro del artículo, pero no se derivaron del análisis exhaustivo de las instancias del concepto. Los contraejemplos para falsificar las hipótesis se encuentran en la literatura y en el conocimiento de los expertos de dominio. Proponemos encontrar los contraejemplos en dos fases:

1) Usar los motores de búsqueda de literatura científica y palabras clave.
2) Organizar una discusión colectiva con expertos de dominio. En esta fase se presentan los resultados de la fase previa y se aclara cualquier punto que no se haya podido resolver en la fase 1. Esta fase además es importante para llegar a una conceptualización compartida del dominio. Es una fase en la que se requiere llegar a definiciones consensuadas.

Las características cuya necesidad o suficiencia no sean falsificadas formarán la nueva definición propuesta, que, sin embargo seguirá siendo provisional y podrá ser falsificada por los resultados científicos venideros. El método se resume en la imagen 3.

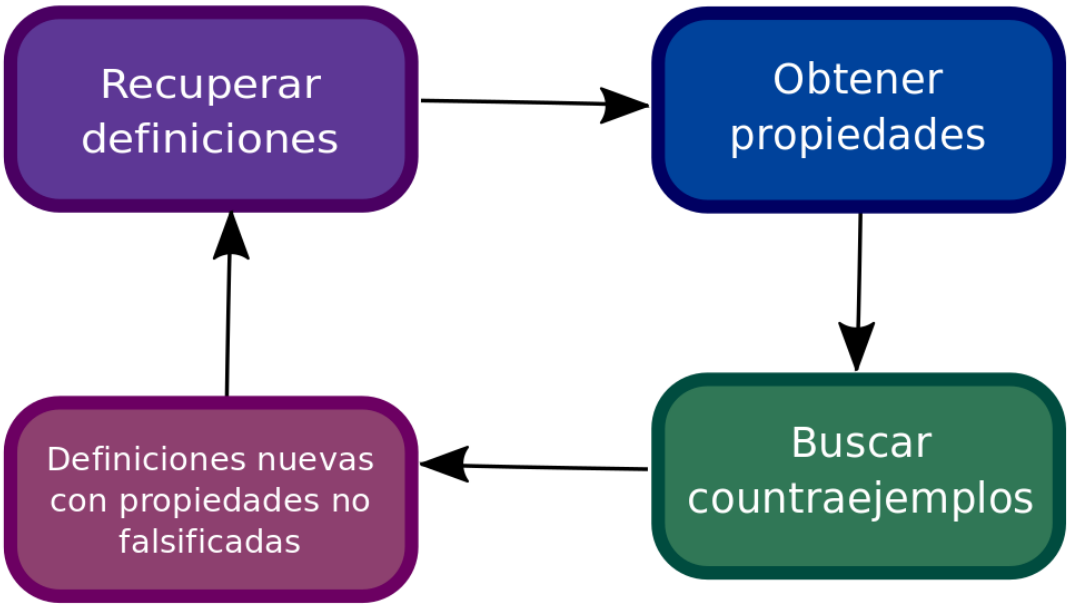

Imagen 3 Flujo de trabajo para derivar definiciones ontológicas mediante el método de falsificación de propiedades especificadas en definiciones existentes.

RESULTADOS

Este trabajo doctoral generó varios resultados. Primero, se propuso un método para derivar definiciones ontológicas a partir de las definiciones existentes en la literatura. Segundo, para aplicar el método a los conceptos de promotor, factor de transcripción, sitio de unión del factor de transcripción, operón, unidad de transcripción, regulón, efector y señal se generó una compilación de definiciones recuperadas de la literatura para cada concepto junto con las características definitorias propuestas en dichas definiciones. Tercero, después de buscar los contraejemplos y de discutir la necesidad y suficiencia de las propiedades de cada concepto con

un panel de expertos, se propusieron definiciones actualizadas. En paralelo, se investigó la posibilidad de generar un modelo ontológico para inferir la relaciones de regulación procariótica a nivel fisiológico a partir de las relaciones de regulación mecanísticas. A continuación se dedica una sección a cada resultado.

Método de falsificación para crear y revisar definiciones ontológicas

Hasta donde sé, el método de falsificación no se había propuesto previamente para derivar definiciones ontológicas, por lo que la propuesta del método es el primer resultado de este trabajo. Una recomendación sobre cómo proceder para encontrar las características esenciales de una entidad biológica se esboza en la referencia 15 de la siguiente manera:

A la luz del conocimiento científico disponible, intenta imaginar la sustracción y la variación de las propiedades de una entidad típica que cae bajo el término correspondiente, revisando en cada etapa si la variación o sustracción considerada provocaría que la entidad en cuestión ya no sea instancia del concepto en cuestión. Si esa variación o eliminación de la propiedad de la entidad imaginada tiene esta consecuencia, entonces es altamente probable que esa propiedad es una de las características esenciales de un tipo dado. De esta manera es posible imaginar sillas de diferentes tamaños, formas, materiales y colores; sin embargo, en cuanto imaginamos una cosa en la que es imposible que un humano normal se siente (por ejemplo, que esté hecha de un material gelatinoso), está claro que, cualquiera que sea la entidad imaginada, no es una silla, por lo que la propiedad de ser algo en lo que un ser humano normal pueda sentarse es al menos una condición necesaria para que algo sea una silla. La solidez es en este sentido una propiedad constante (esencial) de una silla; el color una propiedad variable.

Aunque este procedimiento es bastante parecido al propuesto en este trabajo, nuestro método propone un punto de partida concreto para elegir las propiedades a evaluar y una manera de hacer preguntas específicas para saber si una propiedad no es esencial. Por ejemplo, ¿se conoce algún promotor que no contenga un motivo? Si la respuesta es sí, la propiedad de tener un motivo no es necesaria, o ¿existen sitios de unión de la holoenzima de la polimerasa que no se consideren promotores? Si la respuesta es sí, la condición de ser sitio de unión de la polimerasa no es suficiente para la definición de promotor. Estas preguntas pueden responderse buscando artículos científicos relevantes usando palabras clave en los motores de búsqueda y son una forma más sucinta de involucrar a los expertos de dominio.

El método de falsificación es ampliamente usado en Matemáticas para falsificar proposiciones. Sin embargo, en Biología y ciencias experimentales existe la posibilidad de que un resultado publicado sea un falso positivo, por lo que no necesariamente todo resultado publicado constituye un contraejemplo. Por esta razón, es importante involucrar a expertos que tengan la capacidad de juzgar la precisión en la aplicación de los métodos experimentales y analíticos mediante los que se derivan las observaciones que pueden constituir un contraejemplo, sobretodo cuando se trata de resultados que no han sido replicados.

El método se presentó en la Conferencia Internacional de Ontologías Biomédicas 2018 y se publicó en las actas del congreso. A continuación se anexa el acta correspondiente.

A Falsification approach to create and check ontology definitions

Citlalli Mejía-Almonte, Julio Collado-Vides
Computational Genomics Program
Center for Genomic Sciences, UNAM
Cuernavaca, México

***Abstract*— One of the most important features of ontological representation of knowledge is the possibility of creating formal definitions that allow automatic reasoning. Reasoning in ontologies is based on symbolic logic representation. This requires that ontological definitions state either necessary conditions or necessary and sufficient conditions. Here we propose a manual approach to review the necessity and sufficiency of ontological definitions that can be used to analyze the most prominent concepts of a domain.**

Keywords—falsification; ontology definition; necessary and sufficient conditions

I. INTRODUCTION

Since the publication of the Gene Ontology, Biomedical ontologies have thrived. As a result, a growing number of ontologies are created to represent all aspects of the biological world. Currently there are 182 ontologies in OntoBee [1] and 716 in BioPortal [2], the OBO foundry [3] and the NCBO [4] ontology repositories respectively. Some of these ontologies are foundational, for they are species-independent models aimed to be reused in or extended by species-specific ontologies. Although categorization of ontologies into species dependent and species independent is not straightforward if authors have not established it in the scope description, we found 57 species-independent, 36 taxonomically restricted (at higher taxonomic ranges), 19 whose scope does not include biological entities, and 63 species-specific ontologies in OntoBee. When authors did not specify taxonomic range, this classification was based on the next criteria: species-independent if the ontology includes classes representing organisms of more than one kingdom, and species-specific if the ontology is human-centric.

This large set of computational models can provide the means for automatic reasoning to generate mechanistic hypothesis for the biomedical research [5]. However, foundational, species-independent ontologies must have formal definitions general enough to support pertinent inferences throughout all kingdoms of life.

Here we present a manual approach to check the suitability of necessity and sufficiency of ontological definitions for the current state of affairs in biological sciences. This allowed us to find out that if we consider natural language definitions of extant foundational ontologies as necessary and sufficient conditions, some prokaryotic instances may be left out.

II. METHODS

Ontological primitive classes are described only by necessary conditions, whereas defined classes are described by necessary and sufficient conditions [6]. Necessary and sufficient conditions are explained in terms of the conditional logical relation. Let A be a class or concept and let P be some property. There are many language items to refer to this [7]:

- A only if P; if A, then P; P is necessary for A; and A is sufficient for P.

Any of these statements means that all instances of A satisfy property P, or that for all objects of the universe, if some satisfies P then it is an instance of A. When this logical condition holds in both directions, that is:

- A is necessary and sufficient condition for B and B is necessary and sufficient condition for A

We say that A means B, or A is equivalent to B. This relation of equivalency is the one we look for to make ontological definitions.

Necessity of P is proved by demonstrating that all instances of A have property P. However, demonstration of necessity is epistemologically impossible in experimental sciences, even assuming an agent with the complete knowledge of the current state of affairs. Thus, we took a falsification approach [8].

- We can disprove sufficiency by finding some object that has property P and does not belong to A.
- We can disprove necessity by finding some instance of A that does not hold property P.

Based on this, we propose the following workflow to analyze necessity and sufficiency of proposed definitions:

- Retrieve definitions from diverse sources such as the literature and extant ontologies.
- Based on the retrieved definitions, generate a list of the commonly used properties to define these concepts.
- Search counter examples for definitions to discard necessity or sufficiency of the defining properties.

- Keep those properties that were not falsified to generate a new definition.

III. RESULTS

As a matter of example, we apply this approach to the definition of bacterial promoter in the sequence ontology (SO) [9]. The following are the two relevant definitions extracted from this ontology in July 2018:

- Promoter: A regulatory_region composed of the TSS(s) and binding sites for TF_complexes of the basal transcription machinery
    - Bacterial RNA-polymerase promoter: A DNA sequence to which bacterial RNA polymerase binds, to begin transcription.

Bacterial RNA-polymerase promoter is a subclass of promoter. Thus, the list of properties that define a Bacterial RNA-polymerase promoter is:

- has part some TSS
- has part some basal TF binding sites
- initiates some transcription
- binds some RNA polymerase

If we assume that basal transcription factor (TF), which is a term most commonly used in the domain of eukaryotic gene regulation, is equivalent to the most common sense in which transcription factor term is used in the domain of prokaryotic gene regulation, then "has part basal TF binding site" is not a necessary condition, since we can find counter examples in constitutive promoter sequences [10] that transcribe without the need of any transcription factor, and promoters of endosymbionts, whose reduced genome has been found to have lost most of the regulation by means of transcription factors [11]. On the other hand, from the biological point of view the closest to those "basal TFs" would be sigma factors. In this case, definition is correct and just have to be more explicitly specified in the definition.

A. Automatic logical consistency check is not suitable to detect these lack of generality

We are aware that logical consistency is one of the main applications of automatic reasoning [12]. However, the necessity of a restriction is more an issue of ontological commitment [13] that would be dropping out some class instances, owing to the lack of generality of definitions.

That is, if, in the first assumption scenario (i.e., basal transcription factors are bacterial transcription factors), we reuse the current conceptualization of SO and then create an instance or a subclass representing a specific promoter lacking the TF binding site constraint, either no logical inconsistency will rise owing to the open world assumption [6] or the reasoner will fail to infer the subsuming relation and we are going lose track of this entity as a promoter.

We are currently applying this approach to generate an ontology on prokaryotic gene regulation. In the process, we are reviewing the applicability of definitions of the existing ontologies. This step-by-step workflow can ease up the involvement of domain-experts in the generation of logically-sound ontological definitions based on ontological realism. However, we have not planned any training session to help other groups to check their ontological definitions.

IV. LIMITATIONS

This approach can be useful to apply OBO principle of maintenance [3]. However, as it requires huge human effort, we believe it could be applied in a top-down approach to check for the necessity and sufficiency of the most general or prominent concepts of a domain.

ACKNOWLEDGMENT

C.M.A. is a doctoral student from Programa de Doctorado en Ciencias Biomédicas, UNAM, receives fellowship 576333 from CONACYT and received financial aid from Programa de Apoyos para Estudios de Posgrado (PAEP) for this conference.

REFERENCES

[1] Xiang Z, Mungall C, Ruttenberg A, He Y. Ontobee: A Linked Data Server and Browser for Ontology Terms. *Proceedings of the 2nd International Conference on Biomedical Ontologies (ICBO)*, July 28-30, 2011, Buffalo, NY, USA. Pages 279-281. URL: http://ceur-ws.org/Vol-833/paper48.pdf.

[2] Noy, N. F., Shah, N. H., Whetzel, P. L., Dai, B., Dorf, M., Griffith, N., & Musen, M. A. (2009). BioPortal: ontologies and integrated data resources at the click of a mouse. *Nucleic acids research*, *37*(suppl_2), W170-W173.

[3] Smith, B., Ashburner, M., Rosse, C., Bard, J., Bug, W., Ceusters, W., ... & Leontis, N. (2007). The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration. *Nature biotechnology*, *25*(11), 1251.

[4] Musen, M. A., Noy, N. F., Shah, N. H., Whetzel, P. L., Chute, C. G., Story, M. A., ... & NCBO team. (2011). The national center for biomedical ontology. *Journal of the American Medical Informatics Association*, *19*(2), 190-195.

[5] Hunter, L. E. (2018). Mechanistic hypothesis generation in molecular biology: A grand challenge for knowledge-based reasoning.

[6] Horridge, M., Knublauch, H., Rector, A., Stevens, R., & Wroe, C. (2004). A practical guide to building OWL ontologies using the Protégé-OWL plugin and CO-ODE tools edition 1.0. *University of Manchester*.

[7] Devlin, K. Introduction to Mathematical Thinking. [Week 2: Equivalence] MOOC offered by Stanford University through Coursera. Retrieved June 30th, 2018 from https://www.coursera.org/learn/mathematical-thinking/lecture/A5msF/lecture-4-equivalence

[8] Popper, Karl. *The logic of scientific discovery*. Routledge, 2005.

[9] Mungall, Christopher J., Colin Batchelor, and Karen Eilbeck. "Evolution of the Sequence Ontology terms and relationships." *Journal of biomedical informatics* 44.1 (2011): 87-93.

[10] Liang, S-T., et al. "Activities of constitutive promoters in Escherichia coli1." *Journal of molecular biology* 292.1 (1999): 19-37.

[11] Miravet-Verde, S., Lloréns-Rico, V., & Serrano, L. (2017). Alternative transcriptional regulation in genome-reduced bacteria. *Current opinion in microbiology*, *39*, 89-95.

[12] Hunter, L. E. Knowledge-based biomedical Data Science. *Data Science*, (Preprint), 1-7.

[13] Guarino, N., Oberle, D., & Staab, S. (2009). What is an ontology?. In *Handbook on ontologies* (pp. 1-17). Springer, Berlin, Heidelberg

Las definiciones recuperadas de la literatura

Cuando hablamos sobre la regulación génica, usamos términos que tratan de generalizar hechos biológicos. Algunos de estos términos fueron propuestos inicialmente por una generación de científicos quienes son considerados los precursores de la biología molecular y que estudiaron los genes y su expresión en sistemas particulares. En otras palabras, los términos generales se derivaron del estudio de algunos casos particulares. Con el paso del tiempo, se ha acumulado conocimiento nuevo y esto ha provocado que los términos se usen con significados diferentes o, algunas veces, los hechos nuevos no caben en las definiciones originales. A continuación se presenta la compilación de definiciones recuperadas de la literatura que ilustran dicha variación. Junto a estas, se presentan las propiedades definitorias usadas en cada una de ellas en una especie de pseudocódigo lógico. Estas propiedades se sometieron al análisis de necesidad suficiencia anteriormente descrito.

SUPPLEMENTARY INFORMATION

Table S1. Ontological definitions for "transcription unit" and "operon" and their sources. Definitions vary in their generality and there are ambiguities between the concepts of transcription units and operons. The most general definition is simply a set of adjacent co-transcribed genes and this may define either, whilst more specific definitions add restrictions. The ontological pseudocode describes the logics of the definitions. For instance, the original definition of an operon as a "unit of coordinated expression constituted by an operator and the group of closely linked structural genes coordinated by it, these genes corresponding to enzymes that belong to the same sequence of biochemical reactions" implies that it defines a set of adjacent co-transcribed genes from a single operator, and its genes are involved in a metabolic pathway. In the same way we translated all other definitions into their logic requirements.

DEFINITION	ONTOLOGICAL PSEUDOCODE
Operon: The regulator genes that have been identified so far present the remarkable property of exerting a pleitropic coordinated effect, each one governing the expression of many closely linked structural genes that correspond to proteins-enzymes that are involved in the same biochemical sequence. To explain this effect, it seems necessary to assume the existence of a new genetic entity, called "operator", that: a) is adjacent to a group of genes and it commands their activity; b) is sensitive to the repressor produced by a particular regulator gene… the units of coordinated expression (operons) are constituted by an operator and the group of structural genes coordinated by it (Jacob, Perrin, Sanchez, & Monod, 1960)	'set of adjacent cotranscribed genes' and ('is regulated by' exactly 1 operator) and ('involved in' exactly 1 'metabolic pathway')
Operon: The DNA segment whose transcription is thus "coordinated" by a given operator may involve one or several genes (or cistrons); it constitutes a unit of genetic expression called an operon(Monod et al., 1963)	DNA segment and ('is coordinated by' exactly 1 operator)
Operon: The operon, as defined by Jacob & Monad, is a unit of primary transcription coordinated by a single operator (Maas & Clark, 1964)	'set of adjacent cotranscribed genes' and ('is regulated by' exactly 1 operator)
Transcription unit (TU): a contiguous DNA region corresponding to many polypeptide chains of related or different functions(Gilbert, 1978).	Set of adjacent co-transcribed genes

This definition is also found with slightly different wording but with the exact same ontological content in the following references: used to define TUs (X. Mao et al., 2015; Okuda et al., 2007); used to define operon,(Bockhorst et al., 2003; Okuda, Katayama, Kawashima, & Goto, 2006; Pertea, Ayanbule, Smedinghoff, & Salzberg, 2009)	
Transcription unit: the complete extent of a sequence of DNA that is transcribed to produce a single mRNA transcript. That is, each of our predicted TUs consists of a sequence of co-transcribed genes along with the associated polymerase binding site and transcription termination signal (Bockhorst et al., 2003)	'set of adjacent cotranscribed genes' and ('begins in' exactly 1 promoter) and ('ends in' exactly 1 terminator)
Transcription unit: refers to operons according to the original definition of Jacob and Monod, i.e. sequences of consecutive genes that each encode a single RNA molecule along with their own promoters and terminators. The typical relationship between operons and TUs is that TUs tend to be sub-units of operons, while in some cases, a TU may span more than one operon (X. Mao et al., 2014)	'set of adjacent cotranscribed genes' and ('begins in' exactly 1 promoter) and ('ends in' exactly 1 terminator) and ('part of' some operon')
Transcription unit: a set of one or more genes transcribed from a single promoter. A TU may also include regulatory protein binding sites affecting this promoter and a terminator. Note: A complex operon with several promoters contains, therefore, several transcription units. According to the definition of operon, at least one transcription unit must include all the genes in the operon (Santos-Zavaleta et al., 2018)	'set of adjacent cotranscribed genes' and ('begins in' exactly 1 promoter) and ('ends in' exactly 1 terminator) and ('part of' some operon') and ('has part' some operator)
Transcription unit: bacterial transcription unit is defined as having one or more ORFs that are transcribed from one promoter into a single mRNA. Multiple transcription units can be obtained from a single modular unit, if it contains multiple TSSs (Cho et al., 2009)	'set of adjacent cotranscribed genes' and ('has part' exactly 1 'transcription start site') and ('begins in' exactly 1 promoter)
Transcription unit: A TU consists of a promoter, a transcriptional start site, a coding region containing one or multiple genes, and a transcription terminator. When a TU is expressed, its gene(s) are transcribed into a single RNA molecule(Chou et al., 2015)	'set of adjacent cotranscribed genes' and ('has part' exactly 1 promoter) and ('has part' exactly 1 'transcription start site') and ('has part' exactly 1 terminator)

Operon: While the classical definition of operons is the same as that of TUs, a common practice in the past two decades, popularized by operon databases such as ODB, DBTBS, OperonDB and DOOR, has been that operons do not overlap with each other and hence can be derived in general based on genomic sequence information alone(Chou et al., 2015)	'set of adjacent cotranscribed genes' and (not (overlaps some operon)) and ('has part' exactly 1 promoter) and ('has part' exactly 1 'transcription start site') and ('has part' exactly 1 terminator)
Operon: a group of genes arranged in tandem on the same strand of a genome sharing a common promoter and terminator. Genes in the same operon are transcribed together as one messenger RNA(F. Mao, Dam, Chou, Olman, & Xu, 2009) The same definitions is used in DBTBS (Sierro, Makita, Hoon, & Nakai, 2008)	'set of adjacent cotranscribed genes' and ('is regulated by' some promoter) and ('has part' some terminator)
Operon: we use operons to refer to static non-overlapping 'transcriptional units'(X. Mao et al., 2014)	'set of adjacent cotranscribed genes' and (not (overlaps some 'operon'))
Transcription unit: We define transcriptome structure as the collection of TSSs and TTSs that together characterize transcriptional units (mono- and polycistronic mRNAs, tRNAs, rRNAs, and other ncRNAs) (Koide et al., 2009) The same definition is used in (Li, Dong, & Su, 2013) to define operon.	'set of adjacent cotranscribed genes' and ('begins in' exactly 1 'transcription start site') and ('ends in' exactly 1 'transcription termination site')
Large operon or complex operon: Under different conditions, large operons could be transcribed as smaller sub-operons (Li et al., 2013)	'set of adjacent cotranscribed genes' and ('has part' some operon)
Operon: A group of contiguous genes transcribed as a single (polycistronic) mRNA from a single regulatory region (retrieved in October 2018 from (http://www.sequenceontology.org/browser/current_svn/term/SO:0000178)(Eilbeck et al., 2005)	'set of adjacent cotranscribed genes' and ('is regulated by' exactly 1 operator) and ('is regulated by' exactly 1 promoter) and ('is regulated by' exactly 1 terminator)
Operon: set of one or several genes and their associated regulatory elements, which are transcribed as a single unit. We extended the definition in order to include the possibility of operons with only one gene. Note: An operon is, therefore, one or more contiguous genes	'set of adjacent cotranscribed genes' and ('has part' some operator) and ('has part' some promoter)

transcribed in the same direction. Please note that, according to this definition, an operon must contain a promoter upstream of all genes and a terminator downstream. It is also relatively common to find operons with several promoters, some of them internally located, thus, transcribing a partial group of genes. In all cases so far, one gene belongs to only one operon(Santos-Zavaleta et al., 2018)	
Transcription unit: sequence of nucleotides in DNA that codes for a single RNA molecule, along with the sequences necessary for its transcription; normally contains a promoter, an RNA-coding sequence, and a terminator [https://www.nature.com/scitable/definition/transcription-unit-260].	DNA region and (codes for exactly 1 RNA molecule) and ('has part' exactly 1 promoter) and ('has part' exactly 1 terminator)
Transcription unit: a linear sequence of DNA that extends from a transcription start site to a transcription stop site(Pray, 2008)	DNA region and ('begins in' exactly 1 transcription start site) and ('ends in' exactly 1 transcription stop site)

Table S2. Different definitions of "promoter".

DEFINITIONS	ONTOLOGICAL PSEUDOCODE
A sequence located between the operator and the beginning of the operon indispensable for operon expression (Jacob, Ullman, & Monod, 1964)	DNA region and (Locates after exactly 1 operator) (Locates before exactly 1 operon start) Indispensable for exactly operon expression
A site which serves to initiate transcription of an operon (Epstein & Beckwith, 1968)	DNA region and (Initiates transcription of exactly 1 operon)
In functional and molecular terms, a promoter is a region along a DNA template at which RNA polymerase first "recognizes" some sequence, then "melts out" a section of the DNA (42), forming a tight binary complex, and then initiates an RNA chain with ATP or GTP (29) as it begins transcribing the template (Pribnow, 1975)	DNA region and (is recognized by RNA polymerase) (melts out some DNA region) (initiates some RNA chain transcription)
Promoters contain, as an essential part, a site for the oriented and stable attachment of the enzyme RNA polymerase, thus ensuring the strand- and site-specific initiation of the mRNA chain. This promoter DNA fragment, pDNA-I, contains all the information necessary for	DNA region and has part some RNA polymerase attachment site initiates some site specific mRNA transcription

maintaining the enzyme in the stable promoter complex (3), and for the specific initiation of transcription pDNA-I contains three regions with 2-fold rotational symmetry. These are located around symmetry axes at positions +10, +4, and -10 (Schaller, Gray, & Herrmann, 1975)	
A DNA sequence at which RNA polymerase binds and initiates transcription (Golbeck et al., 2003)	DNA region and Binds some RNA polymerase Initiates some transcription
Promoter: A regulatory_region composed of the TSS(s) and binding sites for TF_complexes of the basal transcription machinery (Mungall, Batchelor, & Eilbeck, 2011) Bacterial RNA-polymerase promoter: A DNA sequence to which bacterial RNA polymerase binds, to begin transcription (Mungall et al., 2011)	DNA region and Has part TSS Has part basal TF binding sites Promoter and Initiates some transcription Binds some bacterial RNA polymerase
A gene promoter is a region of DNA that initiates transcription of a particular gene (Dumontier et al., 2014)	DNA region and Initiates transcription of exactly 1 gene

Table S3. Definitions of transcription factor (TF) and transcription factor binding site (TFBS). We include the definition for TF activity of the Gene Ontology (GO). We can see that the definitions for TF vary in the kind of entity it is regarded as (any entity, a gene product, a protein, a molecular function, or a role); the function of a TF (repressor of gene expression, repressor of transcription, regulator of transcription, or regulator of the binding of the RNA polymerase); and the mechanism by which TFs effect their function (by specifically binding to DNA, by binding to a promoter, by binding to a TFBS, or by binding either to a promoter or an enhancer). On the other hand, according to these definitions, a TFBS site is either any entity, a genetic entity, a nucleotide region, a regulatory region, or a DNA region, whose function is either to regulate the expression of an operon, regulate the binding of the RNA polymerase, repress transcription, or regulate transcription; some definitions state that TFBS location must overlap promoter or be adjacent to an operon.

DEFINITION	**ONTOLOGICAL PSEUDOCODE**
Transcription factor	
The product of a regulator gene that controls the rate of information transfer from structural genes to proteins, acting as a cytoplasmic repressor, without contributing any information to the protein themselves (Jacob et al., 1960)	gene product and (represses some gene expression) and (present in the cytoplasm)

A protein that blocks transcription from DNA to RNA by directly binding to DNA (Ptashne, 1967)	Protein and (Represses some transcription) and (binds some DNA)
transcription regulator activity: A molecular function that controls the rate, timing and/or magnitude of transcription of genetic information. The function of transcriptional regulators is to modulate gene expression at the transcription step so that they are expressed in the right cell at the right time and in the right amount throughout the life of the cell and the organism. (Ashburner et al., 2000) DNA binding transcription factor activity: Interacting selectively and non-covalently with a specific DNA sequence (sometimes referred to as a motif) within the regulatory region of a gene in order to modulate transcription. (Ashburner et al., 2000) bacterial-type RNA polymerase transcription factor activity, sequence-specific DNA binding: Interacting selectively and non-covalently with a specific DNA sequence in order to modulate transcription by bacterial-type RNA polymerase. The transcription factor may or may not also interact selectively with a protein or macromolecular complex (Ashburner et al., 2000)	Molecular function and (regulates some transcription) Transcription regulatory activity and (interact specifically with some DNA) not (binds covalently to DNA) DNA binding transcription regulatory activity and (regulates some bacterial RNAP activity)
Transcription factors are a diverse group of proteins that bind to DNA at specific promoter or enhancer regions. They also bind to DNA-associated proteins to initiate, stimulate, inhibit or terminate transcription. The proteins are often physically associated in a preinitiation complex and contain structural motifs (Golbeck et al., 2003)	Protein and (regulates some transcription) and (binds specifically to some promoter or some enhancer or some DNA-bound protein)
Proteins that bind to promoters and either up- or down-regulate transcription (Browning & Busby, 2004)	Protein and (regulates some transcription) and (binds to some promoter)
Transcription regulator: Protein that has transcription regulator activity. (Beisswanger et al., 2008)	Protein and (has function some transcription regulator)

Transcription factor: A transcription factor that binds to a specific DNA sequence in order to modulate transcription. The transcription factor may or may not also interact selectively with a protein (other transcription factors of cofactors) or protein or macromolecular complex (Beisswanger et al., 2008)	Transcription regulator and (binds specifically to DNA motif)
Transcription factors bind to transcription factor binding sites and modulate the binding of RNA polymerase (Balleza et al., 2009a)	Protein and (regulates some RNAP binding) and (binds to some TFBS)
Transcription factor: proteins that either promote (as an activator) or block (as a repressor) the binding to DNA of RNA polymerase. Proteins that bind to DNA and control transcription of DNA to mRNA (transcription factors) (Ison et al., 2013)	Protein and (regulates some binding of RNAP to DNA) (binds to DNA)
Transcription factor: A role played by a protein that binds to specific DNA sequences, thereby controlling the transcription of genetic information from DNA to mRNA (Dritsou, Topalis, Mitraka, Dialynas, & Louis, 2014)	Role and Has agent some DNA binding protein Regulates some transcription
A factor which affects transcription (Borukhov, Lee, & Laptenko, 2005; Visweswariah & Busby, 2015)	Entity and (regulates some transcription)
Transcription factor binding site	
A new genetic entity, called operator, that is adjacent to a group of genes, commands their activities and is sensible to repressor produced by a particular regulator gene (Jacob et al., 1960)	Genetic entity and Is adjacent to some operon Sensible to some repressor Regulates some operon expression
Operator: specific DNA binding site with high affinity for the repressor, which union blocks transcription from DNA to RNA(Ptashne, 1967)	DNA region and Has high affinity for some repressor Represses some transcription
Operator : DNA regions specifically recognized by repressor overlapped by RNA polymerase binding sites (Maniatis et al., 1975)	DNA region and Spefically recognized by some repressor Overlaps some RNA polymerase binding site

TFBS: A region of a nucleotide molecule that binds a Transcription Factor or Transcription Factor complex Operator: A regulatory element of an operon to which activators or repressors bind thereby effecting translation of genes in that operon (Eilbeck et al., 2005)	Nucleotide molecule region and Binds some TF or TF complex Regulatory region and Binds some TF and Regulates translation
Transcription factor-binding sites (TFBS) – where transcription factors (TFs) bind to modulate the binding of the RNA polymerase (Balleza et al., 2009b)	Entity and binds some TF Regulates some RNAP binding

Table S4. Regulon definitions extracted from the literature and existing ontologies.
Definitions of regulons vary along four dimensions: the kind of regulatory entity (a repressor substance, a TFBS, a TF, a sRNA, a TF set, or a regulatory signal), the kind of regulated entity (a set of enzymes, a group of operons, or a group of genes), the kind of regulation (positive, negative, or both), and the level of gene expression in which they are involved.

DEFINITION	ONTOLOGICAL PSEUDOCODE
A system in which the production of all enzymes (of a metabolic pathway) can be controlled by a single repressor substance, this substance may consist of several entities, but whatever nature, it acts in a unitary fashion (Maas & Clark, 1964)	system of enzymes and regulated by exactly one repressor substance
A group of genes, whether linked as a cluster or not, that respond to a common regulatory signal (Eilbeck et al., 2005)	group of genes, whether linked as a cluster or not, that respond to a common regulatory signal (Eilbeck et al., 2005)
A group of operons that are transcriptionally co-regulated by the same regulatory machinery, consisting of trans regulators (transcription factors or simply TFs) and cis regulatory binding elements in the promoters of the operons they regulate. Operationally, a regulon contains operons regulated by one same transcription factor (Zhang, Yin, Olman, & Xu, 2012)	group of operons and transcriptionally regulated by exactly one TF

Genes and operons directly co-regulated by the same TF (or by RNA motifs from the same structural family) (Novichkov et al., 2013)	group of operons and regulated by some (TF or riboswitch)
Maximal set of operons in a genome sharing conserved cis regulatory motifs around the promoter region (TFBS) (Liu et al., 2016)	group of operons and regulated by some regulatory motif
Simple regulon: a set of genes subject to regulation of one and only one transcription factor Complex regulon: a group of genes subject to regulation by two or more regulators, where all genes are subject to the regulation of exactly the same transcription factors. Strict complex regulon: a set of genes subject to regulation by two or more transcription factors, where the effect of each regulator (activator or repressor) is the same for all the regulated genes (Santos-Zavaleta et al., 2018)	set of genes and regulated by exactly one TF set of genes and regulated by some transcription factor set set of genes and (positively regulated by some transcription factor set) or (negatively regulated by some transcription factor set)

Table S5. Definitions and ontological pseudocode for "Signal" and "Effector" retrieved from the literature and existent ontologies.
Signal definitions vary along four kinds of properties: the origin of the signal, the kind of entity, the effect of the signal, and the sensing mechanism. Whereas the effector definition varies only concerning the kind of entity an effector is, and the kind of entity subject to its effects.

DEFINITION	ONTOLOGICAL PSEUDOCODE
Signal	
Regulatory signal: frequently a small regulatory molecule like cAMP or ppGpp that indicates an environmental change, induces a cell response recognizable as the induction or repression of particular operons, and can be rapidly synthesized and degraded (Gottesman, 1984)	Small molecule and indicates some environmental change induces some gene expression change
Molecular signal: a molecule originated from their environment or produced by metabolism to which an appropriate cellular response must be mounted. In the majority of cases the	Molecule and originated in (environment or internal

response is transcriptional activation of genes whose products specifically cope with a given molecular signal (Hoch, 2000)	metabolism) induces some cellular response
Signalling molecules, also referred to as autoinducers, bind to receptors on, or in, the bacterial cell, which leads to changes in gene expression at some threshold concentration (Keller & Surette, 2006) Signal: any act, structure or chemical emission that alters the behaviour and gene expression of other organisms which evolved because of that effect, and that is effective because the receiver's response has also evolved. (Keller & Surette, 2006) Cue: an act, structure or chemical emission that alters the behavior and gene expression of other organisms. However, contrary to a signal, it did not evolve specifically for that effect (Keller & Surette, 2006)	Molecule and binds to some receptor and induces gene expression change Entity and emitted by some organism A and alters gene expression of some organism B and organism A is different from organism B is evolved to some specific communication Entity and emitted by some organism A and alters gene expression of some organism B and organism A is different from organism B not (is evolved to some specific communication)
A signal is an object that initiates a sequence of events (SIO ontology) (Dumontier et al., 2014)	object and initiates some sequence of events
Signaling molecule: A molecular messenger in which the molecule is specifically involved in transmitting information between cells. Such molecules are released from the cell sending the signal, cross over the gap between cells by diffusion, and interact with specific receptors in another cell, triggering a response in that cell by activating a series of enzyme controlled reactions which lead to changes inside the cell (CHEBI ontology) (Hastings et al., 2016)	Molecule and emitted by some emitting cell and interacts specifically with some cell receptor and triggers a response in some receptor cell and
Effector	
Compounds that bind specifically and reversibly to an allosteric site (a site other than the catalytic one) and do not participate in the reaction, but bring about a discrete reversible alteration of the molecular structure of the protein (mostly enzymes, but possibly	Compound and binds specifically and reversibly to some allosteric site induces some reversible alteration of protein molecular structure changes some kinetic parameter

also genetic repressors) or allosteric transition, which modifies the properties of the active site, changing one or several of the kinetic parameters which characterize the biological activity of the protein. (Monod et al., 1963) Positive effectors or co-repressor (activating the repressor and thereby blocking messenger and protein synthesis) (Monod et al., 1963) Negative effectors or inducer (inhibiting the repressor and thereby inducing the synthesis of the messenger and of the protein(s)) (Monod et al., 1963)	 Effector and activates some repressor Effector and inhibits some repressor
Allosteric effector: a ligand that elicits an allosteric response upon binding to a target molecule (MI ontology) (http://purl.obolibrary.org/obo/mi.owl)	Ligand and elicits some allosteric response
The effector (i.e., the molecule that induces a conformational change) can be a small molecule such as a metabolite, but in other cases, the effector is a large molecule—another protein or even DNA, for example. In still other cases, the effector is an enzyme that covalently modifies the target protein—a kinase, for example (Ptashne & Gann, 2002)	Molecule and induces some conformational change
A small molecule which increases (activator) or decreases (inhibitor) the activity of an (allosteric) enzyme by binding to the enzyme at the regulatory site (which is different from the substrate-binding catalytic site) (CHEBI ontology) (Hastings et al., 2016)	Small molecule and binds to some enzyme regulatory site regulates some enzyme activity

References

Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., … Sherlock, G. (2000). Gene ontology: Tool for the unification of biology. *Nature Genetics*, *25*(1), 25–29. https://doi.org/10.1038/75556

Balleza, E., López-Bojorquez, L. N., Martínez-Antonio, A., Resendis-Antonio, O., Lozada-Chávez, I., Balderas-Martínez, Y. I., … Collado-Vides, J. (2009a). Regulation by transcription factors in bacteria: Beyond description. *FEMS Microbiology Reviews*. https://doi.org/10.1111/j.1574-6976.2008.00145.x

Balleza, E., López-Bojorquez, L. N., Martínez-Antonio, A., Resendis-Antonio, O., Lozada-Chávez, I., Balderas-Martínez, Y. I., … Collado-Vides, J. (2009b). Regulation by transcription factors in bacteria: Beyond description. *FEMS Microbiology Reviews*, *33*, 133–151. https://doi.org/10.1111/j.1574-6976.2008.00145.x

Beisswanger, E., Lee, V., Kim, J.-J., Rebholz-Schuhmann, D., Splendiani, A., Dameron, O., … Hahn, U. (2008). Gene Regulation Ontology (GRO): design principles and use cases. *EHealth Beyond the Horizon: Get IT There*, *136*, 9–14. https://doi.org/10.1111/j.1365-313X.2005.02352.x

Bockhorst, J., Qiu, Y., Glasner, J., Liu, M., Blattner, F., & Craven, M. (2003). Predicting bacterial transcription units using sequence and expression data, *19*(1). https://doi.org/10.1093/bioinformatics/btg1003

Borukhov, S., Lee, J., & Laptenko, O. (2005). Bacterial transcription elongation factors: new insights into molecular mechanism of action. *Molecular Microbiology*, *55*(5), 1315–1324. https://doi.org/10.1111/j.1365-2958.2004.04481.x

Browning, D. F., & Busby, S. J. W. (2004). The regulation of bacterial transcription initiation. *Nature Reviews Microbiology*, *2*(1), 1–9. https://doi.org/10.1038/nrmicro787

Cho, B., Zengler, K., Qiu, Y., Park, Y. S., Knight, E. M., Barrett, C. L., … Palsson, B. Ø. (2009). The transcription unit architecture of the Escherichia coli genome. *Nature Biotechnology*, *27*(11), 1043–1049. https://doi.org/10.1038/nbt.1582

Chou, W., Ma, Q., Yang, S., Cao, S., Klingeman, D. M., Brown, D., & Xu, Y. (2015). Analysis of strand-specific RNA-seq data using machine learning reveals the structures of transcription units in Clostridium thermocellum. *Nucleic Acids Research*, *43*(10), 1–8. https://doi.org/10.1093/nar/gkv177

Dritsou, V., Topalis, P., Mitraka, E., Dialynas, E., & Louis, C. (2014). miRNAO: An Ontology Unfolding the Domain of microRNAs. In *IWBBIO* (pp. 989–1000). Granada. Retrieved from https://pdfs.semanticscholar.org/de8a/a2405a149c931ad60cc141eaba95bbbb7ffb.pdf

Dumontier, M., Baker, C. J. O., Baran, J., Callahan, A., Chepelev, L., Cruz-Toledo, J., … Hoehndorf, R. (2014). The semanticscience integrated ontology (SIO) for biomedical research and knowledge discovery. *Journal of Biomedical Semantics*, *5*(1), 1–11. https://doi.org/10.1186/2041-1480-5-14

Eilbeck, K., Lewis, S. E., Mungall, C. J., Yandell, M., Stein, L., Durbin, R., & Ashburner, M. (2005). The Sequence Ontology: a tool for the unification of genome annotations. *Genome Biology*, *6*(5), R44.1-R44.12. https://doi.org/10.1186/gb-2005-6-5-r44

Epstein, W., & Beckwith, J. R. (1968). Regulation of Gene Expression. *Annual Review of Biochemistry*, *37*(1), 411–436.

Gilbert, W. (1978). Why genes in pieces? *Nature*, *271*(9), 501. Retrieved from https://www.nature.com/articles/271501a0

Golbeck, J., Fragoso, G., Hartel, F., Hendler, J., Oberthaler, J., & Parsia, B. (2003). The National Cancer Institute ' s Thésaurus and Ontology. *Journal of Web Semantics First Look 1_1_4*.

Gottesman, S. (1984). Bacterial Regulation: Global Regulatory Networks. *Annual Review of Genetics*, *18*(1), 415–441.

Hastings, J., Owen, G., Dekker, A., Ennis, M., Kale, N., Muthukrishnan, V., … Steinbeck, C. (2016). ChEBI in 2016: Improved services and an expanding collection of metabolites. *Nucleic Acids Research*, *44*(D1), D1214–D1219. https://doi.org/10.1093/nar/gkv1031

Hoch, J. A. (2000). Two-component and phosphorelay systems Signal transduction systems. *Current Opinion in Microbiology*, *3*, 165–170. Retrieved from http://www2.hawaii.edu/~scallaha/SMCsite/475 Lectures/475Lecture36.pdf

Ison, J., Kalaš, M., Jonassen, I., Bolser, D., Uludag, M., McWilliam, H., … Rice, P. (2013). EDAM: An ontology of bioinformatics operations, types of data and identifiers, topics and formats. *Bioinformatics*, *29*(10), 1325–1332. https://doi.org/10.1093/bioinformatics/btt113

Jacob, F., Perrin, D., Sanchez, C., & Monod, J. (1960). L'opéron: groupe de gènes à expression coordonnée par un opérateur. *Comptes Rendus de l'Academie Des Sciences*, *250*, 1727–1729.

Jacob, F., Ullman, A., & Monod, J. (1964). Le promoteur, élément génétique nécesaire à l'expression d'un opéron, *258*, 3125–3128.

Keller, L., & Surette, M. G. (2006). Communication in bacteria: An ecological and evolutionary perspective. *Nature Reviews Microbiology*, *4*(4), 249–258. https://doi.org/10.1038/nrmicro1383

Koide, T., Reiss, D. J., Bare, J. C., Pang, W. L., Facciotti, M. T., Schmid, A. K., … Baliga, N. S. (2009). Prevalence of transcription promoters within archaeal operons and coding sequences. *Molecular Systems Biology*, *5*(285), 1–16. https://doi.org/10.1038/msb.2009.42

Li, S., Dong, X., & Su, Z. (2013). Directional RNA-seq reveals highly complex condition-dependent transcriptomes in E . coli K12 through accurate full-length transcripts assembling. *BMC Genomics*, *14*(1), 1–24.

Liu, B., Zhou, C., Li, G., Zhang, H., Zeng, E., & Liu, Q. (2016). Bacterial regulon modeling and prediction based on systematic cis regulatory motif analyses. *Nature Publishing Group*, 1–11. https://doi.org/10.1038/srep23030

Maas, W. K., & Clark, A. J. (1964). Studies on the mechanism of repression of arginine biosynthesis in Escherichia coli: II. Dominance of repressibility in diploids. *Journal of Molecular Biology*, *8*(3), 365–370. https://doi.org/10.1016/S0022-2836(64)80200-X

Maniatis, T., Ptashne, M., Backman, K., Kleid, D., Flashman, S., Jeffrey, A., & Maurer, R. (1975). Recognition sequences of repressor and polymerase in the operators of bacteriophage lambda. *Cell*, *5*(2), 109–113. https://doi.org/10.1016/0092-8674(75)90018-5

Mao, F., Dam, P., Chou, J., Olman, V., & Xu, Y. (2009). DOOR : a database for prokaryotic operons, *37*(November 2008), 459–463. https://doi.org/10.1093/nar/gkn757

Mao, X., Ma, Q., Liu, B., Chen, X., Zhang, H., & Xu, Y. (2015). Revisiting operons: An analysis of the landscape of transcriptional units in E. coli. *BMC Bioinformatics*, *16*(356), 1–9. https://doi.org/10.1186/s12859-015-0805-8

Mao, X., Ma, Q., Zhou, C., Chen, X., Zhang, H., Yang, J., … Xu, Y. (2014). DOOR 2 . 0 : presenting operons and their functions through dynamic and integrated views, *42*(November 2013), 654–659. https://doi.org/10.1093/nar/gkt1048

Monod, J., Changeux, J. P. J. P. E., Jacob, F. F., Ioxod, J., Changeux, J. P. J. P. E., Jacob, F. F., & Biochimie, S. De. (1963). Allosteric proteins and cellular control systems. *Journal of*

Molecular Biology, *6*(4), 306–329. https://doi.org/10.1016/S0022-2836(63)80091-1
Mungall, C. J., Batchelor, C., & Eilbeck, K. (2011). Evolution of the Sequence Ontology terms and relationships. *Journal of Biomedical Informatics*, *44*(1), 87–93. https://doi.org/10.1016/j.jbi.2010.03.002
Novichkov, P. S., Kazakov, A. E., Ravcheev, D. A., Leyn, S. A., Kovaleva, G. Y., Sutormin, R. A., … Rodionov, D. A. (2013). RegPrecise 3 . 0 – A resource for genome-scale exploration of transcriptional regulation in bacteria.
Okuda, S., Katayama, T., Kawashima, S., & Goto, S. (2006). ODB : a database of operons accumulating known operons across multiple genomes, *34*, 358–362. https://doi.org/10.1093/nar/gkj037
Okuda, S., Kawashima, S., Kobayashi, K., Ogasawara, N., Kanehisa, M., & Goto, S. (2007). Characterization of relationships between transcriptional units and operon structures in Bacillus subtilis and Escherichia coli. *BMC Genomics*, *8*, 1–12. https://doi.org/10.1186/1471-2164-8-48
Pertea, M., Ayanbule, K., Smedinghoff, M., & Salzberg, S. L. (2009). OperonDB : a comprehensive database of predicted operons in microbial genomes, *37*(October 2008), 479–482. https://doi.org/10.1093/nar/gkn784
Pray, L. A. (2008). What is a gene? Colinearity and transcription units. *Nature Education*, *1*(1).
Pribnow, D. (1975). Nucleotide sequence of an RNA polymerase binding site at an early T7 promoter. *Proceedings of the National Academy of Sciences of the United States of America*, *72*(3), 784–788. https://doi.org/10.1073/pnas.72.3.784
Ptashne, M. (1967). Specific Binding of the λ Phage Repressor to λ DNA. *Nature*, *214*(5085), 232–234. https://doi.org/10.1038/214232a0
Ptashne, M., & Gann, A. (2002). *Genes & Signals*. (S. Curtis, D. Brown, & S. Schaefer, Eds.). New York: Cold Spring Laboratory Press.
Santos-Zavaleta, A., Sánchez-Pérez, M., Salgado, H., Velázquez-Ramírez, D. A., Gama-Castro, S., Tierrafría, V. H., … Collado-Vides, J. (2018). A unified resource for transcriptional regulation in Escherichia coli K-12 incorporating high-throughput-generated binding data into RegulonDB version 10.0. *BMC Biology*, *16*(1), 1–12. https://doi.org/10.1186/s12915-018-0555-y
Schaller, H., Gray, C., & Herrmann, K. (1975). Nucleotide Sequence of an RNA Polymerase Binding Site from the DNA of Bacteriophage fd. *Proceedings of the National Academy of Sciences*, *72*(2), 737–741.
Sierro, N., Makita, Y., Hoon, M. De, & Nakai, K. (2008). DBTBS : a database of transcriptional regulation in Bacillus subtilis containing upstream intergenic conservation information, *36*(October 2007), 93–96. https://doi.org/10.1093/nar/gkm910
Visweswariah, S. S., & Busby, S. J. W. (2015). Evolution of bacterial transcription factors: How proteins take on new tasks, but do not always stop doing the old ones. *Trends in Microbiology*, *23*(8), 463–467. https://doi.org/10.1016/j.tim.2015.04.009
Zhang, H., Yin, Y., Olman, V., & Xu, Y. (2012). Genomic Arrangement of Regulons in Bacterial Genomes, *7*(1). https://doi.org/10.1371/journal.pone.0029496

Actualización de las definiciones

Como resultado del análisis de la necesidad y suficiencia de las definiciones anteriores, se propusieron definiciones actualizadas para los conceptos de promotor, factor de transcripción, sitio de unión del factor de transcripción, operón, unidad de transcripción, regulón, efector y señal. Las definiciones fueron incluidas en la ontología *Sequence Ontology* que forma parte de la *OBO Foundry*. La definiciones y los contraejemplos usados para falsificar la necesidad y la suficiencia de las propiedades se explicaron, ilustraron y publicaron en el artículo (REF 16) que se anexa a continuación.

Redefining fundamental concepts of transcription initiation in bacteria

Citlalli Mejía-Almonte[1], Stephen J. W. Busby[2], Joseph T. Wade[3], Jacques van Helden[4,5], Adam P. Arkin[6], Gary D. Stormo[7], Karen Eilbeck[8], Bernhard O. Palsson[9], James E. Galagan[10] and Julio Collado-Vides[1,10]✉

Abstract | Despite enormous progress in understanding the fundamentals of bacterial gene regulation, our knowledge remains limited when compared with the number of bacterial genomes and regulatory systems to be discovered. Derived from a small number of initial studies, classic definitions for concepts of gene regulation have evolved as the number of characterized promoters has increased. Together with discoveries made using new technologies, this knowledge has led to revised generalizations and principles. In this Expert Recommendation, we suggest precise, updated definitions that support a logical, consistent conceptual framework of bacterial gene regulation, focusing on transcription initiation. The resulting concepts can be formalized by ontologies for computational modelling, laying the foundation for improved bioinformatics tools, knowledge-based resources and scientific communication. Thus, this work will help researchers construct better predictive models, with different formalisms, that will be useful in engineering, synthetic biology, microbiology and genetics.

Gene expression — the transcription of DNA into RNA and the translation of RNA into a polypeptide chain — and its regulation encompass a collection of genetic and molecular programmes that underlie the major biological capabilities of eukaryotic and prokaryotic cellular differentiation and development. Gene expression is regulated in response to environmental conditions, which is critical for bacterial fitness and survival. Any step in the gene expression pathway can be regulated, from transcription initiation to mRNA stability and translation. The foundations of our understanding of gene expression regulation rely on terminology and models derived from research in *Escherichia coli* and bacteriophage λ[1,2]. These fundamental studies were followed by decades of research in a number of regulated bacterial systems, and more recently by studies using high-throughput genomic methodologies and advanced biophysical single-molecule approaches[3–5], which have led to discoveries that could not have been imagined when the original concepts were proposed. Thus, some terms have acquired meanings that differ from their original ones. As experimental biology becomes a data science, mainly because of the advent of genomics, computational models that rigorously organize our knowledge become essential. Databases and ontologies are the two major tools that underpin the computational representation of knowledge. Because these tools are specified in formal language, they require definitions at a level of detail that is beyond what is common in communications among experts.

Here, we focus on transcription initiation, the most studied step in the regulation of gene expression. To discuss the limitations of existing definitions that constitute our core understanding of the regulation of transcription initiation in bacteria, we searched the literature for original and more recent definitions (Supplementary Tables 1–5), as well as for examples of regulatory systems that do not conform to these definitions. A group of experts on the regulation of transcription initiation in bacteria organized a collective process of evaluating the necessity and sufficiency of the different features used to characterize the concepts involved and their relation to other concepts. More precisely, a feature X is not sufficient to define a class A if there is an object that does not belong to A and has feature X. Conversely, if a member of class A does not have feature X, then X is not necessary to the definition of A. On the basis of an initial draft, the authors engaged in systematic discussions, one concept at a time, of how to better expand each concept, until final agreement was reached. Most of the reviewed concepts fall under the scope of the Sequence Ontology, which aims to define the sequence features used in biological sequence annotation, and in which final definitions were added or updated[6]. These updated definitions are also being incorporated into RegulonDB, a database that has curated knowledge on transcriptional regulation

✉*e-mail:* colladojulio@gmail.com

https://doi.org/10.1038/s41576-020-0254-8

Activators
Gene products that increase transcription, indicating that their function is to enhance promoter activity.

Repressors
Gene products that decrease transcription, indicating that their function is to hamper promoter activity.

Operator
A genetic entity adjacent to a group of genes that regulates their expression and is sensitive to a repressor.

Operon
A set of adjacent co-transcribed genes.

in *E. coli*[7] for the past three decades and that populates gene regulation data in EcoCyc[8], a scientific database for *E. coli* K-12 MG1655.

Below, we begin with a brief overview of bacterial transcription initiation. Next, we contrast the classic concepts and terms relating to bacterial transcription initiation and its regulation with their current use, in light of the current body of knowledge on transcriptional regulation in *E. coli*[7] and other bacterial organisms. We aim to construct up-to-date and precise definitions that support a logically consistent conceptualization of gene regulation. We believe that this will provide a reference for knowledge representation, for modelling and for future thinking about bacterial gene regulation. Indirectly, these concepts may also influence understanding of gene regulation frameworks beyond bacteria.

Overview of transcription initiation

The first step in transcription is the formation of the RNA polymerase (RNAP) holoenzyme (Eσ)[9], a molecular complex composed of the core RNAP plus a σ factor, which is capable of initiating gene transcription at specific DNA positions (FIG. 1a). The σ factor interacts with different promoter elements to position the Eσ to unwind the double-stranded DNA in the region of the transcription start site (TSS), which corresponds to the first base of the transcript[6]. Most bacteria rely on different σ factors to take Eσ to different sets of promoters in response to changes in growth conditions[10]. Alternative σ factors are classified into two evolutionarily distinct families: σ^{54} and σ^{70}. The σ^{54} family has a single member, whereas σ^{70} typically has several members, whose number varies among bacterial species. For example, σ^{24}, σ^{28}, σ^{32}, σ^{38} and σ^{70} are all members of the σ^{70} family in *E. coli*[11].

Eσ initially binds to promoters in a closed complex (RPc) (FIG. 1b), whereby it covers DNA from approximately −55 bp to approximately +15 bp relative to the TSS (positive numbers represent bases downstream of the TSS, whereas negative ones represent upstream positions), according to DNA footprints[12]. This binding triggers a series of conformational changes, both in the DNA and in Eσ, that create an open complex (RPo), culminating in separation of the DNA strands from approximately positions −11 to +3 bp (REF.[13]). The region where the DNA strands are separated is often referred to as the 'transcription bubble', and it includes the base on the template strand, designated +1, that will act as the template for the first nucleotide of the transcript[13,14]. Transcription initiates with a short, unstable region potentially subject to abortive transcription, with Eσ synthesizing short products before transitioning to a stable elongation complex[15] (FIG. 1b). The transcription cycle then proceeds with elongation and termination steps, which have been reviewed in detail elsewhere[16].

The role of promoter elements is primarily to interact with Eσ in order to dock the DNA–Eσ complex that is competent for the subsequent steps of transcription; hence, promoter elements determine the autonomous activity of the promoter[17]. The activity of different promoters varies by many orders of magnitude, ranging from promoters that produce less than one RNA copy per cell generation to promoters that generate tens of thousands of RNA copies[18,19].

Essentially all promoters are subject to regulation, either indirectly, by changes in Eσ concentration[20] or substrate concentrations[21,22], or directly, by specific regulators. A subclass of regulators includes activators and repressors, collectively known as transcription factors (TFs), which act by binding to specific DNA targets. Promoters prone to activation are often intrinsically weak, owing to their low affinity for Eσ, with activators compensating for this weakness by recruiting Eσ[23]. By contrast, repressors prevent Eσ from transcribing, often by directly occluding the promoter or preventing some isomerization step[24].

The activity of most DNA-binding transcription regulators is coupled to outside signals by a variety of mechanisms that facilitate quick responses and make gene expression sensitive to environmental changes[25]. Parallel mechanisms sensitive to internal and external changes support our molecular understanding of genetic developmental programmes in eukaryotes[26,27].

Promoter

The original definition of promoter, proposed by Jacob and Monod in 1964, is a sequence located between the operator and the beginning of the operon that is indispensable for operon initiation of gene expression[28]. This definition deserves to be revised; the following discussion shows that there is no well-defined sequence that can define any promoter.

Core RNAP transcription. Although core RNAP can transcribe single-stranded DNA in nicked regions or from DNA ends, and although it has been shown that RNAP can initiate transcription from double-stranded, circular DNA[29,30], sites of core RNAP transcription initiation are not promoters, because they are not specific. Thus, a promoter is not necessary for random transcription, although it is essential for transcript initiation at specific TSSs[30].

Promoter sequence motifs. One of the most prominent characteristics of promoter sequences is the presence of Eσ recognition elements. There is a long history of promoter sequence comparisons and mutations, from which the base sequence consensus motifs

Author addresses

[1]Programa de Genómica Computacional, Centro de Ciencias Genómicas, Universidad Nacional Autónoma de México, Morelos, Cuernavaca, México.
[2]School of Biosciences, University of Birmingham, Birmingham, UK.
[3]Division of Genetics, Wadsworth Center, New York State Department of Health, Albany, NY, USA.
[4]Aix-Marseille University, INSERM UMR S 1090, Theory and Approaches of Genome Complexity (TAGC), Marseille, France.
[5]CNRS, Institut Français de Bioinformatique, IFB-core, UMS 3601, Evry, France.
[6]Department of Bioengineering, University of California, Berkeley, Berkeley, CA, USA.
[7]Department of Genetics, Washington University School of Medicine, St Louis, MO, USA.
[8]Department of Biomedical Informatics, University of Utah School of Medicine, Salt Lake City, UT, USA.
[9]Department of Bioengineering, University of California, San Diego, La Jolla, CA, USA.
[10]Department of Biomedical Engineering, Boston University, Boston, MA, USA.

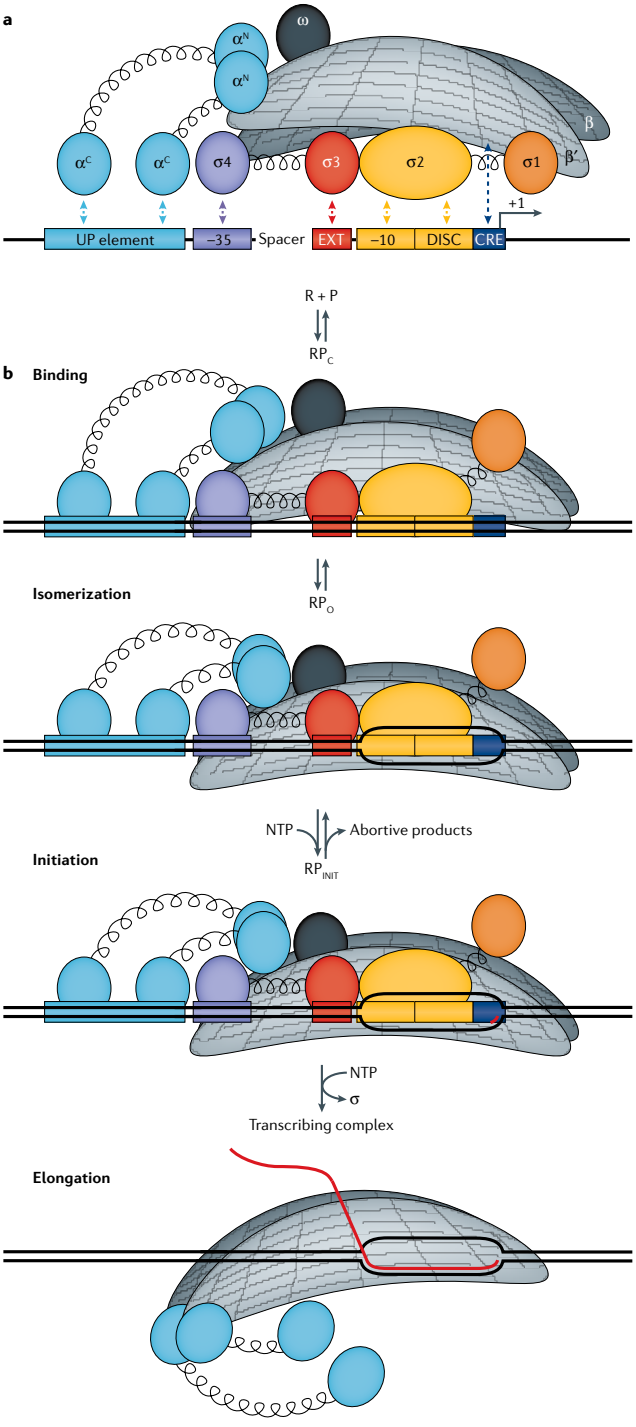

Fig. 1 | **Schematics of bacterial transcription initiation.** **a** | Bacterial RNA polymerase (RNAP) is a multisubunit enzyme. The core RNAP is composed of five subunits: α_2 (cyan bubbles), β and β′ (grey) and ω (black). The core RNAP contains the active site that catalyses the formation of the phosphodiester bond of nascent RNA. The RNAP α subunits interact with the upstream promoter (UP) element, which consists of two distinct subsites located upstream of the −35 element[40]. The RNAP core enzyme interacts in a sequence-specific manner with the template-strand positions −4 to +2, which constitute the core recognition element (CRE) (navy blue)[38]. To form the holoenzyme (Eσ), RNAP is associated with a σ factor. There are two structurally and evolutionarily distinct families of σ factors: σ^{54} (not shown) and σ^{70}. The σ^{70}-related factors contain up to four functional domains (σ1–4; the spirals represent linkers between the domains). The σ2 domain recognizes and interacts with the −10 element, and the σ4 domain interacts with the −35 element. The extended −10 element interacts with σ3; this interaction is crucial in promoters whose −35 and −10 elements show a poor match to consensus sequences[35]. Some promoters, particularly the ones that respond to amino acid starvation, have an element called a discriminator (DISC), which is recognized by and interacts with the σ2 domain (conserved region σ1.2; not shown)[206,207]. The σ^{70} factor bound to the non-template strand captures the −10 region in an open complex and allows the single-strand template DNA to enter the active site. For this purpose, σ^{70} does not need an energy source such as ATP or GTP[208]. **b** | The RNAP holoenzyme (R) and promoter DNA (P) interact to form the closed complex (RP_C). The duplex DNA around the transcription start site is separated (represented by a 'bubble' in the DNA that is bound by the Eσ) in order to form the open complex (RP_O). The initiating complex (RP_{INIT}) is formed and begins synthesis of the RNA chain (shown as a small red line), directed by the DNA template strand. In initiation of the synthesis of RNA, a phosphodiester bond is formed between the initiating and adjacent phosphodiester nucleoside triphosphates (NTPs). The final stage of transcription initiation is elongation. In the elongation phase, the RNA chain length increases, shown as a solid red line. Part **a** adapted from REF.[9], CC BY 4.0 (https://creativecommons.org/licenses/by/4.0/). Part **b** adapted from REF.[24], Springer Nature Limited.

for −35 (REF.[31]), −10 (REFS[31–33]), extended −10 (REFS[34,35]), −12 and −24 (REF.[36]), as well as the discriminator[37], the core recognition element (CRE)[38] and the upstream promoter (UP)[39,40] elements, have been identified. Instances of these motifs interact with the σ and α subunits of RNAP (FIG. 1a).

A variety of old and recent evidence points to the existence of sequences that conform to the motifs but do not support transcriptional activity, such as the presence of promoter-like sequences involved in transcriptional pausing[41], the presence of a high density of σ^{70} promoter-like sequences in intragenic regions[42–44] and, in some bacteria, the overrepresentation of the −10 element in coding sequences[45]. Thus, the presence of motifs is not a sufficient condition for promoter activity.

Furthermore, the presence of a sequence in a DNA segment matching a motif is not necessary for us to regard a segment as a promoter. It is known that weak promoters may lack one or several functional

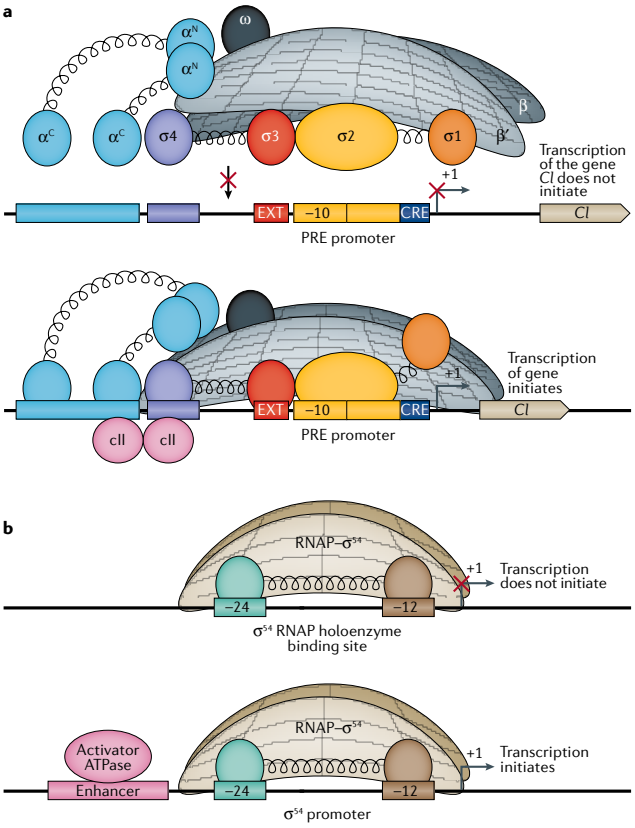

Fig. 2 | **Eσ intrinsic recognition by a sequence is neither necessary nor sufficient for that sequence to be a promoter. a** | A schematic representing the PRE promoter, which does not have the autonomous capability of binding RNA polymerase (RNAP) holoenzyme (Eσ). The Eσ and promoter boxes are the same as in Fig. 1, except that boxes with little similarity to the consensus sequence have no label. This promoter shows little to no interaction with Eσ in the absence of the activator cII (pink bubbles). When the cII protein is bound to its site, which overlaps the −35 element, this compensates for the lack of consensus of this element, thereby allowing Eσ to interact at PRE and allowing transcription of the *Cl* gene (brown box)[209]. Because there are promoters that are not bound autonomously by Eσ, this is not a necessary feature in the definition of a promoter. **b** | The σ^{54} factor is divided into three conserved regions. Region I comprises a domain that inhibits polymerase isomerization and initiation in the absence of activation and that stimulates initiation in response to activation. Owing to this inhibiting element, σ^{54} requires an ATP-dependent activator (pink oval)[210]. Region II is variable and is implicated in DNA melting. Region III (brown and green ovals) is the primary DNA-binding region and recognizes the −12 element and −24 element, to which an inactive $E\sigma^{54}$ (beige) binds. Transcription initiates upon binding of an ATP-dependent activator to a sequence called an enhancer. Binding sites of $E\sigma^{54}$ from which no transcription occurs, possibly because there is not an enhancer nearby, are not promoters. Thus, the binding of Eσ is not sufficient to define a promoter. To be a promoter, a sequence must allow Eσ both to bind and to initiate transcription.

Motifs
Representations of a collection of binding sites that summarize the binding-site characteristics.

elements, which can be compensated for by activators[23]. Furthermore, there may be sites that have poor matches to motifs but are recognized efficiently by Eσ; it has been shown that a single mutation over random 100 bp DNA stretches can generate a promoter[46,47].

Eσ binding does not define a promoter. Eσ binding sites that drive transcription are defined as promoters. For both families of σ factors, there is evidence that Eσ binding is not sufficient to define a promoter. For example, there are sequences that conform to the motifs but do not support transcriptional activity, such as sequences leading to unproductive binding of $E\sigma^{70}$ (REF.[48]). Similarly, it has been suggested that $E\sigma^{70}$ binds in an inactive conformation under salicylic acid stress, because it was observed bound adjacent to strongly downregulated genes[49].

$E\sigma^{54}$ can bind to promoters in a transcriptionally inactive state, and its activation requires an enhancer-binding protein (EBP)[50]. Genome-wide studies have greatly increased the number of known $E\sigma^{54}$ binding sites in *E. coli*[51,52] and other bacteria[53,54]. Many of the newly found sites are intragenic, and most of them are not conserved in other species; thus, not all of them are likely to be functional. Hence, the binding of Eσ to a site is not a sufficient criterion to regard that site as a promoter (FIG. 2).

Some promoters, comprising elements that are very different from the consensus sequence or that lack some recognition element, cannot bind Eσ alone; they require additional factors for their function. For example, the λ phage PRE promoter has −10 and −35 elements that differ from those in the consensus sequence, and it requires the cII protein for Eσ binding[35]. Therefore, autonomous binding of Eσ to a sequence — that is, without the need for other molecules — is not a prerequisite for a sequence to be a promoter (FIG. 2).

Although transcription initiation by Eσ may not lead to a functional RNA, it nevertheless requires a promoter. Some promoters generate short, non-functional transcripts 2–15 bp long as a result of abortive transcription; this phenomenon plays a role in the regulation of transcription[55]. Non-functional RNAs also result from so-called TSS-associated RNAs of around 35–50 bp (REF.[56]) and other pervasively transcribed spurious RNAs[57,58].

Regulatory binding sites are not part of the promoter. Some Eσ binding sites require the binding of activators in order to initiate transcription[35,59], which raises the question of whether to annotate both the Eσ binding site and the required activator sites as part of the promoter. We propose not to do so, because not all promoters are activator-dependent, and it is thus not a necessary feature. In cases in which the activator site overlaps the promoter, the annotated promoter will include the activator site; currently, RegulonDB contains at least 40 activator sites of 22 different TFs that overlap their corresponding promoter. However, the overlapped sequence should be annotated again, independently, as a regulatory site (discussed below). In cases in which the activator site does not overlap the promoter, the region annotated as a promoter should not include the activator site. Although this promoter sequence is not competent for transcription on its own, being annotated as a promoter indicates that it is the sequence recognized by RNAP, with the help of the activator to initiate transcription. Similarly, repressor sites and overlapping

Transcriptional pausing
A process through which the RNA polymerase slows down transcription during elongation.

5′-Rapid amplification of cDNA ends
(5′-RACE). A method to amplify mRNA between a defined internal site and its initiation site.

promoters are separate entities, as the repressor sites are not necessary for promoter activity, although they are necessary for regulation. The existence of overlapping elements in DNA is a recurring theme in the modelling of transcriptional regulation.

Different σ factors initiating at the same TSS define different promoters. Overlapping promoters can initiate transcripts at the same TSS[60,61]. For example, *glmY* expression in *E. coli* is controlled by two overlapping promoters, one recognized by σ^{54} and the other by σ^{70} (REF.[61]). 5′-Rapid amplification of cDNA ends (RACE) analyses have shown that these promoters initiate transcription of the *glmY* gene at the same position. Mutations in the −10 element, recognized by σ^{70}, and in the −24 element, recognized by σ^{54}, abolished the corresponding Eσ activity while leaving the activity of the second Eσ unaffected[61] (FIG. 3a). This finding demonstrates that RNAP holoenzyme can use different recognition elements even when selecting the same base as the TSS. Similarly, $E\sigma^{70}$ can initiate transcription from the majority of σ^{32} promoters at identical TSSs, and $E\sigma^{70}$ transcribes 40% of σ^{24} promoters[62].

FIGURE 3b shows the numbers of TSSs targeted by different σ factors, according to the information in RegulonDB[7]. We propose that overlapping binding sites targeted by different σ factors be considered separate promoters, even if the TSSs are the same. Certainly, such cases may be subject to different regulatory inputs, and different sequence elements will be used, according to the nature of the σ factor.

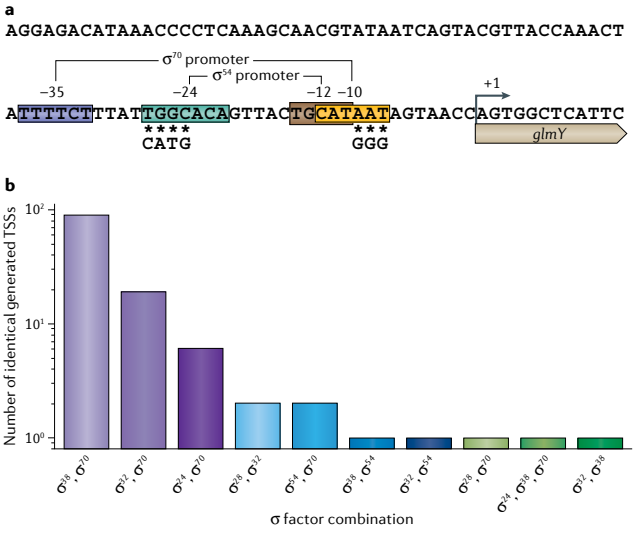

Fig. 3 | **Different σ factors that start transcription at the same TSS define different promoters. a** | The sequence shows overlapping σ^{54}- and σ^{70}-dependent promoters that start *glmY* transcription at the same position. The −24 and −12 sequence motifs of the σ^{54}-dependent promoter are highlighted by green and brown boxes, respectively. The putative −35 and −10 sequences of an overlapping σ^{70}-dependent promoter are boxed in purple and yellow, respectively. Asterisks indicate mutations that abolish the activity of the corresponding promoter but do not affect the secondary promoter. **b** | Relative numbers of transcription start sites (TSSs) generated by two or more σ factors according to the information in RegulonDB. Part **a** adapted with permission from REF.[61], Wiley.

Proposed definition for promoters. In summary, promoters are essential for transcription that is specific; an instance of a sequence motif is not mandatory; autonomous binding of Eσ is neither sufficient nor necessary; and promoters are σ factor specific. On the basis of these considerations, we define a promoter as a DNA segment essential for the specific initiation of transcription at a defined location in a DNA molecule, although this location might not be one single base. It is recognized by a specific Eσ, and this recognition is not necessarily autonomous.

Transcription factors

In the operon model, the product of a regulator gene, the 'cytoplasmic repressor', acts on the operator to affect the synthesis of a set of genes[63,64]. The chemical identity and mechanism of repressors were unknown until the *lac* and the λ phage repressors, proteins with high specificity to a site in the DNA, were isolated[65,66]. Later, an Eσ binding site was found to overlap both the λ repressor and *lac* repressor operator sites, which confirmed the mechanism for a repressor, which prevents RNAP from binding to the promoter[67]. It was assumed that gene regulation was mediated solely by repressors, until genes in the maltose and arabinose operons and in λ phage were found to be positively regulated[68–70]. Now, activators and repressors are collectively called TFs, and their sites of action are called transcription factor binding sites (TFBSs).

Many factors affect transcription. The term TF should not be confused with the more general class of factors that regulate transcription. Regulators that act on RNA, DNA or RNAP throughout the whole transcription cycle — including proteins, small peptides, non-coding RNAs and a variety of small ligands — have also been referred to as TFs[14,71].

Originally, the regulators of transcription initiation were thought to exclusively be proteins; however, there are other kinds of molecules that regulate transcription initiation, such as regulatory RNAs[72,73] and small ligands — for example, ppGpp[74]. Here, we deal only with regulatory gene products as originally conceived by Jacob and Monod. For instance, *E. coli* 6S RNA regulates transcription initiation by directly binding to $E\sigma^{70}$ and preventing its binding to the promoter, leading to an increase in $E\sigma^{38}$ transcription[75,76]. However, the term TF has a traditionally well-established meaning in all domains of life: a protein that binds to DNA in order to regulate transcription initiation[66,77–81].

A large number of protein factors that bind directly to Eσ in order to regulate transcription initiation have been identified over the past 20 years[82–85]. Moreover, some proteins that regulate transcription initiation bind to both DNA and Eσ. Thus, the criterion of mere binding to a molecule, be it DNA or the holoenzyme, makes the definition of a TF imprecise.

By contrast, specificity — the ability to promote or repress the expression of a subset of genes — must be

Global regulators
Transcription factors that affect a large number of genes involved in many different functions.

a feature of any regulator of gene expression, as it is the means for transcription to differentially respond to different conditions. This criterion can be used to decide whether a regulator that binds to both the DNA and any component of the Eσ is a TF. If the specificity of the regulator is determined directly by the sequence of the DNA to which it binds, then it is a TF. If specificity is determined by the regulator's interaction with Eσ, then we propose to use the term 'Eσ-centred regulatory protein'[86].

To refine the TF definition, we focus on proteins that specifically bind to DNA and regulate transcription and ask whether they should be covered by this term.

σ factors. Both TFs and σ factors bind to DNA and lead Eσ to different sets of promoters in response to different environmental conditions. For example, the synthesis of σ^{29} is induced during sporulation in *Bacillus subtilis*, thereby enabling the transcription of the subset of genes required during sporulation[87]. Another example is *E. coli* σ^{32}, whose expression is induced under heat-shock conditions; in turn, $E\sigma^{32}$ induces the expression of the heat-shock response genes[88,89]. All σ factors could be regarded as activators, as they were initially discovered as factors that increase transcription activity in vitro[90]. However, σ factors can be defined as the proteins that regulate and are essential for specific transcription initiation while being part of the RNAP holoenzyme[91]. The features that make σ factors different from TFs are the abilities to confer core RNAP promoter specificity, to open duplex DNA and to facilitate template strand entry into the RNAP active site[92].

Some TFs have activities that are very similar to those of σ factors, such as forming a complex with RNAP in solution or stabilizing the RPo. For example, in *E. coli*, SoxS forms a complex with Eσ, which then scans DNA using SoxS to search for their cognate sites, called Sox boxes[93,94]. Although the sequence specificity of SoxS has been demonstrated[95], it does not aid in DNA opening and template strand entry to the active site; instead, this TF acts by pre-recruitment of Eσ. The CarD and RbpA proteins bind to both Eσ and the promoter just upstream of the −10 element as a means to stabilize the unstable RPo in *Mycobacterium* spp.[96–98]. As the specificity of CarD and RbpA has not yet been shown to be determined by the DNA sequence, these are RNAP-centred regulatory proteins[86] that help σ stabilize the RPo.

Nucleoid-associated proteins. The consideration of TFs and NAPs provides a perfect example of how the initial idea of two different types of molecule, based on genetics and function, changed with the realization that TF and NAP functions are sometimes performed by the same molecules[99]. NAPs are a group of DNA-binding proteins that are believed to play important roles in chromosome structure and compaction[100,101]. Some NAPs have been shown to function as site-specific regulators of transcription initiation[100,102–110], similarly to other TFs, as well as acting at a distance from the target promoter by bending the intermediate DNA; integration host factor (IHF) is a well-known example of such a molecule at σ^{54} promoters[111,112]. Some NAPs tend to bind to many sites with fairly low sequence specificity[113], and most global regulators in *E. coli* are NAPs[99,114]. As NAPs have functions that overlap with those of TFs, for the purposes of the definitions we propose that they not be considered as a separate class.

Proposed definitions for TFs. A comprehensive terminology to describe all types of regulatory gene product that act on the different levels of transcription and/or other mechanisms of gene regulation is beyond the scope of this article; however, we outline the terminology to designate different kinds of regulators. In the words of Jacob and Monod, we can begin by defining the general term 'regulatory gene product' as any gene product that increases or decreases the expression of a specific set of genes (note that gene product complexes such as heteromultimeric TFs are included — for example, IHF in *E. coli*).

We can distinguish two general kinds of regulatory gene products: regulatory RNA and regulatory proteins. These we can further subclassify according to the level at which they act to regulate gene expression or according to the specific mechanism of regulation. In particular, we have to define the class 'DNA-binding regulatory protein of transcription initiation' as the subclass of regulatory proteins that bind to specific DNA sequences in order to regulate transcription initiation. This class includes both TFs — that is, DNA-binding regulatory proteins that bind near a promoter and affect transcript initiation at that promoter — and σ factors — that is, DNA-binding regulatory proteins of transcription initiation that are part of the RNAP holoenzyme and are essential for the specific initiation of transcription. Another subclass of regulatory proteins is 'RNAP-centred regulatory proteins of transcription initiation', defined as the proteins that regulate transcription initiation by interacting with Eσ and whose specificity is not determined directly by the recognition of specific DNA sequences.

Transcription factor binding sites

TFs bind specifically to their binding sites in order to activate or repress adjacent promoters. Recent genome-scale technologies capable of identifying TFBSs anywhere in the genome, such as chromatin immunoprecipitation followed by sequencing (ChIP–seq) or by identification using a DNA array, or chip (ChIP–chip), have shown that sites where TF binding has no direct effect on transcriptional regulation are common (FIG. 4). For instance, when analysing 12 studies of 9 different TFs in *E. coli*, only ~25% of 3,973 TF–gene interactions showed evidence of regulating local gene expression[115]. Similar observations have been made in other bacterial genomes, including *Mycobacterium tuberculosis*[4,116], *Pseudomonas aeruginosa*[117,118], *Salmonella enterica*[119], *Listeria monocytogenes*[120], *Helicobacter pylori*[121] and *Shigella flexneri*[122]. The diverse behaviour of TFs at the genomic level raises the question: which fraction of TFs have binding sites that are mostly only involved in transcriptional regulation, and thus behave like MelR[123] and MarA[124]? Indeed, most TFs have sites that support other functions, such as contributing to the nucleoid structure of the genome, akin to NAPs[125], or have sites with no apparent function.

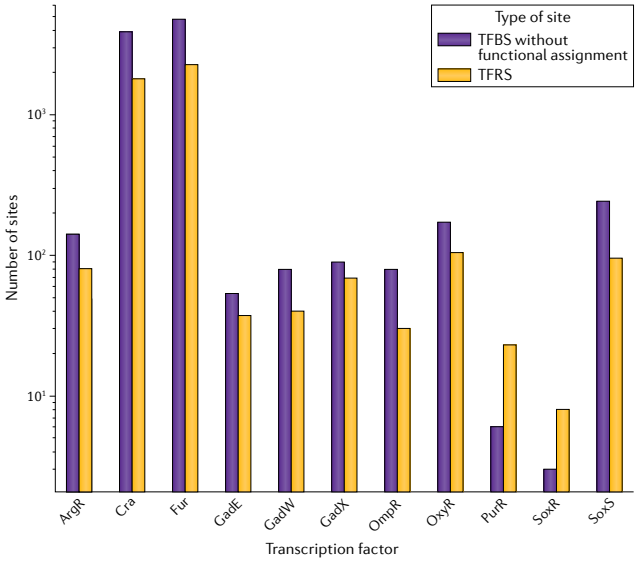

Fig. 4 | **Number of transcription factor binding sites without functional assignment versus number of transcription factor regulatory sites in *Escherichia coli*.** The sum of transcription factor binding sites (TFBSs) without functional assignment (purple) and transcription factor regulatory sites (TFRSs; yellow) represents the complete set of TFBSs. The figure was drawn using data from published chromatin immunoprecipitation (ChIP) experiments[49,126,129,130,211–220] and data from classical experiments. Only transcription factors with non-functional sites found by high-throughput methodologies were included. A logarithmic scale is used to help visualize the disparate numbers of sites known for Fur and Cra versus other TFs, such as PurR.

It will be interesting to understand the different distributions of TFBSs in non-coding versus coding regions. For instance, two-thirds of 96 binding sites of the regulator of iron homeostasis Fur lie in intergenic regions[126]. Moreover, a genome-wide study of uncharacterized TFs in *E. coli* identified binding sites for 10 candidate TFs using ChIP–seq combined with exonuclease treatment (ChIP–exo) and showed that only 41% of 241 sites were in regulatory regions[127]. By contrast, as an extreme case, 70% of binding sites for RutR were found within coding regions in *E. coli*, a tendency conserved in other bacteria[128].

Additionally, a more clear distinction of the subclasses of binding sites is emerging, owing either to the contribution of additional TFs[126,129,130] or to the same TF working differently in varying conditions. For example, Fur was shown by ChIP–seq to bind to lower or higher numbers of TFBSs under aerobic or anaerobic conditions, respectively[126]. In another study, ChIP–seq combined with RNA sequencing (RNA-seq) under nine physiological conditions showed that nutrient levels or growth phase affect the mechanism of action of the TF Lrp[131].

Taken together, these studies distinguish between sites that affect transcription and those that do not. We thus propose that the term TFBS be defined as a DNA site where a TF binds specifically and that the term 'transcription factor regulatory site' (TFRS) be defined as a member of the subset of TFBSs that are involved in transcription regulation (FIG. 4).

Architecture of regulatory regions. TFRSs, initially termed 'operator sites', were conceived as a single entity located near a cognate regulated promoter[28,132]. However, only 37% of all current promoters in RegulonDB are regulated by a single TFRS. Certainly, the steady increase of well-characterized regulated promoters has led to what we now see as the rich architecture of regulatory regions, with promoters subject to the effect of one or several TFs binding to one or several TFRSs.

The diversity of TFRSs close to promoters became clear as a result of studies in the early 1990s in an initial review of 107 σ^{70}-dependent and 12 σ^{54}-dependent regulated promoters[3]. A general principle emerging from this cohort of bacterial regulatory regions was the requirement for a proximal site, defined as an activator or repressor site, located in a position that enables direct interaction of the TF with Eσ^{70}. Only 4 out of 107 σ^{70} promoters lacked a proximal TFRS site[3,133], a finding that has been sustained as more promoters have been characterized[134]. In one study, only 6.9% of σ^{70} promoters (48 out of 692) lacked a proximal site located between −95 and +20 relative to the TSS[7]. Note that promoters were counted individually even if multiple promoters coexisted in the same upstream region.

This principle does not apply to σ^{54}-dependent promoters, the activator sites of which are more distal, with the activator often being brought close to the promoter by DNA looping induced by IHF binding between the enhancer-like TF sites and the promoter[135,136]. This different organization is associated with the capability of Eσ^{54} to form stable but inactive closed complexes and the absolute requirement of an activator for transcript initiation, as opposed to the Eσ^{70} holoenzyme, which is competent to initiate transcription without activators and forms a transient closed complex[3,137].

Regulatory modules or phrases. The collection of sites affecting a promoter can be partitioned into groups of sites that work together, similar to words in a sentence that form syntactic categories. These 'regulatory phrases' or '*cis*-modules'[138] can be homotypic modules, grouping sites for the same TF, or heterotypic modules, grouping sites for different TFs that work jointly in the regulation of a promoter (FIG. 5). These modules constitute the building blocks of a grammatical[139], combinatorial logic[139,140] and of quantitative thermodynamic models of regulated promoter transcriptional activity[141–143]. They are inferred in approaches that seek to understand large amounts of gene expression data[144] and aim at a combinatorial construction of all possible regulatory arrangements in bacterial genomics. For instance, a grammatical model was implemented with a reduced number of rules to generate the whole collection of regulatory architectures of σ^{70}-regulated promoters[139].

We propose to define a bacterial TFRS module, or TFRS phrase, as a combination of one or several TFRSs whose bound TFs work jointly in the regulation of a promoter. A bacterial TFRS collection is defined as all the TFRSs that regulate a promoter.

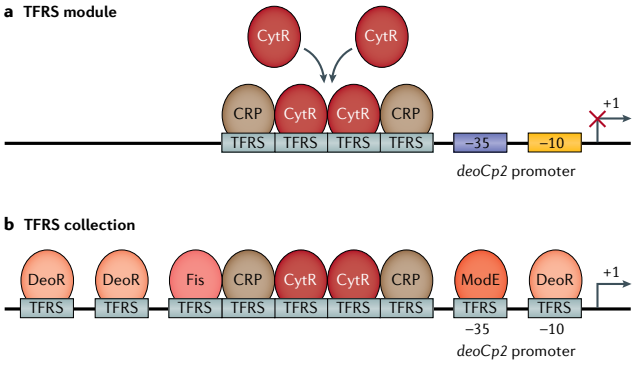

Fig. 5 | **Cis-regulatory architecture of the promoter *deoCp2*.** The line represents the *deoCp2* promoter along with its transcription factor regulatory sites (TFRSs). The −35 and −10 elements and the transcription start site (TSS; +1) are labelled, and the *cis*-regulatory architecture is represented by transcription factors (TFs; bubbles), each bound to its corresponding TFRS. **a** | Cooperative regulatory interaction between CRP and CytR in the *deoCp2* promoter. CytR is recruited as a co-repressor by pre-bound CRP[221]. Since CytR necessarily requires pre-bound CRP in order to repress expression of the transcription unit downstream, these four sites form a TFRS module. **b** | The TFRS collection of promoter *deoCp2* is the complete set of TFRSs known to regulate the *deoCp2* promoter. Although there might be indirect regulatory interactions among DeoR, ModE, Fis, CRP and CytR, the only direct and necessary interaction is the one between CytR and CRP[222–224]. The other TFs act independently under different conditions on *deoCp2*, each with its own proximal sites. Part **a** adapted with permission from REF.[225], Elsevier.

Operon

The classic definition of an operon is the units of coordinated expression constituted by an operator and the group of structural genes coordinated by it[132], thus requiring an operon to have one operator. However, multiple TFRSs organized in a module can regulate a single promoter. Furthermore, transcriptional regulation of some co-transcribed genes may be independent of TFs. For example, the genome of *Mycoplasma genitalium* encodes a limited number of TFs, and it has been suggested that this bacterium depends mostly on DNA supercoiling for gene expression regulation[145,146]. For example, the –10 element along with DNA supercoiling induced by hyperosmolar conditions have been shown to be sufficient for the induction of expression of the *MG_149* gene[147]. Because some co-transcribed genes may not be regulated by TFs, we propose that TFRSs be considered independent of these 'units of coordinated expression' and that their connections be captured by defining their regulatory relationships.

Co-transcribed genes are not limited to one pathway. The first operons studied were those that coordinate the expression of genes whose products are involved in a common pathway (that is, lactose, maltose or arabinose). Hence, it was reasonable to expect that co-transcribed genes of an operon would participate in the same biological pathway. However, not long after the proposal of the operon model, an early example of co-transcribed genes involved in different pathways was found; in *B. subtilis*, the tryptophan gene cluster is expressed coordinately with genes involved in histidine and tyrosine production under tryptophan-limiting conditions[148]. Comprehensive studies of the functional classes of large numbers of transcripts in *E. coli* have now provided us with numbers that underscore the diversity of functions of many co-transcribed genes[149].

Although the operon concept was initially proposed to account for the discovery that some units of transcriptional regulation in bacteria were not single genes, it was later extended to include monocistronic operons by Jacob and Monod themselves[150,151]. Thus, we consider it to be unnecessary for an operon to have more than one gene.

Transcription units versus operons. The most general definition for both 'operon' and 'transcription unit' is a set of adjacent co-transcribed genes (Supplementary Table 1), although there are ambiguities about what distinguishes the two concepts. We think that it is better to define transcription units as physical entities, whereas operons are more complex conceptual entities. At least in theory, every transcription unit in a cell can be determined by measurement (for example, cDNA sequencing). In the literature, there are two ways to define transcription units: they are segments of DNA that extend from the promoter to the terminator[152,153], which are included, or they are the DNA segments that begin at a TSS and end at a transcription termination site (TTS)[154,155] (Supplementary Table 1). The latter, and most common, usage is equivalent to defining transcription units as the DNA that corresponds to a primary transcript.

A promoter can have more than one TSS, and a terminator can have more than one TTS. Promoters were found to have on average 1.6 TSSs in an integrated genome-wide analysis of *E. coli*[156]. If we opt to define transcription units as the DNA sequences that begin at a TSS and end at a TTS, we would be representing multiple units that differ by a few nucleotides, but most of this microvariation can be considered functionally spurious. Studies using single-molecule fluorescence resonance energy transfer (FRET) support a model in which this microvariation arises from the thermodynamics of transcription[5]. Thus, we prefer to think of these units of transcription as having variable ends and only consider the predominant ones for each promoter or terminator (FIG. 6a).

Nonetheless, some regulatory mechanisms alter TSS selection from a single promoter. For example, the TSSs of the *pyrC* and *pyrD* promoters of *E. coli* and *S. enterica* are shifted by nucleotide concentration[21,157] (FIG. 6b). In addition, some promoters can be primed by nanoRNAs in a growth-phase-dependent manner, thereby altering TSS selection[158]. Regarding TTSs, it has long been known that Rho-dependent termination is diffuse, and elongation factors can conditionally allow bypass of terminators, which results in multiple termination points[71,159–161] and complex patterns of expression[162–164]. To include microvariation subject to differential regulation, we propose to define transcription units as the DNA regions delimited by different nonspurious TSS–TTS pairs (FIG. 6a,b).

A widely held distinction between transcription units and operons is that one gene can belong to one or

DNA supercoiling
The writhe of DNA over the double-stranded axis.

Monocistronic operons
Operons that encode a single gene product.

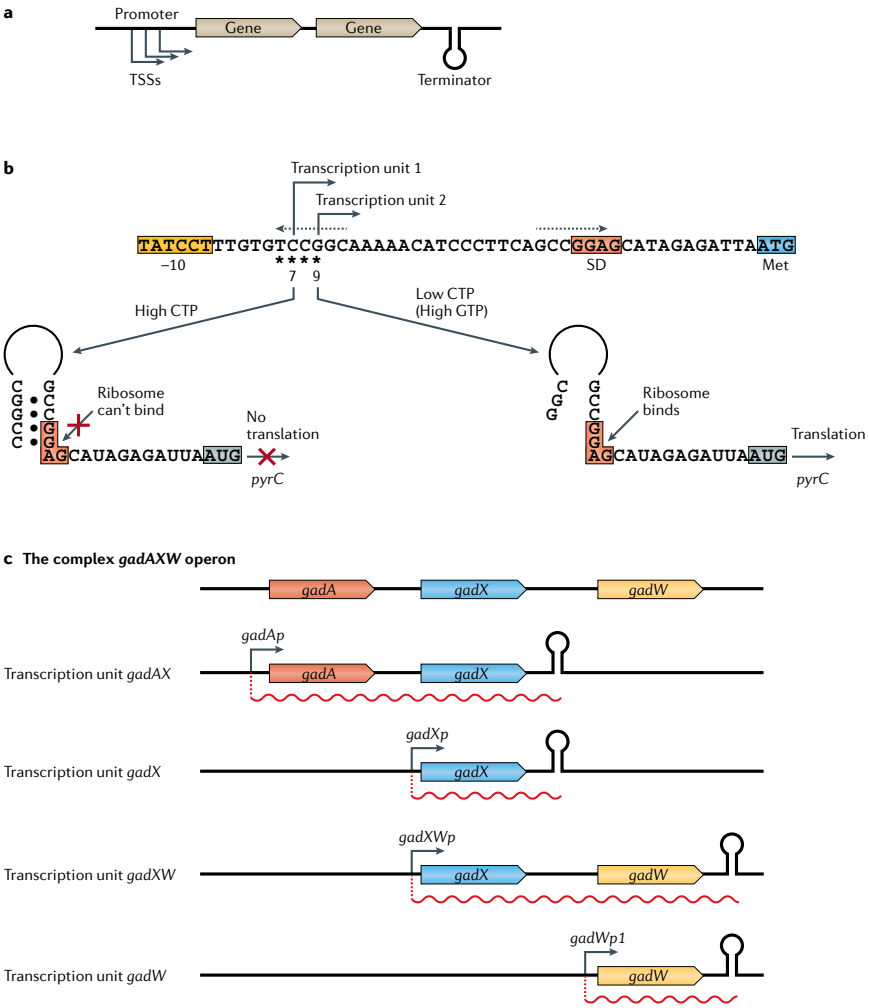

Fig. 6 | **Transcription unit and operon schematic. a** | When different transcription start sites (TSSs) from the same promoter are not differentially regulated, they form a single transcription unit that is limited by, but does not include, a single promoter and a single terminator. **b** | Model for TSS switching and translational control of *pyrC* by CTP concentration. *pyrC* encodes a pyrimidine biosynthetic enzyme in *Escherichia coli* and *Salmonella enterica* subsp. *enterica* serovar Typhimurium. The nucleotide sequence of the *E. coli pyrC* promoter is shown, with the –10 region, Shine–Dalgarno (SD) sequence and *pyrC* initiation (Met) codon boxed and labelled. The transcription of *pyrC* initiates at four adjacent sites (asterisks) — named T6, C7, C8 and G9 — each of which produces a different transcript. Under conditions of pyrimidine excess, the intracellular level of CTP is high, and position C7 is the dominant start site[22,157]. C7 transcripts are not translated, because they form a stable hairpin at their 5′ ends that blocks ribosome binding to the *pyrC* SD sequence. Under conditions of pyrimidine limitation, the GTP level is high, which makes the G9 start site dominant. G9 transcripts do not form the inhibitory hairpin and are readily translated[22]. Thus, C7 and G9 are different non-spurious TSSs from the same promoter (*pyrCp*), because they originate two transcripts that are differentially translated. Thus, these are two transcription units originating from the same promoter. **c** | Schematic of the *gadAXW* complex operon. Several internal promoters and terminators enable different sets of genes to be co-transcribed in different combinations: *gadAX* by the *gadAp* promoter; *gadX* and *gadXW* by the *gadXp* promoter; and *gadW* by the *gadWp1* and *gadWp2* (not shown) promoters. Red wavy lines represent mRNAs. These promoters are subject to regulation by different sets of TFs (not shown), so they are not all subject to the same signals. There are at least four operons in RegulonDB with no evidence of a single polycistronic transcript that includes all the genes of the operon. Part **b** from REF.[22], *Microbiol. Mol. Biol. Rev.*, 2008, **72**, 266–300, https://doi.org/10.1128/MMBR.00001-08 and adapted with permission from American Society for Microbiology.

more transcription units but only to one operon. This distinction comes from the existence of promoters and terminators that are internal to operons. In 1967, dis-coordinated expression of the cluster of five tryptophan synthesis genes of *Salmonella enterica* subsp. *enterica* serovar Typhimurium was reported[165]. Deletions ranging from the operator to the second gene of the cluster suppressed expression of the first two genes only. The last three were silenced after the deletion reached the region between the second and third genes. Deletion of the operator upstream of the five genes deregulated all of them. These five genes were considered an operon regulated by a single operator but subdivided into two parts, determined by promoter-like elements. Similarly, the *glnALG* operon is differentially transcribed in different transcription units depending on the nitrogen source, by means of an internal terminator and alternative promoters[166]. Now, many systems are known in which subsets of the genes of an operon are differentially expressed under different conditions, due to alternative combinations of promoters and terminators[154,156,164,167–171]. This has led to the notion that operons with internal promoters and terminators contain several transcription units.

However, internal promoters are often differentially regulated. Currently, 862 operons with known regulation are listed in RegulonDB. Of these, 143 operons consist of more than one transcription unit. Of the 143 multi-transcription-unit operons, 92 are differentially regulated, whereby at least one pair of their constituent transcription units are regulated by different TFRS collections. Moreover, there are cases in which the genes of an operon are not all co-transcribed. This phenomenon is generally described as an operon containing genes ABC with two transcription units, AB and BC (FIG. 6c).

To preserve both the notion of operons being the set of genes in the same transcription unit and the notion of operons being the set of genes coordinated in the maximal set of overlapping transcription units, we propose to use the term 'simple operon' for the former and 'complex operon' for the latter. Although some complex operons may have genes that are not co-transcribed, there is 'cooperation' of their function (they use the same infrastructure); despite the individual transcription units being differentially regulated, there is likely a complex functional interplay among them that makes it hard to consider them individually.

Whereas operons are defined in terms of genes, we think that transcription units do not necessarily bear genes or have a function. Transcriptome analysis has revealed the widespread production of non-canonical transcripts[57,58] — that is, transcripts that are non-coding and are often antisense; such 'pervasive transcripts' rarely have an assigned function[172,173]. Evolutionarily, we should not assume that transcriptional regulation has been selected as 'optimal' — that is, to express exactly the right genes at the right moment in the right cells[174–176]. Non-functional transcripts may constitute raw material for evolution, given novel mutations. Moreover, pervasive transcription may have a function in itself; that is, a basal level of pervasive transcription means that core RNAP or σ levels have to be higher, effectively buffering their levels in the cell[177].

Co-expression may extend beyond operon limits. Significant co-expression has been observed in regions of 10 kb and even in larger regions, which has been associated with transcriptional read-through, supercoiling and nearby regulons[162,163,178]. Read-through of terminators has been documented[179,180], and a recent transcriptome obtained using single-molecule long-read sequencing has extended 34% of RegulonDB operons by at least one additional gene[164].

Proposed definitions for operons. An operon is a set of adjacent genes whose transcription is coordinated by one or several mutually overlapping transcription units that are transcribed in the same direction and share at least one gene. A simple operon is an operon whose transcription is coordinated from a single transcription unit. A complex operon is an operon whose transcription is coordinated through several mutually overlapping transcription units that are transcribed in the same direction and share at least one gene.

Regulon

Regulons were originally viewed as a system in which the production of all enzymes (of a metabolic pathway) can be controlled by a single repressor substance; this substance may consist of several entities but, whatever its nature, it acts in a unitary fashion[181].

Although regulons were originally proposed by studying the pathway of arginine biosynthesis, they must now be defined exclusively by regulation. From the 149 described *E. coli* regulons that include enzymes that catalyse metabolic reactions, only 21% included enzymes for a single metabolic pathway whose inputs and outputs form a connected network[182].

What kind of regulatory entity defines a regulon? We suggest that the regulator that defines a regulon must be a regulatory gene product. Before this discussion, Sequence Ontology used to refer to the regulator as the 'regulatory signal'[6]. However, the term 'signal' is used to refer to cues that are transformed into an effector that interacts with regulators, thereby providing information about environmental and physiological states so that the cell can adjust gene expression levels (see below). Although most TFs bind to one effector, there are documented cases in which TFs allosterically bind to several metabolites — for example, tryptophan, tyrosine and phenylalanine, in the case of TyrR[183]. Conversely, small molecules may act as effectors for more than one TF, such as Zn^{+2} binding to ZntR and Zur, and tryptophan to TyrR and TrpR. Thus, regulation by an effector does not imply regulation by a specific TF. The initial Sequence Ontology definition of regulons corresponded to that of stimulons[184].

On the basis of the original definition, a regulator entity should be either one regulator that can work independently (with one or multiple TFRSs) or any collection of regulators working in unity, such as complex heteromultimeric proteins. Currently, RegulonDB has documented 598 transcription units regulated by more than one TFRS; of these, 477 are regulated by different TFs. These data motivate the expansion of our notion

Transcriptional read-through
Transcription that allows RNA polymerase to continue transcription beyond termination sites.

Stimulons
Sets of genes whose products are increased in response to a common environmental stimulus.

of a regulon to include groups of genes subject to multiple regulators. These regulators will not be acting in unity but will support a complex multiple input–output regulation. Such sets of multiple regulators may be, for instance, TFs comprising TFRS modules anchored by a proximal site or, alternatively, the set of TFs binding the TFRS collection (FIG. 5). How these groupings will help map mechanisms to physiology is an open question.

Regulated entity and regulated stage of gene expression. The regulator, the regulated entity and the level at which gene expression is regulated are interdependent features. The concentration of the products of genes that are transcriptionally coordinated may be uneven[185], implying that the mapping of transcription units to translation units is complex. Most regulatory RNAs regulate gene expression post-transcriptionally by base-pairing with mRNA. Thus, if the regulon definition is such that it includes RNAs as regulator entities, then the regulated entities should include transcription units and the transcribed coding sequences.

Proposed definition for regulons. In general, units of gene expression are defined as transcription units or transcribed coding sequences. A regulon is a set of units of gene expression directly regulated by a common set of one or more common regulatory gene products. A simple regulon is a regulon defined by considering one regulatory gene product, and a complex regulon is one defined by considering the units of expression regulated by a specified set of regulatory gene products.

Signal and effector

Effectors were originally defined as compounds that bind specifically and reversibly to an allosteric site on a protein and bring about a discrete reversible alteration of the protein's molecular structure that modifies its properties, changing one or several of its kinetic parameters[150,186].

It is more difficult to trace the classic definition of a signal. As we want to define the term signal in the context of gene regulation, one appropriate definition to consider is a molecule originated from the environment or produced by metabolism to which a cellular response must be mounted[187].

Difference between signal and effector. The concept of signal operates at the physiological level, whereas the effector is critical at the mechanistic level of gene expression. The signal is the starting point of an information flux that will use a variety of reactions and mechanisms, ultimately reaching the regulatory machinery of cells. Effectors provide the necessary continuity to information flux by binding to the regulator that modifies gene expression (FIG. 7). Originally, effectors modulated protein activity. We propose to generalize the definition of the targets of effectors to include all kinds of regulatory gene products. Some mRNAs have motifs to which small molecules bind in order to modify secondary structure, thereby regulating gene expression[188].

In this flux, an effector action has been considered a reversible allosteric transition that changes some chemical (kinetic or affinity) parameter of a regulatory molecule[150]. The effector concept must now be extended. Reversibility and allosteric features are no longer necessary; some effector-induced changes result in proteolysis, such as for *E. coli* LexA[189,190]; and covalently attached groups that irreversibly change TFs are also considered as making a chemical change produced by effectors. Phosphotriester and 6-*O*-methylguanine DNA lesions irreversibly methylate Cys residues of Ada protein, a TF that induces its own expression[191]. Furthermore, the activation of this TF is not due to a conformational change but to changes in the electrostatic repulsion between the protein and the DNA[192].

Signal and effector are not disjointed concepts. An example that fits perfectly with the 1963 definition of effector is that of an extracellular molecule that binds to a transmembrane sensor that reversibly modifies its conformation. This effector also acts as a signal, because the conformational change of the transmembrane sensor triggers an intracellular cascade of protein–protein interactions — that is, the signal transduction pathway — that ends in the activation or inactivation of a TF, which will in turn repress or activate the transcription of its target genes (FIG. 7).

What kinds of entities are signals, and which are effectors?. Some signals are not material in nature. Environmental changes are mostly complex and elicit a plethora of changes in metabolic fluxes that generate internal signals in the form of changes in the concentration and ratios of metabolites[193].

We propose to extend Ptashne's view, that any kind of molecule can be an effector[27], to include other kinds of entities, such as temperature and light, as they can induce conformational changes in DNA, in proteins and in RNA, thereby regulating gene expression. It has been proposed that changing from a cooler environment to a warm host triggers virulence factors in bacteria[194–196]. For example, in *S.* Typhimurium, high temperature unfolds the autorepressor TlpA, preventing dimerization[197,198]. TlpA monomers are unable to bind to DNA. High temperature can directly melt mRNA-inhibitory structures that prevent binding of the ribosome, thereby inducing translation initiation[197,198]. As RNA melting is a conformational change, temperature can play the role of effector (for example, in the induction of the heat-shock σ factors *rpoH* of *E. coli* and *prfA* of *L. monocytogenes*[188]). An example of a transcriptional regulator that is activated by light is the antirepressor AppA of *Rhodobacter sphaeroides*. Light sensed through the BLUF domain of AppA causes reduction of the affinity for DNA of the complex it forms with the repressor PpsR, thereby regulating expression of photosynthetic genes[199,200].

To take into account non-material signals and effectors, we make use of the conceptualization of the Basic Formal Ontology (BFO), an upper-level, domain-independent ontology that describes the most general classes under which domain-specific classes can be located. The BFO partitions reality into two general classes: continuants and occurrents. A continuant is defined as an entity that persists, endures or continues through time while maintaining its identity. It was

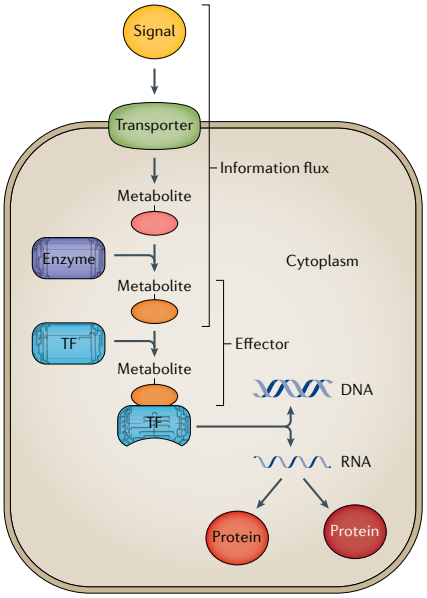

Fig. 7 | **Schematic of signal and effector.** A cell reacting to the environment is represented. The rectangle represents the cell membrane, separating cytoplasm and the environment. A signal is represented as an environmental molecule (yellow bubble) that elicits a cellular response. The environmental molecule is introduced into the cell (light red oval) through a membrane transporter (green) and transformed by an enzyme (purple) into another molecule (orange oval) that plays the role of effector, by binding a transcription factor (TF; cyan square) and modifying its ability to recognize its DNA-binding sites. The concept of genetic sensory response units (GENSOR units) captures all the elements, from the signal via the effector and regulation to the final regulated gene products, as the response[182] (for more information see RegulonDB Gensor Unit Groups).

defined in contrast to occurrents, which are processes or entities that occur in time[201].

Proposed definitions for signals and effectors. Signals and effectors are any kind of continuant, where a signal is defined as a continuant that is the first step in a flow of information that causes a change in gene expression. An effector is defined as a continuant that produces a change in a molecule and modifies its activity and/or specificity. More precisely, we propose the definition that an effector of a gene expression regulator is an effector that acts on a regulatory component of a genetic switch.

Conclusions

Terms that historically made sense face limitations with new discoveries. As a consequence, ambiguities in their use emerge, requiring that concepts be revised and refined as natural science progresses. Problems arise when attempting to define classes of objects that include all and only the intended objects, given the abundance at all levels of unusual cases. Here, we have analysed the adequacy of the properties used to define the elements involved in bacterial transcription initiation, considering possible extreme cases and generalizing definitions accordingly. We envision that an additional strategy to resolve ambiguities will be to define terms that capture the object's most general behaviour and to allow any member of a class to have a different role in specific circumstances. For example, RNAP can have a repressor role at a convergent adjacent promoter[202].

The original concepts regarding sequence features were defined in genetic or physiological terms. The discovery of a functional sequence normally was followed by its characterization. However, sequence motifs cannot ensure the identity of a function and not all functional sequences are motif compliant. Motifs are not obligatory elements in the new definitions.

Deriving universally robust functional definitions of sequences is complicated by the fact that evolution in bacteria happens rapidly, and many of the revealed features might serve no biological purpose. It is likely that some features are just pawns in the evolution game, as evidenced by the fact that nearly all DNA segments of a bacterial genome can be transcribed[58,172,173,203]. In fact, we suggest that the arrangement of DNA elements and regulatory architectures in a bacterial genome derives from the functional and anatomical properties of RNAPs, TFs and NAPs, which are more conserved in evolution than DNA regulatory sequences. After analysing the different types of concept, we can grasp some common themes. The distinction of functional versus spurious sequences is suggested as a guiding principle to define the level of signal versus noise at which some sequences are to be classed as functional elements, and some are not. Thus, promoters can be defined if their activity can be measured, even if they transcribe a non-functional transcript. Similarly, when defining elements, a guiding principle to considering two entities as separate is that the entities are subject to different regulation. We have followed this guidance in order to describe variable TSSs or TTSs as one element, and to prefer modelling the same TSS transcribed by different σ factors as different promoters. The same is true for the same binding-site sequence being recognized by two different TFs. Certainly, differential regulation has to be captured in any model of gene regulation.

A recurrent theme is that overlapping of different DNA elements does not imply that these elements are a single entity. For example, different promoters can overlap and can use the same TSS but different σ factors; similarly, activator and repressor sites can overlap a promoter region. In this sense, the correspondence between DNA and concepts is not a one-to-one relationship.

Overall, we have expanded a number of definitions while trying to retain their essence. We are aware, however, that different generalizations are feasible; for instance, we here focused on transcriptional regulators that are gene products, but they could be expanded to include small ligands such as ppGpp, in order to consider the ppGpp regulon[204].

The challenge in biology is that, experimentally, we may understand a few cases but evolution enables a much

larger combination of possibilities. Ideally, a corpus of well-known cases would generate principles that would predict the universe of all possible combinations. For example, we can think of TFRS modules anchored at a proximal site in σ^{70}-dependent promoters[3,133] as an initial working hypothesis that restricts the range of possible promoter architectures that can be validated, or corrected, by bioengineering and synthetic approaches[205]. The challenge is to implement methods and strategies that will test the validity of our current definitions, on the one hand, and advance our quantitative and qualitative integrated understanding of microbial gene regulation, on the other.

Published online: 14 July 2020

1. Miller, J. H. & Reznikoff, W. S. (eds) *The Operon* (Cold Spring Harbor Laboratory, 1980).
2. Beckwith, J. in *Escherichia coli and Salmonella: Cellular and Molecular Biology* (eds Neidhardt, F. et al.) 1227–1231 (ASM Press, 1996).
3. Collado-Vides, J., Magasanik, B. & Gralla, J. D. Control site location and transcriptional regulation in *Escherichia coli*. *Microbiol. Rev.* **55**, 371–394 (1991). **This landmark review from the pre-genomics era stresses the importance of promoter architecture rather than sequence**.
4. Galagan, J. E. et al. The *Mycobacterium tuberculosis* regulatory network and hypoxia. *Nature* **499**, 178–183 (2013). **Extensive binding of transcription factors has been found using high-throughput techniques**.
5. Robb, N. C. et al. The transcription bubble of the RNA polymerase-promoter open complex exhibits conformational heterogeneity and millisecond-scale dynamics: implications for transcription start-site selection. *J. Mol. Biol.* **425**, 875–885 (2013).
6. Eilbeck, K. et al. The Sequence Ontology: a tool for the unification of genome annotations. *Genome Biol.* **6**, R44.1–R44.12 (2005).
7. Santos-Zavaleta, A. et al. RegulonDB v 10.5: tackling challenges to unify classic and high throughput knowledge of gene regulation in *E. coli* K-12. *Nucleic Acids Res.* **47**, D212–D220 (2019).
8. Karp, P. D. et al. The EcoCyc database. *EcoSal Plus* https://doi.org/10.1128/ecosalplus.ESP-0006-2018 (2018).
9. Ruff, E. F., Thomas Record, M. & Artsimovitch, I. Initial events in bacterial transcription initiation. *Biomolecules* **5**, 1035–1062 (2015).
10. Losick, R. & Pero, J. Cascades of sigma factors. *Cell* **25**, 582–584 (1981).
11. Paget, M. & Helmann, J. Protein family review — the sigma^{70} family of sigma factors. *Genome Biol.* **4**, 1–6 (2003).
12. Li, X.-Y. & McClure, W. R. Characterization of the closed complex intermediate formed during transcription initiation by *Escherichia coli* RNA polymerase. *J. Biol. Chem.* **273**, 23549–23557 (1998).
13. Saecker, R. M., Record, M. T. & Dehaseth, P. L. Mechanism of bacterial transcription initiation: RNA polymerase–promoter binding, isomerization to initiation-competent open complexes, and initiation of RNA synthesis. *J. Mol. Biol.* **412**, 754–771 (2011).
14. Haugen, S. P., Ross, W. & Gourse, R. L. Advances in bacterial promoter recognition and its control by factors that do not bind DNA. *Nat. Rev. Microbiol.* **6**, 507–519 (2008). **This review describes regulators of transcription initiation that do not bind to DNA**.
15. Revyakin, A., Liu, C., Ebright, R. H. & Strick, T. R. Abortive initiation and productive initiation by RNA polymerase involve DNA scrunching. *Science* **314**, 1139–1143 (2006).
16. Greive, S. J. & Von Hippel, P. H. Thinking quantitatively about transcriptional regulation. *Nat. Rev. Mol. Cell Biol.* **6**, 221–232 (2005).
17. Helmann, J. D. & Pieter, L. Protein–nucleic acid interactions during open complex formation investigated by systematic alteration of the protein and DNA binding partners. *Biochemistry* **38**, 5959–5967 (1999).
18. Schneider, D. A., Ross, W. & Gourse, R. L. Control of rRNA expression in *Escherichia coli*. *Curr. Opin. Microbiol.* **6**, 151–156 (2003).
19. Gourse, R. L. et al. Strength and regulation without transcription factors: lessons from bacterial rRNA promoters. *Cold Spring Harb. Symposia Quant. Biol.* **63**, 131–140 (1998).
20. Grigorova, I. L., Phleger, N. J., Mutalik, V. K. & Gross, C. A. Insights into transcriptional regulation and σ competition from an equilibrium model of RNA polymerase binding to DNA. *Proc. Natl Acad. Sci. USA* **103**, 5332–5337 (2006).
21. Sorensen, K. I., Baker, K. E., Kelln, R. A. & Neuhard, J. Nucleotide pool-sensitive selection of the transcriptional start site in vivo at the *Salmonella typhimurium pyrC* and *pyrD* promoters. *J. Bacteriol.* **175**, 4137–4144 (1993).
22. Turnbough, C. L. Jr. & Switzer, R. L. Regulation of pyrimidine biosynthetic gene expression in bacteria: repression without repressors. *Microbiol. Mol. Biol. Rev.* **72**, 266–300 (2008).
23. Adhya, S., Gottesman, M., Garges, S. & Oppenheim, A. Promoter resurrection by activators — a minireview. *Gene* **132**, 1–6 (1993).
24. Browning, D. F. & Busby, S. J. W. The regulation of bacterial transcription initiation. *Nat. Rev. Microbiol.* **2**, 1–9 (2004). **This review proposes definitions for DNA- and RNAP-centred transcriptional regulators**.
25. Libis, V., Delépine, B. & Faulon, J. L. Sensing new chemicals with bacterial transcription factors. *Curr. Opin. Microbiol.* **33**, 105–112 (2016).
26. Davidson, E. H. *The Regulatory Genome: Gene Regulatory Networks in Development and Evolution* (Elsevier, 2010).
27. Ptashne, M. & Gann, A. *Genes & Signals* (Cold Spring Laboratory Press, 2002).
28. Jacob, F., Ullman, A. & Monod, J. Le promoteur, élément génétique nécessaire à l'expression d'un opéron. *C. R. Acad. Sci.* **258**, 3125–3128 (1964). **This is the first mention of a bacterial promoter, which is presented almost as a by-product of the operon model**.
29. Vogt, V. Breaks in DNA stimulate transcription by core RNA polymerase. *Nature* **223**, 854–855 (1969).
30. Dausse, J. P., Sentenac, A. & Fromageot, P. Interaction of RNA polymerase from *Escherichia coli* with DNA: influence of DNA scissions on RNA–polymerase binding and chain initiation. *Eur. J. Biochem.* **31**, 394–404 (1972).
31. Takanami, M., Sugimoto, K., Sugisaki, H. & Okamoto, T. Sequence of promoter for coat protein gene of bacteriophage fd. *Nature* **260**, 297–302 (1976).
32. Schaller, H., Gray, C. & Herrmann, K. Nucleotide sequence of an RNA polymerase binding site from the DNA of bacteriophage fd. *Proc. Natl Acad. Sci. USA* **72**, 737–741 (1975).
33. Pribnow, D. Nucleotide sequence of an RNA polymerase binding site at an early T7 promoter. *Proc. Natl Acad. Sci. USA* **72**, 784–788 (1975). **This first promoter sequence analysis reveals that promoters contain instances of conserved sequence motifs**.
34. Ponnambalam, S., Webster, C., Bingham, A. & Busby, S. Transcription initiation at the *Escherichia coli* galactose operon promoters in the absence of the normal-35 region sequences. *J. Biol. Chem.* **261**, 16043–16048 (1986).
35. Keilty, S. & Rosenberg, M. Constitutive function of a positively regulated promoter reveals new sequences essential for activity. *J. Biol. Chem.* **262**, 6389–6395 (1987).
36. Morett, E. & Buck, M. In vivo studies on the interaction of RNA polymerase-σ^{54} with the *Klebsiella pneumoniae* and *Rhizobium meliloti* nifH promoters. *J. Mol. Biol.* **210**, 65–77 (1989).
37. Travers, A. A. Promoter sequence for stringent control of bacterial ribonucleic acid synthesis. *J. Bacteriol.* **141**, 973–976 (1980). **This study highlights the role of the discriminator sequence at bacterial promoters and lays the basis for work on detailed sequence rules**.
38. Zhang, Y. et al. Structural basis of transcription initiation. *Science* **338**, 1076–1080 (2012). **Following earlier structures, this publication reveals new details about interactions by the transcription bubble during initiation**.
39. Ross, W. et al. A third recognition element in bacterial promoters: DNA binding by the a subunit of RNA polymerase. *Science* **262**, 1407–1413 (1993). **At the culmination of many publications describing the contributions of upstream sequences to promoter activity, this study presents a unified model**.
40. Estrem, S. T. et al. Bacterial promoter architecture: subsite structure of UP elements and interactions with the carboxy-terminal domain of the RNA polymerase a subunit. *Genes. Dev.* **13**, 2134–2147 (1999).
41. Harden, T. T. et al. Bacterial RNA polymerase can retain σ^{70} throughout transcription. *Proc. Natl Acad. Sci. USA* **113**, 602–607 (2016).
42. Sun, Z. et al. Density of σ^{70} promoter-like sites in the intergenic regions dictates the redistribution of RNA polymerase during osmotic stress in *Escherichia coli*. *Nucleic Acids Res.* **47**, 3970–3985 (2019).
43. Huerta, A. M. & Collado-Vides, J. Sigma70 promoters in *Escherichia coli*: specific transcription in dense regions of overlapping promoter-like signals. *J. Mol. Biol.* **333**, 261–278 (2003).
44. Huerta, A. M., Francino, M. P., Morett, E. & Collado-Vides, J. Selection for unequal densities of σ^{70} promoter-like signals in different regions of large bacterial genomes. *PLoS Genet.* **2**, 1740–1750 (2006).
45. Froula, J. L. & Francino, M. P. Selection against spurious promoter motifs correlates with translational efficiency across bacteria. *PLoS One* **2**, 1–11 (2007).
46. Yona, A. H., Alm, E. J. & Gore, J. Random sequences rapidly evolve into de novo promoters. *Nat. Commun.* **9**, 1–10 (2018).
47. Urtecho, G. et al. Genome-wide functional characterization of *Escherichia coli* promoters and regulatory elements responsible for their function. Preprint at *bioRxiv* https://doi.org/10.1101/2020.01.04.894907 (2020).
48. Jones, B. B., Chan, H., Rothstein, S., Wells, R. D. & Reznikoff, W. S. RNA polymerase binding sites in λplac5 DNA. *Proc. Natl Acad. Sci. USA* **74**, 4914–4918 (1977).
49. Grainger, D. C., Hurd, D., Harrison, M., Holdstock, J. & Busby, S. J. W. Studies of the distribution of *Escherichia coli* cAMP-receptor protein and RNA polymerase along the *E. coli* chromosome. *Proc. Natl Acad. Sci. USA* **102**, 17693–17698 (2005).
50. Wigneshweraraj, S. et al. Modus operandi of the bacterial RNA polymerase containing the σ^{54} promoter-specificity factor. *Mol. Microbiol.* **68**, 538–546 (2008).
51. Bonocora, R. P., Smith, C., Lapierre, P. & Wade, J. T. Genome-scale mapping of *Escherichia coli* σ^{54} reveals widespread, conserved intragenic binding. *PLoS Genet.* **11**, 1–30 (2015).
52. Schaefer, J., Engl, C., Zhang, N., Lawton, E. & Buck, M. Genome wide interactions of wild-type and activator bypass forms of σ^{54}. *Nucleic Acids Res.* **43**, 7280–7291 (2015). **This is an authoritative study of the omics of sigma factor 54**.
53. Bono, A. C. et al. Novel DNA binding and regulatory activities for σ^{54} (RpoN) in *Salmonella enterica* serovar Thyphimurium 14028s. *J. Bacteriol.* **199**, 1–24 (2017).
54. Shao, X. et al. RpoN-dependent direct regulation of quorum sensing and the type VI secretion system in *Pseudomonas aeruginosa* PAO1. *J. Bacteriol.* **200**, 1–17 (2018).
55. Goldman, S. R., Ebright, R. H. & Nickels, B. E. Direct detection of abortive RNA transcripts in vivo. *Science* **324**, 927–928 (2009).
56. Yus, E. et al. Transcription start site associated RNAs in bacteria. *Mol. Syst. Biol.* **8**, 1–7 (2012).
57. Dornenburg, J. E., DeVita, A. M., Palumbo, M. J. & Wade, J. T. Widespread antisense transcription in *Escherichia coli*. *MBio* **1**, 1–4 (2010).
58. Raghavan, R., Sloan, D. B. & Ochman, H. Pervasive transcription is widespread but rarely conserved in Enteric bacteria. *MBio* **3**, 1–7 (2012).

59. Sasse-dwight, S. & Gralla, J. A. Y. D. Probing the *Escherichia coli* glnALG upstream activation mechanism in vivo. *Proc. Natl Acad. Sci. USA* **85**, 8934–8938 (1988).
60. Domínguez-Cuevas, P., Marín, P., Ramos, J. L. & Marqués, S. RNA polymerase holoenzymes can share a single transcription start site for the Pm promoter: critical nucleotides in the −7 to −18 region are needed to select between RNA polymerase with σ^{38} or σ^{32}. *J. Biol. Chem.* **280**, 41315–41323 (2005).
61. Reichenbach, B., Göpel, Y. & Görke, B. Dual control by perfectly overlapping σ^{54}- and σ^{70}-promoters adjusts small RNA GlmY expression to different environmental signals. *Mol. Microbiol.* **74**, 1054–1070 (2009).
62. Wade, J. T. et al. Extensive functional overlap between σ factors in *Escherichia coli*. *Nat. Struct. Mol. Biol.* **13**, 806–814 (2006).
63. Jacob, F. & Monod, J. Genetic regulatory mechanisms in the synthesis of proteins. *J. Mol. Biol.* **3**, 318–356 (1961).
64. Pardee, A. B., Jacob, F. & Monod, J. The genetic control and cytoplasmic expression of 'inducibility' in the synthesis of β-galactosidase by *E. coli*. *J. Mol. Biol.* **1**, 165–178 (1959).
65. Gilbert, W. & Müller-Hill, B. Isolation of the Lac repressor. *Proc. Natl Acad. Sci. USA* **56**, 1891–1898 (1966).
The first characterization of a transcription factor defines these entities as proteins that specifically bind to DNA regions.
66. Ptashne, M. Specific binding of the λ phage repressor to λ DNA. *Nature* **214**, 232–234 (1967).
67. Maniatis, T. et al. Recognition sequences of repressor and polymerase in the operators of bacteriophage lambda. *Cell* **5**, 109–113 (1975).
This first characterization of a transcription factor binding site describes them as specific regions in the DNA.
68. Englesberg, E., Irr, J., Power, J. & Lee, N. Positive control of enzyme synthesis by gene C in the L-arabinose system. *J. Bacteriol.* **90**, 946–957 (1965).
This landmark paper undermines the belief that regulation must be due to repressors.
69. Schwartz, M. Aspects biochimiques et génétiques du metabolisme du maltose chez *Escherichia coli* K12. *Comptes Rendus Hebd. Des. Seances De L Acad. Des Sci.* **260**, 2613 (1965).
70. Thomas, R. Control of development in temperate bacteriophages. I. Induction of prophage genes following hetero-immune super-infection. *J. Mol. Biol.* **22**, 79–95 (1966).
71. Borukhov, S., Lee, J. & Laptenko, O. Bacterial transcription elongation factors: new insights into molecular mechanism of action. *Mol. Microbiol.* **55**, 1315–1324 (2005).
72. Storz, G., Opdyke, J. A. & Wassarman, K. M. Regulating bacterial transcription with small RNAs. *Cold Spring Harb. Symp. Quant. Biol.* **71**, 269–273 (2006).
73. Schwenk, S. & Arnvig, K. B. Regulatory RNA in *Mycobacterium tuberculosis*, back to basics. *Pathog. Dis.* **76**, 1–12 (2018).
74. Gourse, R. L. et al. Transcriptional responses to ppGpp and DksA. *Annu. Rev. Microbiol.* **72**, 163–184 (2018).
75. Wassarman, K. M. 6S RNA: a small RNA regulator of transcription. *Curr. Opin. Microbiol.* **10**, 164–168 (2007).
76. Wassarman, K. M., Repoila, F., Rosenow, C., Storz, G. & Gottesman, S. Identification of novel small RNAs using comparative genomics and microarrays. *Genes. Dev.* **15**, 1637–1651 (2001).
77. Eckweiler, D., Dudek, C.-A., Hartlich, J., Brötje, D. & Jahn, D. PRODORIC2: the bacterial gene regulation database in 2018. *Nucleic Acids Res.* **46**, D320–D326 (2018).
78. Fornes, O. et al. JASPAR 2020: update of the open-access database of transcription factor binding profiles. *Nucleic Acids Res.* **48**, D87–D92 (2020).
79. Hu, H. et al. AnimalTFDB 3.0: a comprehensive resource for annotation and prediction of animal transcription factors. *Nucleic Acids Res.* **47**, D33–D38 (2019).
80. Jin, J. et al. PlantTFDB 4.0: toward a central hub for transcription factors and regulatory interactions in plants. *Nucleic Acids Res.* **45**, D1040–D1045 (2016).
81. Teixeira, M. C. et al. YEASTRACT: an upgraded database for the analysis of transcription regulatory networks in *Saccharomyces cerevisiae*. *Nucleic Acids Res.* **46**, D348–D353 (2018).
82. Paul, B. J. et al. DksA: a critical component of the transcription initiation machinery that potentiates the regulation of rRNA promoters by ppGpp and the initiating NTP. *Cell* **118**, 311–322 (2004).
83. Gregory, B. D. et al. A regulator that inhibits transcription by targeting an intersubunit interaction of the RNA polymerase holoenzyme. *Proc. Natl Acad. Sci. USA* **101**, 4554–4559 (2004).
84. Pratt, L. A. & Silhavy, T. J. Crl stimulates RpoS activity during stationary phase. *Mol. Microbiol.* **29**, 1225–1236 (1998).
85. Srivastava, D. B. et al. Structure and function of CarD, an essential mycobacterial transcription factor. *Proc. Natl Acad. Sci. USA* **110**, 12619–12624 (2013).
86. Browning, D. F. & Busby, S. J. W. Local and global regulation of transcription initiation in bacteria. *Nat. Rev. Microbiol.* **14**, 638–650 (2016).
87. Haldenwang, W. G., Lang, N. & Losick, R. A sporulation-induced sigma-like regulatory protein from *B. subtilis*. *Cell* **23**, 615–624 (1981).
88. Grossman, A. D., Erickson, J. W. & Gross, C. A. The htpR gene product of *E. coli* is a sigma factor for heat-shock promoters. *Cell* **38**, 383–390 (1984).
89. Taylor, W. E. et al. Transcription from a heat-inducible promoter causes heat shock regulation of the sigma subunit of *E. coli* RNA polymerase. *Cell* **38**, 371–381 (1984).
90. Burgess, R. R., Travers, A. A., Dunn, J. J. & Bautz, E. K. Factor stimulating transcription by RNA polymerase. *Nature* **221**, 43–46 (1969).
91. Feklístov, A., Sharon, B. D., Darst, S. A. & Gross, C. A. Bacterial sigma factors: a historical, structural, and genomic perspective. *Annu. Rev. Microbiol.* **68**, 357–376 (2014).
This is a must-read, beautifully written review of bacterial sigma factors, starting with their history.
92. Campagne, S., Marsh, M. E., Capitani, G., Vorholt, J. A. & Allain, F. H. T. Structural basis for −10 promoter element melting by environmentally induced sigma factors. *Nat. Struct. Mol. Biol.* **21**, 269–276 (2014).
93. Griffith, K. L., Shah, I. M., Myers, T. E., O'Neill, M. C. & Wolf, R. E. Evidence for 'pre-recruitment' as a new mechanism of transcription activation in *Escherichia coli*: The large excess of SoxS binding sites per cell relative to the number of SoxS molecules per cell. *Biochem. Biophys. Res. Commun.* **291**, 979–986 (2002).
94. Shah, I. M. & Wolf, R. E. Novel protein–protein interaction between *Escherichia coli* SoxS and the DNA binding determinant of the RNA polymerase α subunit: SoxS functions as a co-sigma factor and redeploys RNA polymerase from UP-element-containing promoters to SoxS-dependent promoters during oxidative stress. *J. Mol. Biol.* **343**, 513–532 (2004).
95. Li, Z. & Demple, B. Sequence specificity for DNA binding by *Escherichia coli* SoxS and Rob proteins. *Mol. Microbiol.* **20**, 937–945 (1996).
96. Kaur, G. et al. *Mycobacterium tuberculosis* CarD, an essential global transcriptional regulator forms amyloid-like fibrils. *Sci. Rep.* **8**, 1–13 (2018).
97. Hubin, E. A. et al. Structure and function of the mycobacterial transcription initiation complex with the essential regulator RbpA. *eLife* **6**, 1–40 (2017).
98. Rammohan, J., Manzano, A. R., Garner, A. L., Stallings, C. L. & Galburt, E. A. CarD stabilizes mycobacterial open complexes via a two-tiered kinetic mechanism. *Nucleic Acids Res.* **43**, 3272–3285 (2015).
99. Dorman, C. J., Schumacher, M. A., Bush, M. J., Brennan, R. G. & Buttner, M. J. When is a transcription factor a NAP? *Curr. Opin. Microbiol.* **55**, 26–33 (2020).
100. Schneider, R. et al. An architectural role of the *Escherichia coli* chromatin protein FIS in organising DNA. *Nucleic Acids Res.* **29**, 5107–5114 (2001).
101. Dillon, S. C. & Dorman, C. J. Bacterial nucleoid-associated proteins, nucleoid structure and gene expression. *Nat. Rev. Microbiol.* **8**, 185–195 (2010).
102. Dame, R. T. The role of nucleoid-associated proteins in the organization and compaction of bacterial chromatin. *Mol. Microbiol.* **56**, 858–870 (2005).
103. Blot, N., Mavathur, R., Geertz, M., Travers, A. & Muskhelishvili, G. Homeostatic regulation of supercoiling sensitivity coordinates transcription of the bacterial genome. *EMBO Rep.* **7**, 710–715 (2006).
104. Rimsky, S., Zuber, F., Buckle, M. & Buc, H. A molecular mechanism for the repression of transcription by the H-NS protein. *Mol. Microbiol.* **42**, 1311–1323 (2001).
105. Opel, M. L. et al. Activation of transcription initiation from a stable RNA promoter by a Fis protein-mediated DNA structural transmission mechanism. *Mol. Microbiol.* **53**, 665–674 (2004).
106. Ihara, K. et al. Expression of the alaE gene is positively regulated by the global regulator Lrp in response to intracellular accumulation of L-alanine in *Escherichia coli*. *J. Biosci. Bioeng.* **123**, 444–450 (2017).
107. Finkel, S. E. & Johnson, R. C. The Fis protein: it's not just for DNA inversion anymore. *Mol. Microbiol.* **7**, 1023–1023 (1993).
108. Brandi, A., Giangrossi, M., Giuliodori, A. M. & Falconi, M. An interplay among FIS, H-NS, and guanosine tetraphosphate modulates transcription of the *Escherichia coli* cspA gene under physiological growth conditions. *Front. Mol. Biosci.* **3**, 1–12 (2016).
109. Govantes, F., Orjalo, A. V. & Gunsalus, R. P. Interplay between three global regulatory proteins mediates oxygen regulation of the *Escherichia coli* cytochrome d oxidase (cydAB) operon. *Mol. Microbiol.* **38**, 1061–1073 (2000).
110. Meenakshi, S., Karthik, M. & Munavar, M. H. A putative curved DNA region upstream of rcsA in *Escherichia coli* plays a key role in transcriptional regulation by H-NS. *FEBS Open. Bio* **8**, 1209–1218 (2018).
111. Carmona, M., Claverie-Martin, F. & Magasanik, B. DNA bending and the initiation of transcription at σ^{54}-dependent bacterial promoters. *Proc. Natl Acad. Sci. USA* **94**, 9568–9572 (1997).
112. Ninfa, A. J., Reitzer, L. J. & Magasanik, B. Initiation of transcription at the bacterial glnAp2 promoter by purified *E. coli* components is facilitated by enhancers. *Cell* **50**, 1039–1046 (1987).
113. Azam, T. A. & Ishihama, A. Twelve species of the nucleoid-associated protein from *Escherichia coli*. *J. Biol. Chem.* **274**, 33105–33113 (1999).
114. Martinez-Antonio, A. & Collado-Vides, J. Identifying global regulators in transcriptional regulatory networks in bacteria. *Curr. Opin. Microbiol.* **6**, 482–489 (2003).
115. Santos-Zavaleta, A. et al. A unified resource for transcriptional regulation in *Escherichia coli* K-12 incorporating high-throughput-generated binding data into RegulonDB version 10.0. *BMC Biol.* **16**, 1–12 (2018).
This reports on the most recent update of the RegulonDB database on regulation of transcription initiation and operon organization in *E. coli*.
116. Galagan, J., Lyubetskaya, A. & Gomes, A. ChIP-Seq and the complexity of bacterial transcriptional regulation. *Curr. Top. Microbiol. Immunol.* **363**, 43–68 (2013).
117. Babin, B. M. et al. SutA is a bacterial transcription factor expressed during slow growth in *Pseudomonas aeruginosa*. *Proc. Natl Acad. Sci. USA* **113**, E597–E605 (2016).
118. Jones, C. J. et al. ChIP-Seq and RNA-Seq reveal an AmrZ-mediated mechanism for cyclic di-GMP synthesis and biofilm development by *Pseudomonas aeruginosa*. *PLoS Pathog.* **10**, e1003984 (2014).
119. Perkins, T. T. et al. ChIP-seq and transcriptome analysis of the OmpR regulon of *Salmonella enterica* serovars Typhi and Typhimurium reveals accessory genes implicated in host colonization. *Mol. Microbiol.* **87**, 526–538 (2013).
120. Lobel, L. & Herskovits, A. A. Systems level analyses reveal multiple regulatory activities of CodY controlling metabolism, motility and virulence in *Listeria monocytogenes*. *PLoS Genet.* **12**, 1–27 (2016).
121. Vannini, A. et al. Comprehensive mapping of the *Helicobacter pylori* NikR regulon provides new insights in bacterial nickel responses. *Sci. Rep.* **7**, 1–14 (2017).
122. Vergara-Irigaray, M., Fookes, M. C., Thomson, N. R. & Tang, C. M. RNA-seq analysis of the influence of anaerobiosis and FNR on *Shigella flexneri*. *BMC Genomics* **15**, 1–22 (2014).
123. Grainger, D. C. et al. Genomic studies with *Escherichia coli* MelR protein: applications of chromatin immunoprecipitation and microarrays. *J. Bacteriol.* **186**, 6938–6943 (2004).
124. Sharma, P. et al. The multiple antibiotic resistance operon of enteric bacteria controls DNA repair and outer membrane integrity. *Nat. Commun.* **8**, 1444 (2017).
125. Visweswariah, S. S. & Busby, S. J. W. Evolution of bacterial transcription factors: how proteins take on new tasks, but do not always stop doing the old ones. *Trends Microbiol.* **23**, 463–467 (2015).
126. Beauchene, N. A. et al. Impact of anaerobiosis on expression of the iron-responsive Fur and RyhB regulons. *MBio* **6**, e01947-15 (2015).
127. Gao, Y. et al. Systematic discovery of uncharacterized transcription factors in *Escherichia coli* K-12 MG1655. *Nucleic Acids Res.* **46**, 10682–10696 (2018).
128. Shimada, T., Ishihama, A., Busby, S. J. W. & Grainger, D. C. The *Escherichia coli* RutR transcription factor binds at targets within genes as well as intergenic

regions. *Nucleic Acids Res.* **36**, 3950–3955 (2008). **This first genome-wide study shows a large fraction of intergenic TF sites**.
129. Wade, J. T., Reppas, N. B., Church, G. M. & Struhl, K. Genomic analysis of LexA binding reveals the permissive nature of the *Escherichia coli* genome and identifies unconventional target sites. *Genes Dev.* **19**, 2619–2630 (2005).
130. Myers, K. S. et al. Genome-scale analysis of *Escherichia coli* FNR reveals complex features of transcription factor binding. *PLoS Genet.* **9**, e1003565 (2013).
131. Kroner, G. M., Wolfe, M. B. & Freddolino, P. L. *Escherichia coli* Lrp regulates one-third of the genome via direct, cooperative, and indirect routes. *J. Bacteriol.* **201**, e00411–18 (2019).
132. Jacob, F., Perrin, D., Sanchez, C. & Monod, J. L'opéron: groupe de gènes à expression coordonnée par un opérateur. *C. R. Acad. Sci.* **250**, 1727–1729 (1960). **This report provides the original definition of an operon as a set of co-transcribed genes whose expression is coordinated by an operator**.
133. Gralla, J. D. & Collado-Vides, J. in *Escherichia coli and Salmonella: Cellular and Molecular Biology* (eds Neidhardt, F. & Curtiss, R.) 1232–1245 (ASM Press, 1996).
134. Collado-Vides, J. et al. Bioinformatics resources for the study of gene regulation in bacteria. *J. Bacteriol.* **91**, 23–31 (2009).
135. Reitzer, L. J. & Magasanik, B. Transcription of glnA in *E. coli* is stimulated by activator bound to sites far from the promoter. *Cell* **45**, 785–792 (1986).
136. Claverie-Martin, F. & Magasanik, B. Role of integration host factor in the regulation of the glnHp2 promoter of *Escherichia coli*. *Proc. Natl Acad. Sci. USA* **88**, 1631–1635 (1991).
137. Gralla, J. D. Promoter recognition and mRNA initiation by *Escherichia coli* $E\sigma^{70}$. *Methods Enzymol.* **185**, 37–54 (1990).
138. Hancock, J. M. & Zvelebil, M. J. *Concise Encyclopaedia of Bioinformatics and Computational Biology* (Wiley, 2014).
139. Collado-Vides, J. The search for a grammatical theory of regulation is formally justified by showing the inadequacy of context-free grammars. *Bioinformatics* **7**, 321–326 (1991).
140. Buchler, N. E., Gerland, U. & Hwa, T. On schemes of combinatorial transcription logic. *Proc. Natl Acad. Sci. USA* **100**, 5136–5141 (2003).
141. Bintu, L. et al. Transcriptional regulation by the numbers: applications. *Curr. Opin. Genet. Dev.* **15**, 125–135 (2005).
142. Phillips, R. et al. Figure 1 theory meets figure 2 experiments in the study of gene expression. *Annu. Rev. Biophys.* **48**, 121–163 (2019).
143. Bintu, L. et al. Transcriptional regulation by the numbers: models. *Curr. Opin. Genet. Dev.* **15**, 116–124 (2005). **This step-by-step explanation describes the thermodynamic quantitative modelling of regulatory arrangements**.
144. Segal, E. et al. Module networks: identifying regulatory modules and their condition-specific regulators from gene expression data. *Nat. Genet.* **34**, 166–176 (2003).
145. Fraser, C. M. et al. The minimal gene complement of *Mycoplasma genitalium*. *Science* **270**, 397–404 (1995).
146. Dorman, C. J. Regulation of transcription by DNA supercoiling in *Mycoplasma genitalium*: global control in the smallest known self-replicating genome. *Mol. Microbiol.* **81**, 302–304 (2011).
147. Zhang, W. & Baseman, J. B. Transcriptional regulation of MG_{149}, an osmoinducible lipoprotein gene from *Mycoplasma genitalium*. *Mol. Microbiol.* **81**, 327–339 (2011).
148. Roth, C. W. & Nester, E. W. Co-ordinate control of tryptophan, histidine and tyrosine enzyme synthesis in *Bacillus subtilis*. *J. Mol. Biol.* **62**, 577–589 (1971).
149. Salgado, H., Moreno-Hagelsieb, G., Smith, T. F. & Collado-Vides, J. Operons in *Escherichia coli*: genomic analyses and predictions. *Proc. Natl Acad. Sci. USA* **97**, 6652–6657 (2000).
150. Monod, J., Changeux, J. P. & Jacob, F. Allosteric proteins and cellular control systems. *J. Mol. Biol.* **6**, 306–329 (1963).
151. Jacob, F. Genetics of the bacterial cell. *Science* **152**, 1470–1478 (1966).
152. Bockhorst, J. et al. Predicting bacterial transcription units using sequence and expression data. *Bioinformatics* **19**(Suppl. 1), i34–i43 (2003).
153. Mao, X. et al. DOOR 2.0: presenting operons and their functions through dynamic and integrated views. *Nucleic Acids Res.* **42**, 654–659 (2014).
154. Koide, T. et al. Prevalence of transcription promoters within archaeal operons and coding sequences. *Mol. Syst. Biol.* **5**, 1–16 (2009).
155. Pray, L. A. What is a gene? Colinearity and transcription units. *Nat. Educ.* **1**, 97 (2008).
156. Cho, B. et al. The transcription unit architecture of the *Escherichia coli* genome. *Nat. Biotechnol.* **27**, 1043–1049 (2009).
157. Liu, J. & Turnbough, C. L. Effects of transcriptional start site sequence and position on nucleotide-sensitive selection of alternative start sites at the pyrC promoter in *Escherichia coli*. *J. Bacteriol.* **176**, 2938–2945 (1994).
158. Goldman, S. R. et al. NanoRNAs prime transcription initiation in vivo. *Mol. Cell* **42**, 817–825 (2011).
159. Ciampi, M. S. Rho-dependent terminators and transcription termination. *Microbiology* **152**, 2515–2528 (2006).
160. Lau, L. F. & Roberts, J. W. Rho-dependent transcription termination at lambda R1 requires upstream sequences. *J. Biol. Chem.* **260**, 574–584 (1985).
161. Richardson, L. V. & Richardson, J. P. Rho-dependent termination of transcription is governed primarily by the upstream rho utilization (rut) sequences of a terminator. *J. Biol. Chem.* **271**, 21597–21603 (1996).
162. Jeong, K. S., Ahn, J. & Khodursky, A. B. Spatial patterns of transcriptional activity in the chromosome of *Escherichia coli*. *Genome Biol.* **5**, (2004).
163. Junier, I., Unal, E. B., Yus, E., Lloréns-Rico, V. & Serrano, L. Insights into the mechanisms of basal coordination of transcription using a genome-reduced bacterium. *Cell Syst.* **2**, 391–401 (2016).
164. Yan, B., Boitano, M., Clark, T. A. & Ettwiller, L. SMRT-cappable-seq reveals complex operon variants in bacteria. *Nat. Commun.* **9**, 3676 (2018). **This study reports the identification of transcription units using long-read sequencing**.
165. Bauerle, R. H. & Margolin, P. Evidence for two sites for initiation of gene expression in the tryptophan operon of *Salmonella typhimurium*. *J. Mol. Biol.* **26**, 423–436 (1967).
166. Ueno-Nishio, S., Backman, K. C. & Magasanik, B. Regulation at the glnL-operator-promoter of the complex glnALG operon of *Escherichia coli*. *J. Bacteriol.* **153**, 1247–1251 (1983).
167. Conway, T. et al. Unprecedented high-resolution view of bacterial operon architecture revealed by RNA sequencing. *MBio* **5**, e01442-14 (2014).
168. Li, S., Dong, X. & Su, Z. Directional RNA-seq reveals highly complex condition-dependent transcriptomes in *E. coli* K12 through accurate full-length transcripts assembling. *BMC Genomics* **14**, 1–24 (2013).
169. Mao, X. et al. Revisiting operons: an analysis of the landscape of transcriptional units in *E. coli*. *BMC Bioinformatics* **16**, 1–9 (2015).
170. Ju, X., Li, D. & Liu, S. Full-length RNA profiling reveals pervasive bidirectional transcription terminators in bacteria. *Nat. Microbiol.* **4**, 1907–1918 (2019).
171. Sharma, C. M. et al. The primary transcriptome of the major human pathogen *Helicobacter pylori*. *Nature* **464**, 250–255 (2010).
172. Lybecker, M., Bilusic, I. & Raghavan, R. Pervasive transcription: detecting functional RNAs in bacteria. *Transcription* **5**, e944039 (2014).
173. Wade, J. T. & Grainger, D. C. Pervasive transcription: illuminating the dark matter of bacterial transcriptomes. *Nat. Rev. Microbiol.* **12**, 647–653 (2014).
174. Price, M. N. et al. Indirect and suboptimal control of gene expression is widespread in bacteria. *Mol. Syst. Biol.* **9**, 1–18 (2013).
175. Price, M. N., Wetmore, K. M., Deutschbauer, A. M. & Arkin, A. P. A comparison of the costs and benefits of bacterial gene expression. *PLoS One* **11**, 1–22 (2016).
176. Shao, W., Price, M. N., Deutschbauer, A. M., Romine, M. F. & Arkin, A. P. Conservation of transcription start sites within genes across a bacterial genus. *MBio* **5**, 1–13 (2014).
177. Wade, J. T. & Grainger, D. C. Spurious transcription and its impact on cell function. *Transcription* **9**, 182–189 (2018).
178. Pannier, L., Merino, E., Marchal, K. & Collado-Vides, J. Effect of genomic distance on coexpression of coregulated genes in *E. coli*. *PLoS One* **12**, e0174887 (2017).
179. Stringer, A. M. et al. Genome-scale analyses of *Escherichia coli* and *Salmonella enterica* AraC reveal noncanonical targets and an expanded core regulon. *J. Bacteriol.* **196**, 660–671 (2014).
180. Chen, Y.-J. et al. Characterization of 582 natural and synthetic terminators and quantification of their design constraints. *Nat. Methods* **10**, 659–664 (2013).
181. Maas, W. K. & Clark, A. J. Studies on the mechanism of repression of arginine biosynthesis in *Escherichia coli*: II. Dominance of repressibility in diploids. *J. Mol. Biol.* **8**, 365–370 (1964).
182. Ledezma-Tejeida, D., Altamirano-Pacheco, L., Fajardo, V. & Collado-Vides, J. Limits to a classic paradigm: most transcription factors in *E. coli* regulate genes involved in multiple biological processes. *Nucleic Acids Res.* **47**, 6656–6667 (2019).
183. Pittard, J. & Yang, J. Biosynthesis of the aromatic amino acids. *EcoSal Plus* **3**, 1–39 (2008).
184. Smith, M. W. & Neidhardt, F. C. Proteins induced by aerobiosis in Escherichia coli. *J. Bacteriol.* **154**, 336–343 (1983).
185. Schaefer, E. M., Hartz, D., Gold, L. & Simoni, R. D. Ribosome-binding sites and RNA-processing sites in the transcript of the *Escherichia coli* unc operon. *J. Bacteriol.* **171**, 3901–3908 (1989).
186. Monod, J., Wyman, J. & Changeux, J.-P. On the nature of allosteric transitions: a plausible model. *J. Mol. Biol.* **12**, 88–118 (1965). **This study describes the classic, elegant model of allosteric transitions**.
187. Hoch, J. A. Two-component and phosphorelay signal transduction. *Curr. Opin. Microbiol.* **3**, 165–170 (2000).
188. Grundy, F. J. & Henkin, T. M. Regulation of gene expression by effectors that bind to RNA. *Curr. Opin. Microbiol.* **7**, 126–131 (2004).
189. Horii, T. et al. Regulation of SOS functions: purification of *E. coli* LexA protein and determination of its specific site cleaved by the RecA protein. *Cell* **27**, 515–522 (1981).
190. Jenal, U. & Hengge-Aronis, R. Regulation by proteolysis in bacterial cells. *Curr. Opin. Microbiol.* **6**, 163–172 (2003).
191. Uphoff, S. et al. Stochastic activation of a DNA damage response causes cell-to-cell mutation rate variation. *Science* **351**, 1094–1097 (2016).
192. Takinowaki, H., Matsuda, Y., Yoshida, T., Kobayashi, Y. & Ohkubo, T. The solution structure of the methylated form of the N-terminal 16-kDa domain of *Escherichia coli* Ada protein. *Protein Sci.* **15**, 487–497 (2006).
193. Kotte, O., Zaugg, J. B. & Heinemann, M. Bacterial adaptation through distributed sensing of metabolic fluxes. *Mol. Syst. Biol.* **6**, 1–9 (2010).
194. Mekalanos, J. J. Environmental signals controlling expression of virulence determinants in bacteria. *J. Bacteriol.* **174**, 1–7 (1992).
195. Maurelli, A. T. Temperature regulation of virulence genes in pathogenic bacteria: a general strategy for human pathogens? *Microb. Pathog.* **7**, 1–10 (1989).
196. Miller, J. F., Mekalanos, J. J. & Falkow, S. Coordinate regulation and sensory transduction in the control of bacterial virulence. *Science* **243**, 1355–1362 (1989).
197. Hurme, R., Berndt, K. D., Normark, S. J. & Rhen, M. A proteinaceous gene regulatory thermometer in *Salmonella*. *Cell* **90**, 55–64 (1997).
198. Piraner, D. I., Abedi, M. H., Moser, B. A., Lee-Gosselin, A. & Shapiro, M. G. Tunable thermal bioswitches for in vivo control of microbial therapeutics. *Nat. Chem. Biol.* **13**, 75–80 (2017).
199. Lindner, R. et al. Photoactivation mechanism of a bacterial light-regulated adenylyl cyclase. *J. Mol. Biol.* **429**, 1336–1351 (2017).
200. Winkler, A. et al. A ternary AppA-PpsR-DNA complex mediates light regulation of photosynthesis-related gene expression. *Nat. Struct. Mol. Biol.* **20**, 859–867 (2013).
201. Smith, B., Kumar, A. & Bittner, T. *Basic Formal Ontology for Bioinformatics* (IFOMIS Reports, 2005).
202. Strainic, M. G., Sullivan, J. J., Collado-Vides, J. & DeHaseth, P. L. Promoter interference in a bacteriophage lambda control region: effects of a range of interpromoter distances. *J. Bacteriol.* **182**, 216–220 (2000).
203. Scherrer, K. Primary transcripts: from the discovery of RNA processing to current concepts of gene expression—review. *Exp. Cell Res.* **373**, 1–33 (2018).
204. Sanchez-Vazquez, P., Dewey, C. N., Kitten, N., Ross, W. & Gourse, R. L. Genome-wide effects on *Escherichia coli* transcription from ppGpp binding to its two sites on RNA polymerase. *Proc. Natl Acad. Sci. USA* **116**, 8310–8319 (2019).
205. Browning, D. F., Butala, M. & Busby, S. J. W. Bacterial transcription factors: regulation by Pick 'N' Mix. *J. Mol. Biol.* **431**, 4067–4077 (2019).

206. Haugen, S. P. et al. rRNA promoter regulation by nonoptimal binding of σ region 1.2: an additional recognition element for RNA polymerase. *Cell* **125**, 1069–1082 (2006).
207. Josaitis, C. A., Gaal, T. & Gourse, R. L. Stringent control and growth-rate-dependent control have nonidentical promoter sequence requirements. *Proc. Natl Acad. Sci. USA* **92**, 1117–1121 (1995).
208. Davis, M. C., Kesthely, C. A., Franklin, E. A. & MacLellan, S. R. The essential activities of the bacterial sigma factor. *Can. J. Microbiol.* **63**, 89–99 (2017).
209. Ho, Y. Sen, Wulff, D. L. & Rosenberg, M. Bacteriophage λ protein cII binds promoters on the opposite face of the DNA helix from RNA polymerase. *Nature* **304**, 703–708 (1983).
210. Buck, M. & Cannon, W. Specific binding of the transcription factor sigma-54 to promoter DNA. *Nature* **358**, 422–424 (1992).
211. Seo, S. W. et al. Revealing genome-scale transcriptional regulatory landscape of OmpR highlights its expanded regulatory roles under osmotic stress in *Escherichia coli* K-12 MG1655. *Sci. Rep.* **7**, 1–10 (2017).
212. Cho, B. K., Federowicz, S., Park, Y. S., Zengler, K. & Palsson, B. Deciphering the transcriptional regulatory logic of amino acid metabolism. *Nat. Chem. Biol.* **8**, 65–71 (2012).
213. Cho, B. K. et al. The PurR regulon in *Escherichia coli* K-12 MG1655. *Nucleic Acids Res.* **39**, 6456–6464 (2011).
214. Kim, D. et al. Systems assessment of transcriptional regulation on central carbon metabolism by Cra and CRP. *Nucleic Acids Res.* **46**, 2901–2917 (2018).
215. Seo, S. W. et al. Deciphering Fur transcriptional regulatory network highlights its complex role beyond iron metabolism in *Escherichia coli*. *Nat. Commun.* **5**, 1–10 (2014).
216. Seo, S. W., Kim, D., O'Brien, E. J., Szubin, R. & Palsson, B. O. Decoding genome-wide GadEWX-transcriptional regulatory networks reveals multifaceted cellular responses to acid stress in *Escherichia coli*. *Nat. Commun.* **6**, 1–8 (2015).
217. Seo, S. W., Kim, D., Szubin, R. & Palsson, B. O. Genome-wide reconstruction of OxyR and SoxRS transcriptional regulatory networks under oxidative stress in *Escherichia coli* K-12 MG1655. *Cell Rep.* **12**, 1289–1299 (2015).
218. Grainger, D. C., Aiba, H., Hurd, D., Browning, D. F. & Busby, S. J. W. Transcription factor distribution in *Escherichia coli*: studies with FNR protein. *Nucleic Acids Res.* **35**, 269–278 (2007).
219. Partridge, J. D., Bodenmiller, D. M., Humphrys, M. S. & Spiro, S. NsrR targets in the *Escherichia coli* genome: new insights into DNA sequence requirements for binding and a role for NsrR in the regulation of motility. *Mol. Microbiol.* **73**, 680–694 (2009).
220. Grainger, D. C., Hurd, D., Goldberg, M. D. & Busby, S. J. W. Association of nucleoid proteins with coding and non-coding segments of the *Escherichia coli* genome. *Nucleic Acids Res.* **34**, 4642–4652 (2006).
221. Sogaard-Andersen, L., Mellegaard, N. E., Douthwaite, S. R. & Valentin-Hansen, P. Tandem DNA-bound cAMP–CRP complexes are required for transcriptional repression of the deoP2 promoter by the CytR repressor in *Escherichia coli*. *Mol. Microbiol.* **4**, 1595–1601 (1990).
222. Tao, H., Hasona, A., Do, P. M., Ingram, L. O. & Shanmugam, K. T. Global gene expression analysis revealed an unsuspected deo operon under the control of molybdate sensor, ModE protein, in *Escherichia coli*. *Arch. Microbiol.* **184**, 225–233 (2005).
223. González-Gil, G., Bringmann, P. & Kahmann, R. FIS is a regulator of metabolism in *Escherichia coli*. *Mol. Microbiol.* **22**, 21–29 (1996).
224. Valentin-Hansen, P., Albrechtsen, B. & Løve Larsen, J. E. DNA–protein recognition: demonstration of three genetically separated operator elements that are required for repression of the *Escherichia coli* deoCABD promoters by the DeoR repressor. *EMBO J.* **5**, 2015–2021 (1986).
225. Barnard, A., Wolfe, A. & Busby, S. Regulation at complex bacterial promoters: how bacteria use different promoter organizations to produce different regulatory outcomes. *Curr. Opin. Microbiol.* **7**, 102–108 (2004).

Acknowledgements
C.M.-A. is a doctoral student from the Programa de Doctorado en Ciencias Biomédicas, Universidad Nacional Autónoma de México (UNAM), and has received CONACyT fellowship 576333. J.C.-V. acknowledges funding by Universidad Nacional Autónoma de México (UNAM) and the National Institutes of Health (5R01GM110597-04, 1RO1GM131643-01A1 and R01GM077678). J.C.-V. acknowledges being on sabbatical leave at the Center for Genomic Regulation, Barcelona, Spain. B.O.P. acknowledges the support of the Galletti Endowment at UC San Diego. The authors acknowledge J. Soffer, S. Gama-Castro and H. Salgado for useful discussions, and D. W. Sant for updating the definitions in Sequence Ontology. The authors also acknowledge the highly valuable suggestions from the referees.

Author contributions
C.M.-A. researched the literature. C.M.-A., S.J.W.B., J.T.W., J.v.H., A.P.A., G.D.S., K.E., B.O.P., J.E.G. and J.C.-V. provided substantial contributions to discussions of the content. C.M.-A., S.J.W.B., J.T.W., J.v.H., A.P.A., G.D.S. and J.C.-V. wrote the article. C.M.-A., S.J.W.B., J.T.W., J.v.H., A.P.A., G.D.S., K.E., B.O.P. and J.C.-V. reviewed and/or edited the manuscript before submission.

Competing interests
The authors declare no competing interests.

Peer review information
Nature Reviews Genetics thanks A. S. Ribeiro and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

Publisher's note
Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Supplementary information
Supplementary information is available for this paper at https://doi.org/10.1038/s41576-020-0254-8.

RELATED LINKS
Basic Formal Ontology: https://basic-formal-ontology.org/
EcoCyc: https://ecocyc.org
RegulonDB: http://regulondb.ccg.unam.mx/
RegulonDB Gensor Unit Groups: http://regulondb.ccg.unam.mx/central_panel_menu/integrated_views_and_tools/gensor_unit_groups
Sequence Ontology: http://www.sequenceontology.org

Modelo ontológico parcial

Paralelamente al análisis conceptual del dominio, se experimentó con la expresividad del lenguaje ontológico OWL para representar la regulación procariótica con el propósito específico de clasificar sistemas de genes en inducibles y reprimibles a partir de las interacciones moleculares de regulación directas. La ontología resultante puede considerarse un primer esbozo de una ontología de aplicación: no es una ontología de referencia y no incluye las definiciones propuestas finales. Esta ontología se presentó en la 11ª Conferencia Internacional sobre Ingeniería del Conocimiento y Desarrollo de Ontologías y está publicada en las actas del congreso.

Towards the Prokaryotic Regulation Ontology: An Ontological Model to Infer Gene Regulation Physiology from Mechanisms in Bacteria

Citlalli Mejía-Almonte[a] and Julio Collado-Vides[b]
Center for Genomic Science, UNAM, Av. Universidad, Cuernavaca, Mexico

Keywords: Formal Ontology, Domain Ontology, Gene Regulation, Bacteria.

Abstract: Here we present a formal ontological model that explicitly represents regulatory interactions among the main objects involved in transcriptional regulation in bacteria. These formal relations allow the inference of gene regulation physiology from gene regulation mechanisms. The automatically instantiated classes can be used to assist in the mechanistic interpretation of gene expression experiments done at the physiological level, such as RNA-seq. This is the first step to develop a more comprehensive ontology focused on prokaryotic gene regulation. The ontology is available at https://github.com/prokaryotic-regulation-ontology

1 INTRODUCTION

Since the success shown by the Gene Ontology as a controlled vocabulary, bio-ontologies are increasingly important tools in bio-informatics. However, little has been explored regarding formal ontological representation in the domain of bacterial gene regulation. There are two granularity levels at which gene regulation can be studied. At the physiological level, transcript concentration or gene product activity is directly measured under some condition, normally adding or depleting certain chemicals to growth media (Burstein et al., 1965). At the mechanistic level, the effect of specific mutations on gene expression is studied to discover the precise regulators involved in some system (Ptashne, 1967). At this level, the most studied mechanisms are those of transcription initiation mediated by transcription factors. These proteins can adjust gene expression to environmental requirements using their two main functional domains: the effector-binding domain that senses the environmental signal and the DNA binding domain. Transcription factors bind to DNA in sites called transcription factor binding sites, thereby increasing or decreasing the activity of a promoter. Promoters are the DNA regions where transcription of transcription units (TUs) begins; TUs in turn contain one or more genes. Therefore, the expression of a TU is regulated by regulation of the promoter activity. Here, we develop an ontological model that can infer the physiology from mechanisms of gene regulation.

The result of transcriptome analysis are sets of genes that are either underexpressed or overexpressed under a given condition, including the addition of chemicals to growth media. The observation of underexpression corresponds to the observation of gene inhibition, whereas the observation of overexpression corresponds to the observation of gene induction. This means that transcriptome analysis gives us physiological insights, rather than mechanistic ones. The model presented here, automatically instantiates sets of genes that are induced or repressed by some molecule based on the mechanisms of induction or repression. The final terms will encode both the physiology and the mechanisms of gene regulation (see below). Thus, this ontology can help in the mechanistic interpretation of gene expression experiments that are done at the physiological level, such as transcriptome analysis.

No ontology explicitly states the aim of modeling gene regulation in the obo-foundry repository (Smith et al., 2007); whereas a search in bioportal (Noy et al., 2009; Noy et al., 2001) only retrieves the Gene Regulation Ontology (GRO) (Beisswanger et al., 2008). This ontology includes object properties to define *agents* and *patients* of regulation, but it focuses on the mechanistic description of gene regulation and it does not distinguishes the two granularity levels of gene regulation described in this paper. Thus, here we develop an ontology to represent both mechanisms and physiology of gene regulation, the later inferred from the former.

[a] https://orcid.org/0000-0002-0142-5591
[b] https://orcid.org/0000-0001-8780-7664

Mejía-Almonte, C. and Collado-Vides, J.
Towards the Prokaryotic Regulation Ontology: An Ontological Model to Infer Gene Regulation Physiology from Mechanisms in Bacteria.
DOI: 10.5220/0008387804950499
In *Proceedings of the 11th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management (IC3K 2019)*, pages 495-499
ISBN: 978-989-758-382-7

2 DEVELOPMENT PROCESS

We are using top-down ontology development approach (Noy et al., 2001). First, we included the most general and important entities involved in regulation of transcription initiation: transcription factor, transcription factor binding site, promoter, transcription unit, effector, etc. Second, we included the corresponding biological relations among them. Third, we created the classes that will be automatically instantiated: *TF bound to the DNA* and *regulated system*. Fourth, we formally defined these classes taking advantage of the biologically relations included in the second step (figure 2). Lastly, most specific terms have to be generated for each specific TF, TU, promoter, etc. along with their relations. The model will automatically classify these specific entities into the defined classes (see glycolate example). RegulonDB can be used to instantiate the ontology with knowledge about *Escherichia coli K-12* (Santos-Zavaleta et al., 2018).

We are following the OBO-foundry principles. For this, we are taking advantage of the OBO tools ROBOT (Overton et al., 2015) and the Ontology Development Kit (https://github.com/INCATools/ontology-development-kit). The first one is mainly used to extract terms and modules from existing ontologies, while the later is designed for standardized ontology documentation and release of OBO ontologies, taking care of quality control issues. We are using the Basic Formal Ontology as upper-level ontology. So far, we have reused terms from six OBO-foundry ontologies: CHEBI, GO, MSO, NCIT, OGG, and SO (Ashburner et al., 2000; de Matos et al., 2010; Mungall et al., 2011; Sioutos et al., 2007; He et al., 2014) The creation of new classes and axioms was done using Protégé version 5.5. (Musen et al., 2015)

3 MODEL DESCRIPTION

In this paper, classes are written in italics and object properties are written in bold face. Hierarchy is represented as indentation of bulleted lists.

3.1 An n-ary Relation to Represent the Central Transcriptional Regulatory Interaction

Figure 1 depicts the main elements involved in transcriptional regulation along with the relations that exist among them. These were ontologically represented as follows. *Transcription factor* (TF), *TF binding site* (TFBS), *effector*, and *functional conformation* classes were created. Then, an n-ary relation design pattern was used to link these four elements (Noy and Rector, 2004). **TF bound to TFBS** class was created with four properties: **has binding transcription factor**, **has target TFBS**, **is realized in functional conformation**, and **has effector** (Figure 1).

3.2 A Property Chain to Infer Regulation from Anatomy

Figure 1 also depicts how the two key relations that distinguish physiology from mechanisms of transcriptional regulation were ontologically represented. The mechanistic level describes the direct effect that a TF bound to a TFBS has over its cognate promoter, while the physiological level describes the effect that the environmental condition (in our current model represented by the effector molecule) has over the expression of genes in a transcription unit. *Promoter* and *transcription unit* classes were created. Then transcription unit was related with promoter using the property **is transcribed from**, whereas promoter was related to the class *TF bound to TFBS* with the property **has activity regulated by**. The **has expression regulated by** property was created along with the following rule chain expressed in functional syntax (Figure 1) (Hitzler et al., 2009):

```
SubObjectPropertyOf(
   ObjectPropertyChain( :is transcribed from
   :has activity regulated by )
   :has expression regulated by
 )
```

This rule chain represents the fact that if a TU is transcribed from a promoter, and this promoter has its activity regulated by a TF bound to a TFBS, then this TF bound to a TFBS regulates the expression of the TU.

3.3 Automatic Classification of Regulated Systems

At the physiological level, there are only two possibilities: induction or inhibition of gene expression. At the mechanistic level, there are four possibilities. Transcription factors bind to their cognate TFBSs and regulate transcription only when they are in functional conformation. Induction can be achieved by activation when the binding of the effector activates a transcription factor that increases the expression of a TU (active conformation of TF is holo), or by derepression when the binding of the effector deactivates a transcription factor that decreases the expression of a TU (active conformation of TF is apo). Inhibition can be achieved by repression when the bind-

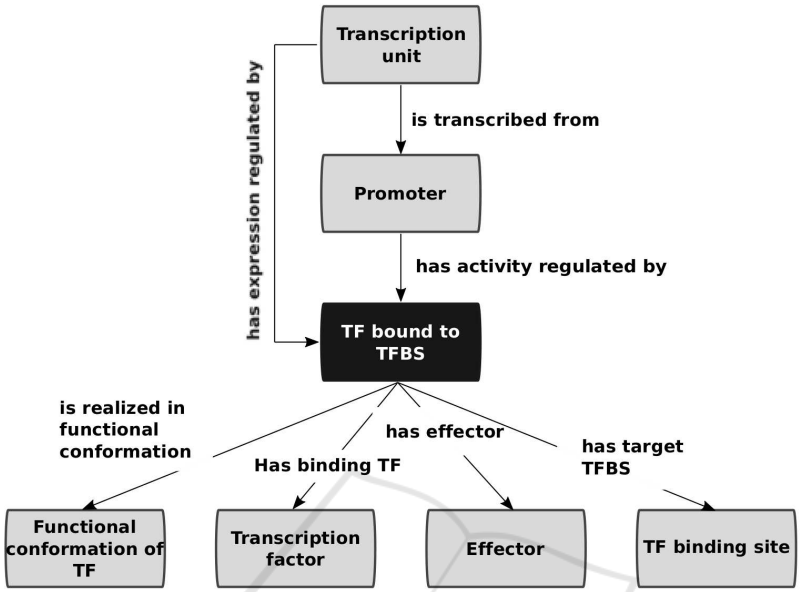

Figure 1: An n-ary relation and a property chain to represent the central regulatory interaction.

ing of the effector activates a transcription factor that decreases the expression of a TU (active conformation is holo), or by de-activation when the binding of the effector deactivates a transcription factor that increases the expression of a TU (active conformation is apo) (Balderas-Martínez et al., 2013). All of these cases describe the physiological response to the appearance of the effector. The disappearance of the effector reverses the response. We will treat these cases later.

Therefore, to automatically classify TUs that are induced or inhibited by an effector we have have created the following subclasses of *TF bound to TFBS* (Figure 2). Equivalent class axioms are shown.

- *TF bound to TFBS in apo conformation* **is realized in functional conformation** some *apo functional conformation of TF*
  - *TF-glycolate active in apo* **has effector** some *glycolate*
- *TF bound to TFBS in holo conformation* **is realized in functional conformation** some *holo functional conformation of TF*
  - *TF-glycolate active in holo* **has effector** some *glycolate*

The classes *inducible system* and *inhibitable system* were created with the following subclasses. Equivalent class axioms are shown.

- *System induced by activation* **has expression increased by** some *transcription factor bound to TFBS in holo conformation*
  - *System induced by activation by glycolate* **has expression increased by** some *TF-glycolate active in holo*
- *System induced by derepression* **has expression decreased by** some *transcription factor bound to TFBS in apo conformation*
  - *System induced by derepression by glycolate* **has expression decreased by** some *TF-glycolate active in apo*
- *System inhibited by repression* **has expression decreased by** some *transcription factor bound to TFBS in holo conformation*
  - *System inhibited by repression by glycolate* **has expression decreased by** some *TF-glycolate active in holo*

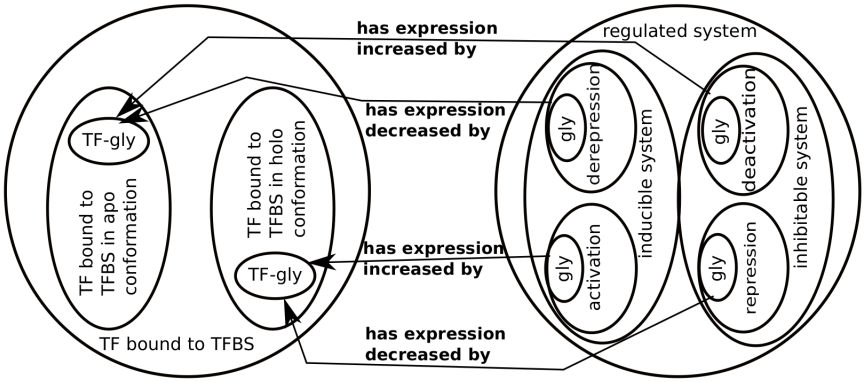

Figure 2: Defined classes to infer physiology from mechanisms. The outer circles represent the most general classes and inner ovals more specific classes. On the left, the hierarchy of the molecular complex TF-TFBS-effector classes is shown. In the text, the most specific classes are named *TF-glycolate active in holo* and *TF-glycolate active in apo*; in the figure, the terms were shortened as TF-gly due to space issues. These classes are automatically instantiated due to the n-ary relation shown in Figure 1. On the right, the hierarchy of effector-induced or effector-repressed systems are shown. The terms were shortened due to space issues: activation is short for *system induced by activation*, derepression is short for *system induced by derepression*, deactivation is short for *system inhibited by deactivation*, and repression is short for *system inhibited by repression*, whereas gly is short for *system induced by activation by glycolate*, *system induced by derepression by glycolate*, *system inhibited by deactivation by glycolate*, and *system inhibited by repression by glycolate*, depending on the superclass. These classes can be automatically instantiated due to the property chain shown in Figure 1.

- *System inhibited by deactivation* **has expression increased by** some *transcription factor bound to TFBS in apo conformation*
  - *System inhibited by deactivation by glycolate* **has expression increased by** some *TF-glycolate active in apo*

In this listing of formal definitions, we included only examples of classes defined by the specific effector glycolate. The final ontology have to be extended to include classes for all known effectors. We plan to do this extension using *E. coli* information retrieved from RegulonDB.

4 CONCLUSIONS

An ontological model that can automatically classify transcription units as effector-dependent repressible or inducible systems was developed. This adds a layer of formal knowledge to the mechanistic representation of bacterial gene regulation included in databases like RegulonDB.

ACKNOWLEDGEMENTS

C.M.A. is a Ph.D. student from the Programa de Doctorado en Ciencias Biomedicas, Universidad Nacional Autonoma de Mexico, receives fellowship 576333 from CONACYT and received financial aid from Programa de Apoyos para Estudios de Posgrado (PAEP) for this conference. JCV acknowledges support by UNAM and by NIH-NIGMS grant RO1-GM110597.

REFERENCES

Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., Davis, A. P., Dolinski, K., Dwight, S. S., Eppig, J. T., et al. (2000). Gene ontology: tool for the unification of biology. *Nature genetics*, 25(1):25.

Balderas-Martínez, Y. I., Savageau, M., Salgado, H., Pérez-Rueda, E., Morett, E., and Collado-Vides, J. (2013). Transcription factors in escherichia coli prefer the holo conformation. *PloS one*, 8(6):e65723.

Beisswanger, E., Lee, V., Kim, J.-J., Rebholz-Schuhmann, D., Splendiani, A., Dameron, O., Schulz, S., Hahn, U., et al. (2008). Gene regulation ontology (gro): design principles and use cases. In *MIE*, pages 9–14.

Burstein, C., Cohn, M., Kepes, A., and Monod, J. (1965). Role du lactose et de ses produits metaboliques dans

l'induction de l'operon lactose chez escherichia coli. *Biochimica et Biophysica Acta (BBA)-Nucleic Acids and Protein Synthesis*, 95(4):634–639.

de Matos, P., Dekker, A., Ennis, M., Hastings, J., Haug, K., Turner, S., and Steinbeck, C. (2010). Chebi: a chemistry ontology and database. *Journal of cheminformatics*, 2(1):P6.

He, Y., Liu, Y., and Zhao, B. (2014). Ogg: a biological ontology for representing genes and genomes in specific organisms. In *ICBO*, pages 13–20. Citeseer.

Hitzler, P., Krötzsch, M., Parsia, B., Patel-Schneider, P. F., and Rudolph, S. (2009). Owl 2 web ontology language primer. *W3C recommendation*, 27(1):123.

Mungall, C. J., Batchelor, C., and Eilbeck, K. (2011). Evolution of the sequence ontology terms and relationships. *Journal of biomedical informatics*, 44(1):87–93.

Musen, M. A. et al. (2015). The protégé project: a look back and a look forward. *AI matters*, 1(4):4.

Noy, N. and Rector, A. (2004). Defining n-ary relations on the semantic web: Use with individuals. *W3C Working Draft*, 21:102.

Noy, N. F., McGuinness, D. L., et al. (2001). Ontology development 101: A guide to creating your first ontology.

Noy, N. F., Shah, N. H., Whetzel, P. L., Dai, B., Dorf, M., Griffith, N., Jonquet, C., Rubin, D. L., Storey, M.-A., Chute, C. G., et al. (2009). Bioportal: ontologies and integrated data resources at the click of a mouse. *Nucleic acids research*, 37(suppl_2):W170–W173.

Overton, J. A., Dietze, H., Essaid, S., Osumi-Sutherland, D., and Mungall, C. J. (2015). Robot: A command-line tool for ontology development. In *ICBO*.

Ptashne, M. (1967). Specific binding of the λ phage repressor to λ dna. *Nature*, 214(5085):232.

Santos-Zavaleta, A., Salgado, H., Gama-Castro, S., Sánchez-Pérez, M., Gómez-Romero, L., Ledezma-Tejeida, D., García-Sotelo, J. S., Alquicira-Hernández, K., Muñiz-Rascado, L. J., Peña-Loredo, P., et al. (2018). Regulondb v 10.5: tackling challenges to unify classic and high throughput knowledge of gene regulation in e. coli k-12. *Nucleic acids research*, 47(D1):D212–D220.

Sioutos, N., de Coronado, S., Haber, M. W., Hartel, F. W., Shaiu, W.-L., and Wright, L. W. (2007). Nci thesaurus: a semantic model integrating cancer-related clinical and molecular information. *Journal of biomedical informatics*, 40(1):30–43.

Smith, B., Ashburner, M., Rosse, C., Bard, J., Bug, W., Ceusters, W., Goldberg, L. J., Eilbeck, K., Ireland, A., Mungall, C. J., et al. (2007). The obo foundry: coordinated evolution of ontologies to support biomedical data integration. *Nature biotechnology*, 25(11):1251.

DISCUSIÓN

Idoneidad del método de falsificación

Algunas características no se pueden seleccionar contradiciendo su necesidad o suficiencia, sino que requieren la estandarización prescriptiva. En los siguientes casos no se pudo analizar la necesidad y la suficiencia para determinar si la característica era definitiva.

Invariablemente las definiciones de promotor recuperadas de la literatura requieren que sea un sitio de unión de la polimerasa, lo cual nos indica que es una condición necesaria. Sin embargo en ninguna de estas definiciones se reconoce la variedad de holoenzimas dependiendo del factor sigma, ni se especifica la cardinalidad de la relación entre el promotor y las variedades de holoenzima. Recientemente se han encontrado que transcritos que inician en precisamente el mismo sitio son generados por más de una factor sigma. Al no haber una especificación previa de esta característica no podíamos falsificarla, sino que tuvimos que tomar una decisión con base en otros criterios. En este caso decidimos que la relación promotor-factor sigma debe ser uno a uno, dado que estructuralmente, las holoenzimas deben interactuar con diferentes subsitios del promotor, además de que el promotor puede estar sujeto a mecanismos de regulación distintos.

¿Los factores de transcripción (TF por transcripción factor) son proteínas, productos génicos o cualquier molécula? ¿Actúan en el inicio, la elongación, la terminación de la transcripción o en todo el ciclo? ¿Es necesario que se unan al ADN? ¿Es suficiente que se una al ADN?

Las definiciones de TF varían en el tipo de entidad considerada como TF: proteína (en la mayoría de los casos), producto génico o cualquier entidad. Existen reguladores transcripcionales que son

ARN, proteínas y moléculas pequeñas y se ha usado el término factor de transcripción para referirse a tales reguladores. Ninguna definición especifica a qué nivel de la transcripción debe actuar un TF y hay reguladores de todos los niveles. Todas especifican que se debe unir a ADN; además, Gene Ontology[17] especifica que debe reconocer un motivo, lo cual significa que su especificidad debe estar dada por el ADN. Pero existen proteínas que se unen a la holoenzima y son consideradas TF por algún autor (CarD). Si tomamos diferentes combinaciones de estas características, seguramente cada clase definida por una combinación de características no estará vacía. En otras palabras si definimos la clase TF con las características más generales: cualquier entidad, a cualquier nivel de la transcripción que interactúe con cualquiera de los componentes moleculares básicos de la transcripción, podríamos tener instancias que son proteínas que se unen a ADN y que regulan la elongación, ARN pequeños que se unen a ADN y que regulan la terminación, moléculas pequeñas que se unen a ADN y que regulan el inicio de la transcripción, etc. En este caso podríamos haber falsificado las características más específicas al encontrar una instancia de un TF que funciona de otra manera y proponer la definición más general que incluiría todo tipo de molécula actuando a cualquier nivel de la transcripción mediante cualquier mecanismo. Sin embargo, en este caso no usamos la falsificación para derivar la definición, sino que nos apegamos a la idea que creemos que es la más extendida: proteínas que se unen a ADN de manera específica y regulan el inicio de la transcripción. Pero esto no es suficiente, dado que los factores sigma también presentan estas características. En esta caso lo que hicimos fue añadir características para definir los factores sigma y decir que los TF no presentan estas características adicionales, para evitar que la clase factor sigma sea subclase de factor de transcripción.

La definición de los extremos estructurales de las unidades de transcripción (TU por *transcription unit*) no se definieron mediante el método de falsificación. Según las definiciones recuperadas, hay dos alternativas para definir los extremos de las TUs: 1) los pares de promotor y terminador, y 2) los pares de sitio de inicio de la transcripción (TSS por sus siglas en inglés) y sitio de terminación de la transcripción (TTS por sus siglas en inglés). La existencia de transcritos que no se generan a partir de un promotor[18,19] podría haber falsificado la necesidad de un promotor para la definición de TU. Sin embargo, consideramos necesario que la transcripción de una TU debe iniciar en un promotor y terminar en un terminador y excluir las regiones transcritas en ausencia de un promotor (por el núcleo de la polimerasa sin factor sigma). Por otro lado, está claro que también es necesario que las TU incluyan los TSS y TTS. Por lo que un enfoque de falsificación para la desambiguación de la definición de esta entidad respecto a esta característica no fue útil. En lugar de eso, propusimos por consenso de varios expertos, que los límites físicos de las TU son los pares TSS-TTS no espurios (que tienen un efecto en la regulación), mientras que sus relaciones (necesarias) con el promotor y el terminador son de regulación, i.e., el promotor regula la selección del TSS y el terminador regula la selección del TSS.

La distinción entre TU y operón. La definición más general de ambos que se encontró en la literatura es la de un conjunto de genes que se cotranscriben. Sin embargo, recientemente se han encontrado transcritos sin función aparente[20]. El concepto de TU está estrechamente relacionado con el de transcrito e incluso también podría haber ambigüedades entre ellos. Por consenso de expertos proponemos que la TU es la región en el ADN equivalente a la del transcrito primario generado a partir de un promotor. Entonces podemos falsificar la necesidad de contenido génico para las TUs. Mientras que el concepto original de operón siempre ha estado presente en la

conciencia colectiva necesariamente como un conjunto de genes cuya expresión está coordinada. Por lo tanto, para desenredar las relaciones entre las definiciones de TU y operón, primero se tuvo que proponer una desambiguación por consenso entre los conceptos de TU y transcrito y, con base en esta, falsificar la necesidad de contenido génico de la TU, dejando así la definición de operón en función de genes y la de TU en términos estructurales. Además, la definición de operón se extendió para poder considerar que varias TU que se sobrelapan forman un operón, sin exigir coregulación ni cotranscripción.

La definición original de regulón se especificaba en términos de una sustancia represora, la cual podía consistir en varias entidades, pero cualquiera que sea su naturaleza actúa de manera unitaria[21]. En este caso no falsificamos necesidad ni suficiencia, sino que, a la luz del conocimiento nuevo proponemos por consenso del grupo de expertos precisar que si la entidad reguladora compleja consiste en entidades que actúan de manera unitaria como en una proteína heteromultimérica entonces es un regulón simple. Por otro lado, si la entidad compleja reguladora no consiste en entidades actuando de manera unitaria, sino coordinada, como en el caso de combinaciones de diferentes factores de transcripción que se unen a distintos sitios que regulan a un promotor, entonces es un regulón complejo.

Formalización de las definiciones en OWL

La jerarquía de tipos

Las ontologías serializadas en OWL están constituidas por un esqueleto de relaciones «es subclase de» (subClassOf en OWL, $\sqsubseteq$ en DL), también llamada jerarquía esUn (isA) o jerarquía de tipos. Una clase A es subclase de una clase B si todos los elementos de A también son elementos de B. Además de esta relación básica, podemos añadir cualesquier relaciones necesarias para vincular

las clases representadas, es decir, para especificar las condiciones necesarias y suficientes (o sólo necesarias) de pertenencia a la clase.

No existe una forma única de modelar un dominio. La manera más rápida de diseñar una ontología es pensando en una aplicación específica[22]. Sin embargo, queremos hacer una ontología de referencia, independiente de cualquier propósito específico, por lo que nuestro objetivo es desarrollar una ontología completa y correcta. Noy y Mcguiness proponen tres enfoques para el desarrollo de una ontología:

1) de arriba hacia abajo: el proceso de desarrollo empieza con la definición de los conceptos más generales del dominio y la subsecuente especialización de esos conceptos;
2) de abajo hacia arriba: el proceso de desarrollo empieza con la definición de las clases más específicas, las hojas de la jerarquía, seguida de la agrupación de estas clases en conceptos cada vez más generales y
3) un proceso de desarrollo combinando ambos enfoques en el que primero se definen los conceptos más importantes del dominio y subsecuentemente se generalizan y especializan[22].

Las definiciones derivadas en este trabajo son la base para el desarrollo de la ontología usando el tercer enfoque y siguiendo las principios OBO mencionados en el marco conceptual.

Uno de los principios OBO indica que cada clase de la ontología debe contar con definiciones aristotélicas o género-diferencia de la forma: A es un tipo de B (genero) que tiene la propiedad

distintiva C (diferencia). Definir de esta forma las clases ayuda a establecer la jerarquía. Por ejemplo, en el artículo de Nature Reviews Genetics definimos los siguientes conceptos:

Producto génico regulador: producto génico que aumenta o disminuye la expresión de un conjunto de genes específico (incluidos los complejos de productos génicos como los TFs heteromultiméricos como IFH de E. coli),

Proteína reguladora del inicio de la transcripción de unión a ADN: proteína reguladora que se une a secuencias específicas de ADN para regular el inicio de la transcripción.

Factor de transcripción: proteína reguladora del inicio de la transcripción de unión a ADN que se une cerca del promotor y afecta el inicio de la transcripción desde ese promotor, pero no es esencial para el inicio de la transcripción.

Factor σ: proteína reguladora del inicio de la transcripción de unión a ADN que es parte de la holoenzima de la ARN polimerasa (Eσ) y es esencial para la iniciación específica de la transcripción.

Proteína reguladora del inicio de la transcripción centrada en la Eσ^{23}: proteína reguladora de la transcripción que regula mediante su interacción con Eσ y cuya especificidad no está determinada directamente por el reconocimiento de secuencias de ADN.

De estas definiciones se puede derivar directamente la siguiente jerarquía de tipos, representada de tal manera que la clase más externa subsume a la clase sangrada.

Producto génico regulador
    Proteína reguladora
        Proteína reguladora del inicio de la transcripción de unión a ADN
            Factor de transcripción
            Factor σ
        Proteína reguladora del inicio de la transcripción centrada en Eσ

Podemos ver que esta jerarquía tiene huecos conceptuales evidentes. Podemos ver que las primeras dos subclases de la proteína reguladora especializan a la misma de acuerdo con dos criterios: la fase de regulación y el mecanismo o interacción molecular mediante el que regulan. Dado que queremos tener una ontología completa, tendremos que extender esta jerarquía de la siguiente manera (el ejemplo es ilustrativo, mas no exhaustivo).

Producto génico regulador
- Proteína reguladora
  - Proteína reguladora de la transcripción
    - Proteína reguladora del inicio de la transcripción
      - Proteína reguladora del inicio de la transcripción de unión a ADN
        - Factor de transcripción
        - Factor σ
      - Proteína reguladora del inicio de la transcripción centrada en Eσ
    - Proteína reguladora de la elongación transcripcional
    - Proteína reguladora de la terminación de la transcripción
  - Proteína reguladora de la traducción
- ARN regulador

Hay que tener en cuenta que una clase debe ser subclase del género especificado en su definición textual, aunque éste no constituye necesariamente la clase madre inmediata después de razonar la ontología[24]. Además, se recomienda que las definiciones lógicas concuerden en sentido amplio con las definiciones textuales [http://obofoundry.org/principles/fp-006-textual-definitions.html], por lo que se tendrían que añadir los símbolos de rol correspondientes (también llamados relaciones o propiedades de objeto en OWL) a la ontología. Por ejemplo, repasemos la definición de factor de transcripción:

Factor de transcripción: proteína reguladora del inicio de la transcripción de unión a ADN que se une cerca del promotor y afecta el inicio de la transcripción desde ese promotor y que no es esencial para el inicio específico de la transcripción.

Para definir lógicamente la clase factor de transcripción tendríamos que especificar en un lenguaje ontológico los símbolos de propiedad seUneA y regula, así como la clase promotor y la clase inicio de la transcripción y establecer las relaciones apropiadas. En este caso, probablemente será mejor asociar la propiedad de proximidad al promotor a la clase sitio de unión del factor de transcripción. También tenemos que definir las restricciones de cardinalidad de las relaciones. Lo más común es que las definiciones en lenguaje natural se traduzcan a restricciones existenciales (∃ en DL, someValuesFrom en OWL), i.e., existe al menos un miembro de la clase objeto con el que cada miembro de la clase sujeto se relaciona mediante una propiedad dada. La definición formal de TF se vería de la siguiente manera:

factor de transcripción ≡ [proteína reguladora del inicio de la transcripción de unión a ADN
⊓ ∃seUneA.sitio de regulación de factor de transcripción]

A su vez, tendríamos:

sitio de unión de factor de transcripción ≡ [región de ADN
⊓ [∃regula.promotor ⊔ ∃regula.unidad de expresión genética]

Hay que notar que representamos la parte de «afecta el inicio de la transcripción» establecida en la definición textual diciendo que, para que una proteína se considere un TF, es necesario que se conozca al menos un TFRS, que a su vez tendría una relación directa con el promotor o con la unidad de expresión regulada, asumiendo que esto implica que existe al menos un evento de transcripción en el que participa dicha proteína. Por otro lado al decir que el sitio de unión es una subclase de **región de ADN** satisface la condición de que se une a ADN de la definición textual.

Definir el ámbito de la ontología

Otro principio OBO especifica que la ontología debe tener un ámbito claramente especificado y debe adherirse a él. El ámbito de una ontología es la extensión del dominio o del contenido que intenta cubrir. Esto evita el traslape de ontologías y la duplicación de términos, facilita las búsquedas por parte de los usuarios de contenido específico y permite la selección rápida de ontologías de interés, al mismo tiempo de que permite la creación de términos nuevos a través de la combinación de los términos existentes.

Esto significa que antes de empezar una ontología de la regulación genética bacteriana, debemos estudiar las ontologías existentes para reusar los términos pertinentes. Seguiré desarrollando el ejemplo de factor de transcripción para ilustrar la manera que propone la OBO foundry para reusar e integrar las ontologías existentes. Las ontologías que actualmente contienen términos relacionados con el concepto de factor de transcripción son Gene Ontology (GO) y Protein Ontology (PR). Nosotros propusimos la definición de un concepto que incluye entidades moleculares (proteínas) que tienen determinada función, mientras que GO proporciona términos y definiciones para las funciones moleculares en sí y PR para las proteínas de cualquier organismo.

A continuación vemos segmentos de las jerarquías de GO y PR respectivamente relevantes para el ejemplo que estamos desarrollando.

Gene Onotlogy

molecular function regulator
- transcription regulator activity
  - RNA-binding transcription regulator activity
  - transcription antitermination factor activity, RNA binding
  - transcription antitermination factor activity, DNA binding
  - **DNA-binding transcription factor activity**
    - DNA-binding transcription factor activity, RNA polymerase II-specific
    - DNA-binding transcription activator activity
    - DNA-binding transcription repressor activity
    - ligand-activated transcription factor activity
  - transcription elongation regulator activity
  - sigma factor activity
  - anti-sigma factor antagonist activity
  - general transcription initiation factor activity

Protein ontology

- amino acid chain
  - protein
    - integration host factor subunit alpha
      - integration host factor subunit alpha (Aeromonas hydrophila subsp. Hydrophila)
      - integration host factor subunit alpha (Escherichia coli K-12)
      - integration host factor subunit alpha (Pseudomonas syringae pv. tomato…)
      - integration host factor subunit alpha (Xanthomonas campestris pv. campestris)
    - integration host factor subunit beta
      - integration host factor subunit beta (Agrobacterium fabrum str. C58)
      - integration host factor subunit beta (Escherichia coli K-12)

Para cumplir el principio de definir términos nuevos con base en los términos que ya existen en las ontologías OBO debemos encontrar la función molecular que sea equivalente a la definición que proponemos de TF. El concepto más parecido definido en GO es el siguiente:

DNA-binding transcription factor activity: Interacting selectively and non-covalently with a specific double-stranded genomic DNA sequence (sometimes referred to as a motif) within the regulatory region of a gene to modulate transcription. Regulatory regions include promoters (proximal and distal) and enhancers. Genes are transcriptional units, and include bacterial operons.

Suponiendo que la definición propuesta en GO efectivamente describe la función molecular que proponemos en nuestra definición, la definición formal de factor de transcripción en la ontología de la regulación sería:

Factor de transcripción ≡ [proteína ⊓ ∃tieneLaFunciónMolecular.DNA-binding transcription factor activity]

Donde proteína es una clase definida en PR. Al importar un módulo de GO (los términos que describen la actividad de regulación genética) y un módulo de PR (los términos que describen proteína bacterianas) y asociar a cada proteína de PR formalmente su respectivo término de GO con el símbolo de relación ∃tieneLaFunciónMolecular, podremos clasificar automáticamente los TFs, es decir inferir la relación subClassOf entre las proteínas pertinentes y la clase factor de transcripción.

Nuestra jerarquía de TFs podría incluso extenderse con criterios de clasificación por especie, dado que PR incluye esta información.

Creemos que las definición de TF de GO es equivalentes a la que proponemos. Aunque GO impone la restricción de que la unión de un TF al ADN debe ser no covalente y la nuestra no, nuestra definición especifica que los TFs no son necesarios para el inicio de la transcripción. Entonces, proteínas como FimA y FimB se excluyen tanto en la definición de GO como en la nuestra. Estas proteínas se unen de manera específica y covalente al ADN (no es TF según GO) y afectan el inicio de la transcripción al invertir una región de ADN arriba del promotor de los genes

involucrados en la formación de las fimbrias, determinando así si dichos genes se transcriben o no[25,26]. Este tipo de regulación es ON-OFF, es decir, este tipo de proteínas son necesarias para el inicio de la transcripción de las unidades de expresión que regulan, por lo tanto, no son TFs según nuestra definición. Por otro lado, en la jerarquía de GO faltan términos para describir la actividad molecular de, por ejemplo, los reguladores del inicio de la transcripción que se unen a Eσ en lugar de a ADN o la actividad de regulador de la traducción.

Entonces un paso previo a la definición de la jerarquía, sería proponer las ediciones e inclusiones correspondientes a GO o a cualquier ontología existente cuyo ámbito incluya entidades involucradas en la regulación. La razón de este principio OBO es que las ontologías existentes ya se están usando para anotar diferentes recursos (especialmente GO) y la duplicación de conceptos en diferentes ontologías locales es contrario al objetivo de unificación de la biología. Si una primera aplicación de una ontología de referencia de la regulación transcripcional de procariontes es la alineación de RegulonDB con el principio FAIR de interoperabilidad, es necesario usar los términos de GO y de otras ontologías existentes para definir nuestras clases.

Una vez que definamos las entidades en términos de las ontologías ya existentes (o simplemente importemos el término si ya está definido en una ontología existente), en la ontología de la regulación podemos añadir, además, las interacciones de regulación entre entidades reguladoras y reguladas como se muestra en el tercer artículo de la sección de resultados y clasificar, por ejemplo a las unidades de transcripción y regulones como sistemas inducibles o reprimibles. Esta adición de relaciones sería la contribución a la representación explícita y estándar de la biología de la regulación bacteriana.

Implementación de las definiciones para verificar la consistencia lógica del sistema de definiciones

Ontológicamente, en términos de definiciones formales, se habla de dos tipos de clases: primitivas y definidas. Las clases primitivas son clases que sólo tienen condiciones necesarias especificadas, mientras que las clases definidas tienen especificadas condiciones necesarias y suficientes. Los razonadores sólo pueden clasificar clases bajo las clases definidas[27]. Las subclases de las primitivas se afirman; las subclases de las clases definidas se infieren.

A continuación hicimos una implementación de las definiciones formales de todos los conceptos definidos en este trabajo. El objetivo era ver si existe consistencia lógica en el conjunto de definiciones propuestas. La implementación inicial parcial no reveló inconsistencias lógicas. Sin embargo, en una ontología de referencia de la regulación, probablemente no será necesario implementar formalmente cada definición propuesta, sino sólo aquellas cuyas subclases puedan ser calculadas automáticamente. En otras palabras, el razonamiento ontológico se usa para calcular una polijerarquía a partir de una jerarquía simple.

A continuación discutiré algunas preguntas y consideraciones que surgieron al tratar de definir formalmente todos los conceptos analizados en este trabajo.

Aunque la definición de unidad de transcripción propuesta (TU) no depende del contenido génico, tenemos que relacionar la clase TU con la clase gen porque la definición propuesta de operón es un conjunto de genes cuya expresión está coordinada al ser parte de una TU o de un conjunto de TUs que se traslapan.

El problema es que la condición de contener genes no es necesaria en la definición propuesta, ya que consideramos que los segmentos de ADN que corresponden a transcritos primarios sin contenido génico son TUs. Además, tampoco es necesario que los genes sean parte de una TU ya que se ha definido un tipo de genes llamados genes fantasma que son regiones de ADN que alguna vez se pensó que eran genes, por ejemplo, porque un programa computacional para encontrar genes lo identificó como tal, pero que actualmente se cree que no son genes funcionales, porque su expresión nunca se ha observado. Este tipo de genes se mantienen en la bases de datos Ecocyc y RegulonDB para mantener registro de su estado pasado comparado con su estado actual. Por lo tanto, la relación entre gen y TU no debe establecerse a nivel general, sino a nivel de genes y TU específicos.

La definición de sitio regulador de un factor de transcripción propuesta es un sitio de unión de factor de transcripción que efectivamente participa en algún evento de regulación de la transcripción. En la definición formal, la actividad reguladora del sitio se representó con la unión de dos alternativas: la regulación de la actividad de un promotor o la regulación de la expresión de una unidad de expresión. La caracterización completa de un sistema de regulación transcripcional requiere saber el promotor involucrado, pero algunos experimentos reportan solamente la relación entre un TF y su sitio con la expresión de una TU o de un gen (ver más abajo).

La definición propuesta de efector es un continuo que produce un cambio en una molécula y modifica su actividad y/o especificidad. Los continuos, según se definen en *Basic Formal Ontolog*y (BFO), son entidades que continúan o persisten a través del tiempo, incluidos (1) objetos independientes (por ejemplo, cosas como tú y yo); (2) continuos dependientes, incluidas

cualidades (como la temperatura o mi estatura) y funciones (como la función de un interruptor de aprender y apagar la luz); y (3) las regiones espaciales que ocupan estas entidades en cualquier momento. Tal vez sea más fácil entender la clase de los continuos como el complemento de la clase de los ocurrentes, también llamados eventos o procesos, que ocurren o se desarrollan a lo largo del tiempo[15]. Es decir los límites de una entidad tipo ocurrente están definidos temporalmente, mientras que los de una entidad tipo continuo no.

En el caso del concepto de efector surge el dilema de si es necesario representar el tipo de cambio molecular del efector sobre una molécula o si simplemente afirmar que cierta entidad es un efector. Creo que por simplicidad, dentro del ámbito de una ontología de la regulación genética, debemos optar por la segunda opción.

El concepto de efector tiene algo de representación en una ontología existente llamada *Chemical Entities of Biological Interest* (CHEBI)[28]. Esta ontología incluye una clase llamada «efector» definida como una molécula pequeña que aumenta (activador) o disminuye (inhibidor) la actividad de una enzima (alostérica) al unirse al sitio de regulación de la enzima (que es diferente del sitio catalítico donde se une el sustrato). En dicha ontología simplemente se relacionan algunas moléculas (que son efectores de enzimas) mediante una relación has_role con el término efector, por ejemplo, acetil-CoA has_rol effector.

Sin embargo, nosotros proponemos una definición de efector más general en el sentido de que la molécula cuya actividad es modulada por el efector puede ser un regulador genético y proponemos

un concepto especializado para referirnos a esta clase de efectores: **efector del regulador de la expresión génica**.

Creo que es conveniente seguir la lógica de CHEBI para representar a los efectores, pero de una manera más precisa: crear una clase **efector,** definida como un rol, y una clase **entidad efectora,** definida como las entidades que tienen el rol de efector. Las jerarquías de estos términos se verían algo así:

Continuo
- Rol
  - efector
    - Efector de regulador de la expresión génica
      - Efector de regulador de la expresión génica de E coli
- Entidad efectora
  - Entidad efectora de regulador de la expresión génica
    - Entidad efectora de regulador de la expresión génica de E coli

De esta manera podemos importar un módulo de CHEBI que incluya las moléculas que se sabe que son efectores de reguladores de la expresión génica de E. coli curadas en RegulonDB, asociarlas con el rol e inferir las subclases de entidad efectora. Así tendremos un esqueleto de primitivas que incluye dichas moléculas clasificadas por criterios químicos y podremos inferir su clasificación por criterios funcionales.

Pero los efectores no sólo son moléculas, la temperatura y la luz también son tipos de continuos que pueden tener el rol de efector, por lo que habría que buscar dichas entidades en otras ontologías, incluirlas en nuestra ontología y relacionarlas con el rol de efector.

El concepto de regulón es un buen ejemplo de la utilidad de las clasificación automática y sirve como ilustración de un problema común en el modelado ontológico llamado sobrecarga de esUn. esUn es otro símbolo de relación usado para referirse a subClassOf o ⊑ o «tipo de». En breve, el problema de sobrecarga de esUn se refiere a relacionar dos clases mediante la relación subClassOf cuando la clase hija realmente no es un tipo de la clase madre[29].

En el caso de la clase regulón, la definición es simple: conjunto de unidades de expresión reguladas por un producto génico común. Sin embargo, si especificamos la siguiente definición en una ontología:

regulón de LacI ≡ unidad de expresión genética ⊓ ∃tieneExrpesiónReguladaPor.LacI

Y suponiendo que tenemos la jerarquía y las relaciones de regulación necesarias, es decir:

unidad de expresión genética
    unidad de transcripción
        lacZYA
factor de transcripción
    LacI

Donde:

LacZYA ⊑ “unidad de transcripción” ⊓ ∃tieneExpresiónReguladaPor.LacI

se inferirá la siguiente jerarquía de tipos:

regulón de LacI
    lacYZA

Sin embargo, lacZYA es una unidad de transcripción que por definición no es un regulón. Esto no quiere decir que la definición de regulón no admita regulones que contengan una única unidad de transcripción o un único gen. Los regulones están definidos como conjuntos de unidades de expresión y por el axioma de fundación de la teoría de conjuntos de Zermelo-Fraenkel, un conjunto que contiene un solo elemento no es igual a su elemento. Si denotamos a un conjunto listando sus elementos entre corchetes, eso quiere decir que lacZYA ≠ {lacZYA}. Extendamos el ejemplo anterior, para ilustrar de manera práctica este caso. Si incluimos al TF AlsR junto con las TUs que regula, la jerarquía se vería así:

unidad de expresión genética
    unidad de transcripción
        lacZYA
        alsRBACE_1
        alsRBACE_2
        rpiB

factor de transcripción
    LacI
    AlsR

Incluyendo las siguientes definiciones además de las previamente especificadas:

regulón de AlsR ≡ unidad de expresión genética ⊓ ∃tieneExrpesiónReguladaPor.AlsR

alsRBACE_1 ⊑ “unidad de transcripción” ⊓ ∃tieneExrpesiónReguladaPor.AlsR

alsRBACE_2 ⊑ “unidad de transcripción” ⊓ ∃tieneExrpesiónReguladaPor.AlsR

rpiB ⊑ “unidad de transcripción” ⊓ ∃tieneExrpesiónReguladaPor.AlsR

Inferiríamos la siguiente jerarquía

regulón de LacI
    lacYZA
regulón de AlsR
    alsRBACE_1
    alsRBACE_2
    rpiB

Esta jerarquía no sólo significa que lacZYA es un regulon de LacI, sino también que cada TU individual regulada por AlsR es un tipo de regulon de AlsR: alsRBACE_1 es un regulon de AlsR, alsRBACE_2 es un regulon de AlrsR, rpiB es un regulon de AlsR.

En realidad las unidades de transcripción son un miembros del regulón. Entonces, para calcular automáticamente los miembros de un regulón a partir de las relaciones de regulación lo adecuado es definir la siguiente clases:

miembro de regulón de LacI ≡ unidad de expresión genética ∃tieneExrpesiónReguladaPor.LacI

miembro de regulón de AlsR ≡ unidad de expresión genética ∃tieneExrpesiónReguladaPor.AlsR

Para inferir correctamente la siguiente jerarquía de tipos:

miembro del regulón de LacI
    lacYZA
miembro del regulón de AlsR
    alsRBACE_1
    alsRBACE_2
    rpiB

Este caso ilustra la utilidad del razonamiento automático en el mantenimiento de la ontología: basta afirmar las interacciones de regulación y el conocimiento adicional que se deriva de dichas relaciones se calcula automáticamente. Si se descubriera que LacI regula otra unidad de

transcripción, solo hay que establecer la relación de regulación entre LacI y la nueva unidad de transcripción, y los miembros del regulón se calculan automáticamente. Esto es útil en la consulta y visualización de los regulones a través de la ontología.

Necesidad de una nomenclatura

Dijimos previamente que una ontología de referencia debe ser completa y correcta. Para la completitud de la ontología será necesario establecer una nomenclatura para las clases que representen entidades reguladoras y reguladas específicas, con nombres que sean fácilmente interpretados por humanos. Algunas de estas entidades ya están representadas en otras ontologías. Por ejemplo, *Ontology of Genes and Genomes*[30] (OGG) ya incluye los ~4,500 genes de E. coli y *Protein Ontology*[31] (PR), las proteínas. Por otro lado, RegulonDB, ya ha adoptado ciertas convenciones, por ejemplo, para la nomenclatura de promotores, pero para otras no. RegulonDB representa los sitios reguladores de los TF solamente con el identificador alfanumérico único de la base de datos, mientras que la nomenclatura de operones es ambigua con respecto a la usada para unidades de transcripción.

¿Qué beneficios hay en compartir conceptualizaciones?

Las ontologías se usan para mejorar la interoperabilidad entre diferentes bases de datos. Proponer términos preferidos evita el uso de diferentes sinónimos para anotar entidades en diferentes bases de datos, lo que facilita el tratamiento automático de información proveniente de diferentes fuentes.

Que los términos preferidos vayan acompañados con definiciones precisas es de suma importancia para reducir la variabilidad entre anotadores y para el uso adecuado de los datos. Un ejemplo de

confusión de uso de datos se encuentra en la referencia [32]. En los métodos de dicho artículo se lee: para este estudio unificamos las interacciones de diferentes promotores del mismo operón. En este caso, el usuario de la base de datos usó el concepto de operón definido como un conjunto de genes cotranscritos y coregulados, mientras que la base de datos contiene operones con promotores internos diferencialmente regulados. Como mostramos en el artículo de redefinición de conceptos, las interacciones de promotores de un mismo operón no pueden unificarse, dado que cada promotor puede regular diferencialmente distintos subconjuntos de genes del operón. El usuario de los datos tenía una conceptualización distinta a la representada en la base de datos y esto puede haber tenido un efecto en las conclusiones que derivó usando los datos bajo suposiciones erróneas.

Cambios en RegulonDB y EcoCyc motivados por las nuevas definiciones

Se está trabajando en RegulonDB y EcoCyc para que cumpla con las definiciones propuestas en este trabajo. A continuación presento, por concepto, los cambios necesarios e indico los casos que ya se implementaron.

Promotor

Se implementaron cambios para hacer consistente la relación uno a uno de promotor con factor sigma en la versión actual de EcoCyc. Previo a este trabajo, esta relación se introducía en la base de datos según lo reportado en el artículo del que se extraía dicha relación, por lo que existían promotores a los que se asignaba más de un factor sigma y secuencias duplicadas que se registraban como distintos promotores dado que son reconocidas por más de un factor sigma.

Factor de transcripción

En RegulonDB actualmente existen sólo dos tipos de reguladores TFs y smallRNAs. El cambio necesario para que refleje lo propuesto en el artículo sería modificar el esquema para que admita

todo tipo de reguladores y no sólo TFs y sRNAs. Una opción es crear una tabla para todos los reguladores con un atributo llamada tipo de regulador y usar los términos de la jerarquía de regulador. Ontológicamente, podemos crear una superclase de «producto génico regulador» llamada «regulador» que incluya también moléculas pequeñas.

Sitio de unión del factor de transcripción

Se está adaptando la base de datos para incluir información derivada de experimentos de alto rendimiento. Para incluir datos derivados de Chip-Seq, se está haciendo la distinción entre sitios de unión y sitios de regulación de los factores de transcripción.

Frases o módulos de TFRS

Será necesario hacer la distinción entre frases homotípicas y heterotípicas.

Unidad de transcripción

Actualmente, una TU puede tener asociado más de un terminador. Se debe adaptar de tal manera que una TU sólo tenga asociados un promotor y un terminador; la relación inversa puede ser múltiple, es decir un promotor puede estar relacionado con más de una TU y lo mismo para un terminador.

Operon

La definición propuesta de operón describe las entidades que ya están representadas en RegulonDB. Sólo habría que hacer la distinción entre operones simples y complejos, y entre operones complejos hacer la distinción entre corregulados y los no corregulados.

Regulon

Actualmente en RegulonDB, los regulones simples se definen como los genes regulados por uno y solo un TF, lo que significaría que si un gen esta regulado por más de un TF, no sería parte de un regulón simple. La nueva propuesta es que el regulón simple son todos los genes regulados por un TF sin importar si están regulados por más de uno.

Por otro lado, RegulonDB define regulones de ppGpp y en el artículo de redefinición de conceptos decimos que los regulones están definidos por productos génicos reguladores. En la ontología podríamos generalizar la definición propuesta en el artículo, diciendo que los regulones pueden estar definidos por cualquier tipo de regulador *trans*. Quedaría pendiente definir qué clase de moléculas pequeñas definen un regulón o las condiciones en que una molécula pequeña define un regulón.

Finalmente, en RegulonDB, actualmente los regulones se definen en términos de genes o de operones. Sin embargo los regulones no deben definirse en términos de operones porque los operones son conjuntos de TUs que podrían estar reguladas diferencialmente. Por esta razón, en el artículos los definimos en términos de unidades de expresión, i. e., unidades de transcripción y genes.

Efector

Actualmente sólo existen efectores que son moléculas en RegulonDB y la definición propuesta incluye todo tipo de continuos, como concentraciones relativas de metabolitos, luz y temperatura. Esta extensión de la definición no parece implicar un cambio en el esquema, sólo se tendrían que usar los términos como atributos para describir el tipo de efector.

Por otro lado, RegulonDB actualmente considera que los grupos funcionales (fostato, metilo) con que se modifican covalentemente algunos TFs son efectores. Sin embargo, nosotros proponemos que la modificación covalente es un tipo de cambio ocasionado por un efector. En estos casos el efector sería la molécula que ocasiona la fosforilación o la metilación. Por ejemplo en el caso de los sistemas de dos componentes, la quinasa sería el efector del regulador de respuesta, mientras que la fosforilación resultante es el cambio químico que aumenta o disminuye la actividad reguladora del TF. Este cambio implicaría la curación de los efectores que actualmente tienen el atributo de tipo de interacción covalente.

Señal

Actualmente RegulonDB no tiene señales compiladas, por lo que se tendrían que curar *de novo*. Para la curación será necesario definir de manera más precisa la definición propuesta en este trabajo de lo que es una señal interna. En este trabajo definimos señal como el primer paso en un flujo de información que ocasiona un cambio en la expresión genética, pero no precisamos lo que es un primer paso en un flujo de información dentro de la célula. Si lo pensamos como una molécula producto del metabolismo, es fácil imaginar que la red de interacciones metabólicas está muy conectada, lo cual dificulta la detección de los primeros pasos en un flujo de información.

Caracterización incompleta de las entidades

En este trabajo se especifican las propiedades necesarias y suficientes para definir las entidades fundamentales involucradas en el inicio de la transcripción genética. Sin embargo, muchas veces conocer cada una de estas propiedades requiere más de un experimento. En otras palabras, muchas veces un experimento identifica una sola propiedad y, como consecuencia, existen piezas «sueltas»

de conocimiento. La red de regulación completa se reconstruye integrando varias fuentes de información.

A continuación se analiza cómo las definiciones derivadas de este trabajo se pueden desglosar en términos de las piezas de información que hay actualmente en RegulonDB y de las características que pueden faltar y que pueden quedar desconectadas dependiendo del tipo de evidencia. Para los conceptos de promotor, sitio de regulación del factor de transcripción y unidad de transcripción se presentan tablas que representan diferentes combinaciones de presencia y ausencia de las piezas de información: el número 1 represente presencia y el 0 ausencia. De todas las combinaciones posibles de dichas piezas de información, sólo se incluyen las que existen en RegulonDB. Y para cada combinación de piezas de información se presenta una lista de evidencias que respaldan la inclusión de la entidad con ese grado de caracterización en RegulonDB. Es decir, las evidencias listadas están asociadas al promotor, al sitio o la TU, excepto que se indique otra cosa.

Es importante notar que RegulonDB incluye una entidad llamada regulatory interaction (RI), ya que se mencionará en el análisis de los conceptos. Esta entidad es la combinación un TFBS, un TF en su conformación activa y el promotor regulado. También tiene una entidad llamada TF-gene interaction que es la combinación de un TF un TFBS y un gene regulado.

Promotor

Es un sitio de reconocimiento de la polimerasa específico para un factor sigma desde el que se inicia la transcripción. El hecho de que es un sitio de reconocimiento de la polimerasa podría desglosarse en dos datos recopilados por RegulonDB: una holoenzima con un factor sigma específico que se une en una posición del genoma y/o que regula la expresión de una entidad; y

por los motivos de reconocimiento (aunque en el artículo discutimos que los motivos no son condición necesaria ni suficiente, es información que recopila RegulonDB). Por otro lado, el hecho de que una secuencia inicia la transcripción puede desglosarse en dos piezas de información recopiladas en la base de datos: los sitios de inicio de la transcripción y la existencia de una unidad de transcripción asociada al promotor. En la tabla 1 están las combinaciones de piezas de información de promotores parcialmente caracterizados que contiene actualmente la base de datos junto con las evidencias que detectan promotores con esos grados de caracterización.

TSS	σ	motivos	TU	Evidencia
1	0	0	0	RNA-seq using two enrichment strategies for primary transcripts and consistent biological replicates
1	1	1	0	• *Automated inference of promoter position* • *High-throughput transcription initiation mapping* • *Human inference of promoter position* • *Inferred computationally without human oversight* • *Inferred from expression pattern* • *RNA-seq using two enrichment strategies for primary transcripts and consistent biological replicates* • *Transcription initiation mapping*
1	0	1	1	• *Traceable author statement* • *High-throughput transcription initiation mapping* • *Inferred from mutant phenotype* • *Human inference of promoter position* • *Transcription initiation mapping* • *RNA-seq using two enrichment strategies for primary transcripts and consistent biological replicates* • *Mapping of signal intensities*
0	1	0	1	• *Non-traceable author statement* • *Traceable author statement to experimental support* • *Traceable author statement* • *RNA polymerase footprinting* • *Inferred from direct assay* • *Human inference of promoter position* • *Inferred from mutant phenotype* • *Transcription initiation mapping* • *Automated inference of promoter position*

				• *Inferred from expression pattern*
0	1	0	0	• *ChIP-seq*

Tabla 1 Promotores parcialmente caracterizados. Cada renglón especifica una posible combinación de piezas que caracteriza a algunos promotores contenidos en RegulonDB.

Como se puede ver en el primer renglón de la tabla, existen metodologías que encuentran sitios de inicio de la transcripción sin experimentos de unión de la polimerasa. Estos experimentos se basan en encontrar y enriquecer transcritos primarios y usar la retrotranscriptasa para determinar el extremo 5'. Actualmente RegulonDB considera estos sitios como promotores. Sin embargo, como discutimos en el artículo de redefinición de conceptos, se ha observado *in vitro* que el núcleo de la polimerasa puede generar transcritos. Se sabe que los transcritos generados por el núcleo de la polimerasa también tienen el extremo 5' trifosfatado[33], pero también se ha calculado que la concentración del factor σ^{70} es mayor a la del núcleo (core) de la polimerasa y que el núcleo de la polimerasa se une más firmemente al factor σ que a segmentos aleatorios del ADN por lo que los núcleos de la polimerasa libre se unen preferencialmente al factor σ^{70} REF [34]. Entonces, ¿es posible ignorar la transcripción no específica de la polimerasa núcleo *in vivo*? De no ser así, no todos los sitios de inicio de la transcripción son parte de un promotor y deberían representarse como parte del transcrito o como entidades independientes.

Por otro lado, experimentos como Chip-Seq pueden encontrar sitios de unión de la holoenzima específicos para un factor sigma, pero, sin una evidencia de inicio de la transcripción, estos sitios sólo pueden considerarse sitios de unión de la polimerasa y no promotores.

Además, experimentos de fusión con genes reporteros pueden demostrar que determinada secuencia tiene las dos propiedades necesarias y suficientes de un promotor sin identificar el sitio real del inicio de la transcripción. En este caso, tendríamos un promotor parcialmente caracterizado

al que probablemente tendríamos problemas para nombrar, si no sabemos el nombre su gen más cercano. Otra forma de encontrar promotores sin identificar el sitio de inicio de la transcripción es por medio de mutaciones que alteren la expresión de las unidades de expresión, por lo que podríamos tener una relación directa entre el promotor y la TU o el gen que transcribe, sin saber el sitio de inicio de la transcripción.

Sitio regulador del factor de transcripción

La definición propuesta en este trabajo para TFRS es sitio reconocido por un TF que tiene un efecto en la expresión de una TU. Los datos de RegulonDB desglosarían esta definición de la siguiente manera. Para que sea un sitio, debe haber una posición; el efecto en la expresión puede estar evidenciado por el conocimiento de su entidad regulada y el efecto que tiene sobre la entidad regulada. La entidad regulada puede ser un promotor, un gen o una TU.

Posición/ secuencia	TF	Efecto	Promotor regulado	Combinaciones de evidencias para diferentes grados de caracterización
1	1	0	1	• *[A person inferred or reviewed a computer inference of sequence function based on similarity to a consensus sequence, Automated inference based on similarity to consensus sequences, Gene expression analysis]* • *[A person inferred or reviewed a computer inference of sequence function based on similarity to a consensus sequence, Binding of purified proteins]* • *[A person inferred or reviewed a computer inference of sequence function based on similarity to a consensus sequence, Binding of purified proteins, Gene expression analysis]* • *[A person inferred or reviewed a computer inference of sequence function based on similarity to a consensus sequence, Binding of purified proteins, Gene expression analysis, Site mutation]* • *[Automated inference based on similarity to consensus sequences']* • *[Binding of purified proteins, Gene expression analysis]* • ***(La evidencia es de la RI y no hay evidencia para estos sitios)***
1	0	0	0	• *Computational prediction*

0	1	1	1	• *[A person inferred or reviewed a computer inference of sequence function based on similarity to a consensus sequence, Binding of purified proteins],* • *[A person inferred or reviewed a computer inference of sequence function based on similarity to a consensus sequence, Binding of purified proteins, Gene expression analysis'],* • *[A person inferred or reviewed a computer inference of sequence function based on similarity to a consensus sequence, Gene expression analysis'],* • *[Automated inference based on similarity to consensus sequences, Binding of purified proteins, Gene expression analysis],* • *[Automated inference based on similarity to consensus sequences, Gene expression analysis],* • *[Automated inference based on similarity to consensus sequences, Gene expression analysis, Inferred from mutant phenotype],* • *[Binding of cellular extracts, Binding of purified proteins, Gene expression analysis],* • *[Binding of cellular extracts, Gene expression analysis],* • *[Binding of cellular extracts, Non-traceable author statement],* • *[Binding of purified proteins, Gene expression analysis],* • *[Binding of purified proteins, Gene expression analysis, Inferred from genetic interaction],* • *[Binding of purified proteins, Gene expression analysis,* • *cross validation (GEA/GS)',* • *genomic SELEX],* • *[Binding of purified proteins, Inferred from direct assay],* • *[Gene expression analysis, Inferred computationally without human oversight],* • *[Gene expression analysis, Inferred from expression pattern],* • *[Gene expression analysis, Inferred from mutant phenotype],* • *[Gene expression analysis, Site mutation],* • *[Gene expression analysis, Traceable author statement],* • *[Gene expression analysis, cross validation(GEA/GS), genomic SELEX]* **(Las evidencias son de la RI y no hay evidencia para estos sitios)**
0	1	0	1	• *[Gene expression analysis, Binding of purified proteins]* **(La evidencia es de la RI y no hay evidencia para estos sitios)**
Posición/ secuencia	TF	efecto	Gen regulado	Combinaciones de evidencias para diferentes grados de caracterización
0	1	1	1	• *[A person inferred or reviewed a computer inference of sequence function based on similarity to a consensus sequence, Gene expression analysis]* • *[Automated inference based on similarity to consensus sequences, Gene expression analysis]* • *[Binding of purified proteins]* • *[Binding of purified proteins, Gene expression analysis]*

				• *[Binding of purified proteins, Gene expression analysis, Inferred from Biological aspect from Ancestor]* • *[Binding of purified proteins, Gene expression analysis, Inferred from genetic interaction]* • *[Binding of purified proteins, Gene expression analysis, genomic SELEX]* • *[Gene expression analysis]*
0	1	0	1	• *Automated inference based on similarity to consensus sequences*

Tabla 2. Sitios de regulación de factor de transcripción parcialmente caracterizados. Cada renglón especifica una posible combinación de piezas que caracteriza a algunos TFRS contenidos en RegulonDB junto con la combinación de evidencias que respaldan cada grado de caracterización.

Factor de transcripción

La definición propuesta de factor de transcripción es una proteína que se une a ADN para regular promotores específicos. La especificidad está dada por el reconocimiento de secuencias en el ADN. Las piezas de información incluidas en RegulonDB que cuentan para esta definición son: la participación de la proteína en una interacción de regulación y las coordenadas de unión al ADN. Todos los TFs incluidos en RegulonDB tienen tanto una secuencia de unión como una interacción de regulación. Sin embargo, como se mencionó previamente, las interacciones de regulación pueden estar parcialmente caracterizadas. Cuando no se conoce el promotor, se relaciona al TF y su sitio directamente con un gen o una TU.

Unidad de transcripción

La definición propuesta de unidad de transcripción es una región de ADN delimitada por pares TSS-TTS diferentes no espurios; su transcripción inicia en un solo promotor y termina en un solo terminador. Actualmente, RegulonDB no especifica sitios de terminación de la transcripción explícitamente.

Promotor	TSS	Terminador	Evidencia
1	1	1	Automated inference that a single-gene directon is a transcription unit Boundaries of transcription experimentally identified Cross validation(LTED/PM) Inferred by a human based on computational evidence Inferred by curator Inferred computationally without human oversight Inferred from expression pattern Inferred through co-regulation Length of transcript experimentally determined Polar mutation Products of adjacent genes in the same biological process Traceable author statement
1	1	0	Automated inference that a single-gene directon is a transcription unit Boundaries of transcription experimentally identified Cross validation(LTED/PM) Inferred by a human based on computational evidence Inferred by curator Inferred computationally without human oversight Inferred from Biological aspect from Ancestor Inferred from direct assay Inferred from expression pattern Inferred from mutant phenotype Inferred through co-regulation Length of transcript experimentally determined Non-traceable author statement Polar mutation Products of adjacent genes in the same biological process Traceable author statement Traceable author statement to experimental support
1	0	1	Products of adjacent genes in the same biological process Inferred by a human based on computational evidence Boundaries of transcription experimentally identified
1	0	0	Automated inference that a single-gene directon is a transcription unit Boundaries of transcription experimentally identified Inferred by a human based on computational evidence Inferred by curator Inferred computationally without human oversight Inferred through co-regulation Length of transcript experimentally determined

			Polar mutation Products of adjacent genes in the same biological process
0	1	1	Automated inference that a single-gene directon is a transcription unit Boundaries of transcription experimentally identified Inferred by a human based on computational evidence Inferred by curator Inferred computationally without human oversight Inferred from expression pattern Inferred through co-regulation Length of transcript experimentally determined Polar mutation Products of adjacent genes in the same biological process
0	0	1	Automated inference that a single-gene directon is a transcription unit Boundaries of transcription experimentally identified Inferred by curator Inferred computationally without human oversight Inferred through co-regulation Length of transcript experimentally determined
0	0	0	Author hypothesis Automated inference that a single-gene directon is a transcription unit Boundaries of transcription experimentally identified Ccross validation(LTED/PM) Inferred by a human based on computational evidence Inferred by curator Inferred computationally without human oversight Inferred from expression pattern Inferred from mutant phenotype Inferred through co-regulation Length of transcript experimentally determined No biological data available Non-traceable author statement Polar mutation Products of adjacent genes in the same biological process

Tabla 4. Diferentes grados de caracterización de las unidades de transcripción incluidas en RegulonDB. LTED son las siglas de Length of Transcript Experimentally Determined y PM son las siglas de Polar Mutation.

Operon, regulon, efector y señal

La definición de operon propuesta es un conjunto de genes cuya expresión está coordinada por una o varias TUs traslapadas en la misma dirección. Este concepto requiere muy poco análisis

independiente, dado que depende del concepto de TU que se analizó previamente. Para satisfacer la definición, basta con tener conjuntos de TUs que contengan genes. Actualmente, RegulonDB sólo representa unidades de transcripción que sí contienen genes.

De manera similar, el concepto de regulón depende de entidades analizadas previamente: conjunto de unidades de expresión genética reguladas directamente por un conjunto común de uno o más productos génicos reguladores. Actualmente RegulónDB sólo representa información sobre el inicio de la transcripción, por lo que los regulones estarían compuestos por unidades de transcripción reguladas por factores de transcripción comunes y por ppGpp. Entonces el regulón hereda las carencias de información de las entidades que lo definen.

Finalmente, todos los efectores que contiene RegulonDB actualmente están completamente caracterizados, ya que simplemente son moléculas que se sabe que modifican la conformación de los factores de transcripción.

¿Cómo tratar estas piezas de información ontológicamente?

El razonamiento en OWL opera bajo el supuesto del mundo abierto. Esto significa que no podemos suponer que algo no existe hasta que se establece explícitamente que no existe. En otras palabras, que no se haya indicado que algo es verdadero no implica que sea falso, sino que se supone que el conocimiento no se ha añadido a la base de conocimiento[3]. El supuesto del mundo abierto hace que OWL sea adecuado para manejar conocimiento incompleto, como es el caso del conocimiento biológico y el conocimiento científico en general.

Sin embargo, no podemos ignorar el hecho de que el supuesto del mundo abierto contradice las expectativas de muchos usuarios de los sistemas de información y debemos ser pragmáticos y tener cuidado de no confundir a los usuarios. La suposición del mundo abierto es dañina cuando se desvía gravemente de las expectativas de uso. Debe haber una suposición de cobertura razonable. Hay que incluir toda la información que un usuario razonable espera encontrar dado el ámbito de la ontología[35].

Mientras que el supuesto del mundo abierto se asume en cualquier información curada, también ayuda pensar en términos de un contrato implícito predominante entre el proveedor de información y el consumidor de información: proporcionar información lo más completa posible. De lo contrario, habría que documentar claramente los sesgos y suposiciones hechos al momento de recopilar y representar la información en la ontología[35].

Valdría la pena explorar la posibilidad de documentar ontológicamente estos huecos de información clasificando las entidades de acuerdo al grado de caracterización, con la ventaja de que esta clasificación se haría automáticamente.

CONCLUSIÓN

En este trabajo, se aplicó un enfoque de falsificación para la revisión de definiciones de entidades biológicas en cuanto a la necesidad y suficiencia de las características usadas en dichas definiciones. El enfoque fue aplicable en algunos casos, pero en otros se tuvo que recurrir a la estandarización del término por consenso de un grupo de expertos usando ya sea criterios biológicos o de generalidad del uso del término en cuestión. También se discutieron los pasos

subsiguientes para la implementación de las definiciones en una ontología y se mencionan algunos de los efectos que la redefinición de los conceptos básicos de la regulación transcripcional tendrá en el esquema de RegulonDB y de EcoCyc.

REFERENCIAS

1. Hofweber, T. Logic and Ontology. in *The Stanford Encyclopedia of Philosophy* (ed. Zalta, E. N.) (Metaphysics Research Lab, Stanford University, 2020).
2. Suppes, P. *Introduction to logic*. (Courier Corporation, 1999).
3. Horridge, M., Knublauch, H., Rector, A., Stevens, R. & Wroe, C. *A practical guide to building OWL ontologies using the Protégé-OWL plugin and CO-ODE tools edition 1.0*. *University of Manchester* (2004).
4. Guarino, N., Oberle, D. & Staab, S. What is an ontology? in *Handbook on ontologies* 1–17 (Springer, 2009).
5. Baader, F., Horrocks, I. & Sattler, U. Description logics. in *Handbook on ontologies* 3–28 (Springer, 2004).
6. Schulz, S. & Jansen, L. Formal ontologies in biomedical knowledge representation. *Yearb. Med. Inform.* **8**, 132–46 (2013).
7. Popper, K. *The logic of scientific discovery*. (Routledge, 2005).
8. Smith, B. *et al.* The OBO Foundry: Coordinated evolution of ontologies to support biomedical data integration. *Nat. Biotechnol.* **25**, 1251–1255 (2007).
9. Burgun, A. Desiderata for domain reference ontologies in biomedicine. *J. Biomed. Inform.* **39**, 307–313 (2006).
10. Brinkley, J. F., Suciu, D., Detwiler, L. T., Gennari, J. H. & Rosse, C. A framework for using reference ontologies as a foundation for the semantic web. *AMIA Annu. Symp. Proc.* 96–100 (2006).

11. Golbreich, C., Grosjean, J. & Darmoni, S. J. The Foundational Model of Anatomy in OWL 2 and its use. *Artif. Intell. Med.* **57**, 119–132 (2013).

12. Gkoutos, G. V. *et al.* Entity/quality-based logical definitions for the human skeletal phenome using PATO. *Proc. 31st Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. Eng. Futur. Biomed. EMBC 2009* 7069–7072 (2009). doi:10.1109/IEMBS.2009.5333362

13. Amann, R. & Rosselló-móra, R. After All, Only Millions? **7**, 1–2 (2016).

14. Mejía-Almonte, C. & Collado-Vides, J. A Falsification approach to create and check ontology definitions. in *Proceedings of the 9th International Conference on Biological Ontology (ICBO 2018)* (eds. Jaiswal, P., Cooper, L., Haendel, M. A. & Mungall, C. J.) **2285**, 10–11 (CEUR Workshop Proceedings, 2018).

15. Arp, R., Smith, B. & Spear, A. D. *Building Ontologies with Basic Formal Ontology*. *Building Ontologies with Basic Formal Ontology* (MIT Press, 2015). doi:10.7551/mitpress/9780262527811.001.0001

16. Mejía-Almonte, C. *et al.* Redefining fundamental concepts of transcription initiation in bacteria. *Nat. Rev. Genet.* **21**, 699–714 (2020).

17. Ashburner, M. *et al.* Gene ontology: Tool for the unification of biology. *Nat. Genet.* **25**, 25–29 (2000).

18. Vogt, V. Breaks in DNA stimulate Transcription by Core RNA Polymerase. *Nature* **223**, 854–855 (1969).

19. Dausse, J. -P, Sentenac, A. & Fromageot, P. Interaction of RNA Polymerase from Escherichia coli with DNA: Influence of DNA Scissions on RNA-Polymerase Binding and Chain Initiation. *Eur. J. Biochem.* **31**, 394–404 (1972).

20. Raghavan, R., Sloan, D. B. & Ochman, H. Pervasive transcription is widespread but rarely

conserved in Enteric bacteria. *MBio* **3**, 1–7 (2012).

21. Maas, W. K. & Clark, A. J. Studies on the mechanism of repression of arginine biosynthesis in Escherichia coli: II. Dominance of repressibility in diploids. *J. Mol. Biol.* **8**, 365–370 (1964).

22. Noy, N. F. & McGuinness, D. L. *Ontology development 101: A guide to creating your first ontology*. *Stanford Knowledge Systems Laboratory Technical Report KSL-01- 05 and Stanford Medical Informatics Technical Report SMI-2001-0880* **32**, (2001).

23. Browning, D. F. & Busby, S. J. W. Local and global regulation of transcription initiation in bacteria. *Nat. Rev. Microbiol.* (2016). doi:10.1038/nrmicro.2016.103

24. Mungall, C. OntoTip: Write simple, concise, clear, operational textual definitions. *https://douroucouli.wordpress.com/2019/07/08/ontotip-write-simple-concise-clear-operational-textual-definitions/* (2019).

25. Abraham, J. M., Freitag, C. S., Clements, J. R. & Eisenstein, B. I. *An invertible element of DNA controls phase variation of type 1 fimbriae of Escherichia coli (genomic rearrangement/transcription/operon fusion)*. *Proc. Nati. Acad. Sci. USA* **82**, (1985).

26. Hallet, B. Playing Dr Jekyll and Mr Hyde: combined mechanisms of phase variation in bacteria. *Curr. Opin. Microbiol. 2001,* **4**, 570–581 (2001).

27. Rector, A. *et al.* OWL Pizzas: Practical Experience of Teaching OWL-DL: Common Errors & Common Patterns. *Proc. 14th Int. Conf. Knowl. Acquis. Model. Manag. (EKAW 2004)* 63–81 (2004). doi:10.1007/978-3-540-30202-5_5

28. Hastings, J. *et al.* ChEBI in 2016: Improved services and an expanding collection of metabolites. *Nucleic Acids Res.* **44**, D1214–D1219 (2016).

29. Guarino, N., Guarino, N., Welty, C. & Welty, C. Towards a Methodology for Ontology

Based Model Engineering. *Proc. Int. Work. Model Eng. Nice, Fr. 2000, June, 13* (2000).

30. He, Y., Liu, Y. & Zhao, B. OGG: A biological ontology for representing genes and genomes in specific organisms. *CEUR Workshop Proc.* **1327**, 13–20 (2014).
31. Natale, D. A. *et al.* The Protein Ontology: A structured representation of protein forms and complexes. *Nucleic Acids Res.* **39**, 539–545 (2011).
32. Shen-orr, S. S., Milo, R., Mangan, S. & Alon, U. Network motifs in the transcriptional regulation network of Escherichia coli. *Nat. Genet.* **31**, 64–68 (2002).
33. Sugiura, M., Okamoto, T. & Takanami, M. RNA Polymerase σ-Factor and the Selection of Initiation Site. *Nature* **225**, 598–600 (1970).
34. Grigorova, I. L., Phleger, N. J., Mutalik, V. K. & Gross, C. A. Insights into transcriptional regulation and σ competition from an equilibrium model of RNA polymerase binding to DNA. *Proc. Natl. Acad. Sci. U. S. A.* **103**, 5332–5337 (2006).
35. Mungall, C. The Open World Assumption Considered Harmful. *https://douroucouli.wordpress.com/2020/09/04/the-open-world-assumption-considered-harmful/* (2019).