UN ACERCAMIENTO A LA ONTOLOGÍA DE GENES Y SUS APLICACIONES

AN APPROACH TO GENE ONTOLOGY AND ITS USES

Autor(es):

Lic. Ivette Camayd Viera[1], MSc. Miguel Sautié Castellanos [2], MSc. María A. Zardón Navarro[1], MSc. Carlos Martínez Ortiz[2], DrC. José Luis Hernández Cáceres[2]

[1)] Centro Nacional de Genética Médica, Universidad de Ciencias Médicas, MINSAP.
[2)] Centro de Cibernética Aplicada a la Medicina, Universidad de Ciencias Médicas, MINSAP.

Email de contacto: msc@infomed.sld.cu

RESUMEN:

Las ontologías desempeñan un papel importante en las investigaciones biomédicas. La Ontología de Genes (GO) es la ontología más aceptada y utilizada. Tal ontología es el fruto de una colaboración entre las bases de datos de organismos modelos para generar vocabularios estructurados con fines de anotación. Si bien la GO se diseñó como una ontología dirigida a la estandarización de la anotación de productos génicos, muchas aplicaciones la utilizan como herramienta para el cálculo semántico. Este artículo se centra en la descripción general de los componentes de la Ontología de Genes y en su relación con otras tecnologías avanzadas como las relacionadas con la Web semántica; asimismo, ilustramos su utilidad a través de algunas de sus aplicaciones más comunes en genómica funcional, minería de textos biomédicos y predicción de funciones de proteínas. Finalmente, consideramos algunas de las tendencias actuales en el desarrollo de la GO.

PALABRAS CLAVE:
Ontología, Ontología de genes, anotación funcional, genómica funcional, Web semántica, minería de textos biomédicos.

ABSTRACT:

The ontologies play an important role in biomedical research. The Gene Ontology (GO) is the most widely accepted and used ontology. This ontology is the result of a

collaboration among model organisms databases to generate structured vocabularies with annotation purposes. While GO was designed as a vocabulary for standardization of gene products annotations, many others applications also use it as a tool for semantic computation. This paper is focused on a general description of the constituent parts of GO and on its relationship with other cutting-edge technologies such as the ones known jointly as semantic Web technologies. Furthermore, we show the usefulness of GO by providing some examples of its applications in functional genomics, biomedical text mining and protein function prediction. Finally, we consider some current trends in GO development.

KEY WORDS:
Ontology, Gene Ontology, functional annotation, Functional genomics, Semantic Web, biomedical text mining.

1. INTRODUCCIÓN

Los avances en las tecnologías experimentales dentro de las ciencias biológicas han permitido la adquisición de datos a gran escala. Las tareas de investigación más comunes requieren comparar estos datos experimentales con resultados ya publicados y con resultados de otros laboratorios utilizando diversas herramientas de análisis. Uno de los problemas más importantes en las ciencias biomédicas consiste en encontrar información relevante en bases de datos públicas, lo que tiene su origen en la falta de estandarización de términos y en la diversidad de interfaces y lenguajes de consulta, en un momento en el que además es muy popular el acceso programático a bases de datos sin que exista un lenguaje estandarizado. Existen dificultades también en el intercambio de datos, tanto entre las bases de datos como entre las herramientas de análisis. Se impone, por tanto, establecer un nivel de estandarización en la comunidad científica para manejar y utilizar la información.

La necesidad de estándares no es reciente. Uno de los primeros vocabularios estandarizados de la medicina, conocido como Estadísticas de Mortalidad (*Bills of Mortality*), surgió en Londres en el siglo XVII (posteriormente integrado en la Clasificación Internacional de Enfermedades[1]). La biología, a su vez, ha utilizado diversos sistemas de clasificación. Uno de los ejemplos más notables es la clasificación taxonómica de las especies de Linneo[2], que además de definir sus clases incorporaba la relación entre ellas. La última década ha estado marcada por la creación de múltiples sistemas para formalizar el conocimiento biomédico. Las estructuras utilizadas van desde vocabularios controlados y tesauros hasta las ontologías. Uno de los vocabularios controlados más conocidos es el UMLS[3] (del inglés *Unified Medical Language System*), un compendio de vocabularios que integra unas 130 fuentes

diferentes mediante un proceso de mapeo retrospectivo, que se basa fundamentalmente en la identificación de relaciones de sinonimia entre los términos constituyentes. El UMLS es muy útil en la indexación y recuperación de documentos; pero los vocabularios que relaciona no tienen una arquitectura común[4] y el sistema no logra homogenizar las fuentes en una terminología única[5]. Otro vocabulario muy utilizado es el Encabezados de Temas Médicos[6] (MeSH, en inglés es el acrónimo de *Medical Subject Headings*), que organiza alrededor de 25000 descriptores en una estructura jerárquica, creada para la indización y búsqueda de documentos. El MeSH tiene una probada utilidad en este sentido, pero no define relaciones entre sus términos.

A diferencia de los vocabularios controlados, las ontologías constituyen representaciones formalizadas del conocimiento de gran utilidad en las ciencias de la información y de la computación que incluyen además del significado de los términos, las relaciones lógicas que los conectan. Una de las ontologías médicas actuales más relevante es el SNOMED-CT que agrupa a alrededor de 300 000 términos y que está construida a partir del sistema Apelon basado en el lenguaje formal dirigido a la representación del conocimiento conocido como *Description Logic*. Por otro lado, muchas de las más notables ontologías relacionadas con las ciencias de la vida están agrupadas dentro de la organización OBO (The Open Biomedical Ontologies) entre las que se destacan, *Sequence Ontology* (SO), *Foundational Model of Anatomy*, *Human disease, Cell type, Mammalian Phenotype, Protein Ontology, Mouse Pathology, Systems Biology, Molecular Biology Ontology*, conjuntamente con otras que formalizan la anatomía, la biología del desarrollo, los rasgos fenotípicos, las patologías y en general, el conocimiento sobre diversos organismos como *Caenorhabditis elegans*, *Drosophila sp*, plantas, etc. Una de las ontologías más utilizadas en estos momentos es, sin lugar a dudas, la Ontología de genes[7] (GO, del inglés *Gene Ontology*). La GO desarrolla vocabularios biológicos, aplicables a todas las especies, con el propósito de anotar los productos génicos de forma consistente en las diferentes bases de datos. Con el surgimiento de GO, la creación y uso de ontologías ha crecido en importancia dentro de la comunidad de las ciencias biológicas [8][9].

Los investigadores deben comprender cómo las ontologías están estructuradas y cómo se realiza la anotación funcional de productos génicos. En este artículo se describe el diseño de la GO y su papel en el proceso de anotación funcional de genes. Asimismo, se explican algunas de las aplicaciones, teniendo en cuenta las limitantes de las metodologías actuales, y varias tendencias que caracterizan el desarrollo y las aplicaciones de GO.

2. CONCEPTO E IMPORTANCIA DE LAS ONTOLOGÍAS

Los sistemas de clasificación juegan un papel crucial en las actividades del hombre. En las ciencias, y en particular en las ciencias biomédicas, los investigadores describen los diferentes fenómenos mediante el uso de categorías y sub-categorías. Una ontología, sin embargo, es algo más que un sistema de clasificación; en una ontología, cada una de estas categorías, clases, términos o conceptos quedan definidos por una serie de aserciones que los conectan a otros términos.

Los filósofos llaman Ontología a un sistema particular de categorías que representan determinadas visiones del mundo. En la inteligencia artificial, y en las ciencias de la computación en general, una ontología es un modelo construido a partir de un vocabulario específico que se utiliza para describir determinada realidad, además de un conjunto de supuestos explícitos referentes al significado que se desea atribuir a los términos del vocabulario. En el caso más sencillo, una ontología describe una jerarquía de conceptos conectados mediante relaciones mereológicas y de atribución; en casos más sofisticados, se utilizan además axiomas para expresar relaciones mucho más complejas entre los conceptos y así, adecuar más la interpretación científica a determinados dominio de la realidad[10]. En general, los aportes más significativos de las ontologías son los siguientes:

1) Al proporcionar una terminología unificada y lógicamente consistente dentro de determinado dominio del conocimiento garantizan: 1.1) Interoperabilidad entre Bases de datos (BDs), es decir, permiten una mayor eficiencia en la conexión semántica y programática entre las BDs, 1.2) La posibilidad de un lenguaje común que favorezca la comunicación entre los científicos dentro de determinada disciplina. 1.3) Sientan las bases para la creación de sistemas de gestión del conocimiento y de la información dirigidos al descubrimiento.

2) Al formalizar las definiciones y las conexiones lógicas entre los conceptos dentro de un dominio dado constituyen: 2.1) Una armazón lógica totalmente independiente del código fuente e incluso, del lenguaje de programación en que está implementada, 2.2) Facilitan el desarrollo de software dirigido a la interpretación automática, o a la minería de textos; lo que se aprecia en su enorme utilidad para la asignación inferencial de conceptos o cadenas de conceptos a un texto o dato determinado. Los motores de inferencia semántica realizan estas operaciones teniendo en cuenta las propiedades lógicas de las relaciones codificadas en las ontologías. 2.3) Facilitan el análisis de datos, por ejemplo, lo que se puede ver en el rol que juega actualmente la GO en el análisis estadístico de los datos de genómica funcional, en la anotación automática de las funciones de los productos génicos sobre la base de la similitud de secuencia, estructural o de perfiles de co-expresión. 2.4) Las ontologías sirven de base para su propio crecimiento y actualización, hecho que es posible gracias a la existencia de los editores y generadores de ontologías. Estas aplicaciones les permiten a los expertos

añadir nuevos conceptos, y sustituir o eliminar otros, sin que se afecte la consistencia lógica de una ontología, que puede llegar a agrupar decenas de miles de conceptos y cientos de miles de relaciones lógicas entre estos.

Las ontologías actuales están dedicadas a dominios específicos, muchas veces relacionados entre sí; las operaciones de comparación y mapeo entre estas, basadas en algoritmos de alineamiento de grafos, permitirían un mutuo enriquecimiento o la fusión, en nuevas ontologías más generales. Así, el desarrollo de las ontologías se puede considerar una etapa, o una alternativa, de importancia crucial en la formalización del lenguaje y del conocimiento acumulado en el estudio de objetos y fenómenos con tantos niveles de complejidad como los que usualmente se encuentran dentro de las ciencias biomédicas.

3. SURGIMIENTO Y EVOLUCIÓN DE LA ONTOLOGÍA DE GENES

A finales del siglo pasado, los genomas de cuatro microorganismos (*Haemophilus influenzae*, *Mycoplasma genitalium*, *Saccharomyces cerevisiae*, *Escherichia coli*) y de un organismo pluricelular (*Caenorhabditis elegans*) estaban secuenciados; el primer borrador del genoma humano estaba a punto de publicarse[11]. La disponibilidad de secuencias biológicas cambió radicalmente muchos puntos de vista en la biología. La comparación de secuencias nucleotídicas y proteicas ha demostrado que, incluso entre organismos muy diferentes, puede encontrarse una fracción considerable de genes ortólogos[12]. Muchas de las proteínas que codifican estos genes participan en procesos celulares comunes, como pueden ser la replicación del ADN, la traducción y el metabolismo; dichas proteínas están conservadas en la mayoría de las células vivientes. Un grado de conservación secuencial y funcional tan elevado representa la oportunidad de transferir, automáticamente, las anotaciones biológicas entre organismos diferentes, a partir de las similitudes a nivel de secuencias.

En esos momentos, la comunidad biológica contaba con varios algoritmos para la comparación de secuencias, incluso para la comparación de estructuras tridimensionales; la comparación de las anotaciones funcionales de productos génicos era, sin embargo, una tarea más compleja. La información almacenada en bases de datos cuenta con una amplia variedad de formas lexicales; consecuentemente, existe gran diversidad en cuanto a terminologías (sinónimos, alias, fórmulas), sintaxis (estructuras de ficheros, separadores, ortografía) y semántica (homónimos intra- e interdisciplinarios). El conocimiento relacionado con estos recursos es, efectivamente, muy difícil de recuperar y procesar, tanto para seres humanos como para las computadoras.

La GO es un esfuerzo colaborativo para establecer descripciones consistentes de los productos génicos anotados en bases de datos diferentes. En 1998, representantes

de las bases de datos de tres organismos 'modelo' –*Drosophila* (FlyBase), *Saccharomyces Genome Database* y *Mouse Genome Database*– fundaron el Consorcio GO. El trabajo del grupo tiene tres objetivos fundamentales: (1) desarrollar y mantener las ontologías; (2) anotar los productos génicos, lo que implica establecer asociaciones entre las ontologías y los genes y productos génicos en las bases de datos; y (3) desarrollar herramientas que faciliten la creación, mantenimiento y empleo de las ontologías. Actualmente, la cobertura de GO es muy superior[7] (Tabla 1).

3.1 Términos de la Ontología de Genes

Todos los términos de la GO tienen un nombre y un identificador único de la forma GO:nnnnnnn, la mayoría con una definición textual, con referencia a la fuente donde fue descrito. Cualquier observación necesaria se incluye en un campo de comentarios. GO utiliza sinónimos en un sentido amplio, pues no es necesario que los nombres dentro del campo 'sinónimo' signifiquen exactamente lo mismo que el término al que están vinculados; esta flexibilidad resulta muy útil en búsquedas y varias aplicaciones como la minería de texto.

Muchas funciones, procesos y componentes no son comunes a todas las formas vivientes; sin embargo, el objetivo de GO es desarrollar un vocabulario capaz de describir cualquier organismo. Con este propósito, GO acordó incluir cualquier término aplicable a más de una clase taxonómica. Para especificar la clase de organismo en cuestión, GO utiliza el conector *sensu,* ("en el sentido de").

Los términos que se consideran obsoletos se marcan como '*obsolete*', pero tanto el término como su identificador se mantienen en la base de datos de GO, por lo general, se añade un comentario que explica su caducidad y se sugiere un término actual para remplazar el término obsoleto.

3.2 Divisiones de la Ontología de Genes

La Ontología de Genes agrupa realmente tres ontologías que se corresponden con tres aspectos diferentes de la biología celular: función molecular, proceso biológico y componente o localización sub-celular. Aunque la GO incluye fundamentalmente conceptos que se refieren al nivel sub-celular y celular, abarca también niveles superiores, como los correspondientes a sistemas órganos y a organismo.

Proceso biológico (PB): Los PBs implican generalmente transformaciones químicas o físicas que ocurren por la acción de un conjunto de funciones moleculares organizadas; es decir, el objeto que va a un PB sufre transformaciones que lo convierten en algo diferente. Los PBs pueden ser de un nivel más elevado o abstracto, como son el "crecimiento celular" o la "transducción de señales", o de un nivel menor o más específico como son el "metabolismo de pirimidinas" o la "biosíntesis de AMPc".

Función molecular (FM): Describe actividades que ocurren a nivel molecular; sus términos representan a las actividades y no a las entidades (moléculas o complejos

moleculares) que llevan a cabo las acciones, sin especificar cuándo, dónde, o en qué contexto ocurren. Para evitar confusiones entre los nombres de los productos génicos y las FMs, muchos términos incorporan la palabra *activity* (actividad).

Componente celular (CC): Se refiere al espacio celular donde se encuentra el producto génico. Un componente celular puede ser una estructura anatómica, como el retículo endoplasmático, el núcleo celular, o una estructura molecular más simple formada por productos génicos, como un ribosoma o un dímero proteico.

3.3 Estructura de las ontologías

Los términos en GO se organizan en un grafo acíclico dirigido (GAD) en el cual los términos son vértices o nodos y las relaciones entre ellos son los arcos. En un GAD, los arcos son unidireccionales, no existen ciclos y un nodo "hijo" puede relacionarse con diferentes nodos "padres". Los términos heredan las relaciones y propiedades de sus nodos padres.

En un inicio, las relaciones entre los términos de GO eran de dos tipos fundamentales: *is_a* y *part_of*[13]. En el 2008, el consorcio agregó tres relaciones: *regulates, positively_regulates* y *negatively_regulates* para representar la relación entre procesos que afectan otros procesos sin ser parte de ellos[14]. Un año después, se incorporó la relación *has_part* que representa una relación parte-todo, pero desde la perspectiva de un nodo padre y es por tanto, un complemento lógico de la relación *part_of.* GO no relacionaba las tres ontologías entre sí, recientemente se han establecido relaciones entre PB y FM: existen relaciones *part_of* entre FM y PB y *regulates* entre FM y PB[15].

Los curadores de GO y de SO utilizan OBO-Edit[16], un editor de ontologías que permite filtrar, editar, razonar y detectar errores y posibles inconsistencias lógicas, asegura definiciones sintácticamente correctas y mantiene un camino *is_a* completo para cada nodo de las ontologías.

Gran parte del éxito de GO se debe a la adopción de una arquitectura simple – nótese que en un inicio solo contaba con dos tipos de relaciones y no existían vínculos directos entre las tres ontologías. La incorporación de nuevos términos y definiciones, en principio, no requiere cumplir con protocolos complicados; un biólogo puede hacerlo intuitivamente. En un estudio reciente, Alterovitz *et al* demuestran que, a pesar de que la GO ha mejorado estructuralmente desde su creación, todavía existen ineficiencias. Uno de los problemas descritos se relaciona con la variabilidad del contenido de información entre los términos dentro de un mismo nivel de la ontología. Otra de las dificultades aparece cuando la información decrece de un nivel a otro, es decir, pocos términos transmiten la información del nivel superior al siguiente, lo que se traduce en un decrecimiento de la especificidad. Una tercera deficiencia estructural radica en la variabilidad topológica, producto de una organización no óptima de las ramas del grafo. Las deficiencias estructurales afectan marcadamente la interpretación de los resultados

de experimentos de alto flujo, que parten del supuesto de que la profundidad de los niveles dentro del grafo es una medida de su especificidad[17].

3.4 Aspectos fundamentales de las anotaciones de la Ontología de Genes

Una anotación asocia un gen con determinados términos de las ontologías. Los genes se asocian con tantos términos como sea necesario, siempre que los términos en cuestión reflejen lo que realmente se conoce acerca del gen. Los curadores en las diferentes bases de datos asocian manualmente los términos de GO a los productos génicos de interés mediante la extracción de anotaciones a partir de datos experimentales publicados y/o a la inferencia de anotaciones basada en la homología con productos génicos que tienen datos experimentales publicados. Existen además métodos automáticos basados en similitud de secuencias y composición de dominio que se utilizan para anotar productos génicos sin la intervención del curador[18].

Toda anotación basada en la GO requiere de un código de evidencia que registra las condiciones en que se registra se realizó la anotación. Los códigos de evidencia caen en cuatro categorías fundamentales: experimental, computacional, derivado indirectamente de cualquiera de las categorías anteriores, o desconocido. En estos momentos, se utilizan 17 códigos diferentes[19]. Si se desconoce el proceso, función o localización de un gen, se anota en el nodo raíz con el código de evidencia ND (‘*no biological data available’*). Las anotaciones ND permiten diferenciar genes no anotados de genes no caracterizados.

Otra característica de GO es su actualización constante, incluyendo los mapeos entre términos GO y otros descriptores. Por ello, es imprescindible emplear las anotaciones de las últimas versiones de GO y citar la versión empleada, con vistas a garantizar la reproducibilidad de los resultados[20].

4. APLICACIONES DE LA ONTOLOGÍA DE GENES

4.1 Anotación de genes en bases de datos

GO es una de las herramientas más utilizadas para la anotación funcional de genes. El Proyecto de Anotación basada en la Ontología de Genes[21] (GOA, siglas del inglés *Gene Ontology Annotation*), dirigido por el Instituto Europeo de Bioinformática (EBI, siglas del inglés) surge en el 2001 con el objetivo de utilizar los términos GO en la descripción funcional de los productos génicos dentro de UniProtKB[22]. El desarrollo del GOA ha sido paralelo al crecimiento de anotaciones y secuencias disponibles en UniProtKB; actualmente con alrededor de 43 millones de proteínas que describen 32 millones de anotaciones[21]. GOA utiliza métodos manuales y electrónicos para asociar las entradas de UniProtKB con términos de GO y proporciona varios recursos para

garantizar el acceso a estas anotaciones. Muchas BDs hacen uso de las anotaciones de GOA, como son UniProtKB, Ensembl y Entrez-Gene[23].

El Proyecto Genoma de Referencia dentro del Proyecto de Ontología de Genes proporciona anotaciones de GO exhaustivas para el genoma humano, así como para otros 11 organismos[24]. Los organismos seleccionados representan un rango amplio dentro del espectro filogenético, generan una parte significativa de la literatura científica, son estudiados por grandes comunidades científicas, y son importantes como sistemas experimentales en el estudio de las enfermedades humanas o en actividades económicas como la agricultura. En todos los casos, las BDs de estos organismos son el resultado del trabajo de equipos de curadores con experiencia en la anotación funcional de genes utilizando GO.

4.2 Uso de la Ontología de Genes en estudios de genómica funcional[1]

El análisis de los datos obtenidos en experimentos de alto flujo (como son los microarreglos de ADN) es una de las aplicaciones más comunes de GO. Los resultados de estos experimentos son generalmente los conjuntos de genes que se expresan en condiciones experimentales específicas. GO permite determinar los procesos, funciones y localizaciones que están sobre-/sub-representadas en un grupo de genes. La metodología más simple consiste en el procedimiento conocido como *análisis de enriquecimiento*, calculado como el porcentaje de genes de una categoría GO determinada que esta sobre-/sub-representada en un conjunto de genes, en relación con un conjunto de genes de referencia. Una limitante del método es que un valor de enriquecimiento puede ocurrir por mera casualidad; por tanto, el enriquecimiento *per se* no debe interpretarse como una evidencia inequívoca de la implicación de un término GO en el fenómeno estudiado, sin una prueba estadística apropiada. Los modelos estadísticos que más se utilizan con estos fines son la distribución hipergeométrica, la distribución binomial, la prueba exacta de Fisher y la prueba de chi cuadrado (χ^2). La distribución binomial se recomienda en grandes conjuntos de genes y no si se trabaja con un conjunto pequeño específico para un fenómeno determinado. Uno de los aspectos más importantes es el conjunto de genes de referencia que se selecciona, y debe incluir solamente los genes que se monitorean en el estudio. Muchas de las herramientas disponibles utilizan todos los genes del genoma como referencia; tal es el caso de GOTollBox[24], GOstat[25], GoMiner, FatiGO y GOTM[26].

[1] La genómica funcional se basa en el uso de técnicas de alto rendimiento, o de alto flujo, para el estudio de la abundancia relativa de numerosos productos génicos (sean ARNm o proteínas) expresados bajo diferentes condiciones, o en diferentes tejidos o etapas del desarrollo. Entre las plataformas tecnológicas más usadas están los microarreglos de expresión basados en ADN complementario, así como la electroforesis bidimensional en gel y la espectrometría de masas para proteínas. La bioinformática es crucial en el análisis de los resultados, dada la gran cantidad de datos que se generan.

Las tecnologías de alto flujo permiten determinar qué categoría de genes se encuentra sobre-representada, o trabajar con una categoría seleccionada *a priori* en respuesta a una hipótesis específica. En el primer caso, es muy importante utilizar un método de corrección para múltiples experimentos, a fin de disminuir significativamente el total de falsos positivos. Muchas herramientas que desarrollan este análisis, como son GoMiner, DAVID[27] y CLENCH[28] no implementan esta corrección. Entre las herramientas que implementan la corrección para hipótesis múltiples están FunAssociate[29], Onto-Express[30], EasyGO[31], GeneMerge[32], GOstat[25], GOToolBox[24], GOEAST[33], GoSurfer[34] y GOrilla[35]. La mayoría de estas herramientas permite analizar simultáneamente las tres divisiones de GO, mientras que otras relacionan además los resultados con recursos externos, como pueden ser rutas metabólicas y de señalización[32] o la localización cromosómica de los genes de interés.

Los estudios que han comparado algoritmos diferentes atribuyen las variaciones en los resultados a factores como son el método utilizado para mapear los identificadores génicos y de secuencia, las fuentes y las versiones de los archivos de anotación, el método de propagación de la anotación (anotaciones directas contra anotaciones propagadas a los padres), los métodos estadísticos (por ejemplo, de una o dos colas), la fórmula matemática que realmente se emplea en el cálculo y el método de corrección de múltiples hipótesis [17]. En cuanto al tiempo de ejecución, Khatri *et. al.* encontraron que las herramientas más rápidas eran GoSurfer, GeneMerge y Onto-Express; resulta interesante que las dos últimas son recursos web[36]. El sitio de GO enumera algunas de las herramientas disponibles, con vínculos a publicaciones y los sitios donde están implementadas[37].

4.3 Categorización funcional

Es posible realizar categorizaciones funcionales más generales mediante las *GOslim*, conjuntos reducidos de términos de GO de alto nivel. Estas categorizaciones se obtienen mapeando las anotaciones de los genes de interés a un subconjunto de términos de GO. Su empleo es muy poderoso para resumir las anotaciones de todo un genoma[38], una colección de ESTs[39], cDNAs[40] o de patrones diferentes de expresión[41]. GO ofrece la implementación *map2slim* con estos fines. En el sitio de GO están disponibles varios *GOslim* que algunos expertos crearon y actualizan periódicamente –planta, levadura, *Homo sapiens* y GOA (genérica).

La estructura de GAD de GO debe tenerse en cuenta en cualquiera de las *GOSlim* que vaya a utilizarse. Es muy común que un producto génico esté anotado con varios términos que reflejan las funciones, localizaciones y procesos en los que participa; cada uno de estos términos puede tener varios padres. Por tal razón, las anotaciones de un gen equivalen frecuentemente a varios términos de una *GOslim* y además, los gráficos de pastel, muy utilizados para ilustrar la distribución funcional de los genes, no

son las mejores representaciones, pues la suma de las anotaciones sobrepasa el 100%[20].

Los usuarios crean *GOslims* según sus necesidades, específicas para las especies o para determinadas áreas de GO. La creación *de novo* de una *GOslim* requiere la selección manual de los términos de interés, tarea que puede realizarse con OBO-Edit[16]. Las *GOslims* han permanecido desde su introducción en GO como un conjunto de herramientas *ad hoc*; sus definiciones no describen su contenido con claridad y precisión. Asimismo, no existen criterios que especifiquen la inclusión de los términos; por ejemplo, es posible que en una *GOslim* esté excluido un término a pesar de que sus ancestros y sucesores sí aparezcan. Otra de las dificultades en el uso de las *GOslims* específicas para grupos taxonómicos está en la falta de relaciones explícitas entre ellos; como consecuencia, por ejemplo, la *GOslim* de Plantas contiene solo algunos de los elementos de la *GOslim* de arroz, y viceversa[42]. Las *GOslims,* por tanto, deben utilizarse con mucho cuidado, pues pueden influir perjudicialmente en los resultados del análisis.

4.4 Predicción de funciones de los productos génicos

La predicción de funciones génicas sobre la base de las anotaciones GO ofrece ventajas incuestionables. GO abarca todos los procesos biológicos, en contraste con esquemas anteriores que se limitaban, por ejemplo, a rutas metabólicas o de señalización. Las anotaciones manuales, de conjunto con las anotaciones de alta calidad que aporta el Proyecto Genoma de Referencia, aportan estándares de anotación consistentes con todas las formas vivientes e incrementan considerablemente la calidad de los análisis computacionales[20].

Las anotaciones de productos génicos ya caracterizados, referidas a localización, función y procesos en los que están involucrados, pueden transferirse a partir de la similitud de secuencias y estructural a nuevas proteínas. Otros métodos se basan en el comportamiento de los genes en estudios de expresión génica o en las redes de interacción proteína-proteína. El enriquecimiento de términos de la GO se determina por alguno de los métodos anteriores; se presume entonces que los genes no caracterizados tienen funciones biológicas similares a los genes con los cuales se agrupan. Las funciones génicas también pueden inferirse a través del análisis semántico de una matriz de asociación de funciones génicas, sin necesidad de un agrupamiento previo de genes[43].

La anotación manual de alta calidad es, por tanto, un prerrequisito indispensable para inferir funciones de los genes no caracterizados. El uso de las anotaciones para predecir función génica debe incluir un estudio de los códigos de evidencia asociados a las anotaciones utilizadas: propagar funciones en base a anotaciones que no están verificadas experimental o manualmente genera probablemente un número substancial de falsos positivos. Otro punto importante está en el aspecto particular que se analiza

dentro de la GO (dígase función molecular, proceso biológico o localización celular). Las inferencias basadas en estudios de co-expresión, por ejemplo, pueden tener menos sentido para una función molecular o una localización celular que para un proceso biológico más abarcador[20].

4.5 Aplicaciones de la Ontología de Genes en la Minería de texto

La indexación parte de la asignación de entradas de un vocabulario controlado a documentos, por ejemplo, de la literatura biomédica. El proceso se realiza de forma manual en muchas de las colecciones principales. Encontrar los resúmenes relevantes en grandes colecciones de resúmenes indexados, como PubMed, es una tarea compleja. El motor de búsqueda que utiliza PubMed es el sistema Entrez[44], que desarrolla la búsqueda en dos etapas: en un primer momento localiza los términos del MeSH en la consulta hecha por el usuario; posteriormente localiza los términos encontrados en todos los resúmenes a partir de un proceso de alineamiento de caracteres que no considera la semántica de la frase que se consulta. En consecuencia, si el proceso de recuperación de resúmenes de bases de datos se realiza solo a partir de palabras claves, los sinónimos quedarán fuera del análisis. Por otro lado, PubMed organiza los resúmenes cronológicamente, no hay opciones para refinar los resultados de una búsqueda y existe una posibilidad muy elevada de encontrar resúmenes no relevantes.

El componente terminológico de las ontologías biomédicas es un recurso importante en los sistemas de procesamiento del lenguaje natural[2] y en tareas de gestión del conocimiento, como la anotación o indexación de recursos, el acceso y la recuperación de información y el mapeo entre recursos diferentes[45]. Por ejemplo, GoPubMed[46] introduce el concepto de navegación basada en ontologías. El sistema recupera los resúmenes relevantes utilizando Entrez y los estructura según la jerarquía que proporciona GO. GO-KDS[47] utiliza aprendizaje automático para encontrar los resúmenes relevantes, utilizando resúmenes que contienen términos de GO que se encuentran en bases de datos como SwissProt, GeneBank, FlyBase, etc. Otras propuestas similares son Ali-Baba, PubFinder y Chilibot. Los sistemas anteriores no utilizan la semántica de los resúmenes y las consultas en sus métodos de búsqueda. Por su parte, SEGOPubmed adapta el concepto de análisis semántico latente para enlazar los resúmenes de PubMed a GO. Textpresso[48] utiliza una ontología diseñada específicamente para consultar información sobre conceptos biológicos específicos en una colección bibliográfica.

La minería de texto también puede ser una herramienta para analizar la importancia que tienen los términos de una ontología dentro del domino que se quiere describir.

[2] El procesamiento del lenguaje natural consiste en la comprensión, análisis, manipulación y/o generación de lenguaje natural (humano) mediante el uso de computadoras.

Tsoi *et. al.* proponen un método para evaluar la relevancia de los términos pertenecientes a una ontología mediante un análisis de enriquecimiento de los conceptos en cuestión dentro de PubMed[49].

4.6 Análisis semántico basado en la Ontología de Genes

Las medidas de similitud semántica permiten obtener valores numéricos en función de la cercanía del significado entre términos de una ontología, o entre los conjuntos de anotados a determinas entidades. La aplicación de las medidas de similitud semántica entre las anotaciones de GO proporciona una medida de su similitud funcional. En la actualidad, están disponibles diversas propuestas para cuantificar la similitud semántica[50].

Los términos en una ontología con estructura de grafo como GO pueden compararse mediante dos vías fundamentales, dependiendo de que se utilicen los nodos o los arcos como fuente de datos. Los métodos que emplean los arcos se basan fundamentalmente en contar el número de arcos entre dos nodos del grafo[51]. Estas metodologías, aunque intuitivas; se basan en dos supuestos que son ciertos en muy raras ocasiones en la biología: (1) los nodos y los arcos están distribuidos uniformemente y (2) los arcos en el mismo nivel de la ontología se corresponden con igual distancia semántica entre los términos.

Los métodos basados en nodos utilizan las propiedades de los términos implicados, que pueden relacionarse con los propios términos, sus ancestros o descendientes. Uno de los conceptos que se emplea con más frecuencia en estos casos es el Contenido de Información[3] (CI), el cual puede calcularse en base a la ocurrencia de un término en una base de datos[52], o a partir del número de hijos que tiene un término en GO, aunque esta variante es menos empleada. Los métodos basados en el CI son menos sensibles que los métodos basados en arcos a la variabilidad de distancia semántica y densidad de nodos, pues el CI es una medida de la especificidad de un término independiente de su profundidad en la ontología. No obstante, el CI está sesgado por las tendencias actuales de la investigación biomédica, pues aquellos términos de interés científico tienen más probabilidad de estar anotados. El uso del CI todavía tiene sentido desde el punto de vista probabilístico, pues es más probable (y menos significativo) que dos productos génicos compartan términos usados con mucha frecuencia, independientemente de que el termino sea común por ser genérico o por estar relacionado a una temática de investigación activa[50].

Los productos génicos se pueden anotar con muchos términos de GO de cada una de las tres ontologías; por tanto, la evaluación de la similitud funcional (dentro de una

[3] El *contenido de información* es una medida de cuan específico e informativo es un término determinado. El CI de un término *c* se cuantifica como la probabilidad logarítmica negativa *-log p(c),* dónde p(c) es la probabilidad de ocurrencia de c en un corpus especifico.

categoría particular de GO) implica comparar conjuntos de términos en lugar de términos independientes. Un grupo de métodos determina la similitud funcional entre dos productos génicos a partir de la similitud semántica entre los términos a los cuales están anotados. En algunos casos se determina cada combinación de pares de términos anotados[53], mientras que en otros casos solo se consideran las mejores combinaciones.

Existen variantes que no se basan en la combinación de similitudes entre términos individuales para calcular la similitud semántica entre productos génicos, sino que la calculan directamente, por ejemplo, a partir de las anotaciones directas (no las heredadas). Estas metodologías son muy poco comunes, pues muy pocas medidas de similitud semántica consideran solo anotaciones directas[50]. Otros métodos representan los productos génicos como los subgrafos de GO correspondientes a todas sus anotaciones (directas y heredadas). La similitud funcional se calcula utilizando técnicas de alineamiento de grafos[54] o considerando los subgrafos como conjuntos de términos y utilizan alguna medida de similitud semántica[50]. Las metodologías vectoriales representan el producto génico como un espacio vectorial, donde cada término corresponde a una dimensión; la similitud semántica se determina a partir de medidas de similitud vectorial[55].

La diversidad de metodologías y medidas de similitud semántica genera una interrogante fundamental: ¿Qué medida captura mejor la similitud de función entre dos productos génicos? No existe manera de saber la verdadera similitud funcional entre dos productos génicos, por tanto, no es trivial determinar la mejor medida de similitud semántica. Algunas alternativas utilizan métodos de similitud de secuencias y datos de co-expresión para evaluar las medidas de similitud semántica. Desafortunadamente, la mayoría de los estudios de aplicación utilizan una sola medida y no comparan los resultados[50].

Las medidas de similitud semántica se han implementado como herramientas web, herramientas autónomas y paquetes de R (como por ejemplo, GOSim), estos últimos son parte del proyecto Bioconductor[56] y permiten la integración de las herramientas de similitud semántica con otros paquetes para la visualización o el análisis estadístico. La similitud semántica se ha utilizado en múltiples aplicaciones como son la comparación de productos génicos, la predicción de funciones y la evaluación y validación de métodos automáticos de predicción de funciones. Otra de las aplicaciones de estos métodos está en la predicción y validación de interacciones entre productos génicos y de redes de interacciones; dentro de este contexto, permiten además extraer módulos funcionales de redes de interacción, alinear rutas biológicas y generar subconjuntos de redes significativos. En el análisis de datos de transcriptómica y proteómica, el uso de la similitud semántica permite mejorar la agrupación de productos génicos expresados teniendo en cuenta su similitud funcional, comparar resultados de experimentos diferentes, mejorar la calidad de los datos y validar la

selección de genes con fines biomédicos. Otras aplicaciones incluyen evaluar el significado biológico de dominios cromosómicos co-expresados, la predicción de la localización celular y la integración de la búsqueda semántica[50].

4.7 La Ontología de Genes y las tecnologías de la Web semántica

Una de las principales áreas de aplicación de las ontologías en el futuro es en el contexto de la Web Semántica[4]. La investigación en el campo de la Web Semántica se concentra en el diseño de lenguajes, métodos y herramientas que permitirán la construcción de recursos de conocimiento distribuido, de forma similar a la *World Wide Web*, pero con una precisión suficiente como para facilitar algún nivel de razonamiento automático[57].

El Consorcio de la *World Wide Web* (W3C)[58] crea la infraestructura de la Web Semántica mediante la producción de formalismos para representar documentos, recursos y ontologías. Estas especificaciones se conocen colectivamente como tecnologías de la Web Semántica, e incluyen el XML, RDF/S y OWL.

XML (eXtensible Markup Language) es un lenguaje extensible de marcas, es decir, que utiliza etiquetas para definir la semántica de los datos que encapsula. XML ganó popularidad con gran rapidez una vez creado, en diez años ya domina los medios de representación de la información en entornos de redes, todos los navegadores populares soportan XML, y los sistemas de bases de datos más importantes importan y exportan datos en este formato. Si bien el vocabulario terminológico que aplica XML favorece la interoperabilidad entre bases de datos, presenta algunas limitantes. Los esquemas XML no especifican sintaxis ni semánticas, no indican por ejemplo como interpretar porciones de un árbol estructurado en términos de aserciones, y no son suficientes para definir relaciones lógicas entre conceptos. Tal información suele archivarse en la documentación en lenguaje natural que ofrece el esquema, inaccesible a un procesamiento computarizado. Los problemas anteriores quedan resueltos con una tecnología de representación del conocimiento, o tecnologías de la web semántica, capaces de describir explícitamente la semántica de los datos o recursos representados[59].

La tecnología básica de la web semántica, RDF (del inglés, *Resource Description Framework*) extiende las posibilidades que ofrece XML. RDF proporciona un modelo de

[4]La Web semántica (del inglés *semantic web*) es la "Web de los datos". Se basa en la adición de metadatos semánticos y ontológicos a la *World Wide Web*. Esas informaciones —que describen el contenido, el significado y la relación de los datos— deben proporcionarse de manera formal, para que sea posible evaluarlas automáticamente por máquinas de procesamiento. El objetivo es mejorar Internet ampliando la interoperabilidad entre los sistemas informáticos usando "agentes inteligentes" o programas de computación, sin la intervención de operadores humanos. Nótese que el concepto de *web semántica* es diferente del de *red semántica,* propio de la psicología cognitiva. En inglés la Web Semántica se traduce como S*emantic Web* y Red semántica como *Semantic Network*.

datos sencillo para describir las relaciones entre los recursos; la estructura resultante es un grafo donde los nodos con un identificador único (URI, del inglés *Uniform Resource Identifier*) son recursos o datos; los arcos del grafo son relaciones o propiedades. Como grafo, el modelo RDF integra reglas de inferencia limitadas, que permiten, por ejemplo, definir subclases y sub-propiedades. RDFS (del inglés *RDF Schema*) hace más poderoso el modelo al permitir vistas ontológicas sobre las sentencias de RDF, es decir, permite que recursos nuevos sean especializaciones de recursos ya existentes. OWL (del inglés *Web Ontology Language*) se construye encima de RDFS y permite definir nuevas relaciones, a la vez que añade restricciones que aumentan la precisión del modelo que se representa[60].

Uno de los beneficios del uso de las ontologías es que la información que describen puede publicarse en la Web Semántica si la ontología se representa en RDF, RDFS o en OWL, de manera que las herramientas de la Web Semántica puedan utilizarse en el análisis y la integración de los datos[60]. GO y SO están disponibles en estos formatos. GO ofrece además accesos a través de la Web Semántica mediante un SPARQL[5] experimental[61].

4.8 Importancia de la Ontología de Genes y en general, de las ontologías para la integración de bases de datos

Las principales bases de datos biológicos utilizan diferentes formatos, se ubican en sitios distintos y tienen diferentes interfaces de usuario. La mayoría de los bancos de datos permiten realizar consultas de textos completos, mientras que todas las bases de datos soportan consultas basadas en la ocurrencia de una cadena de caracteres en campos predefinidos. Los usuarios en estos bancos de datos, en general, utilizan la interfaz web que le proporciona el sitio, donde las consultas pueden realizarse a través de la línea de comandos en el lenguaje de consulta del sistema.

El uso de los sistemas disponibles presenta algunas dificultades. El usuario que accede a determinado recurso necesita conocimientos básicos de la fuente de datos, así como del lenguaje de consulta y la interfaz del sistema. En varias ocasiones es necesario consultar varias fuentes, y resulta laborioso aprender lo necesario sobre cada recurso.

El problema de la interoprabilidad dado por la carencia de estándares para la comunicación entre datos y aplicaciones se hace cada vez más importante dentro de la Bioinformática. Uno de los métodos para lograr la interoperabilidad entre bases de datos propone utilizar vínculos estáticos entre los registros de los datos en distintas fuentes. El servidor SRS (siglas del inglés *The Sequence Retrieval System*)[62] dentro

[5] SPARQL, acrónimo del inglés *SPARQL Protocol and RDF Query Language,* es de un lenguaje estandarizado para la consulta de grafos RDF, normalizado por el DAWG (*RDF Data Access Working Group*) del W3C. Es una tecnología clave en el desarrollo de la Web Semántica.

del EBI analiza sintácticamente ficheros planos o BDs, creando y almacenando un índice para cada campo. Estos índices se utilizan en las consultas para recuperar las entradas relevantes. El resultado es una simple agregación de registros que satisfacen las condiciones de la búsqueda; los registros contienen vínculos que permiten obtener más información sobre los resultados. SRS actúa como un recurso central para las bases de datos biológicas, incorpora el fichero de asociación del proyecto GOA y un espejo del repositorio de GO. SRS constituye en realidad un almacén de datos basado en la indexación de ficheros planos que se actualiza periódicamente a partir de servidores remotos. Otros ejemplos notorios en Bioinformática de almacenes de datos son el UCSC Genome Browser, Ensemble, BioMolQuest, entre otros.

Las ontologías también permiten la recuperación de datos mediante los sistemas de almacenes de datos (*data warehouse*). Los datos provenientes de servidores diferentes se copian en un servidor local que proporciona un punto de acceso único. El usuario necesita una interfaz única para realizar consultas en bases de datos diferentes. Algunos sistemas que han adoptado esta metodología son BioWarehouse[63], Atlas[64] y ArrayExpress[65], todos ellos incorporan las anotaciones de GO. La principal ventaja está en el rendimiento, pues la dependencia con la rapidez de la conexión es mínima, se evita la latencia producto de la comunicación con las fuentes de datos y las consultas se optimizan localmente. Otro beneficio es la posibilidad del usuario de filtrar, validar, modificar y anotar los datos obtenidos, una propiedad muy valiosa en bioinformática. La necesidad de actualizar los datos a partir de las fuentes originales, y reflejar tales cambios en el repositorio local, constituye la principal desventaja de los almacenes de datos.

Los sistemas que quizás logran una mejor integración de bases de datos son aquellos basados en sistemas mediadores; en ellos la consulta del usuario se traduce a una consulta en el lenguaje de la fuente de datos. La reformulación de las consultas es uno de los elementos más importantes en estos métodos, mientras que sus principales ventajas estriban en la posibilidad de recuperar la información a partir de las fuentes originales en lugar de utilizar vínculos estáticos, a la vez que no necesita copiar y actualizar grandes cantidades de datos. Las metodologías basadas en mediadores superan el problema de la heterogeneidad utilizando vocabularios u ontologías. TAMBIS, por ejemplo, utiliza una ontología y un razonador sobre ella. La ontología de TAMBIS (TaO) contiene más de 2000 conceptos que describen tareas de bioinformática y biología molecular[66].

Asimismo, podríamos mencionar los sistemas mucho más flexibles y populares aunque menos eficientes basados en los servicios Web (*Web services*) dirigidos a la comunicación directa entre aplicaciones, bases de datos o servidores remotos. Se basan en estándares especialmente diseñados para describir datos, servicios y comunicación entre aplicaciones. Estos últimos constituyen extensiones del XML, y entre ellos se destacan el SOAP (Simple Object Access Protocol) y WSDL (Web

Services Description Language). En Bioinformática se destacan varios grupos que usan estas tecnologías, así tenemos el DAS, Sistema de Anotación Distribuida de genomas que usan Ensembl, WormBase y FlyBase entre otros, el Sistema de Bases de Datos de Rutas (*Pathway Database System*) y el KEGG API que proporcionan el acceso a rutas a través de interfaces web basadas en SOAP. En la actualidad, ocupa un lugar relevante el proyecto *Open Source*, BioMOBY, los servicios BioMOBY combinan el uso de SOAP y de la ontología BioMOBY que describe estructuras de metadatos, servicios y las relaciones entre éstos. BioMOBY permite el descubrimiento y la generación de tuberías de *web services* que se adaptan a las solicitudes de los clientes, actualmente hay más de 160 servicios procedentes de 35 proveedores que usan esta tecnología híbrida[67].

4.9 Creación de nuevos vocabularios especializados: combinación de ontologías

Una de las principales cuestiones que se debaten en la comunidad científica plantea la creación de una ontología extensiva para describir todo el campo de la biología molecular, en lugar del diseño de varias ontologías más pequeñas, orientadas a áreas específicas[98]. Una ontología que abarque todo el dominio será muy útil, si pudiera concebirse y mantenerse. En la práctica, sin embargo, parece más eficiente y efectivo contar con muchas ontologías para subdominios particulares del conocimiento.

La combinación de los conceptos de dos vocabularios permite expandir una ontología, o crear una ontología más especializada, con aspectos de las ontologías de partida. Las ventajas del uso de tales estrategias combinatorias son más evidentes a medida que los vocabularios crecen en tamaño y complejidad. El uso de reglas definidas permite a los curadores construir vocabularios nuevos, una vez que se enfocan en la descripción de un subdominio particular.

La construcción de vocabularios compatibles con la generación de subproductos ofrece entonces la posibilidad de extender los vocabularios originales. La combinación de ontologías permite definir también relaciones adicionales. Las potencialidades que ofrece la combinación de ontologías demuestra la importancia de coordinar la construcción de vocabularios elementales a la vez que ilustra los beneficios de combinar conceptos de tales vocabularios para crear vocabularios extendidos consistentes[68].

5. TRASCENDENCIA DE LA ONTOLOGIA DE GENES: DESARROLLO COLABORATIVO DE NUEVAS ONTOLOGÍAS

El protagonismo de la comunidad biológica en la conceptualización inicial de GO es uno de los factores que más ha influido en su popularidad. Los fundadores de GO son los líderes de tres de las bases de datos de organismos modelos más importantes; sus decisiones tienen, por tanto, un impacto muy elevado en la comunidad científica[69]. A pesar de sus múltiples aplicaciones, la anotación de genes ha sido el objetivo principal de GO desde su surgimiento en 1998.

El amplio campo de aplicaciones de GO ha fomentado el desarrollo de otras ontologías para la anotación en bases de datos. El consorcio OBO[70] (del inglés *Open Biomedical Ontologies*) se creó para coordinar tales esfuerzos. El sitio oficial de OBO proporciona hipervínculos a 93 ontologías, todas abiertas e independientes. OBO aplica los principios fundamentales de GO, que a su vez han propiciado su éxito en la comunidad biomédica: las ontologías son abiertas, ortogonales, han de estar instanciadas en una sintaxis bien especificada y deben estar concebidas para compartir un espacio común de identificadores carentes de sentido semántico y con definiciones en el lenguaje natural[71]. Las ontologías son abiertas, pues ellas y los datos que describen deben estar disponibles para su uso sin restricciones o licencias, de manera que sean aplicables a cualquier propósito sin restricciones. Además, deben admitir las modificaciones resultantes de los debates de la comunidad, han de ser ortogonales para asegurar la aditividad de las anotaciones y favorecer los beneficios del desarrollo modular. Las dos ontologías más importantes dentro de OBO son GO y SO[72]. Esta última proporciona un vocabulario para describir los diversos componentes y tipos de secuencias biológicas. Simultáneamente, los creadores de OBO iniciaron *OBO Foundry*[71], un experimento colaborativo basado en la aceptación de un conjunto de principios adicionales a los de OBO: basado en el desarrollo colaborativo, en el uso de relaciones definidas sin ambigüedad, en procedimientos de retroalimentación con el usuario dirigidos a identificar las versiones exitosas y en el principio de que cada ontología debe utilizar conceptos bien circunscritos al dominio que describe.

6. TENDENCIAS ACTUALES EN EL DESARROLLO DE LA ONTOLOGÍA DE GENES

GO marcó un punto de inflexión en el desarrollo de las ontologías biomédicas. Su diseño no estuvo exento de críticas[73], que propiciaron las mejoras estructurales que se han introducido y se introducen paulatinamente. Una meta permanente en el Consorcio GO es alentar las contribuciones de la comunidad científica, para aumentar el número de anotaciones, ampliar el rango de especies incluidas y mejorar la calidad de los vocabularios[15].

Otras orientaciones buscan alinear e integrar los vocabularios de GO con vocabularios externos pertenecientes a OBO, en áreas como son los productos y

procesos bioquímicos y los tipos celulares; con ello se extenderían las posibilidades del razonamiento basado en estos vocabularios y se facilitaría la comprobación de errores, así como la incorporación automática de términos durante el desarrollo y crecimiento de GO[68].

Gracias al éxito de GO han surgido otras ontologías e instituciones especializadas en su desarrollo y aplicación, como son IFOMIS (*Formal Ontology and Medical Information Science*), el NCBO (*National Center for Biomedical Ontology*), el NCOR (*National Center for Ontological Research*) y el ECOR (*European Center for Ontological Research*)[9]. Tales instituciones están tomando el liderazgo y, muy probablemente, guíen el desarrollo de la disciplina en el futuro.

La certificación y validación de ontologías constituyen aspectos cada vez más importantes, lo cual supone necesariamente el desarrollo de métricas objetivas para evaluar calidad[9, 74]. *OBO Foundry*, en este sentido, promueve pautas para el desarrollo de ontologías en relación con la familia de ontologías que pertenecen a OBO, a la vez que selecciona ontologías de alta calidad para que la comunidad utilice como referencia.

Las definiciones de los términos de GO, que utilizan sus usuarios “humanos”, son en general opacas y carentes de significados para las computadoras; problema presente en otras ontologías dentro de OBO y que se agrava por la heterogeneidad que existe en cuanto a metodologías para establecer estas definiciones. Una de las direcciones de trabajo se orienta a crear definiciones computables a partir de una colección de productos cruzados, utilizando varias ontologías. El objetivo final es utilizar razonadores para automatizar el mantenimiento de GO, un proceso tedioso y propenso a errores. Las definiciones computables pueden utilizarse además en consultas entre varias ontologías y en la visualización de los resultados[75]. Las ontologías dentro de OBO son ortogonales; sin embargo, existen relaciones implícitas entre varias ontologías diferentes. La formalización de estas relaciones potenciaría la consulta y análisis de los datos a la vez que ayudaría en la construcción y el mantenimiento de las propias ontologías[9].

La calidad de las principales bases de datos depende en gran medida del proceso de curado manual. Típicamente, un curador lee los textos completos de los artículos y transfiere su esencia a la base de datos. La diversidad de métodos experimentales y computacionales que hoy se utilizan, de conjunto con el crecimiento exponencial del número de publicaciones científicas, dificulta este proceso. Un uso adecuado de las anotaciones que ofrecen las ontologías dependerá de la contribución de toda la comunidad científica con los curadores de las bases de datos. Los autores y los editores de revistas científicas, por ejemplo, deben trabajar juntos para facilitar el intercambio de datos entre las fuentes bibliográficas y las bases de datos.

7. CONCLUSIONES

GO surge como respuesta a la necesidad de vocabularios estandarizados dentro de la comunidad científica. Su utilización inmediata y aceptación general se reflejan en un amplio campo de aplicaciones que incluyen las anotaciones funcionales de genes, minería de textos, la indexación de literatura, el acceso y la gestión de la información, la realización de consultas, actualización y recuperación de información con herramientas especializadas, el intercambio de información y la interoperabilidad semántica entre sistemas y bases de datos.

Los paradigmas para la creación de ontologías están evolucionando. Los retos actuales se centran en la necesidad de coordinar esfuerzos para el desarrollo y perfeccionamiento de ontologías y recursos. GO será una herramienta más poderosa a medida que las ontologías, herramientas y anotaciones evolucionen.

Tabla 1: Estadísticas de GO en septiembre de 2009

Total de términos anotados		
Términos anotados en *procesos biológicos*	19069	
Términos anotados en *función molecular*	8637	
Términos anotados en *componente celular*	2432	
Términos de Ontología de Secuencias	1603	
Bases de datos anotadas*	52	
Especies con anotaciones	197439	
Productos génicos anotados		
Total	44545253	
Electrónico	43665159	
Manual	890094	

* Las bases de datos en su mayoría representan a una sola especie; Gramene, TIGR, UniProt GOA y UniProt PDB representan múltiples especies

8. REFERENCIAS BIBLIOGRÁFICAS

1. International Classification of diseases. < http://www.who.int/classifications/icd/en/> [Consulta: 28 de agosto de 2010].
2. McCray AT. Conceptualizing the world: lessons from history. J Biomed Inform 2006; 39: 267-73.
3. Bodenreider O. The Unified Medical Language System (UMLS): integrating biomedical terminology. Nucleic Acids Res 2004; 32: D267-70.
4. Ceusters W, Smith B, Goldberg L. A terminological and ontological analysis of the NCI Thesaurus. Methods Inf Med 2005; 44: 498-507.
5. Campbell KE, Oliver DE, Shortliffe EH. The Unified Medical Language System: toward a collaborative approach for solving terminologic problems. J Am Med Inform Assoc 1998; 5: 12-6.
6. Medical Subject Headings. <www.ncbi.nlm.nih.gov/mesh> [Consulta: 25 de agosto de 2010].
7. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, et al. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat Genet 2000; 25: 25-9.
8. Bodenreider O. Biomedical ontologies in action: role in knowledge management, data integration and decision support. Yearb Med Inform 2008: 67-79.
9. Bodenreider O, Stevens R. Bio-ontologies: current trends and future directions. Brief Bioinform 2006; 7: 256-274.
10. Guarino N. Formal ontology and information systems. Proc FOIS'98 1998: 3-15.
11. Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, Sutton GG, et al. The sequence of the human genome. Science 2001; 291: 1304-51.
12. Batzoglou S. The many faces of sequence alignment. Brief Bioinform 2005; 6: 6-22.
13. The Gene Ontology (GO) project in 2006. Nucleic Acids Res 2006; 34: D322-6.
14. The Gene Ontology project in 2008. Nucleic Acids Res 2008; 36: D440-4.
15. The Gene Ontology in 2010: extensions and refinements. Nucleic Acids Res 2010; 38: D331-5.

16. Day-Richter J, Harris MA, Haendel M, Lewis S. OBO-Edit--an ontology editor for biologists. Bioinformatics 2007; 23: 2198-200.

17. Alterovitz G, Xiang M, Hill DP, Lomax J, Liu J, Cherkassky M, et al. Ontology engineering. Nat Biotechnol; 28: 128-30.

18. Daraselia N, Yuryev A, Egorov S, Mazo I, Ispolatov I. Automatic extraction of gene ontology annotation and its correlation with clusters in protein networks. BMC Bioinformatics 2007; 8: 243.

19. Guide to GO Evidence Codes. <http://www.geneontology.org/GO.evidence.shtml> [Consulta: 26 de agosto de 2010].

20. Rhee SY, Wood V, Dolinski K, Draghici S. Use and misuse of the gene ontology annotations. Nat Rev Genet 2008; 9: 509-15.

21. Barrell D, Dimmer E, Huntley RP, Binns D, O'Donovan C, Apweiler R. The GOA database in 2009--an integrated Gene Ontology Annotation resource. Nucleic Acids Res 2009; 37: D396-403.

22. Camon E, Magrane M, Barrell D, Binns D, Fleischmann W, Kersey P, et al. The Gene Ontology Annotation (GOA) project: implementation of GO in SWISS-PROT, TrEMBL, and InterPro. Genome Res 2003; 13: 662-72.

23. Camon E, Barrell D, Lee V, Dimmer E, Apweiler R. The Gene Ontology Annotation (GOA) Database--an integrated resource of GO annotations to the UniProt Knowledgebase. In Silico Biol 2004; 4: 5-6.

24. Martin D, Brun C, Remy E, Mouren P, Thieffry D, Jacq B. GOToolBox: functional analysis of gene datasets based on Gene Ontology. Genome Biol 2004; 5: R101.

25. Beissbarth T, Speed TP. GOstat: find statistically overrepresented Gene Ontologies within a group of genes. Bioinformatics 2004; 20: 1464-5.

26. Zhang B, Schmoyer D, Kirov S, Snoddy J. GOTree Machine (GOTM): a web-based platform for interpreting sets of interesting genes using Gene Ontology hierarchies. BMC Bioinformatics 2004; 5: 16.

27. Dennis G, Jr., Sherman BT, Hosack DA, Yang J, Gao W, Lane HC, et al. DAVID: Database for Annotation, Visualization, and Integrated Discovery. Genome Biol 2003; 4: P3.

28. Shah NH, Fedoroff NV. CLENCH: a program for calculating Cluster ENriCHment using the Gene Ontology. Bioinformatics 2004; 20: 1196-7.

29. Gabriel F. Berriz GB, King OD, Bryant B, Sander C, Roth PF. Characterizing gene sets with FuncAssociate. Bioinformatics 2003; 19: 2502.2504.

30. Khatri P, Draghici S, Ostermeier GC, Krawetz SA. Profiling gene expression using onto-express. Genomics 2002; 79: 266-70.

31. Zhou X, Su Z. EasyGO: Gene Ontology-based annotation and functional enrichment analysis tool for agronomical species. BMC Genomics 2007; 8: 246.

32. Castillo-Davis CI, Hartl DL. GeneMerge--post-genomic analysis, data mining, and hypothesis testing. Bioinformatics 2003; 19: 891-2.

33. Zheng Q, Wang XJ. GOEAST: a web-based software toolkit for Gene Ontology enrichment analysis. Nucleic Acids Res 2008; 36: W358-63.

34. Zhong S, Xie D. Gene Ontology analysis in multiple gene clusters under multiple hypothesis testing framework. Artif Intell Med 2007; 41: 105-15.

35. Eden E, Navon R, Steinfeld I, Lipson D, Yakhini Z. GOrilla: a tool for discovery and visualization of enriched GO terms in ranked gene lists. BMC Bioinformatics 2009; 10: 48.

36. Khatri P, Draghici S. Ontological analysis of gene expression data: current tools, limitations and open problems. Bioinformatics 2005; 21: 3587-3595.

37. The Gene Ontology. Tools for Analysis of Data Sets. < http://www.geneontology.org/GO.tools.microarray.shtml> [Consulta: 25 de agosto de 2010].

38. Adams MD, Celniker SE, Holt RA, Evans CA, Gocayne JD, Amanatides PG, et al. The genome sequence of Drosophila melanogaster. Science 2000; 287: 2185-95.

39. Adzhubei AA, Vlasova AV, Hagen-Larsen H, Ruden TA, Laerdahl JK, Hoyheim B. Annotated expressed sequence tags (ESTs) from pre-smolt Atlantic salmon (Salmo salar) in a searchable data resource. BMC Genomics 2007; 8: 209.

40. Kawai J, Shinagawa A, Shibata K, Yoshino M, Itoh M, Ishii Y, et al. Functional annotation of a full-length mouse cDNA collection. Nature 2001; 409: 685-90.

41. Pan G, Tian S, Nie J, Yang C, Ruotti V, Wei H, et al. Whole-genome analysis of histone H3 lysine 4 and lysine 27 methylation in human embryonic stem cells. Cell Stem Cell 2007; 1: 299-312.

42. Kusnierczyk W. Taxonomy-based partitioning of the Gene Ontology. J Biomed Inform 2008; 41: 282-92.

43. Khatri P, Done B, Rao A, Done A, Draghici S. A semantic analysis of the annotations of the human genome. Bioinformatics 2005; 21: 3416-21.

44. Entrez cross-database search page [homepage de Internet] < www.ncbi.nlm.nih.gov/Entrez/> [Consulta: 24 de agosto de 2010].

45. Spasic I, Ananiadou S, McNaught J, Kumar A. Text mining and ontologies in biomedicine: Making sense of raw text. Brief Bioinform 2005; 6: 239-251.

46. Doms A, Schroeder M. GoPubMed: exploring PubMed with the Gene Ontology. Nucleic Acids Res 2005; 33: W783-6.

47. Smith TC, Cleary JG. Automatically linking MEDLINE abstracts to the Gene Ontology. Proc of Bio-Ontologies Meeting 2003.

48. Muller HM, Kenny EE, Sternberg PW. Textpresso: an ontology-based information retrieval and extraction system for biological literature. PLoS Biol 2004; 2: e309.

49. Tsoi LC, Patel R, Zhao W, Zheng WJ. Text-mining approach to evaluate terms for ontology development. J Biomed Inform 2009; 42: 824-30.

50. Pesquita C, Faria D, Falcao AO, Lord P, Couto FM. Semantic similarity in biomedical ontologies. PLoS Comput Biol 2009; 5: e1000443.

51. Yu H, Gao L, Tu K, Guo Z. Broadly predicting specific gene functions with expression similarity and taxonomy similarity. Gene 2005; 352: 75-81.

52. Bodenreider O, Aubry M, Burgun A. Non-lexical approaches to identifying associative relations in the gene ontology. Pac Symp Biocomput 2005: 91-102.

53. Lord PW, Stevens RD, Brass A, Goble CA. Investigating semantic similarity measures across the Gene Ontology: the relationship between sequence and annotation. Bioinformatics 2003; 19: 1275-83.

54. Mistry M, Pavlidis P. Gene Ontology term overlap as a measure of gene functional similarity. BMC Bioinformatics 2008; 9: 327.

55. Chabalier J, Mosser J, Burgun A. A transversal approach to predict gene product networks from ontology-based similarity. BMC Bioinformatics 2007; 8: 235.

56. Bioconductor. Open Source software for Bioinformatics. <http://www.bioconductor.org> [Consulta: 30 de agosto de 2010].

57. Stevens R, Bodenreider O, Lussier YA. Semantic webs for life sciences. Pac Symp Biocomput 2006: 112-5.

58. The World Wide Web Consortium. <http://www.w3.org> [Consulta: 28 de agosto de 2010].

59. Philippi S, Kohler J. Using XML technology for the ontology-based semantic integration of life science databases. IEEE Trans Inf Technol Biomed 2004; 8: 154-60.

60. Wang X, Gorlitsky R, Almeida JS. From XML to RDF: how semantic web technologies will change the design of 'omic' standards. Nat Biotechnol 2005; 23: 1099-103.

61. SPARQL Query Language for RDF. < http://www.w3.org/TR/rdf-sparql-query> [Consulta: 21 de agosto de 2010].

62. Zdobnov EM, Lopez R, Apweiler R, Etzold T. The EBI SRS server-new features. Bioinformatics 2002; 18: 1149-50.

63. Lee TJ, Pouliot Y, Wagner V, Gupta P, Stringer-Calvert DWJ, Tenenbaum JD, et al. BioWarehouse: a bioinformatics database warehouse toolkit. BMC Bioinformatics 2006; 7.

64. Shah SP, Huang Y, Xu T, Yuen MMS, Ling J, Ouellette FBF. Atlas – a data warehouse for integrative bioinformatics. BMC Bioinformatics 2005.

65. Parkinson H, Kapushesky M, Shojatalab M, Abeygunawardena N, Coulson R, Farne A, et al. ArrayExpress—a public database of microarray experiments and gene expression profiles. Nucleic Acids Res 2006; 35: D747–D750.

66. Stevens R, Baker P, Bechhofer S, Ng G, Jacoby A, Paton NW, et al. TAMBIS: transparent access to multiple bioinformatics information sources. Bioinformatics 2000; 16: 184-5.

67. Neerincx PBT, Leunissen JAM. Evolution of web services in bioinformatics. Brief Bioinform 2005; 6: 178-88.

68. Hill DP, Blake JA, Richardson JE, Ringwald M. Extension and integration of the gene ontology (GO): combining GO vocabularies with external vocabularies. Genome Res 2002; 12: 1982-91.

69. Lewis SE. Gene Ontology: looking backwards and forwards. Genome Biol 2005; 6: 103.

70. Open Biomedical Ontologies. <http://obofoundry.org> [Consulta: 29 de agosto de 2010].

71. Smith B, Ashburner M, Rosse C, Bard J, Bug W, Ceusters W, et al. The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration. Nat Biotechnol 2007; 25: 1251-5.

72. Eilbeck K, Lewis SE, Mungall CJ, Yandell M, Stein L, Durbin R, et al. The Sequence Ontology: a tool for the unification of genome annotations. Genome Biol 2005; 6: R44.

73. Smith B, Kumar A. Controlled vocabularies in bioinformatics: a case study in the gene ontology. DDT: Biosilico 2004; 2: 246-252.

74. Zhang S, Bodenreider O. Law and order: assessing and enforcing compliance with ontological modeling principles in the Foundational Model of Anatomy. Comput Biol Med 2006; 36: 674-93.

75. Mungall CJ, Bada M, Berardini TZ, Deegan J, Ireland A, Harris MA, et al. Cross-product extensions of the Gene Ontology. J Biomed Inform (artículo en imprenta).