Universidad Pablo de Olavide
Máster en Análisis Bioinformático Avanzado
19/11/2018
Trabajo de Fin de Máster

Compatax: herramienta computacional para la comparación de anotaciones de clases funcionales desde proteomas completos

Roldán Alés, Francisco J.

Tutorizado por Pérez Pulido, Antonio J.

Resumen: El aumento vertiginoso de proyectos de secuenciación genómica ocurrido a lo largo de los últimos 10 años trae consigo la necesidad de creación de nuevas herramientas bioinformáticas para trabajar con toda esa información. Todas estas secuencias son catalogadas mediante un proceso predictivo de anotación funcional. Pero para una mejor caracterización de los nuevos genomas secuenciados, se hace necesaria la aparición de nuevas herramientas bioinformáticas que sean capaces de obtener nueva información, y que permita la comparación de clases funcionales. El objetivo de este proyecto es el desarrollo de Compatax, una nueva herramienta bioinformática para comparar clases funcionales entre proteomas anotados, la cual se alimenta de la salida del anotador funcional de proteomas y transcriptoma Sma3s v2. La experimentación realizada haciendo uso de Compatax ha demostrado su capacidad para comparar de forma sencilla clases funcionales entre proteomas anotados, siendo de gran utilidad para detectar características taxonómicas diferenciativas. Además, Compatax destaca por su optimización a la hora de aprovechar los recursos de hardware disponibles. Se ha desarrollado una base de datos para almacenar el conjunto de clases funcionales de diversos proteomas que sirva de base para futuros experimentos científicos.

Abstract: The vertiginous increase of genomic sequencing projects that have taken place over the last 10 years brings the need to create new bioinformatics tools to work with all this information. All these sequences are cataloged by a predictive process of functional annotation. But for a better characterization of the new sequenced genomes, it is necessary the arising of new bioinformatic tools that are able to obtain new information, and that allow the comparison of functional classes. The objective of this project is the development of Compatax, a new bioinformatics tool to compare functional classes between annotated proteomes, which feeds from the output of the functional annotator of proteomes and transcriptome Sma3s v2. The experimentation carried out using Compatax has demonstrated its ability to easily compare functional classes between annotated proteomes, being very useful to detect differentiating taxonomic characteristics. In addition, Compatax stands out for its optimization when it comes to taking advantage of available hardware resources. A database has been developed to store the set of functional classes of various proteomes that will serve as the basis for future scientific experiments.

1 Introducción

En la era genómica en la que nos encontramos en la biologia molecular actual, el número de secuencias de genomas completos disponibles desde las bases de datos públicas no para de crecer de forma exponencial a lo largo de los últimos 10 años. Esto ha sido posible gracias a las tecnologías NGS (*Next Generation Sequencing*) que comenzaron a comercializarse en 2005[1] y que provocaron una reducción muy significativa de tiempo, coste y tasa de errores en el proceso de secuenciación genómica respecto a las tecnologías basadas en el método de Sanger[2]. Por ejemplo, el Proyecto Genoma Humano[3] cuyo objetivo era su secuenciación completa, duró 11 años (1990-2001) y requirió una inversión aproximada de 2.400 millones de euros[4]. Hoy día, gracias a las NGS, un genoma humano puede secuenciarse en menos de un día con un coste por debajo de los 1.000 euros[5].

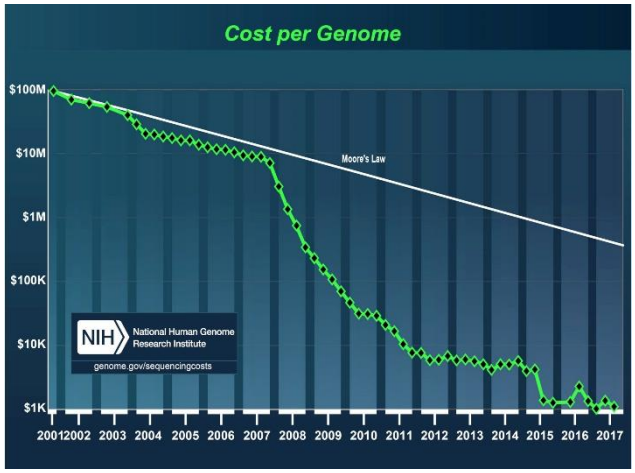

Fig. 1. Evolución del coste en dólares de secuenciación de un genoma completo a lo largo del siglo XXI. Entre el año 2001 y 2007 se aprecia un abaratamiento progresivo con tendencia lineal, la cual se convierte en exponencial a partir del año 2007 y se mantiene hasta hoy. Datos extraídos del Instituto Nacional de Investigación del Genoma Humano[5].

Gracias a esto, la tendencia en los proyectos de secuenciación ha pasado desde solo secuenciar una sección genómica de interés a secuenciar genomas completos de forma masiva.

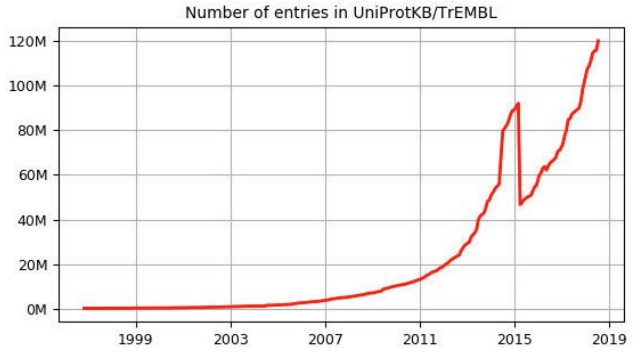

Fig. 2. Evolución del número de secuencias almacenadas en el repositorio de datos sobre proteínas UniProt[6]. El crecimiento de la base de datos pasa a tomar una tendencia exponencial a partir del año 2003, tendencia que mantiene hasta ahora. En 2015 se aprecia una brusca reducción del tamaño provocada por la reducción de secuencias redundantes.

El crecimiento ha sido tan grande que provocó casos como el de la base de datos de referencia de secuencias proteicas, UniProt, que en 2015 decidió aplicar una reducción de secuencias redundantes, eliminando 46.9 millones de entradas procedentes principalmente de proteomas bacterianos[7], sumando éstas casi la mitad del tamaño total de la base de datos. Sin embargo, menos de dos años después ya había recuperado su tamaño previo.

Este aumento vertiginoso de proyectos de secuenciación genómica trae consigo la necesidad de creación de nuevas herramientas bioinformáticas para trabajar con toda esa información. Una vez secuenciado un organismo el resultado final son diferentes secuencias de nucleótidos almacenadas en ficheros de texto en formato FASTA[8], los cuales suponen datos en bruto que han de pasar por diversas fases de análisis *in silico* para extraer información de las mismas. En el proceso de predicción de genes, determinadas herramientas bioinformáticas se encargarán de realizar una predicción de aquellas secuencias que conforman genes codificantes de proteínas y otros elementos biológicamente funcionales. En una bacteria podemos encontrar aproximadamente entre 3000 y 4000 genes mientras que un organismo eucariota puede presentar entre 20000 y 100000. Cada uno de estos genes debe ser catalogado y descrito con una serie de anotaciones funcionales que definan sus características como pueden ser su función biológica, procesos en los que están involucrados, enfermedades relacionadas o su localización dentro de la estructura celular.

Las anotaciones del consorcio GO (*Gene Ontology*)[9] son un estándar desarrollado en 1998 cuya finalidad es ofrecer un soporte para la incorporación de información a genes de forma sistémica e inequívoca. GO propone un vocabulario controlado de términos estructurado en forma de grafo acíclico dirigido, el cual no contiene ciclos ni repeticiones. Esto permite que una anotación de carácter más específico y situada en partes más bajas del grafo pueda estar relacionada con más de una anotación padre, las cuales serán de carácter más genérico y estarán situadas en partes más altas de grafo.

La ontología GO abarca 3 categorías de anotación diferentes: *biological process*, *celular component* y *molecular function*. Asimismo, existen otras fuentes de anotación como las keywords de la base de datos UniProt o los códigos de enzimas de Enzyme.

Anotar fidedignamente cada gen de un genoma requiere de diversos y costosos experimentos en laboratorio. Por ello se suele realizar previamente una predicción de las anotaciones en base a secuencias homólogas anotadas en distintas bases de datos biológicas. Dos secuencias se denominan homólogas cuando tienen un origen evolutivo común, siendo ortólogas cuando provienen de un proceso de especialización o parálogas cuando provienen de un proceso de duplicación. Los métodos de predicción de anotaciones

más precisos realizan búsquedas de secuencias ortólogas mediante un proceso de *BLAST recíproco* que a veces utiliza herramientas exhaustivas como *PSI-BLAST*.

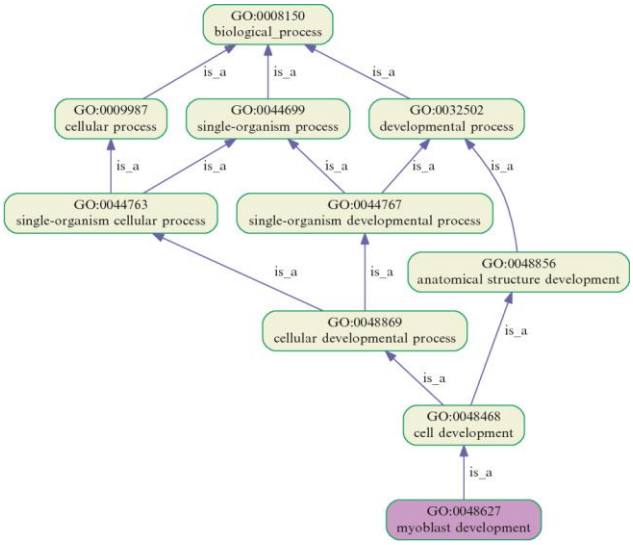

Fig. 3. Ejemplo del grafo de anotaciones resultante partiendo de GO:0048627 (*myoblast development*) como anotación más específica.

El proceso de anotación masiva es muy complejo computacionalmente debido a la enorme cantidad de información que existe en las diversas bases de datos biológicas sobre las cuales se ha de buscar. Existen diferentes aplicaciones bioinformáticas para realizar anotación masiva (tabla 1).

Herramienta	Características
FFPred 3	Especializado en anotar proteínas sin homólogos. Es lenta, tomando más de 30 minutos por secuencia
ARGOT2.5	Anotación de proteínas y secuencias codificantes. Usa BLAST y HMMER para buscar homólogos y luego realiza clustering de los términos GO relacionados con ellos. Sólo permite anotar conjuntos de 500 secuencias. Acepta como entrada resultados de BLAST y HMMER (ficheros < 1GB).
PANNZER	Anotación de proteomas y transcriptomas. Usa BLAST y HMMER y calcula distancias taxónomicas y 'clustering' a partir de los homólogos encontrados.
ESG / PFP	Dos métodos para la anotación de proteínas Ambos usan PSI-BLAST y utilizan los términos GO de los homólogos encontrados. Sólo permite anotar conjuntos de 100 secuencias. Busca secuencias alejadas evolutivamente.
BAR+	Anotación de función y estructura de proteínas. Realiza alineamientos de todos los pares de proteínas en UniProt y proteomas de referencia y construye 'clusteres' de similitudes
Blast2GO	Anotación de proteomas y transcriptomas con múltiples funciones. Realiza una búsqueda de homólogos con BLAST y mapea las anotaciones GO de los resultados. Se necesita una licencia para uso y tiene versión de pago con funciones extra. Puede enlazar las anotaciones de la herramienta BioMart. Es actualmente la herramienta más citada.
Trinotate	Anotación de transcriptomas. Usa diferentes técnicas, como búsquedas de homologías, de señales de secuencia y dominios e información de bases de datos de ortólogos.
FastAnnotator	Anotación de transcriptomas. Es útil para transcriptomas de novo. Actualmente no funciona.
ESTAnnotator	Anotación de secuencias EST. Se ha quedado desactualizado.
Sma3s v2	Anotación de proteínas y secuencias nucleotídicas de forma masiva. Busca la secuencia en la base de datos o a un ortólogo y hace clustering de todos los resultados de un BLAST, asignando anotaciones por enriquecimiento biológico. Funciona tanto en Linux como Windows y Mac. Genera un resumen por procesos biológicos y otros grupos de interés que permite realizar gráficos de los resultados. No tiene limitación de secuencias y es rápido.

Tabla 1. Características de las principales herramientas bioinformáticas sobre anotación masiva.

El proceso de anotación masiva de un organismo da como resultado el conjunto de anotaciones predichas para cada secuencia de su genoma por separado. El objetivo de este proyecto es desarrollar una aplicación bioinformática que aporte un valor adicional al conjunto de anotaciones de un organismo secuenciado en conjunto, permitiendo realizar comparaciones entre parejas o grupos de taxones u organismos. A esta nueva herramienta la llamaremos Compatax. Específicamente se planteó realizar la comparación entre genomas o taxones, utilizando el resumen de clases funcionales de cada anotación genómica. De ese modo, el resultado sería el conjunto de clases funcionales comunes entre los genomas compartidos, asi como aquellas más particulares de uno de ellos y que sirvan para distinguirlo del resto. Esto permitirá dar un mayor valor a la anotación de nuevos genomas, asi como su clasificación respecto a los ya conocidos y presentes en bases de datos actuales.

2 Materiales y métodos

2.1 Anotación funcional

Sma3s (*sequence massive annotation by 3 modules*) es una herramienta bioinformática para la anotación de secuencias proteicas desarrollada por el Grupo UPOBioinfo de la Universidad Pablo de Olavide (bioinfocabd.upo.es). Dado un conjunto de secuencias, Sma3s hace uso del algoritmo de búsqueda por similitud BLAST[10] sobre una base de datos de secuencias aminoacídicas de referencia (UniProt) para encontrar las secuencias más similares a cada una, con el fin extraer de ellas las anotaciones que presente y asociarlas a la secuencia sin anotar. Este proceso se apoya en el hecho de que, si dos secuencias genómicas son muy similares, su función biológica será muy similar. Estudios realizados demuestran que la herramienta presenta un nivel de precisión en los resultados superior al 80%[11].

Compatax se alimenta de uno de los ficheros de salida resultantes de la ejecución de Sma3s, el cual contiene el número de secuencias que tiene el proteoma o transcriptoma anotado en cada clase funcional procedente de dos fuentes distintas: categorías de Gene Ontology y Keywords de UniProt (figura 5).

2.2 Programación informática

Para el desarrollo de la herramienta bioinformática se requiere de un lenguaje de programación que destaque a la hora de tratar con ficheros de textos y gran cantidad de información. Entre los diversos lenguajes existentes el elegido ha sido Perl, lenguaje muy utilizado en proyectos bioinformáticos.

Los requisitos mínimos para la ejecución de Compatax son Perl en su versión 5.20 o superior junto con los módulos Data::Dumper, Getopt::Long, MongoDB, Path::Tiny, DateTime, JSON::MaybeXS, File::Copy, Storable, threads::shared, File::Basename, Cwd y Timer::Runtime. Todos estos módulos se encuentran disponibles en el repositorio CPAN y se recomienda instalar desde ahí en la versión más actual para cada uno de ellos en la fecha de publicación del presente documento.

El código fuente de la herramienta se encuentra publicado en un repositorio público de GitLab: gitlab.com/franroldan/compatax.

2.3 Almacenamiento de información

Con el objetivo de retroalimentar el uso de la herramienta se añade como complemento a la misma un sistema de almacenamiento en el que poder disponer de una serie de organismos de referencia con los cuales realizar comparativas taxonómicas. Este sistema debe ofrecer una gran versatilidad a la hora de almacenar la información sobre las anotaciones dado que éstas no presentan un esquema fijo: no conocemos de antemano todas las anotaciones existentes, un proteoma puede no presentar anotaciones de alguna familia ontológica, una anotación presente en algún proteoma puede no existir en otro… Además, se ha de valorar la velocidad de lectura y la compatibilidad con el lenguaje de programación escogido para el desarrollo de la herramienta.

Entre las diversas tecnologías de bases de datos existentes la elegida fue MongoDB (mongodb.com), un sistema de base de datos NoSQL orientado a documentos, multiplataforma y de código abierto. Su nombre proviene de la palabra *humongous* que significa enorme, dado que está orientado a almacenar grandes cantidades de información de una manera más rápida y sencilla que en los sistemas relacionales clásicos (SQL).

MongoDB almacena internamente la información en ficheros con formato BSON, una versión binaria del lenguaje de notación de objetos JavaScript JSON[12], el cual incorpora algunas características extra como tipos de datos adicionales o índices de ordenación. En la figura 4 se puede observar un pequeño ejemplo del formato interno con el que Compatax organiza la información.


```
{
    "name" : "Escherichia coli",
    "tax_id" : "562",
    "source" : "genbank, NCBI",
    "date" : ISODate("2018-09-04T15:43:24.000Z"),
    "tags" : [
        "prokaryote",
        "bacteria",
        "enterobacteria"
    ],
    "genes" : 5017,
    "summary" : {
        "Cellular component" : {
            "GO:0005886: plasma membrane" : 1066,
            "Membrane" : 1360,
            "GO:0005737: cytoplasm" : 740
        },
        "Biological process" : {
            "GO:0009058: biosynthetic process" : 1096,
            "GO:0044281: small molecule metabolic process" : 692,
            "GO:0034641: cellular nitrogen compound metabolic process" : 821
        },
        "UniProt Pathways" : {
            "Amino-acid biosynthesis" : 84,
            "Glycan metabolism" : 18,
            "Cofactor biosynthesis" : 83
        },
        "Molecular function" : {
            "GO:0022857: transmembrane transporter activity" : 463,
            "GO:0043167: ion binding" : 827,
            "GO:0003677: DNA binding" : 642
        },
        "Disease" : {
            "Disease mutation" : 4
        },
        "Developmental stage" : {
            "Late protein" : 48,
            "Early protein" : 6
        }
    }
}
```


Fig. 4. Información en formato JSON del conjunto de anotaciones de un proteoma almacenado en la base de datos de Compatax, en este caso perteneciente al organismo *E. coli*. En la figura se representan solo las categorías funcionales más importantes de cada grupo.

Compatax es compatible con bases de datos MongoDB cuya versión sea 2.0 o superior.

2.4 Informes estadísticos

La herramienta que se desarrolla debe mostrar de forma sencilla la comparación taxonómica entre organismos, los cuales presentan multitud de anotaciones diferentes. Por tanto, es necesario elegir tecnologías gráficas que faciliten la representación de la información.

Para desarrollar gráficos estadísticos se utilizó Google Chart (developers.google.com/chart), una potente aplicación para realizar gráficos estadísticos en entorno web e integrable con diferentes lenguajes de programación. Dispone de una galería con una amplia gama de gráficos y su uso es totalmente gratuito.

Dado que los gráficos de Google Charts están diseñados para su integración en un entorno web, Compatax usa HTML como formato de salida para representar los resultados. Además, esto ofrece la posibilidad de integrar otras tecnologías web como Bootstrap (getbootstrap.com), un framework de código abierto para el diseño en entornos web. Contiene plantillas de diseño con tipografías, formularios, botones, menús de navegación y otros elementos de diseño basados en HTML5 y CSS3, así como extensiones JavaScript adicionales. Facilita en gran medida el desarrollo de interfaces responsivas y adaptables a diferentes formatos y tamaños de pantalla.

La primera versión de Compatax cuenta inicialmente con dos informes estadísticos: uno destinado a comparar pares de proteomas anotados entre sí y otro destinado a comparar un proteoma anota contra un conjunto de tamaño variable proveniente de la base de datos. El primero representa la información en gráficos de barras y tablas, mientras que el segundo hace uso principalmente de diagramas de cajas (*boxplot*).

2.5 Algoritmo Compatax

Para realizar comparaciones taxonómicas sobre organismos haremos uso de un algoritmo basado en una función de distancia. Este algoritmo calcula las diferencias en el número de cada una de las anotaciones presentes en un par de organismos. El valor de distancia se obtiene del sumatorio de todas estas diferencias las cuales serán ponderadas según el número de genes anotados en cada organismo. Es necesario ponderar las diferencias en base al número de genes dado que, aunque normalmente en un grupo de organismos taxonómicamente similares los organismos que lo componen tendrán un número de genes anotados muy similar, en ciertos casos podemos encontrar organismos poliploides que podrían añadir ruido a los resultados obtenidos. Los organismos poliploides son aquellos cuyo genoma se reparte en un número de juegos cromosómicos por encima de 2, existiendo por tanto parte del código genético repetido.

La función de distancia queda definida como:

$$d(A,B) = \sum_i \left| \frac{A_i}{n} - \frac{B_i}{m} \right|$$

Donde d es el valor de distancia, A y B serían los proteomas anotados a comparar, i recorrería todas las anotaciones de los organismos siendo Ai el número de proteínas en A que presenta la anotación i. n representa el número de genes anotados dentro del proteoma A y m el número de genes anotados dentro del proteoma B.

Al ser una función de distancia, dos organismos serán taxonómicamente más cercanos entre sí a menor distancia presenten entre ellos, siendo la distancia siempre un valor positivo mayor o igual a 0. Valores por debajo de 0,5 en el valor de d indicarían que ese par de proteomas son muy similares, y por encima de 2 indicarían que son muy diferentes.

2.6 Recursos de hardware

Para el desarrollo de este trabajo, el clúster de supercomputación de la Universidad Pablo de Olavide C3UPO (pvcbacteria.org/c3upo/web) nos ha facilitado acceso a su plataforma. Se ha instalado Compatax junto con una base de datos MongoDB en uno de los nodos del clúster el cual cuenta con las siguientes características:

- Sistema operativo CentOS Linux v7.2
- 40 cores de procesamiento
- 62 GB de memoria RAM
- 207 GB de almacenamiento físico

Hay que tener en cuenta que estos recursos se encuentran compartidos con el resto de aplicaciones que se estén ejecutando en la plataforma dentro de ese nodo, por lo que la disponibilidad de los recursos es variable y raramente superará el 80% del total.

Usaremos este entorno para realizar la experimentación, la cual nos sirve además como prueba de rendimiento de la herramienta.

2.7 Organismos de estudio

Para validar el funcionamiento y demostrar el potencial de Compatax se realizaron dos experimentos. En uno de ellos se trabajó con un grupo de 52 organismos bacterianos del género *Bacillus*, obtenidos de la base de datos Ensembl, ya utilizados en Casimiro-Soriguer et al.[14]. En otro de ellos trabajaremos con un grupo de 64 proteomas anotados sobre plantas donde se encuentran desde pequeñas algas unicelulares a grandes arbóreos. El listado completo de organismos usado se lista en el fichero adjunto al presente documento como ANEXO 2 y las anotaciones fueron realizadas por el grupo del Dr. Antonio Muñoz, co-autor del artículo de Sma3s[14].

2.8 Obtención de anotaciones para genomas completos

Para poder trabajar con Compatax debemos previamente realizar un proceso de anotación masiva sobre las secuencias de aquellos organismos que nos sean de interés. Usaremos para ello la version 2 de la herramienta bioinformática Sma3s la cual se caracteriza por su facilidad de uso, una alta precisión en sus predicciones y por ofrecer un resumen de la anotación completa contabilizando genes implicados en procesos biológicos. Sma3s requiere de dos ficheros de entrada para su ejecución: un fichero FASTA con el listado de secuencias a anotar, las cuales en este caso serían el listado completo de secuencias del organismo de interés, y una base de datos sobre la cual realizar las búsquedas por homología mediante el paquete Blast+. Se puede utilizar por ejemplo alguna de las variantes de UniProt disponibles para su descarga vía FTP (ftp.uniprot.org/pub/databases/uniprot/current_release/uniref).

Una vez finalizado el proceso de anotación para un conjunto de secuencias, Sma3s da como resultado dos ficheros en formato de texto tabulado TSV. Estos ficheros pueden visualizarse y editarse mediante software de hojas de cálculo.

El primero de los ficheros contiene las anotaciones para cada secuencia del proteoma que se ha podido anotar. Entre estas anotaciones se incluyen anotaciones GO, anotaciones sobre rutas metabólicas de UniProt (uniprot.org/help/pathway) y Swiss-Prot Keywords (uniprot.org/keywords). Este fichero es el que incluye información más completa y específica de los dos, siendo útil para analizar la predicción de anotaciones para proteínas concretas.

El segundo fichero de salida representa la información del proceso de anotación con un nivel de abstracción mayor, siendo este fichero una de las características principales que diferencian a Sma3s con el resto de herramientas de anotación existentes. Este fichero expone la información de la anotación en forma de resumen, agrupando las anotaciones en diferentes categorías funciones. Este fichero es usado por Compatax como fuente de entrada de información y de él se extraen todas las categorías funcionales junto con el número de genes codificantes a proteína que presentan alguna anotación incluida dentro de la categoría. En la figura 5 se muestra un ejemplo del contenido de este fichero resultante de la anotación masiva del organismo *Bacillus subtilis*, simplificado de forma que solo representamos unas pocas categorías, las más importantes, por cada grupo de anotaciones.

#Annotation statistics	
Number of query sequences:	3940
With Annotations	3583
With GENENAME	3539
With DESCRIPTION	3514
With ENZYME	1904
With GO	3429
With KEYWORD	3555
With PATHWAY	581
#GO Slim	
#Category "Molecular function"	
GO:0043167 ion binding	581
GO:0003677 DNA binding	432
GO:0016491 oxidoreductase activity	393
GO:0022857 transmembrane transporter activity	288
#Category "Cellular component"	
GO:0005886 plasma membrane	978
GO:0005737 cytoplasm	563
GO:0005622 intracellular	178
GO:0005829 cytosol	173
#Category "Biological process"	
GO:0009058 biosynthetic process	912
GO:0034641 cellular nitrogen compound metabolic process	652
GO:0044281 small molecule metabolic process	476
GO:0006810 transport	326
#UniProt Pathways	
Amino-acid biosynthesis	87
Cofactor biosynthesis	66
Cell wall biogenesis	44
Purine metabolism	37
#UniProt Keyword categories	
#Category "Biological_process"	
Transport	478
Transcription	310
Sporulation	255
#Category "Cellular_component"	
Membrane	1118
Cytoplasm	485

Fig. 5. Ejemplo de salida de un proceso de anotación masiva sobre todas las secuencias pertenecientes al organismo *Bacillus subtilis* mediante la herramienta Sma3s. La información representada corresponde solo al fichero de resumen. Se representan solo las categorías funcionales con mayor número de genes anotados por cada grupo.

La estructura de este fichero comienza con una cabecera introductoria que da información sobre proceso de anotación con datos de interés como el número de secuencias que

se han intentado anotar y el número de éstas para las cuales ha tenido éxito.

Tras la cabecera se encuentra el listado de categorías funcionales dividas en varios grupos según su proveniencia (*Uniprot Pathways, Uniprot Keyword, GO Slim*) y subdivididas según su categoría (*cellular component, molecular function…)*. En la columna siguiente a cada categoría, con ordenación de mayor a menor, se enumera el número de anotaciones correspondientes a la categoría con los que se hayan anotado las diferentes secuencias del organismo y que se exponen en detalle en el fichero "completo" anteriormente citado.

3 Experimentación

Para dar un valor añadido al conjunto de anotaciones de genomas completos vamos a desarrollar una herramienta bioinformática que permita almacenarlos y realizar comparativas entre pares o grupos de ellos. Con esta herramienta, a la cual hemos llamado Compatax, intentaremos encontrar que características a nivel taxonómico son diferenciativas de un organismo respecto a aquellos más similares a él.

3.1 Funcionamiento de la herramienta

La herramienta bioinformática Compatax se desarrolla inicialmente sin interfaz gráfica siguiendo el modelo de *script* a ejecutar mediante terminal de comandos. La ejecución de la herramienta se resume en el siguiente comando:

```
perl compatax.pl -o <opción> [argumentos]
```

En los siguientes puntos se exponen en detalle las opciones con las que cuenta la primera versión de Compatax:

3.1.1 Subida de organismos modelo a la base de datos

Dado que Compatax se nutre de una base de datos, las primeras opciones a describir son las que corresponden a la subida de organismos a la misma. Hay dos opciones disponibles para ello: *upload* y *uploadmultiple*.

La funcionalidad de la opción *upload* es subir un solo organismo modelo a la base de datos. Su ejecución sería la siguiente:

```
perl compatax.pl -o upload -file <resumen.tsv> -name <name> -taxid <taxid> -source <source> [-tags <tag1,tag2,…tagn>]
```

Los argumentos de esta opción, los cuales no requieren de un orden concreto en la ejecución, corresponden a:

- **file**: ruta al archivo resumen (*summary*) proveniente de la salida de un proceso de anotación mediante Sma3s.
- **name**: nombre del organismo al que pertenece la anotación.
- **taxid**: número que corresponde con el identificador taxonómico del organismo. Este argumento es muy importante dado que la base de datos adjunta a Compatax almacenará una única anotación por organismo. Así, en el caso de especificar en este argumento un taxid ya existente, la información sobre su anotación se actualizará. En caso contrario se creará un nuevo registro.
- **source**: texto que describe de donde provienen de las secuencias del genoma al que se le ha realizado la anotación funcional. Por ejemplo "UniProt".
- **tags**: este argumento opcional sirve para asociar al organismo una o varias etiquetas, las cuales han se escribirse separadas por coma, y ayudan a catalogar el organismo. Estas etiquetas sirven de apoyo posteriormente para realizar búsquedas de organismos con alguna característica (etiqueta) concreta. Algunos ejemplos del valor de las etiquetas podrían ser "prokaryote", "bacteria", "bacillus", etc.

La otra de las opciones cuya utilidad es subir organismos a la base de datos, *uploadmultiple*, nos permite realizar la subida de múltiples organismos en una sola ejecución. El único argumento que recibe esta opción es *file*, el cual debe tomar una ruta a un fichero CSV. El contenido de este fichero corresponde, por cada fila, a los mismos argumentos que recibe la opción *upload* con un orden concreto. Por cada organismo se debe especificar (figura 6):

```
name ; taxid ; source ; tags ; file
```

	A	B	C	D	E
1	Emergomyces pasteuriana	1447872	genbank, NCBI	fungi	Emergomyces_pasteuriana_Ep9510_GCA_001883825.1_Emmo_past_UAMH9510_V1_protein_uniref90_go_goslim_statistics.tsv
2	Fibularhizoctonia sp.	436010	genbank, NCBI	fungi	Fibularhizoctonia_sp._CBS_109695_GCA_001630335.1_Fibsp1_protein_uniref90_go_goslim_statistics.tsv
3	Absidia glauca	4829	genbank, NCBI	fungi	Absidia_glauca_GCA_900079185.1_AG_v1_protein_uniref90_go_goslim_statistics.tsv
4	Phycomyces blakesleeanus	4837	genbank, NCBI	fungi	Phycomyces_blakesleeanus_NRRL_1555_GCA_001638985.2_Phybl2_protein_uniref90_go_goslim_statistics.tsv

```
1 Emergomyces pasteuriana;1447872;genbank, NCBI;fungi;Emergomyces_pasteuriana_Ep9510_GCA_001883825.1_Emmo_past_UAMH9510_V1_protein_uniref90_go_goslim_statistics.tsv
2 Fibularhizoctonia sp.;436010;genbank, NCBI;fungi;Fibularhizoctonia_sp._CBS_109695_GCA_001630335.1_Fibsp1_protein_uniref90_go_goslim_statistics.tsv
3 Absidia glauca;4829;genbank, NCBI;fungi;Absidia_glauca_GCA_900079185.1_AG_v1_protein_uniref90_go_goslim_statistics.tsv
4 Phycomyces blakesleeanus;4837;genbank, NCBI;fungi;Phycomyces_blakesleeanus_NRRL_1555_GCA_001638985.2_Phybl2_protein_uniref90_go_goslim_statistics.tsv
```

Fig. 6. Ejemplo del contenido de un fichero CSV pasado como argumento a la opción uploadmultiple de Compatax con el objetivo de subir a la base de datos 4 organismos modelo. En la parte superior se visualiza su contenido mediante una herramienta de gestión hojas de cálculo y en la parte inferior mediante un editor de texto.

3.1.2 Listar organismos modelo almacenados en la base de datos

Compatax implementa una opción llamada *list* cuya función es mostrar en pantalla el listado de organismos cuyos proteomas anotados se encuentran almacenados en la base de datos. La salida se estructura comenzando con una línea de cabecera con los nombres de los campos que se muestran y usa como separador de campos el par de caracteres "<>" (menor que junto con mayor que). La información que se muestra para cada organismo es su nombre, identificador taxonómico, proveniencia de las secuencias usadas para su anotación, fecha de subida a la base de datos de Compatax, número de genes anotados y listado de etiquetas asociadas.

Esta opción puede usarse para conocer el listado de organismos modelo, saber el identificador taxonómico de alguno en concreto, analizar cuáles contienen una etiqueta concreta, etc.

La opción *list* no recibe más argumentos. Un ejemplo de salida a la ejecución de esta opción sería, para una base de datos que solo almacenara dos organismos:

```
NAME <> TAX ID <> SOURCE <> DATE <> GENES <> TAGS
Homo sapiens <> 9606 <> Ensembl <> 2018-09-04 <> 51153 <> mammal
Tyto alba <> 56313 <> genbank, NCBI <> 2018-09-04 <> 11013 <> vertebrate
```

3.1.3 Comparar el resultado de una anotación frente a un organismo almacenado en Compatax

Tras realizar un proceso de anotación funcional sobre el conjunto completo de secuencias de un organismo con Sma3s, haciendo uso del fichero con el resumen de la anotación completa, podemos comparar esta anotación con la de cualquier organismo almacenado en la base de datos de Compatax. Para realizar esta comparación se ha implementado la opción *compare* cuya ejecución se realiza así:

```
perl compatax.pl -option compare -file <resumen.tsv>
-taxid <taxid>
```

Los argumentos que recibe corresponden a:

- **file**: ruta al archivo que contiene el resumen de la anotación, proveniente de Sma3s, del organismo que queremos comparar.
- **taxid**: identificador taxonómico del organismo almacenado en la base de datos contra el cual se va realizar la comparación.

Esta opción genera como información de salida un fichero en formato HTML el cual contiene el resultado de la comparación en detalle. Al visualizarse este fichero mediante un navegador web lo primero que se observa es una cabecera con información sobre el organismo modelo sobre el cual hemos realizado la comparativa: su nombre e identificador taxonómico, lugar de obtención de las secuencias anotadas y listado de etiquetas asociado. A la derecha de esta información podemos observar el número de genes anotados del organismo modelo frente al número de genes encontrados en la anotación que se ha usado para la comparativa, además del valor de distancia entre ambas anotaciones calculado según el algoritmo detallado en el punto 2.5 del presente documento.

Fig. 7. Visualización mediante un navegador web del fichero de salida HTML proveniente de la ejecución de la opción compare usando la anotación funcional del organismo *Bacillus altitudinis* frente a la de *Bacillus subtilis*, estando la segunda presente en la base de datos de Compatax.

Bajo la cabecera aparecen listados los grupos de anotaciones encontrados a la hora de realiza la comparativa. Si se selecciona alguno de ellos se despliega la información en detalle: en primer lugar, aparece un gráfico de barras que representa el número de genes con presencia de alguna anotación de cada una de las categorías dentro del grupo, siendo ponderado el valor entre el número de genes de cada organismo para evitar problemas con organismos poliploides. En color azul se representa el valor correspondiente al organismo modelo almacenado en la base de datos de Compatax y en amarillo el correspondiente a la anotación usada para la comparativa cuyos detalles aparecen en la cabecera del informe.

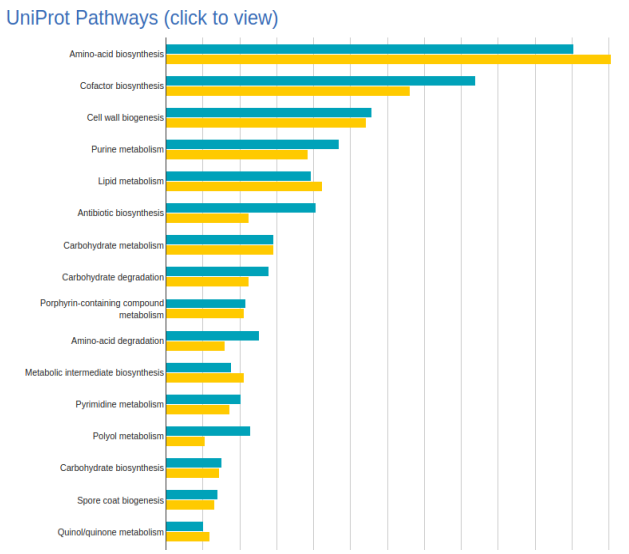

Fig. 8. Parte del gráfico de barras del grupo *Uniprot Pathways* contenido en la comparativa representada en la figura 7.

En la zona inferior de cada gráfico de barras aparece una tabla que representa los valores en mayor detalle. Esta tabla es interactiva: puede ordenarse según los valores de

cualquiera de sus columnas haciendo click en el título de la misma. La primera columna, la cual representa el nombre de la categoría funcional de anotaciones, toma su valor en color verde si los valores de la comparativa para cada registro son similares, anaranjado si son algo dispares y rojo si son muy diferentes. Esto facilita la tarea de detectar que categorías funcionales son las que presentan mayores diferencias entre los dos organismos.

Annotation	Genes in model organism	Genes in your organism	% in model organism	% in your organism	Similitude
Amino-acid biosynthesis	87	91	2.208	2.411	99.797
Cofactor biosynthesis	66	50	1.675	1.325	99.65
Cell wall biogenesis	44	41	1.117	1.086	99.97
Lipid metabolism	31	32	0.787	0.848	99.939
Purine metabolism	37	29	0.939	0.768	99.829
Carbohydrate metabolism	23	22	0.584	0.583	99.999
Carbohydrate degradation	22	17	0.558	0.45	99.892
Antibiotic biosynthesis	32	17	0.812	0.45	99.638
Metabolic intermediate biosynthesis	14	16	0.355	0.424	99.931
Porphyrin-containing compound metabolism	17	16	0.431	0.424	99.992
Pyrimidine metabolism	16	13	0.406	0.344	99.938
Amino-acid degradation	20	12	0.508	0.318	99.81
Carbohydrate biosynthesis	12	11	0.305	0.291	99.987
Spore coat biogenesis	11	10	0.279	0.265	99.986
Quinol/quinone metabolism	8	9	0.203	0.238	99.965
Amine and polyamine biosynthesis	6	8	0.152	0.212	99.94
Ketone degradation	6	8	0.152	0.212	99.94
One-carbon metabolism	6	8	0.152	0.212	99.94
Polyol metabolism	18	8	0.457	0.212	99.755
Protein modification	8	7	0.203	0.185	99.982
Siderophore biosynthesis	5	6	0.127	0.159	99.968
Isoprenoid biosynthesis	6	6	0.152	0.159	99.993
Carbohydrate acid metabolism	8	6	0.203	0.159	99.956
Sulfur metabolism	9	6	0.228	0.159	99.931
Nucleotide-sugar biosynthesis	5	5	0.127	0.132	99.994
tRNA modification	5	5	0.127	0.132	99.994
Glycan degradation	6	5	0.152	0.132	99.98
Glycan metabolism	10	5	0.254	0.132	99.879
Carotenoid biosynthesis	0	4	0	0.106	0

Fig. 9. Parte de la tabla de resultados para la gráfica de barras representada en la figura 8. Se aprecia como la categoría funcional *Carotenoid biosynthesis* aparece en color rojo dado que presenta grandes diferencias entre ambos organismos: *Bacillus altitudinis* presenta 4 genes anotados mientras que *Bacillus subtilis* ninguno.

3.1.4 Buscar organismos en Compatax cuyos conjuntos de anotaciones son similares a un conjunto de anotaciones propuesto

Tras anotar mediante Sma3s el conjunto de secuencias de un organismo es posible obtener una estimación de donde se encuentra ese organismo respecto a los almacenados en Compatax: conocer cuales tienen un conjunto de anotaciones más similar y visualizar en detalles las similitudes y diferencias. Esto serviría para clasificar filogenéticamente un proteoma, utilizando para ellos las clases funcionales de su anotación. Para ello se implementa la opción *find* cuya ejecución es:

```
perl compatax.pl -option find -file <resumen.tsv>
[-tags <tag1,tag2…,tagn> -num <5>]
```

Esta opción busca en la base de datos a aquellos organismos almacenados que presenten un mayor grado de similitud en su anotación respecto a la anotación objetivo en base al algoritmo descrito en el punto 2.5 del presente documento. Estos organismos se muestran por pantalla ordenados de forma ascendente según distancia (menor distancia significa mayor similitud en las anotaciones) y genera un fichero de salida en formato HTML. El detalle de los argumentos que recibe la opción sería el siguiente:

- **file**: ruta al archivo resumen (*summary*) de la anotación completa de las secuencias de un organismo, proveniente de Sma3s.
- **tags** (argumento opcional): conjunto de etiquetas, separadas por coma, las cuales sirven como filtro de búsqueda para los organismos modelo almacenados en base de datos, los cuales han de tener asociadas todas ellas.
- **num** (argumento opcional): número máximo de organismos que devuelve la opción. Si no se especifica la opción devolverá hasta 5 organismos como salida.

Además del listado de organismos junto con su distancia respecto a la anotación objetivo, información que se imprime visualmente sobre el terminal, la ejecución de esta opción genera un fichero de resultados donde aparecen los detalles de la comparativa de la anotación objetivo respecto a las del grupo de salida. Al visualizarlo con un navegador web observamos que este informe se esquematiza de forma similar al de la opción *compare,* aunque cambiando gráficos de barras por gráficos de cajas (*boxplot*). Estos gráficos representan la distribución de valores, en número de genes anotados, que toma cada categoría funcional: las partes exteriores llamadas comúnmente “bigotes” representan desde el valor mínimo al primer cuartil (Q1) el izquierdo, y desde el tercer cuartil (Q3) al valor máximo el derecho; el cuadro central representa la distribución de valores más comunes los cuales se encuentran entre Q1 y Q3; sobre el gráfico se representa con un punto blanco la mediana (Q2) y con uno oscuro el valor correspondiente a la anotación objetivo para esa categoría funcional. El color del gráfico es verde cuando el valor correspondiente a la anotación objetivo se encuentra entre Q1 y Q3 del grupo, anaranjado cuando el valor se encuentra entre el mínimo y Q1 o entre Q3 y el máximo, y en rojo si el valor se sale del rango del grupo (menor que el mínimo o mayor que el máximo), es decir, si se encuentra en la región de outliers. Esta gama de colores ayuda a detectar rápidamente y de forma visual las categorías funcionales de anotaciones donde el organismo al cual pertenece la anotación objetivo se diferencia del grupo de organismos modelo y, por tanto, presentar posibles características particulares con respecto al resto de organismos de su grup taxonómico.

Los valores representados en el informe de salida de esta opción pueden ser representados en número de genes anotados o en porcentaje de genes sobre el total del organismo. Se implementa la representación en valor porcentual para aquellos casos donde el grupo de organismos modelo presente un número total de genes anotados muy heterogéneo, caso que puede ocurrir si en el grupo aparecen organismos poliploides.

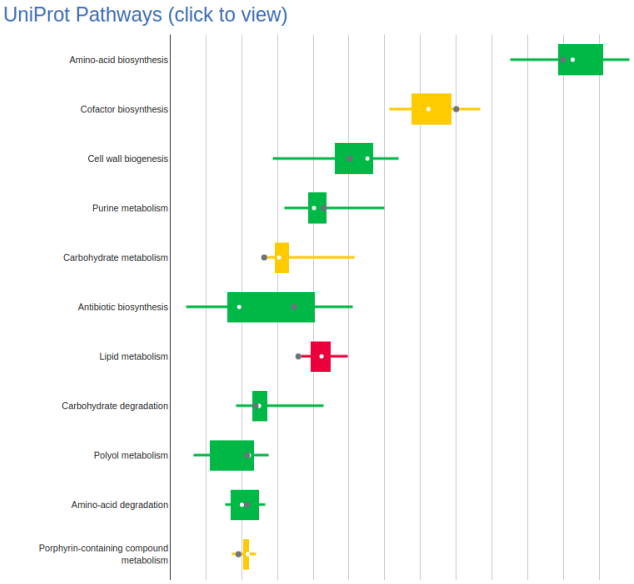

Fig. 10. Parte del informe de resultados obtenido al comparar una anotación completa del organismo *Bacillus subtilis* contra los 20 organismos modelo más similares almacenados en Compatax que tienen asociados la etiqueta "*bacillus*". En la figura se representan los primeros resultados para el grupo de anotaciones provenientes de *UniProt Pathways.* Podemos observar los distintos colores que toman los diagramas de cajas dependiendo del número de genes anotados encontrado en la anotación del proteoma. Una categoría funcional a tener en cuenta sería *Lipid metabolism* dado que aparece en color rojo debido a que el número de genes anotados en *Bacillus subtilis* es inferior al rango encontrado en el grupo de organismos modelo de Compatax. Esto podría significar que en *Bacillus subtilis* el proceso de metabolismo de lípidos esta menos desarrollado que en otros *Bacillus* similares.

3.2 Evaluación de la movilidad celular y resistencia a antibióticos en *Bacillus*

Haciendo uso de la versión 2 del anotador funcional de proteomas Sma3s, Casimiro-Soriguer et al.[14] realizaron la anotación funcional del conjunto de proteomas de 52 bacterias del género *Bacillus*. Una vez anotados los proteomas, estudiaron los resultados obtenidos en dos categorías funcionales en las que las especies de *Bacillus* se diferenciaban: *cell motility* proveniente de GO Slim y *antibiotic biosynthesis* de UniProt keywords. La primera representa la movilidad celular, la cual es una característica que se presenta de forma muy similar en todos los organismos del grupo de bacterias de estudio, excepto en 2 que carecen de ella. La segunda representa la capacidad de biosíntesis de antibióticos de estas bacterias, característica que se presenta de forma heterogénea en este grupo de organismos. Contabilizando el número de genes anotados para estas categorías funcionales en los 52 organismos de estudio, obtuvieron una comparativa que valida el buen funcionamiento del anotador Sma3s en base a las premisas esperadas.

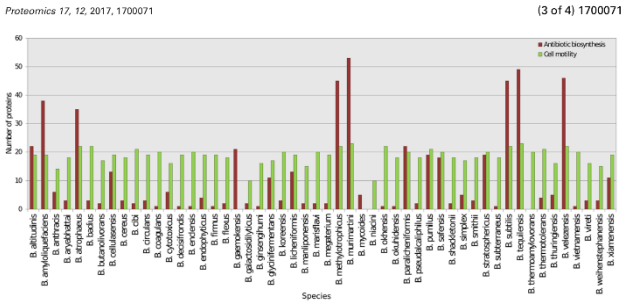

Fig. 11. Análisis de diferencias entre 2 categorías funcionales de anotaciones en un grupo de 52 organismos del género *bacillus* cuyos resultados han sido extraídos mediante una anotación masiva haciendo uso de la herramienta Sma3s[14]. Puede visualizarse con mayor resolución en http://digital.csic.es/bitstream/10261/163609/9/Sma3sfig5.pdf

En los resultados comprobaron como todos los organismos presentan un número muy similar de genes anotados con *cell motility* a *Bacillus subtilis*, organismo modelo para este género, para el cual aparece esa categoría funcional con 22 de sus genes, destacando sólo algunas bacterias como *B. gaemokensis*, la cual no presentaba genes de movilidad. Por el contrario, *antibiotic biosynthesis* presenta valores muy heterogéneos, encontrando organismos con solo 2 genes anotados a otros con más de 40, y destacando otros como *B. niacini* que no presentaba genes de síntesis de antibióticos.

Para comprobar si Compatax era capaz de encontrar fácilmente estos mismos resultas, utilizamos esta nueva herramienta para analizar el conjunto de anotaciones de los mismos organismos esperando obtener unos resultados equivalentes.

El grupo de organismos de estudio presenta un número de genes anotados por organismo bastante homogéneo con una media de 4188. Aun así, existen unos pocos organismos del grupo con un tamaño alejado de esta media y que podrían llevar a error a la hora de interpretar los resultados obtenidos. Por ejemplo, el organismo *Bacillus niacini* que solo dispone de 1922 genes anotados presenta un valor para *cell motility* porcentualmente en torno al valor medio del grupo pero en número de genes anotados se encuentra en torno a la mitad de la media, lo cual puede llevar a la confusión de determinar que este organismo presenta mucha menos movilidad celular que el resto, además de explicar el porqué esta especie carecía de genes de síntesis de antibióticos. Por ello, interpretaremos los resultados del estudio haciendo uso solo de la gráfica que representa los valores de forma porcentual.

Los datos obtenidos en el siguiente informe de resultados de confirman las mismas premisas planteadas en la publicación sobre Sma3s.

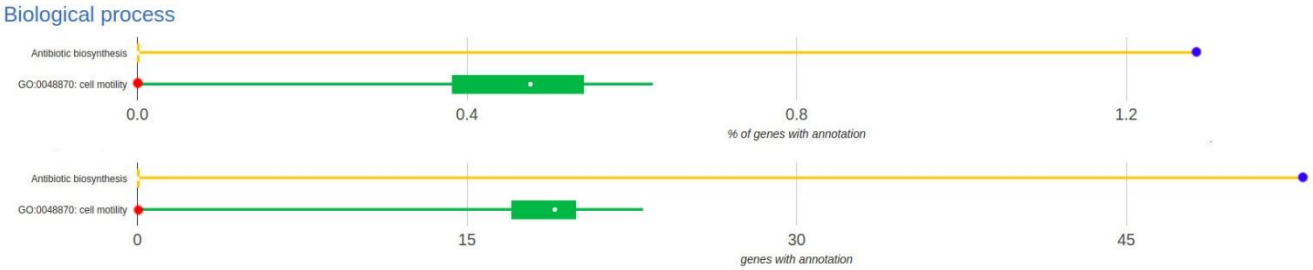

Fig. 12. Análisis de diferencias entre 2 anotaciones en 52 organismos del género *bacillus* con Compatax. Sobre la figura se han marcado con un punto de color rojo el valor correspondiente al organismo *Bacillus gaemokensis* y con un punto de color azul el valor correspondiente al organismo *Bacillus murimartini*, dado que ambos presentan valores atípicos respecto al resto de organimos del grupo.

El número de anotaciones incluidas en la categoría funcional *antibiotic biosynthesis* que presentan los genes del grupo de organismos *Bacillus* estudiado presenta un grado de dispersión mucho mayor que las anotaciones de la categoría funcional *cell motility* cuyo conjunto de valores obtenido se sitúa acotado en un rango mucho más pequeño en torno al valor medio. Esto se traduce en que el conjunto de especies presenta un grado de movilidad celular muy similar pero la síntesis de antibióticos difiere de una a otra (figura 12).

Si entramos más en detalle, el organismo *Bacillus gaemokensis* aparece como un *outlier* (valor atípico) en los resultados dado que no presenta ningún gen anotado con alguna anotación incluida en la categoría funcional *cell motility*, lo cual significa que no presenta movilidad celular alguna. Por lo tanto, Compatax ayuda a encontrar en las anotaciones completas de organismos características diferenciativas a nivel taxonómico, obteniendo los mismos resultados ya publicados con Sma3s, de una forma más específica y rápida.

3.3 Glicosilación en microalgas como factor clave en el paso a la multicelularidad

Tras la formación de los ácidos nucléicos y posteriormente las proteínas, la aparición de los glicanos se considera la tercera revolución ocurrida en la evolución biológica de la vida[15]. Los glicanos son polisacáridos que se unen a las proteínas en un proceso llamado glicosilación, el cual las modifica. Las proteínas y ácidos nucleicos se fabrican directamente a partir de plantillas de ADN. En cambio, los glicanos requieren una compleja ruta biosintética para su formación, la cual puede verse afectada por numerosos factores genéticos y ambientales. Estos factores que afectan a la formación de glicanos se ven reflejados en las proteínas a las que se unen. Por tanto, los glicanos pueden aportar a las proteínas una respuesta adaptativa a cambios en el entorno. La glicosilación es un proceso muy importante biológicamente ya que posibilita la creación de nuevas estructuras protéicas sin necesidad de que existan modificaciones en la información genética[16] [17].

Los glicanos aportan tal variedad biológica que no se contempla la existencia de organismos pluricelulares sin el proceso de glicosilación, ya que se encuentran abundantemente en proteínas de la membrana plasmática y permiten así la interacción y comunicación entre células. Para evaluar este hecho por medio de la comparación de clases funcionales, utilizamos Compatax para detectar diferencias en las anotaciones entre organismos unicelulares y pluricelulares. Para restringir la distancia evolutiva, se utilizó un conjunto de pequeñas microalgas unicelulares relacionadas evolutivamente.

TaxID	Organismo	Genes anotados
296587	*Micromonas commoda*	10.137
41875	*Bathycoccus prasinos*	7.900
554065	*Chlorella variabilis*	9.780
70448	*Ostreococcus tauri*	7.662
248742	*Coccomyxa subellipsoidea*	9.839
242159	*Ostreococcus lucimarinus*	7.603
130081	*Galdieria sulphuraria*	7.174
3055	*Chlamydomonas reinhardtii*	14.412
38833	*Micromonas pusilla*	10.269
145388	*Monoraphidium neglectum*	16.755
105231	*Klebsormidium nitens*	16.283
3075	*Auxenochlorella protothecoides*	7.014
45157	*Cyanidioschyzon merolae*	4.803

Tabla 2. Listado de algas unicelulares contra las que se va a comprar al organismo *Volvox carteri* en busca de diferencias en sus catergorías funcionales que expliquen su multicelularidad.

Compararemos el conjunto de anotaciones de los organismos de este grupo contra uno muy similar aunqeu de naturaleza pluricelular: *Volvox carteri*. Este organismo es una especie eucariótica móvil multicelular de alga verde compuesta por unas 2.000 células somáticas pequeñas y 16 células reproductoras grandes que interactúan en una matriz extracelular para formar colonias huecas y esféricas[18], siendo las células somáticas muy similares a las unicelulares iniciales, especialmente *Chlamydomonas*.

Dado que *Volvox carteri* es un organismo pluricelular esperamos encontrar en su conjunto de anotaciones ciertas características diferenciadoras que destaquen frente al grupo de organismos unicelulares, esperando encontrar entre

estas características algún indicio relacionado con el proceso de glicosilación y otros relacionados a la aparición de la pluricelularidad.

Fig. 13. Imagen microscópica de una colonia del organismo *Volvox carteri*[19]. Se puede observar como el organismo está compuesto de un grupo numeroso de pequeñas células somáticas junto con unas pocas células reproductoras de mayor tamaño.

Tal y como se esperaba, *Volvox carteri* presenta un número de genes anotados con la categoría funcional *Glycan biosynthesis* por encima del grupo de algas unicelulares (figura 14). Además, de entre las más de 460 categorías funcionales de anotaciones diferentes que ha comparado la herramienta, han destacado algunas que pueden estar relacionadas igualmente con la pluricelularidad.

muchos de los cambios químicos de los compuestos que no son necesarios para el crecimiento y el mantenimiento de las células, y son a menudo exclusivos de un taxón. En organismos multicelulares generalmente se lleva a cabo en tipos celulares específicos y puede ser útil para el organismo en su conjunto. En organismos unicelulares se usa a menudo para la producción de antibióticos o para la utilización y adquisición de nutrientes inusuales". Las anotaciones que se engloban dentro de esta categoría funcional aparecen de forma más numerosa en organismos pluricelulares donde existen células de tipos específicos, siendo menos comunes en organismos unicelulares donde el metabolismo secundario tiene un funcionamiento más básico. *Volvox carteri* presenta células específicas para el proceso reproductivo y por tanto su proceso biológico de metabolismo secundario debe ser más complejo.

Otra categoría de anotaciones que destaca en el resultado obtenido de la comparativa realizada con Compatax es la movilidad celular (*GO:0048870 – cell motility*). El grupo de unas 2.000 células somáticas que encontramos en el interior de una colonia de *Volvox carteri* le confieren de mayor soporte y movilidad[20] del esperado en los organismos del grupo de algas unicelulares.

La categoría funcional sobre la que *Volvox carteri* presenta mayores diferencias frente al grupo de organismos unicelulares en lo que a número de genes anotados se refiere es la transducción de señal (*GO:0007165: signal transduction*). Las células en organismos multicelulares coordinan sus funciones usando ciertas moléculas como señales, para las cuales existen receptores específicos en el entorno extracelular que las captan y amplifican al entorno intracelular[21]. Este proceso biológico, con menor frecuencia, puede funcionar de forma similar para la comunicación celular con otros organismos presentes en el entorno[22]. Por tanto,

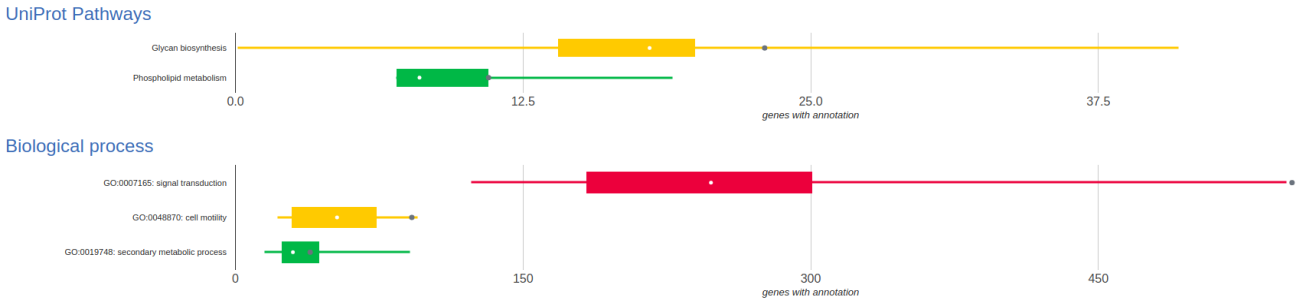

Fig. 14. Resultados más relevantes de la comparativa entre *Volvox carteri* y un grupo de algas unicelulares realizada con la herramienta Compatax. El punto de color oscuro marcado sobre los boxplot representa el valor que toma el organismo *Volvox carteri* para esa categoría funcional de anotaciones.

Una de estas categorías funcionales en la que *Volvox carteri* presenta un número de genes anotado por encima de la media del grupo de algas unicelulares es el proceso de metabolismo secundario (*GO:0019748 - secondary metabolic process*) el cual se define en QuickGO (https://www.ebi.ac.uk/QuickGO/term/GO:0019748) como "las reacciones químicas y las vías que resultan en

aunque este proceso biológico pueda observarse en organismos unicelulares cuya función en ellos es la comunicación con células de otros organismos, será en organismos multicelulares y pluricelulares donde aparezca con mayor importancia y complejidad dada la necesidad de comunicación entre las células del propio organismo, existiendo una mayor diversidad de tipos de células y señales de

comunicación necesarias para coordinas ciertos procesos biológicos.

La última de las categorías funcionales destacada en el resultado comparativo para *Volvox carteri* es la relacionada con el metabolismo de fosfolípidos (*phospholipid metabolism*). Los fosfolípidos son un tipo de lípidos anfipáticos (sus moléculas tienen un extremo hidrofílico y otro hidrófobo) compuestos por una molécula de alcohol a la que se unen dos ácidos grasos y un grupo fosfato. Los fosfolípidos son unos de los componentes principales de la membrana celular. Dado que las proteínas de la membrana controlan las interacciones entre las células de los organismos multicelulares[23], la existencia de un mayor número de genes anotados sobre esta ruta metabólica en *Volvox carteri* es un indicador más de que una de las características biológicas en que se diferencia del resto de organismos de la comparativa es la multicelularidad.

4 Conclusiones

Gracias a las nuevas tecnologías de secuenciación genómica que son cada vez más económicas, y a la alta precisión de las predicciones que ofrecen diversas herramientas bioinformáticas sobre anotación funcional, nos encontramos con la necesidad de analizar los datos provenientes de multitud de procesos de anotación funcional de forma conjunta. Por ello, se desarrolló una nueva herramienta bioinformática que, nutriéndose de la información proveniente de estos procesos de anotación funcional, ofreciera la posibilidad de realizar comparaciones y encontrar características diferenciativas a nivel taxonómico en genomas ya anotados. A esta herramienta se le ha dado el nombre Compatax.

La nueva herramienta ha sido desarrollada teniendo como objetivo la sencillez tanto en su ejecución como la interpretación de la información resultante. Por ello, se decidió trabajar con categorías funcionales de anotaciones provenientes del informe resumen que ofrece como salida la herramienta para la anotación funcional de secuencias Sma3s.

Tras una serie de experimentos realizados sobre la nueva herramienta, se confirmó su utilidad para detectar características taxonómicas diferenciativas en genomas anotados con Sma3s. Se ha conseguido analizar con éxito las diferencias en movilidad celular y resistencia ante antibióticos en un grupo de 52 bacterias del género *Bacillus*. Además, ha ayudado a detectar diversas diferencias taxonómicas relacionadas con la pluricelularidad en el organismo *Volvox carteri* al compararlo frente a un grupo de microalgas unicelulares.

Durante el desarrollo de la herramienta se ha creado una base de datos con el fin de ofrecer a la comunidad científica una serie organismos anotados que sirvan como modelo en futuros experimentos. En esta base de datos se han publicado inicialmente las anotaciones de 456 organismos diversos grupos taxonómicos como plantas (68), bacterias (155), mamíferos (18), protozoos (71), vertebrados (55) e invertebrados (89).

En términos de rendimiento Compatax destaca por su optimización, la cual le permite aprovechar en gran medida los recursos de hardware de los que disponga, ofreciendo la posibilidad de usarse incluso en ordenadores convencionales. Aun así, disponer de una base de datos con un gran número de organismos almacenados es de gran ayuda y por ello el proceso de experimentación con la herramienta se ha desarrollado haciendo uso del cluster de supercomputación C3UPO.

El tiempo de carga de los 456 organismos anotados con los que cuenta inicialmente la base de datos tuvo una duración de 261,344 segundos, lo cual equivale a una media de 573 ms para subir la anotación completa de cada organismo. Listar la información de todos los organismos almacenados tiene una duración de 481 ms. Comparar 2 el conjunto de anotaciones de organismos entre si (*Escherichia coli* vs *Klebsiella aerogenes*) tiene una duración de 142 ms, incluyendo la generación del fichero que contiene el informe de resultados. La opción más compleja computacionalmente es la de buscar el conjunto de organismos modelo más cercano a uno objetivo, la cual toma 58,817 segundos en listar la distancia de todos los organismos de la base de datos con respecto a la bacteria de *Escherichia coli*. Dado que esta opción tomará cada vez más tiempo para su ejecución a medida que la base de datos crezca en tamaño, se ha creado un sistema de etiquetado de organismos con el objetivo de reducir el número de organismos que participan en cada búsqueda.

Los tiempos de ejecución obtenidos para la ejecución de las diferentes opciones de las que dispone Compatax demuestran el gran rendimiento de la herramienta. Este rendimiento se consigue gracias al uso de computación paralela[13], permitiendo realizar tantas comparativas entre organismos de forma simultánea como núcleos de procesamiento se encuentren disponibles.

5 Agradecimientos

En primer lugar quiero agradecer su dedicación a Antonio Pérez Pulido, tutor de este Trabajo Fin de Máster, quien me ha brindado todas las herramientas necesarias y me ha

guiado de la mejor forma posible. Ha sido un honor trabajar junto a alguien con tan valiosos conocimientos.
Debo agradecer el apoyo ofrecido por el C3UPO que ha facilitado los recursos del clúster de supercomputación (*We thank C3UPO for the HPC support*).

Este trabajo ha sido posible gracias al esfuerzo y profesionalidad mostrada por todo el equipo docente que ha participado tanto en el Máster en Análisis Bioinformático Avanzado como en el Cursos de Especialización en Análisis Bioinformático.

6 Referencias y bibliografía

[1] Totty Michael (2005). "A Better Idea". The Wall Street Journal. Avaiable at: https://www.wsj.com/articles/SB112975757605373586

[2] Sanger F; Coulson AR (1975). "A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase". Journal of Molecular Biology, volume 94, issue 3, pages 441-448. doi: 10.1016/0022-2836(75)90213-2

[3] Internatinal Human Genome Sequencing Consortium (2001). "Initial sequencing and analysis of the human genome". Nature, volume 409, pages 860–921. doi: 10.1038/35057062

[4] National Human Genome Research Institute. "The Human Genome Project Completion: Frequently Asked Questions". Available at: https://www.genome.gov/11006943/human-genome-project-completion-frequently-asked-questions

[5] Wetterstrand KA. DNA Sequencing Costs. "Data from the NHGRI Genome Sequencing Program (GSP)". Available at: www.genome.gov/sequencingcostsdata

[6] UniProt. "Current release estatistics". Avaiable at: https://www.ebi.ac.uk/uniprot/TrEMBLstats

[7] UniProt. "Reducing proteome redundancy". Avaiable at: https://www.uniprot.org/help/proteome_redundancy

[8] NCBI-NIH. "FASTA format". Avaiable at: https://blast.ncbi.nlm.nih.gov/Blast.cgi?CMD=Web&PAGE_TYPE=BlastDocs&DOC_TYPE=BlastHelp

[9] The Gene Ontology Consortium, Ashburner M, Ball CA, et al. (May 2000). "Gene ontology: tool for the unification of biology". Nature genetics, volume 25, pages 25-29. doi: 10.1038/75556

[10] Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) "Basic local alignment search tool". Journal of Molecular Biology, volume 215, issue 3, pages 403-410. doi: 10.1016/S0022-2836(05)80360-2

[11] Antonio Muñoz-Mérida Enrique Viguera M. Gonzalo Claros Oswaldo Trelles Antonio J. Pérez-Pulido (2014). "Sma3s: a three-step modular annotator for large sequence datasets.". DNA Research, volume 21, issue 4, pages 341–353. doi: 10.1093/dnares/dsu001

[12] A. A. Abd El-Aziz and A. Kannan (2014). "JSON encryption". International Conference on Computer Communi-cation and Informatics, Coimbatore, pp. 1-6. doi: 10.1109/ICCCI.2014.6921719

[13] Gottlieb, Allan; Almasi, George S. (1989). "Highly parallel computing". Redwood City, Calif.: Benjamin/Cummings. ISBN 0-8053-0177-1

[14] Carlos S. Casimiro-Soriguer, Antonio Muñoz-Mérida, Antonio J. Pérez-Pulido (2017). "Sma3s: A universal tool for easy functional annotation of proteomes and transcriptomes". Proteomics, volume 17, issue12. doi: 10.1002/pmic.201700071

[15] Gordan Lauc, Jasminka Krištić, Vlatka Zoldoš (2014). "Glycans: the third revolution in evolution". Frontiers in Genetics, 5, 145. doi: 10.3389/fgene.2014.00145

[16] Lauc G., Zoldoš V. (2010). "Protein glycosylation – an evolutionary crossroad between genes and environment". Mol. Biosyst. 6, 2373–2379. doi: 10.1039/c0mb00067a

[17] Lauc G., Huffman, J. E., Pucic M., Zgaga L., Adamczyk B., Muzinic A. (2013). "Loci associated with N-glycosylation of human immunoglobulin G show pleiotropy with autoimmune diseases and haematological cancers". PLoS Genet. 9:e1003225. doi: 10.1371/journal.pgen.1003225

[18] Prochnik, Simon E et al. (2010) "Genomic Analysis of Organismal Complexity in the Multicellular Green Alga Volvox Carteri." Science vol. 329, issue 5988, pages 223-226. doi: 10.1126/science.1188800

[19] Benjamin Klein, Daniel Wibberg, Armin Hallmann (2017). "Whole transcriptome RNA-Seq analysis reveals extensive cell type-specific compartmentalization in Volvox carteri". BMC Biology, volume 15, page 111. doi: 10.1186/s12915-017-0450-y

[20] David L Kirk (1988). "The Ontogeny and Phylogeny in Volvox". TIG Reviews, volume 4, issue 2, pages 32-36

[21] Richard A. Firtel (1991): "Signal transduction pathways controlling multicellular development in Dictyostelium". Trends in Genetics, volume 7, issues 11–12, pages 381-388. doi: 10.1016/0168-9525(91)90260-W

[22] "Cell Signaling". Scitable of Nature Education. Avaiable at: https://www.nature.com/scitable/topicpage/cell-signaling-14047077

[23] Cooper GM (2000): "The Cell: A Molecular Approach. 2nd edition". Sunderland (MA), Sinauer Associates, Cell Membranes. Available at: https://www.ncbi.nlm.nih.gov/books/NBK9928/