Automatic BLAST for Massive Sequencing - ABMS

Nelson Enrique Vera Parra
Grupo de Investigación GICOGE - Docente de Planta
Universidad Distrital Francisco José de Caldas, Bogotá D.C., Colombia,
neverap@udistrital.edu.co

Cristian Alejandro Rojas Quintero
Grupo de Investigación GICOGE - Estudiante
Universidad Distrital Francisco José de Caldas, Bogotá D.C., Colombia,
carojasq@correo.udistrital.edu.co

Steven Sierra Forero
Grupo de Investigación GICOGE - Estudiante
Universidad Distrital Francisco José de Caldas, Bogotá D.C., Colombia,
sssierra@correo.udistrital.edu.co

Trabajo realizado con la colaboración del Centro de Computación de Alto Desempeño (**CECAD**) de la Universidad Distrital Francisco José de Caldas, Bogotá D.C., Colombia. (http://cecad.udistrital.edu.co) y el Instituto de Genética de la Universidad Nacional (**IGUN**), Colombia, (http://www.genetica.unal.edu.co).

RESUMEN

En este artículo se presenta y evalúa ABMS (Automatic BLAST for Massive Sequencing) una herramienta bioinformática libre, en línea, diseñada con el objetivo de automatizar y optimizar el proceso de búsqueda mediante alineamientos locales para la comparación de grandes volúmenes de secuencias desconocidas de nucleótidos o aminoácidos contra bases de datos de secuencias conocidas (Swissprot, Uniprot, Refseq, entre otras). ABMS integra los procesos de: gestión de secuencias de entrada, gestión de bases de datos de proteomas, genómas y transcriptómas, ejecución de BLAST (blastp, blastx, blastn, tblastn) y administración de resultados; y los presenta al biologo cómo un proceso unificado, transparente y muy amigable mediante una interfaz web. ABMS está constituido por los siguientes modulos: SM (Sequence manager), LBS (Local BLAST Server), SDBA (Sequence database administrator), RM (Results manager). Al comparar ABMS frente al servidor de BLAST de NCBI utilizando 10, 20, 40, 100, 300, 500 secuencias pertenecientes al transcriptoma de Diploria Strigosa se evidenció la fortaleza de ABMS para el análisis masivo de secuencias y las limitaciones de NCBI BLAST para data sets de mas de 20 secuencias. También se observaron ventajas de ABMS frente NCBI BLAST en cuanto a administración y almacenamiento de data sets y gestión, descarga y retroalimentación de resultados.

Palabras claves: Anotación de secuencias, Alineamiento de secuencias, BLAST, Biopython, Genomas y Transcriptómas.

ABSTRACT

This article presents and evaluate ABMS (Automatic Blast for massive sequencing) a *free* online bioinformatic tool designed with the purpose to automate and optimize the search process through local aligments for big volumes comparison of unknown nucleotide sequences or aminoacids against well known sequence databases (Swissprot,Uniprot,Refseq, among others). ABMS integrates processes of : management of input sequences, proteome databases management, genome and transcriptome, BLAST execution (blastp,blastx,blastn,tblastn) and

results management and presents to the biologist as an unified process, transparent and friendly through a web interface. ABMS is composed by the following modules: SM(Sequence Manager), LBS(Local Blast Server), SDBA(Sequence database administrator), RM (Results Manager). Comparing ABMS with the local NCBI's BLAST server using 10, 20, 40, 100, 200, 500 sequences belonging to the Diploria Strigosa transcriptome we found the strenght of ABMS for the massive sequence analysis and the limitations of NCBI's BLAST for datasets with more than 20 sequences. We observed advantages from ABMS against NCBI BLAST related to management and storage of datasets, download and feedback of results.

Keywords: Sequence annotation, sequence alignment, BLAST, Biopython, Genome and Transcriptome.

1. INTRODUCCIÓN

Descifrar las secuencias biológicas es esencial para prácticamente todas las ramas de investigación de la Biología. Durante varias décadas el proceso de secuenciación se realizó gracias al método de Sanger (incluyendo el proyecto del genoma humano, donde éste método fue fundamental). Sin embargo sus altos costos y limitantes en cuanto a rendimiento, escalabilidad, velocidad y resolución han forzado a que en los últimos 5 años se lleve un proceso de migración a nuevos procedimientos denominados "secuenciación de nueva generación"(Metzker, 2010),(Martin et al., 2011). Estas nuevas tecnologías permiten una secuenciación mucho más económica y eficiente, lo cual ha generado un crecimiento exponencial en el volumen de datos secuenciados.

Optimizar el proceso de secuenciación no tendría sentido si no se acompañara con el desarrollo y optimización de herramientas informáticas capaces de analizar este inmenso volumnen de datos secuenciados. Una de las principales necesidades correspondientes a la minería de datos genómicos y transcriptómicos es la comparación de secuencias mediante alineamientos para la búsqueda de secuencias similares en bases de datos de secuencias conocidas, esto es denominado anotación (asociación de secuencias no conocidas con secuencias conocidas). La herramienta más utilizada para la comparación de secuencias mediante alineamientos es BLAST - Basic Local Alignment Search Tool (Altschul et al., 1990), (Madden et al., 1996), (Camacho et al., 2008).

Cuando se trabaja con conjuntos pequeños de secuencias el proceso de anotación normalmente se puede realizar sobre el servidor BLAST ofrecido por NCBI (National Center for Biotechnology Information), sin embargo cuando el biólogo requiere hacer procesos de anotación masivos (grandes volúmenes de secuencias) o cuando desea comparar contra bases de datos diferentes a las ofrecidas allí, ya éste servidor no es el adecuado y el biólogo se ve obligado a hacer el proceso de anotación de forma local, lo cual implica que tenga una alta experticia en la instalación, configuración y ejecución (por línea de comandos) de BLAST, en la importación y actualización de bases de datos de nucleótidos y aminoácidos; y por último en la interpretación de los resultados, normalmente en formato XML. La herramienta expuesta en este documento, denominada ABMS (Automatic BLAST for Massive Sequencing) presenta la anotación masiva al biólogo como un proceso unificado de muy fácil manejo mediante una interfaz web. ABMS optimiza el uso de BLAST para hacer anotaciones masivas e integra facilidades como la administración de bases de datos y la administración de los resultados.

Este documento se organiza en 4 secciones: inicialmente se presenta una descripción general de ABMS y se exponen sus funcionalidades, luego se describe su arquitectura, explicando módulo a módulo: SM (Sequence manager), LBS (Local BLAST Server), SDBA (Sequence database administrator), RM (Results manager), posteriormente se realiza una comparación de procesos y funcionalidades frente al servidor BLAST de NCBI y finalmente se obtienen las conclusiones.

2. DESCRIPCIÓN GENERAL

ABMS es una herramienta libre (de acceso en línea y con posibilidad de instalación local) que facilita el proceso de anotación masiva para transcriptomas, genomas y proteínas. ABMS está optimizado para trabajar con grandes cantidades de secuencias y acondicionado con interfaces de usuario y procesos muy intuitivos para el biólogo.

Figura 1: Pantalla de bienvenida ABMS

3. ARQUITECTURA

ABMS está compuesto por 4 módulos los cuales forman un flujo de trabajo. Para la ejecución y la integración de estos módulos se requieren algunas herramientas adicionales que se usan transversalmente en todos los módulos. A continuación se muestran los 4 módulos, el flujo de trabajo y las herramientas transversales.

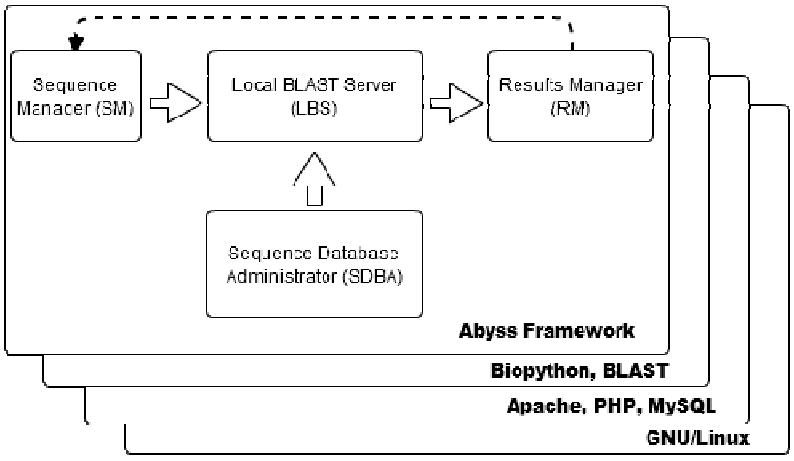

Figura 2: Estructura y flujo de trabajo para ABMS

3.1 COMPONENTES DE SOFTWARE TRANSVERSALES

- **Abbys Framework:** Es un marco de trabajo para desarrollo de aplicaciones web utilizando el lenguaje de programación PHP y bases de datos relacionales como MySQL creado por Steven Sierra Forero estudiante de la Universidad Distrital Francisco José de Caldas. Utiliza arquitectura basada en capas llamada MVC (Modelo, Vista y Controlador), y hace uso de librerias especificas para la interconexión con scripts hechos en otros lenguajes.
- **Biopython:** Proyecto de licencia libre que ofrece varios módulos para hacer más facil el trabajo con datos bioinformáticos.(Cock et al., 2009).
- **Apache:** Es un servidor HTTP de licencia libre.
- **BLAST(Basic Search Alignment Tool**): Herramienta para encontrar regiones locales de similaridad mediante alineamientos de secuencias.
- **MySQL:** Sistema de gestión de bases de datos relacionales, multihilo y multiusuario. También de licencia libre.

- **GNU/Linux:** Sistema operativo libre el cual es óptimo para servidores y para ejecutar herramientas bioinformáticas.

3.2 SM (SEQUENCE MANAGER)

Figura 3: Diagrama del módulo Sequence Manager

Este módulo es el encargado de gestionar los archivos FASTA que contienen las secuencias pertenecientes al data set del usuario. Mediante éste módulo, el usuario puede subir sus secuencias, seleccionar las que requiera suministrar a BLAST y también eliminar las que ya no necesite. Los elementos que componen éste módulo son:

- **FastaToDB.py:** Script que almacena los identificadores y las secuencias en formato FASTA en una nueva tabla MySQL mediante el uso del módulo SeqIO de Biopython.
- **Archivo FASTA de entrada:** Archivo Fasta externo subido por un usuario o un resultado de una ejecución anterior (desde el módulo Results Manager).
- **Resource Manager:** Controlador del core de Abbys Framework que se encarga de gestionar todos los recursos de entrada y salida del sistema, permite la creación, eliminación y actulización de cualquier archivo externo en el sistema.
- **Upload Controller:** Controlador externo de Abbys Framework que se encarga de presentar la interfaz directa al administrador de recursos para que el usuario suba los archivos .fasta al sistema.

3.3 LBS (LOCAL BLAST SERVER)

Este módulo se encarga de ejecutar una búsqueda con BLAST contra una(s) base(s) de dato(s) determinada(s) y almacenar los resultados en una tabla MySQL. Mediante este módulo, el usuario puede escoger los archivos de secuencias que previamente subió al servidor y seleccionar las bases de datos (se incluyen las más populares como RefSeq, UniProt, Swissprot) contra las cuales requiere realizar la búsqueda de sus secuencias. Una vez la búsqueda haya sido ejecutada los resultados se almacenan en una tabla MySQL para posteriormente hacer una gestión más óptima del mismo. Los elementos que componen este módulo son:

- **BlastExecAndSendToDb.py:** Se encarga de ejecutar BLAST (blastp o blastx). En caso de que se seleccionen varias bases de datos para la ejecución de un análisis, este script gestiona por hilos el número de búsquedas al tiempo de tal manera que no exceda la capacidad del servidor. Una vez está la busqueda ejecutada se procede a almacenarla en una tabla mediante el script BlastXMLtoDB.py.

- **BlastXMLtoDB.py:** Se encarga de almacenar la salida de BLAST en una tabla MySQL de tal manera que la visualización sea mucho mas sencilla para el usuario. Este script funciona mediante el uso del módulo Bio.Blast.NCBIXML de Biopython.
- **DBs disponibles:** Listado de bases de datos de sencuencias disponibles actualmente en el servidor.
- **Archivos FASTA:** Archivos de secuencias provenientes del módulo Sequence Manager.
- **NCBI BLAST:** BLAST instalado localmente.

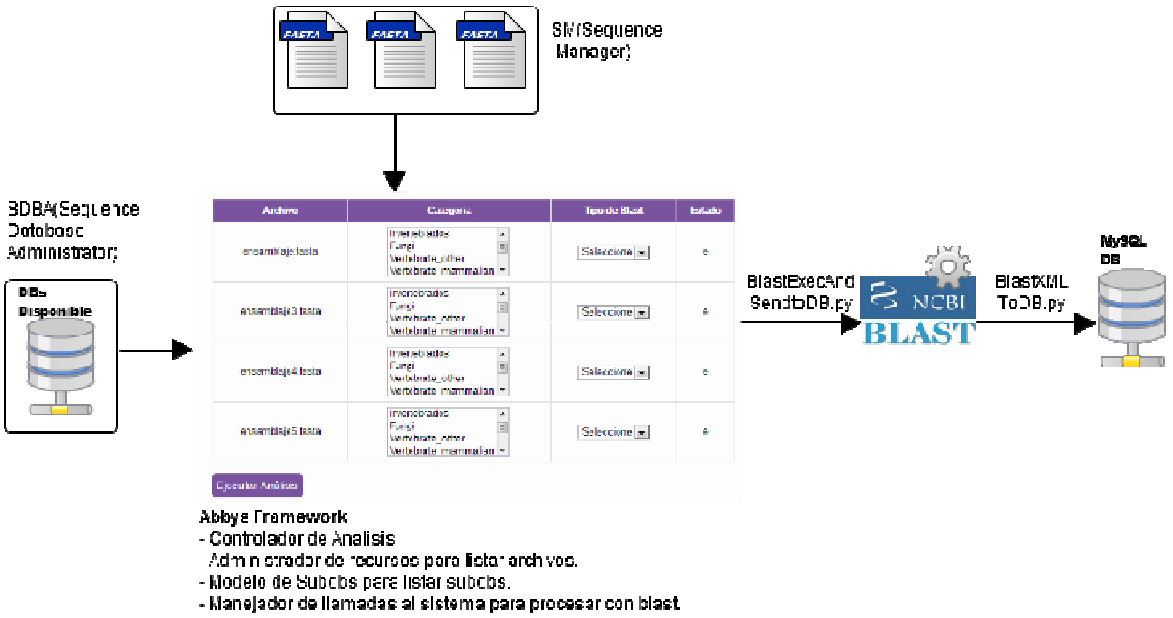

Figura 4: Diagrama del módulo Local BLAST Server

3.4 SDBA (SEQUENCE DATABASE ADMINISTRATOR)

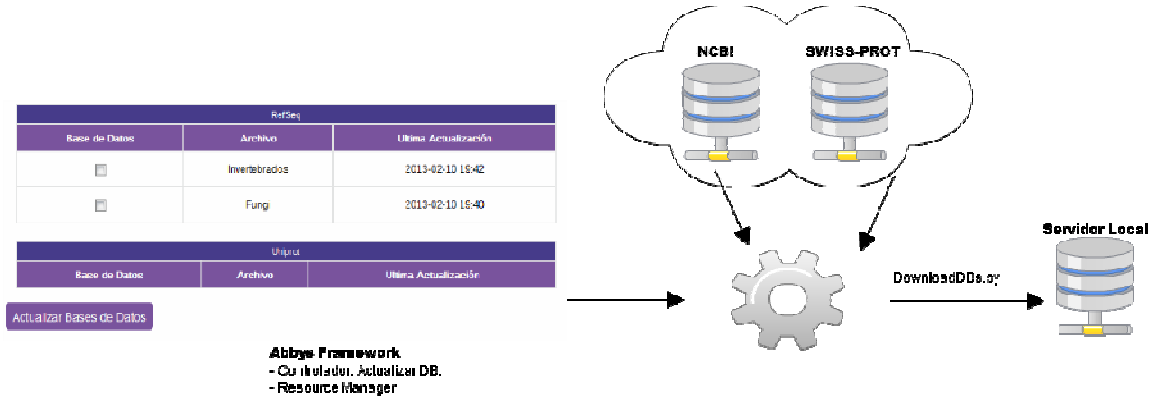

Figura 5: Diagrama del módulo Sequence DataBase Administrator

Este módulo es el encargado de gestionar (adición y actualización) las bases de datos disponibles para que los usuarios puedan ejecutar sus análisis. Mediante este módulo el administrador de ABMS es capaz de agregar nuevas bases de datos de secuencias de una manera muy intuitiva y clasificándolas por su origen y su categoría (taxonomía). Este módulo también es el encargado de actualizar las bases de datos que ya están disponibles en el servidor a medida que el administrador del mismo lo requiera. Los elementos que componen este módulo son:

- **DownloadDbs.py:** Gestiona la actualización (descarga, descompresión, generación de índices para BLAST) de la bases de datos y actualiza las últimas fechas en las que fue actualizado.

- **Controlador de Descarga:** Se encarga de hacer la llamada al script de actualización a través de la interfaz web, toma el listado de bases de datos seleccionados y ejecuta la llamada al sistema.
- **Servidores de NCBI y Swiss-Prot:** Disponibles para descargar bases de datos de proteínas para guardar en el servidor.
- **Controlador de Datos:** Módulo de Abbys Framework encargado de tomar el listado de bases de datos que serán actualizados en el sistema y enviarlo al manejador de llamadas del sistema para ejecutar el script de carga de base de datos.

3.5 RM (RESULTS MANAGER)

Este módulo es el encargado de gestionar (consulta, búsqueda, filtro, descarga) los resultados producto del análisis con BLAST. Mediante este módulo el usuario de ABMS es capaz de filtrar los resultados del análisis, descargar las secuencias en formato FASTA con la posibilidad de reemplazar los identificadores o incluso generar un nuevo archivo FASTA para enviarlo de nuevo al módulo Sequence Manager de tal manera que sea posible volver a ejecutar un análisis sobre un resultado en específico. Los elementos que componen este módulo son:

- **GenFasta.py:** Genera un archivo FASTA a partir de los resultados de una búsqueda. Este script se encarga de almacenar en la ubicación apropiada (para volver a procesar o para descargar un nuevo FASTA), y también proveee la posibilidad de reemplazar identificadores de acuerdo al resultado de la búsqueda hecha por BLAST.
- **GenCSV.py:** Genera un archivo separado por comas a partir de una búsqueda previamente ejecutada por el usuario.

Figura 6: Diagrama del módulo Results Manager

4. METODOLOGÍA DE EVALUACIÓN

4.1 DATA SET

- **Organismo:** Diploria Strigosa.
- **Tipo de secuencias:** Transcriptómicas.
- **Número de secuencias:** 10, 20, 40, 100, 300, 500.
- **Formato:** FASTA.
- **Base de datos para ejecutar búsqueda:** Non-Redundant. (Pruitt et al 2007)
- **Expect Value:** 1e-3.
- **Tipo BLAST:** blastx (Nucleótidos vs Proteínas).

4.2 MÉTRICAS

- **Desempeño:** Tiempo de procesamiento, Número de resultado
- **Funcionalidades:** Gestion de secuencias de entrada, Gestión de resultados, Ejecución de análisis, Gestión de bases de datos (personalización).

4.3 REFERENCIA DE EVALUACIÓN

Actualmente no existe un software público reconocido que ofrezca las mismas funcionalidades que ABMS, sin embargo, se tomó como patrón de referencia el servidor BLAST público de NCBI (Altschul et al., 1997) debido a que es el más usado y el que tiene más funciones similares.

5. RESULTADOS Y ANÁLISIS

5.1 RESULTADOS DE DESEMPEÑO

Número de secuencias		10	20	40	100	300	500
Tiempo (Minutos)	NCBI BLAST	19	47	CPU Limit	CPU Limit	CPU Limit	CPU Limit
	ABMS	26	51	115	259	814	1427
Número de Resultados	NCBI BLAST	5	5	CPU Limit	CPU Limit	CPU Limit	CPU Limit
	ABMS	5	5	20	42	138	247

Tabla 1: Resultados de comparacion de NCBI BLAST y ABMS

En la tabla anterior se evidencia la limitación del servidor público de NCBI BLAST para análisis masivo de secuencias; para el data set usado en esta evaluación la máxima ejecucion permitida fue para un número de secuencias de 20. Por el contrario, ABMS ejecutó análisis hasta 500 secuencias sin presentar problema.

En cuanto al número de resultados se evidenció que para el data set usado la cantidad de hits son similares tanto para NCBI BLAST como para ABMS.

Para data sets de secuencias menores e iguales a 20 se notó que el tiempo de ejecución promedio es el siguiente: 2.1 minutos por secuencia para NCBI BLAST y 3.6 minutos por secuencia para ABMS.

5.2 RESULTADOS DE FUNCIONALIDADES

Gestión de secuencias de entrada: Para la gestión de secuencias de entradas tanto ABMS como el servidor de BLAST de NCBI ofrecen una interfaz de usuario de fácil manejo para subir los archivos FASTA. El servidor NCBI BLAST cuenta con una opción adicional (que ABMS no cuenta), un campo de texto para ingresar manualmente las secuencias. ABMS permite almacenar las secuencias en el servidor de tal manera que el usuario puede contar con su data set para análisis futuros (NCBI BLAST no cuenta con esta funcionalidad).

Ejecución de análisis: El servidor de NCBI BLAST permite la selección de solo un archivo FASTA y una sola base de datos de búsqueda, mientras que ABMS facilita la elección de varios archivos de entrada FASTA y varias bases de datos de búsqueda al tiempo. ABMS habilita al usuario para suministrar sólo un parámetro al algoritmo BLAST: Expect Value; por el contrario, el servidor NCBI BLAST permite introducir una mayor cantidad de parámetros.

Gestión de resultados: ABMS ofrece una interfaz que permite consultar los resultados de la búsqueda de todas las secuencias en una sola interfaz unificada, NCBI BLAST permite la consulta de los resultados de la búsqueda sólo secuencia por secuencia. Adicionalmente ABMS facilita el proceso de exportación y filtrado de las secuencias con las que se encontró similaridad, permitiendo al usuario descargarlas e incluso volver a ejecutar otro análisis sobre las mismas. ABMS permite la exportación de los resultados en un formato CSV el cual es de fácil interpretación para el biólogo.

Gestión de bases de datos (Personalización): ABMS permite al administrador agregar nuevas bases de datos, incluso de organismos no publicados para realizar la búsqueda con BLAST, también provee la posiblidad de

actualizar estas bases de datos sin necesidad de hacer complejas operaciones. Por el contrario, NCBI BLAST sólo provee una lista de bases de datos definidas e inalterables las cuales siempre están actualizadas.

6. CONCLUSIONES

ABMS es una herramienta de anotación automática optimizada para trabajar con grandes cantidades de secuencias y acondicionada con interfaces de usuario y procesos muy intuitivos. Su proyección de utilidad para investigadores biólogos es alta debido a que las herramientas y servidores públicos actuales reconocidos de BLAST presentan limitación a data sets grandes (en este caso NCBI BLAST se limitó a 20 secuencias).

ABMS presenta facilidades al usuario tanto en la administración y almacenamiento de su data set como en la interpretación y descarga de los resultados. Esto evita al biólogo la necesidad de realizar procesos mediante línea de comandos.

Los servidores BLAST públicos reconocidos no ofrecen la posibilidad de agregación de nuevas bases de datos diferentes a las ofrecidas en sus listas. ABMS permite al administrador personalizar el conjunto de bases de datos agregando nuevas, incluso de organismos no publicados para realizar la búsqueda con BLAST.

REFERENCIAS

Altschul, S.F., Gish, W., Miller, W., Myers, E.W. &Lipman, D.J. (1990) "Basic local alignment search tool." J. Mol. Biol. 215:403-410. PubMed

Altschul, Thomas L. Madden, Alejandro A. Schäffer, Jinghui Zhang, Zheng Zhang, Webb Miller, and David J. Lipman (1997), "Gapped BLAST and PSI-BLAST: a new generation of protein database search programs", Nucleic Acids Res. 25:3389-3402.

Camacho C., Coulouris G., Avagyan V., Ma N., Papadopoulos J., Bealer K., & Madden T.L. (2008) "BLAST+: architecture and applications." BMC Bioinformatics 10:421. PubMed

Cock PJ, Antao T, Chang JT, Chapman BA, Cox CJ, Dalke A, Friedberg I, Hamelryck T, Kauff F, Wilczynski B, de Hoon MJ, Biopython: freely available Python tools for computational molecular biology and bioinformatics.Bioinformatics. 2009 Jun 1; 25(11): 1422-3

Madden, T.L., Tatusov, R.L. & Zhang, J. (1996) "Applications of network BLAST server" Meth. Enzymol. 266:131-141. PubMed

Martin, j. & Wang, z. Next-generation transcriptome assembly. Nature Reviews Genetics, 12: 671-682. 2011.

Metzker, M. Sequencing technologies – the next generation. Nature Reviews Genetics; 11:31-46. 2010.

Pruitt, K. D., Tatusova, T., & Maglott, D. R. (2007). NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. Nucleic acids research, 35(suppl 1), D61-D65.

Authorization and Disclaimer

Authors authorize LACCEI to publish the paper in the conference proceedings. Neither LACCEI nor the editors are responsible either for the content or for the implications of what is expressed in the paper.