Búsqueda de Enzimas en Metagenomas

2 de agosto de 2018

1. Resumen

El objetivo de este proyecto es realizar una búsqueda de enzimas en metagenómas. Para esto, se va a seguir un protocolo general de búsqueda y análisis bioinformático que consiste en una serie de pasos en los cuales va a utilizar diferentes recursos bioinformáticos (bases de datos biológicas, herramientas web, programación, herramientas de línea de comandos) y que le van a permitir trabajar y analizar secuencias biológicas de genes, proteínas, y motivos.

El trabajo se va a llevar a cabo de dos maneras: una forma manual a través de aplicaciones y herramientas disponibles en los sitios web de bioinformática; y una forma automática a través de la creación de una tubería o *pipeline* que realice todos los pasos del protocolo usando herramientas de línea de comandos y *scripts* o programas desarrollados por usted.

2. Protocolo

2.1. Búsqueda en Bases de Datos Biológicas

Primero se debe realizar la búsqueda de la enzima en una base de datos (BD) de secuencias biológica, específicamente en la BD del NCBI utilizando los siguientes filtros:

- Solo secuencias de Proteínas
- Secuencias de la BD RefSeq
- Sólo proteínas de bacterias
- Sólo secuencias de tamaño menor que 300 pb

Los resultados se deben guardar en formato Fasta.

2.2. Filtrado Manual de las Secuencias en Línea

Una vez tenga las secuencias en un archivo Fasta, debe realizar un *script* o utilizar una herramienta de línea de comandos para filtrar las secuencias de enzimas cuyo nombre inicie con "WP".

2.3. Multialineamiento de Secuencias

Con las secuencias resultantes del último filtrado, se debe ahora alinear, es decir realizar un multialineamiento con el objetivo de construir la historia de la evolución de estás enzimas. El resultado de este multialineamiento se va a a usar para construir un árbol filogenético que nos describe esta historia y así ver cuál o cuales es son las enzimas ancestrales.

2.4. Descargar Metagenoma

Descargar un metagenoma determinado (https://www.ebi.ac.uk/metagenomics/). El metagenoma puede ser:

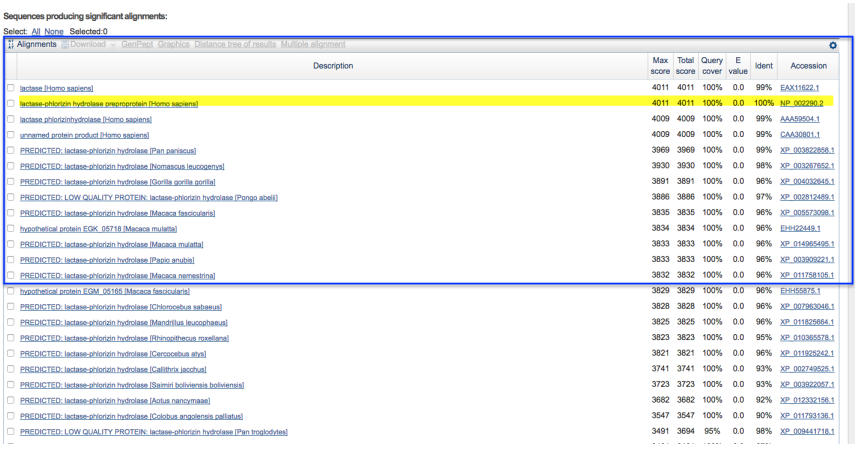
- Terrestre
- De agua dulce

- De agua salada
- Otros

2.5. Alineamientos Locales usando BLAST

Del árbol filogenético anterior se selecciona la enzima más ancestral y se utiliza la herramienta BLAST para buscar regiones similares entre la enzima y las secuencias del metagenoma.

De la lista de resultados de BLAST se van a seleccionar las primeros (50 a 100) resultados.



Sequences producing significant alignments:

Select: All None Selected:0

Alignments | Download | GenPept | Graphics | Distance tree of results | Multiple alignment

Description	Max score	Total score	Query cover	E value	Ident	Accession
lactase [Homo sapiens]	4011	4011	100%	0.0	99%	EAX11622.1
lactase-phlorizin hydrolase preproprotein [Homo sapiens]	4011	4011	100%	0.0	100%	NP_002290.2
lactase phlorizinhydrolase [Homo sapiens]	4009	4009	100%	0.0	99%	AAA59504.1
unnamed protein product [Homo sapiens]	4009	4009	100%	0.0	99%	CAA30801.1
PREDICTED: lactase-phlorizin hydrolase [Pan paniscus]	3969	3969	100%	0.0	99%	XP_003822858.1
PREDICTED: lactase-phlorizin hydrolase [Nomascus leucogenys]	3930	3930	100%	0.0	98%	XP_003267652.1
PREDICTED: lactase-phlorizin hydrolase [Gorilla gorilla gorilla]	3891	3891	100%	0.0	96%	XP_004032645.1
PREDICTED: LOW QUALITY PROTEIN: lactase-phlorizin hydrolase [Pongo abelii]	3886	3886	100%	0.0	97%	XP_002812489.1
PREDICTED: lactase-phlorizin hydrolase [Macaca fascicularis]	3835	3835	100%	0.0	96%	XP_005573098.1
hypothetical protein EGK_05718 [Macaca mulatta]	3834	3834	100%	0.0	96%	EHH22449.1
PREDICTED: lactase-phlorizin hydrolase [Macaca mulatta]	3833	3833	100%	0.0	96%	XP_014965495.1
PREDICTED: lactase-phlorizin hydrolase [Papio anubis]	3833	3833	100%	0.0	96%	XP_003909221.1
PREDICTED: lactase-phlorizin hydrolase [Macaca nemestrina]	3832	3832	100%	0.0	96%	XP_011758105.1
hypothetical protein EGM_05165 [Macaca fascicularis]	3829	3829	100%	0.0	96%	EHH55675.1
PREDICTED: lactase-phlorizin hydrolase [Chlorocebus sabaeus]	3828	3828	100%	0.0	96%	XP_007963046.1
PREDICTED: lactase-phlorizin hydrolase [Mandrillus leucophaeus]	3825	3825	100%	0.0	96%	XP_011825664.1
PREDICTED: lactase-phlorizin hydrolase [Rhinopithecus roxellana]	3823	3823	100%	0.0	95%	XP_010365578.1
PREDICTED: lactase-phlorizin hydrolase [Cercocebus atys]	3821	3821	100%	0.0	96%	XP_011925242.1
PREDICTED: lactase-phlorizin hydrolase [Callithrix jacchus]	3741	3741	100%	0.0	93%	XP_002749525.1
PREDICTED: lactase-phlorizin hydrolase [Saimiri boliviensis boliviensis]	3723	3723	100%	0.0	93%	XP_003922057.1
PREDICTED: lactase-phlorizin hydrolase [Aotus nancymaae]	3682	3682	100%	0.0	92%	XP_012332156.1
PREDICTED: lactase-phlorizin hydrolase [Colobus angolensis palliatus]	3547	3547	100%	0.0	90%	XP_011793136.1
PREDICTED: LOW QUALITY PROTEIN: lactase-phlorizin hydrolase [Pan troglodytes]	3491	3694	95%	0.0	98%	XP_009441718.1

La lista de resultados arrojada por BLAST va a mostrar que secuencias del metagenoma tienen elementos similares con la enzima, las primeras en la lista son las que más elementos tienen.

2.6. Caracterización de las regiones conservadas

Las regiones conservadas encontradas por BLAST pueden corresponder a tres tipos de información: elementos característicos de las enzimas; motivos asociados a factores de trascripción; o simplemente a regiones aleatorios que se repiten y no tienen ningún sentido biológico.

Para conocer si estas regiones representan algo característico de la enzima, estas regiones deben repetirse en más de una secuencia. El problema que se va a presentar es que estas regiones no van a mostrarse totalmente conservadas, es decir van a tener posibles mutaciones, inserciones y deleciones que las van a mostrar como regiones diferentes, sin relación aparente. Para esto, se va a realizar un agrupamiento o *clustering* donde se van a formar grupos de regiones muy similares y con estos grupos se va a construir un perfil o región consenso junto con un *logo* que describa las bases o aminoácidos que se conservan y representan la secuencia.

Para realizar esta búsqueda realice los siguiente pasos:

1. Seleccione cada región conservada y guárdelas en un archivo común (ej. "secuencias_conservadas.txt")
2. Realice un agrupamiento o *clustering* sobre todas estas regiones y seleccione los grupos que tengan mas de una región integrante.
3. Con los grupos más grandes construya sus perfiles y por cada grupo construya su perfil, así: realice un multi-alineamiento de las secuencias; calcule las frecuencias de las bases o aminoácidos; cree la secuencia consenso *y logo*; y finalmente construya la matriz del motivo.

2.7. Búsqueda de las regiones conservadas en Bases de Datos y Artículos de Investigación

Primero realice la búsqueda de las regiones caracterizadas en bases de datos especializadas de motivos. Con las regiones que no aparecen en las BD de motivos, realice una búsqueda en bibliografia especializada (artículos de revistas de investigación).

3. Descripción de los resultados

Para cada región encontrada, va a describir sus características principales de acuerdo a los valores anotados en la BD o en la bibliografia.