

Análisis Filogenético

Carlos Pérez, Abraham Nieto @ ITAM

Abril, 2017

Saccharomyces cerevisiae

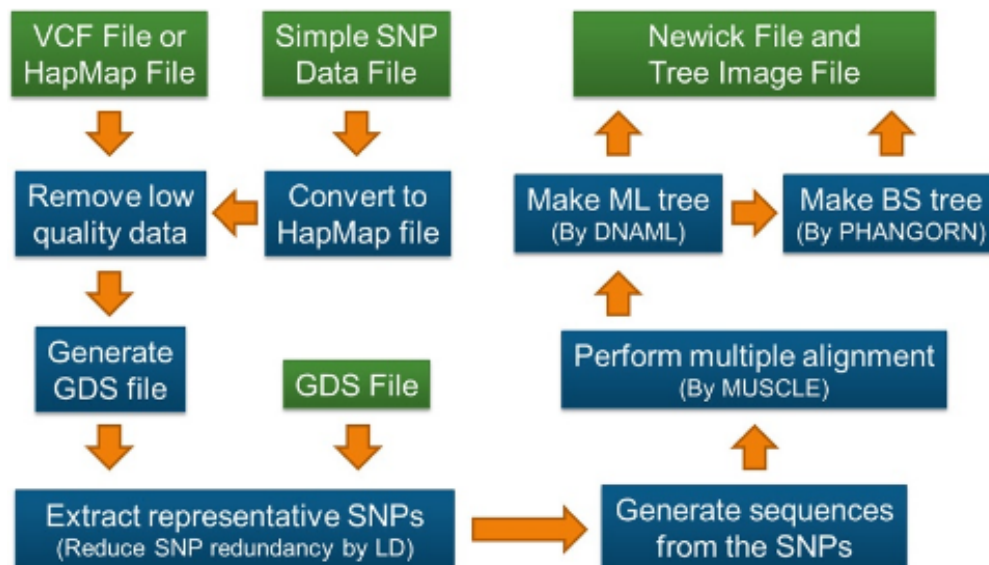
La levadura de cerveza (*Saccharomyces cerevisiae* Meyen ex E.C.Hansen, de Saccharo azúcar, myces hongo y *cerevisiae* cerveza) es un hongo unicelular, un tipo de levadura utilizado industrialmente en la fabricación de pan, cerveza y vino.

Especificación del Problema

Se desea desarrollar el árbol filogenético de *Saccharomyces cerevisiae* utilizando diferentes algoritmos o métodos para su construcción, basados en caracteres (método de parsimonia), basados en distancias (Neighbor joining, UPGMA, Bayesianos) y máxima verosimilitud (pipeline:SNPhylo).

En particular la idea principal es comparar los árboles obtenidos utilizando las distintas técnicas y con ello poder identificar las diferencias (si existieran) de los árboles filogenéticos creados.

En particular de acuerdo con el paper *SNPhylo: a pipeline to construct a phylogenetic tree from huge SNP data* el cual genera los árboles a través de un pipeline creado para dicho fin también habla que la principal diferencia, analíticamente hablando, con el resto de las técnicas es que utiliza el concepto de LD (*Linkage disequilibrium*) para evitar utilizar mutaciones con patrones que violan los supuestos de independencia y por tanto crean un sesgo en la construcción del árbol, de tal modo que el árbol filogenético *SNPhylo* reduce de cierta forma la redundancia de información.



Por tanto se espera corroborar que al menos existan diferencias en la taxonomía de la levadura de cerveza bajo SNPhylo y el resto. En segundo lugar, podremos analizar el tiempo que toma generar el árbol con el pipeline vs. un desarrollo manual en un programa como Python o R.

El pipeline está hecho para soportar grandes volúmenes de datos aunque para este análisis, deberemos evaluar en términos computacionales el número de secuencias para poder crear los árboles filogenéticos a partir los algoritmos que trabajan con distancias y parsimonia.

- Data Gathering: Obtención de archivos .fasta con información de las bases y mutaciones
- Hacer el alineamiento múltiple de las secuencias
- A partir del archivo alineado generar la matriz de distancias entre secuencias.
- Sobre la matriz de distancias aplicar los algoritmos (NJ, UPGMA, etc. ...)
- utilizamos ETEtoolkit para visualizar el árbol filogenético

Fundamentos del Análisis

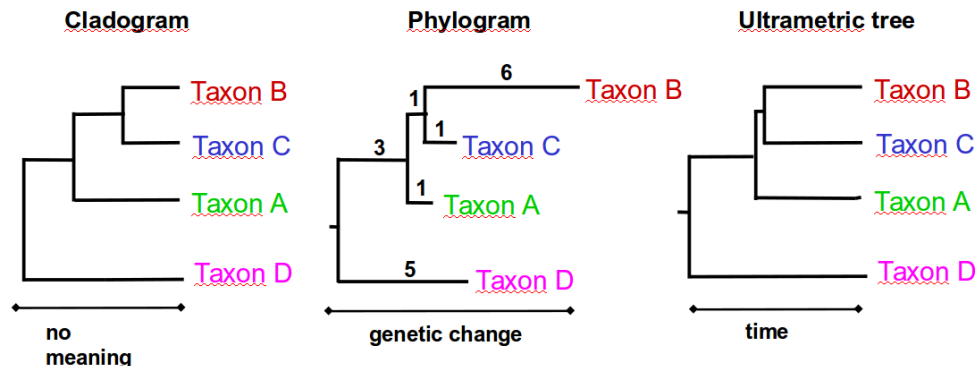
In biology, phylogenetics is the study of the evolutionary history and relationships among individuals or groups of organisms (e.g. species, or populations). These relationships are discovered through phylogenetic inference methods that evaluate observed heritable traits, such as DNA sequences or morphology under a model of evolution of these traits. The result of these analyses is a phylogeny (also known as a phylogenetic tree) – a diagrammatic hypothesis about the history of the evolutionary relationships of a group of organisms. Phylogenetic analyses have become central to understanding biodiversity, evolution, ecology, and genomes.¹

De forma puntual el análisis filogenético tiene como objetivos esenciales

- Encontrar vínculos evolutivos entre organismos, al analizar los cambios en los diferentes organismos durante la evolución.
- Encontrar relaciones entre ancestros y sus descendientes, al analizar familias de secuencias
- Estimar el tiempo de divergencia dentro de un grupo de organismos que comparten el mismo ancestro

Desde un punto de vista operacional, el análisis filogenético tiene dos componentes: La inferencia filogenética (o construcción de árboles) y la aplicación de estas filogenias para entender la evolución de los organismos y sus características.

La forma más común (y natural) de representar las relaciones evolutivas entre un grupo de organismos es un *árbol filogenético*. Este puede ser de distinta configuración dependiendo qué se desea comunicar



¹ <https://en.wikipedia.org/wiki/Phylogenetics>

Especificación del Problema

Se desea desarrollar o aplicar un algoritmo para que dada una secuencia de ADN, se clasifique de forma relativamente precisa² el taxa³ al que pertenece.

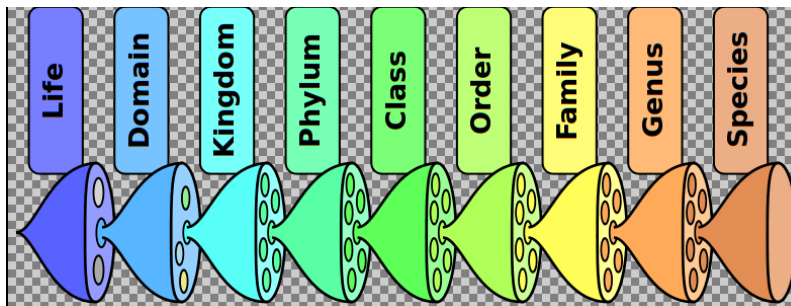
Existen varios supuestos que son necesarios probar, el taxa que se desea estimar tiene una relación directa con la precisión esperada. En particular, a menudo las secuencias de ADN son más similares cuando los organismos son relativamente cercanos en términos evolutivos, por lo que la precisión será menor mientras el taxa a clasificar es más específico.

Por otro lado es importante tener en cuenta que se debe elegir regiones del ADN que vayan acorde a la especificidad del taxa a clasificar. Es decir las cadenas de ADN a menudo son regiones específicas del genoma que son extraídas con base en parámetros de extracción y secuenciación.

Por último, la cantidad de datos disponibles en las bases de datos públicas para poder es crucial realizar un análisis que generalice de forma adecuada.

Inicialmente las hipótesis iniciales son:

- Existe una región o regiones en el genoma de los seres vivos que permite identificar a que taxón pertenecen y esto depende del nivel de especificidad del taxón.
- Clasificar secuencias en niveles más específicos es más difícil, en otras palabras es más difícil atinarle a la Especie que al Reino.
- Mientras más específico es el taxa, los genomas son más homogéneos y por lo tanto más datos son necesarios para realizar una clasificación.





Métodos (Computational Phylogenetics)

El objetivo principal es aplicar algoritmos, métodos computacionales para el análisis filogenético. El objetivo principal es construir un árbol que represente las hipótesis sobre la evolución de un conjunto de genes, especies o taxones.

La construcción de los árboles requiere una medida de homología entre las características observadas (morfológicas o moleculares).

 **Métodos de Distancia:** explicitly rely on a measure of "genetic distance"

 **Métodos de Parsimonia:** identifying the potential phylogenetic tree that requires the smallest total number (or total cost) of evolutionary events to explain the observed sequence data (NP-Hard)

 **Métodos de Máxima Verosimilitud:** standard statistical techniques for inferring probability distributions to assign probabilities to particular possible phylogenetic trees, necessarily requires a substitution model to assess the probability of particular mutations

²Falta determinar como mediremos el desempeño del modelo, típicamente una matriz de confusión

³Choosing among Kingdom, Phylum, Class, Order, Family, Genus, Species

⚙️ **Métodos Bayesianos:** closely related to the maximum likelihood methods with assumptions such as divergence events such as speciation occur as stochastic processes, generally use Markov chain Monte Carlo sampling algorithms

Todos estos métodos dependen de forma implícita o explícita de un modelo matemático que describe la evolución de las características observadas. En particular el concepto de árbol se puede generalizar al concepto de red filogenética, usado para analizar transferencia horizontal de genes.

Más en https://en.wikipedia.org/wiki/Computational_phylogenetics

Data Gathering

Es necesario encontrar un conjunto de datos que sean adecuados para realizar el experimento, esto se realiza usualmente mediante *Homologous Sequences Search*, que realiza un query a las bases públicas de datos genéticos.

Algunas consultas y bases son . . .

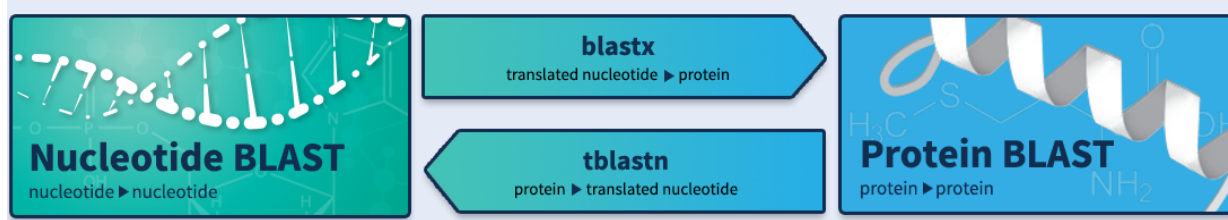
- BlastP (aaquery/aadb) requires as query a protein
- BlastX (ntquery/aadb) requires as query a DNA sequence

– nr (Non redundant protein ncbi) (2017-03-06) – Swissprot from NCBI ftp site (2017-03-06) – Refseq proteins (2017-03-21) – PDB AA (2017-03-06) – Uniprot (2010-03-04)

- BlastN (ntquery/ntdb)
- TblastN (aaquery/ntdb)

–genbank NT (2017-03-03)

Un posible byproduct de este estudio consiste en realizar una investigación e implementación sencilla de este paso para realizarse de forma automatizada para varias secuencias, de modo que se minimice el tiempo utilizado en este paso y las consultas a BLAST puedan hacerse de forma programática e idealmente en paralelo.



Software

Idealmente se utilizarán las paqueterías disponibles en R o Python, dependiendo de las necesidades del proyecto, las capacidades de cada lenguaje y las capacidades de sus usuarios.

Revisión de Literatura

On process ...⁴

- Felsenstein, J. 2004. *Inferring Phylogenies*. Sinauer Associates, Sunderland MA.
- Freeman, S., and Herron, J.C. 2001. *Evolutionary Analysis*, 2nd Edition. Prentice Hall, Upper Saddle River, NJ.
- Hillis, D.M., Moritz, C., and Mable, B.K. 1996. *Molecular Systematics*, 2nd Edition. Sinauer Associates, Sunderland MA.
- Lee, W.-H. 1997. *Molecular Evolution*. Sinauer Associates, Sunderland MA.
- Page, R.D.M., and Holmes, E.C. 1998. *Molecular Evolution, a Phylogenetic Approach*. Blackwell Science, Oxford.

Required Articles

- Baldauf SL (2003) The deep roots of eukaryotes. *Science* 300:1703-1706
- Stewart CB (1993) The Powers and Pitfalls of Parsimony. *Nature* 361:603-607
- Zuckerkandl, E., and L. Pauling. 1965. Molecules as documents of evolutionary history. *J. Theoret. Biol.* 8:357-366.
- Delwiche, C. F. 2004. The genomic palimpsest: Genomics in evolution and ecology. *Bioscience* 54:991-1001.

Ligas

-<http://etetoolkit.org/>
-<http://readiab.org/book/latest/2/4>
-<https://mrnoutahi.com/2016/01/09/Tree-manipulation-with-ETE/>
-<http://etetoolkit.org/docs/2.3/tutorial/index.html>
-<https://pypi.python.org/pypi/phylogenetics/0.3>
-http://evolution.berkeley.edu/evolibrary/article/evo_10
-http://www.phylogeny.fr/one_task.cgi?task_type=blast
-<https://yanailab.org/2016/02/20/mid-developmental-transition/>
-<http://www.life.umd.edu/labs/delwiche/bsci348s/lec/Phylogenetics1.html>
-<http://www.life.umd.edu/labs/delwiche/MSyst/Read2006.html>

Cronograma

- Diseño pipeline e implementación Data Gathering: Semana 24 Abril
- Creación/Selección de árboles filogenéticos: Semana 1 Mayo
- Resultados y Conclusiones: Semana 8 Mayo

⁴Extraído de algún curso de posgrado en Sistemática Molecular