Introducción a la Inferencia Filogenética y Evolución Molecular

23-26 Junio 2008, Fac. C. Biológicas – UANL, México

Pablo Vinuesa (vinuesa@ccg.unam.mx)

Tutor: PDCBM, Ciencias Biológicas, PDCBioq. y Profesor de la Lic. Ciencias Genómicas

Centro de Ciencias Genómicas – UNAM, Campus Morelos, Cuernavaca, México

http://www.ccg.unam.mx/~vinuesa/

Todo el material del curso lo puedes descargar desde:

http://www.ccg.unam.mx/~vinuesa/UANL08

Introducción a la Inferencia Filogenética y Evolución Molecular

23-26 Junio 2008, Fac. C. Biológicas - UANL

- **Tema 4: Introducción a la inferencia Filogenética:**

1. Prácticas de descarga y manipulación de datos de secuencia: Formatos de secuencia, el sistema ENTREZ del NCBI, UNIX y Perl
2. La inferencia filogenética es una forma de inferencia estadística
3. Clasificación de métodos filogenéticos en base al tipo de datos y método de reconstrucción que emplean para recuperar árboles
4. Métodos de búsqueda de árboles vs. reconstrucción algorítmica y "el problema del número de topologías"

Libros de referencia recomendados:

Felsenstein, J., 2004. Inferring phylogenies. Sinauer Associates, INC., Sunderland, MA.
Futuyma, D.J. 2005. Evolution. Sinauer Associates, INC., Sunderland, MA.
Graur, D., Li, W.H., 2000. Fundamentals of Molecular Evolution. Sinauer Associates, Inc., Sunderland.
Nei, M., Kumar, S., 2000. Molecular Evolution and Phylogenetics. Oxford University Press, Inc., NY.
Page, R.D.M., Holmes, E.C., 1998. Molecular Evolution - A Phylogenetic Approach. Blackwell Science Ltd, Oxford.
Swofford, D.L., Olsen, G.J., Waddel, P.J., Hillis, D.M., 1996. Phylogenetic inference. In: Hillis, D.M., Moritz, C., Mable, B.K. (Eds.), Molecular Systematics. Sinauer Associates, Sunderland, MA, pp. 407-514. (Una revisión excelente del campo antes de aparecer los métodos Bayesianos)

Protocolo básico para un análisis filogenético de secuencias moleculares

Colección de secuencias homólogas

⇩ • BLAST, FASTA, ENTREZ ...

Alineamiento múltiple de secuencias

⇩ • Clustal, T-Coffee, muscle, Mafft, POY ...

Análisis evolutivo del alineamiento y selección del modelo de sustitución más ajustado

⇩ • tests de saturación, ModelTest, ...

Estima filogenética

⇩ • NJ, ME, MP, ML, Bayes ...

Pruebas de confiabilidad de la topología inferida

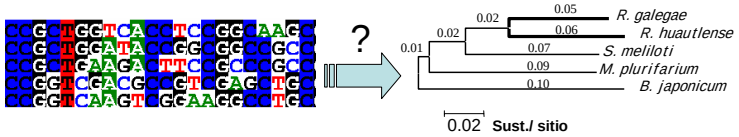
⇩ • proporciones de bootstrap probabilidad posterior ...

Interpretación evolutiva y aplicación de las filogenias

Inferencia Filogenética – introducción

- La inferencia de relaciones filogenéticas a partir de secs. moleculares requiere de la selección de uno de los muchos métodos disponibles

- Con frecuencia la inferencia filogenética es considerada como una "caja negra" en la que "entran las secuencias y salen los árboles"



- **Objetivos de este curso son:**

1. desarrollar un marco conceptual para entender los fundamentos teóricos (filosóficos) que distinguen a los distintos métodos de inferencia (clasificación de métodos)
2. presentar el uso de **modelos y suposiciones** en filogenética
3. manejo empírico de diversos paquetes de software para inferencia filogenética bajo diversos criterios

Métodos de reconstrucción filogenética – introducción

- La inferencia de una filogenia es un **proceso de estimación**; se trata de obtener la mejor estima posible **de una historia evolutiva** basada en la información incompleta y con frecuencia ruidosa contenida en los datos. Estos, por lo general, son moléculas y especies contemporáneas
- En principio, sería posible postular escenarios evolutivos *ad hoc* mediante los cuales cualquier filogenia tomada al azar podría haber producido los datos observados; es esencial por ello contar con un **criterio estadísticamente y biológicamente riguroso para la selección de una o más topologías de entre todas las posibles**
- Los métodos de inferencia filogenética están diseñados para este fin siguiendo una de dos estrategias o caminos:
  1. mediante la definición de un **algoritmo** que determina los pasos a seguir para la reconstrucción de la topología
  2. mediante la definición de un **criterio de optimización** mediante el cual poder decidir cual o qué topología(s) son las mejores (o igualmente favorecidas)

Métodos de reconstrucción filogenética: algoritmos vs. criterios de optimización

- Los métodos algorítmicos combinan la inferencia del árbol y la definición del mejor árbol en una misma operación. Son por ello muy rápidos (NJ y UPGMA)
- Aquellos basados en **criterios de optimización** (CO) tienen en cambio **dos pasos lógicos**.
  1. **definir el criterio de optimización** (descrito formalmente en una **función objetiva**) para evaluar cada posible topología, asignándole una puntuación con la que poder comparar cuantitativamente el mérito de cada árbol en base al criterio de optimización
  2. en un segundo paso se usan **algoritmos de búsqueda** específicos para calcular el valor de la función de objetividad y para encontrar el/los árbol(es) con la mejor puntuación acorde al este criterio (un valor máximo o mínimo, según el caso)
- Los **métodos basados en CO desacoplan** por lo tanto los **supuestos evolutivas** hechos en el primer paso **de las técnicas computacionales** del segundo. El precio de esta claridad lógica es que estos métodos son muchísimo más lentos que los algorítmicos, debido a que tienen que hacer búsquedas en el inmenso espacio de topologías para encontrar la(s) mejor(es)
- Los métodos algorítmicos tratan a los datos de diferente manera que los basados en criterios de optimización: análisis de **distancias vs. caracteres discretos**

Inferencia filogenética molecular – clasificacón de métodos

- Podemos clasificar a los métodos de reconstrucción filogenética en base al tipo de datos que emplean (**caracteres discretos vs. distancias**) y si usan un **método algorítmico o un método de búsqueda basado en un criterio de optimización** para encontrar la topología óptima bajo el criterio seleccionado

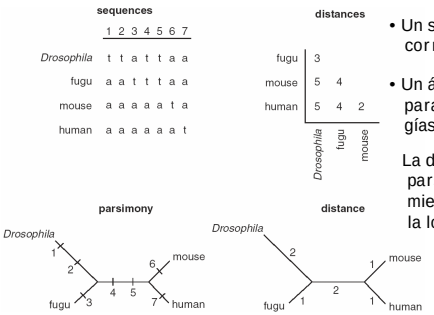
Tipo de datos

Método de reconstrucción	distancias	caracteres discretos
algoritmo de agrupamiento	UPGMA y Neighbor joining	
Búsquedas bajo criterio de optimización	Mínimos cuadrados y Evolución mínima	Máxima parsimonia y Máxima verosimilitud

Métodos de reconstrucción filogenética – una clasificación

I.- Tipos de datos: distancias vs. caracteres discretos

- Los **métodos de distancia** requieren la **transformación de los alineamientos de secuencias en una matriz de distancias genéticas** en base al modelo evolutivo seleccionado, la cual es usada por el método algorítmico de reconstrucción para calcular el árbol (**UPGMA y NJ**)
- Los **métodos discretos** (**MP, ML, Bayesianos**) consideran cada sitio del alineamiento (o una función probabilística para cada sitio) directamente



- Un set de 4 secs. y la matriz de distancias correspondiente
- Un árbol de parsimonia y uno de distancias para este set de datos produce topologías y longitudes de ramas idénticas

La diferencia radica en que el árbol de parsimonia identifica qué sitio del alineamiento contribuye cada paso mutacional en la longitud de cada rama

- El precio: los métodos de agrupamiento algorítmico (UPGMA, NJ) son mucho más rápidos que los métodos basados en búsqueda de árboles

Métodos de reconstrucción filogenética – una clasificación

II. Métodos algorítmicos vs. criterios de optimización

- **Criterios de optimización:** reglas para decidir entre pares de topologías cual es mejor (dados los datos)
- Los métodos de reconstrucción de **MP y ML** utilizan diferentes criterios de optimización para seleccionar el/los árbol(es) entre las de topologías que han de evaluar
- A cada topología se le asigna una puntuación (**score**) que es función del **ajuste existente entre la topología y los datos**
- Los métodos de optimización tienen la gran ventaja de requerir una función probabilística explícita que relaciona los datos con la topología (p. ej. un modelo de sustitución). Ello permite evaluar la calidad de cualquier árbol (topología), permitiendo el uso de distintas técnicas estadísticas para **evaluar la significancia con la que las distintas hipótesis evolutivas (topologías) en competición se ajustan a los datos**!!!
- Ejemplos de métodos de búsqueda de árboles por criterio de optimización son:

- **MP: máxima parsimonia** (menor es mejor)
- **ML: máxima verosimilitud** (mayor es mejor)
- **ME: evolución mínima** (menor es mejor)
- **LS: cuadrados mínimos** (menor es mejor)

- Una limitación potencial de los métodos de optimización es que son computacionalmente muy costosos, requiriendo por lo general implementaciones heurísticas del algoritmo

Métodos de búsqueda de árboles

I.- el problema del número de topologías

El número de topologías posibles incrementa factorialmente con cada nuevo taxon o secuencia que se añade al análisis

No. de árboles no enraizados = $(2n-5)!/2^{n-3}(n-3)$

No. de árboles enraizados = $(2n-3)!/2^{n-2}(n-2)$

Taxa	árboles no enraiz*.	árb. enraiz.
4	3	15
8	10,395	135,135
10	2,027,025	34,459,425
22	$3x10^{23}$	...
50	$3x10^{74}$	...

*por ej. para sólo 15 OTUs tenemos 213,458,046,676,875 topologías

- ¡ si pudiésemos evaluar $1x10^6$ topol./seg. necesitaríamos 6 años y 9 meses para completar la búsqueda! El no. de Avogadro es ~ $6 x10^{23}$ (átomos/mol). Según la teor. de la relatividad de la estructura del universo de Einstein, existen 10^{80} átomos de H_2 en el universo ...

http://en.wikipedia.org/wiki/Observable_universe

Por tanto se requieren de **estrategias heurísticas de búsqueda** árboles cuando se emplean métodos basados en criterios de optimización y n > ~25