Bioinformática estadística o estadística de datos ómicos

Guillermo Ayala Gallego

Guillermo Ayala Gallego

3/14/23

Prólogo

- Un investigador que publique un artículo al mes durante un año tendrá al final de ese año un total de 12 publicaciones.
- Y debiera de ser posiblemente candidato a algún premio importante.
- Después de ese año maravilloso de publicaciones, el siguiente es tan bueno como el anterior.
- Y sigue publicando un trabajo al mes durante este segundo año. Ya reune 24 publicaciones en dos años agotadores.
- Y esto mismo lo hace durante 10 años.

- Al final de esta decada prodigiosa tendrá 120 publicaciones.
- Ya ha justificado su vida como investigador.
- Pero no, incansable sigue otros diez años y alcanza al cabo de 20 años los 240 publicaciones. Y diez años más.
- Exhausto él o ella (y cuantos hayan intentado leer todas estas publicaciones) ya ha llegado a las 360 publicaciones.
- Hay una enorme cantidad de investigadores que superan esta cantidad. La superan ampliamente.
- Es seguro que han participado activamente en todas estas publicaciones.
- Pero el resto no tenemos la culpa.

Crecimiento en el número de publicaciones

- Hay numerosos estudios sobre el crecimiento del número de publicaciones.
- Ahora cada nueve años se dobla el número de publicaciones.

- Con frecuencia leemos que se retiran publicaciones.

- ¿Por qué?

Estas notas

- Tratan de la aplicación de procedimientos estadísticos al **análisis de datos de alto rendimiento**.
- ¿Qué son datos de alto rendimiento?
- Datos que rompen lo que tradicionalmente era un prerrequisito en **Estadística** (multivariante).
- Muchos procedimientos estadísticos empiezan indicando que el número de observaciones, \(n\), ha de ser mayor que el número de variables por observación, \(p\).
- Actualmente los datos tienen dimensiones \(p\) que marean: miles de variables frente a decenas (con suerte algo más de un centenar) de observaciones o muestras.
- ¿Y qué hacemos para analizar esto? Lo que se pueda.
- De esto van estas notas, **de lo que se pueda**.
- En la presente versión utilizamos:
  - Los (viejos) microarrays, en particular, datos de metilación.
  - Técnicas de RNA-Seq (a nivel de muestra de tejido y al nivel de célula única).
  - Estudios de asociación.
- En resumen información de transcripción e información genómica.

De porqué es imposible entender la parte metodológica de en una revista de Biología o Medicina

- Se han de leer artículos de revistas de diversos ámbitos científicos.
- Muchos trabajos tienden a ser una pura ilustración de los métodos propuestos.
- El cómo se hace realmente se relega al final en una sección suplementaria o directamente a un archivo de material suplementario.
- Hay una auténtica aversión a la formulación matemática de las cosas.
- A veces, es imposible saber qué se está haciendo porque no quieren poner una simple formula.
- El artículo es inútil aunque el índice de impacto de la revista sea mayor.

Sobre R/Bioconductor

- Estas notas utilizan sistemáticamente R/Bioconductor.
- Hay algún uso de herramientas de Bash y Python.
- Cuando empecé con esto de la Probabilidad y la Estadística uno leía libros y artículos, entendías aquello y luego intentabas descifrar cómo aplicaba (mejor implementaba) estas técnicas el software comercial (en mi caso **SPSS**).
- Lo primero era bonito. Los autores intentan que les entiendas porque tratan de transmitir ideas.
- Sin embargo, el software comercial no piensa así (y esto no es necesariamente

malo).
- El software comercial intenta dar un producto bueno y fácilmente utilizable para llegar a un máximo de usuarios.
- Solamente suelen considerar temas sobre los que hay mucho interés y muchos (potenciales) usuarios. Y esto es correcto.
- No es la opción elegida en estas notas.
- Hemos elegido trabajar con software libre.
- Tanto R como Bioconductor son el resultado de un gran trabajo coordinado de muchas personas.
- Algunos proceden del mundo académico.
- En otros ocasiones proceden de empresas para las cuales les resulta interesante que se disponga de software que permita utilizar su hardware.

Sobre la investigación reproducible

- Es ingente la cantidad de publicaciones científicas que llevan tratamientos estadísticos.
- Los investigadores suelen saber qué quieren estudiar. Diseñan un experimento y observan datos.
- Lo que se hace después no lo suelen conocer.
- No suelen conocer las técnicas que han utilizado.
- Citan los procedimientos estadísticos y no indican el software utilizado si está disponible en la red o, en el caso de que sea propio de los autores, dónde se puede conseguir.
- Sin duda alguna el control de la calidad de los tratamientos estadísticos descansa en que cada lector de una publicación científica tenga a su disposición el artículo (que básicamente es la explicación de lo que se ha hecho) así como los datos sin ningún tipo de preprocesamiento.
- Se debiera de disponer de todo el código necesario para reproducir todo el tratamiento estadístico realizado con los datos.
- Sin esto, no se puede realizar un control adecuado de un tratamiento estadístico de datos de alto rendimiento (de hecho, de ningún tipo de datos).
- Esto nos lleva a los conceptos de programación literaria o comentada (literate programming) propuesto por Donald K. Knuth: http://en.wikipedia.org/wiki/Literate_programming y, de un modo más genérico, a la investigación reproducible http://en.wikipedia.org/wiki/Reproducibility
- R/Bioconductor incorpora muchas herramientas para realizar investigación reproducible: http://cran.r-project.org/web/views/ReproducibleResearch.html
- En particular, este texto está realizado utilizando **knitr**.

- Todos los datos que se utilizan están disponibles en bases de datos públicas.
- No de todos los datos que se utilizados tenemos los datos sin procesado previo.
- Los usamos por haber sido analizados en otros textos o en ejemplos de R/Bioconductor o, simplemente, porque son bonitos de estudiar.