R na Bioinformática

Aplicacións de R e Bioconductor na secuenciación de segunda xeración

Miguel Blanco Vázquez

Introdución

Os recentes fitos na investigación xenómica e a bioloxía molecular estannos achegando a ser quen de comprender as funcións e os complexos sistemas biolóxicos totalmente a nivel xenómico. A aparición das tecnoloxías de secuenciación de alto rendemento, especialmente a expresión xénica en microarrays, comezou a xerar unha cantidade de datos cada vez máis inxente e, con isto, os investigadores víronse coa necesidade de desenvolver ferramentas informáticas para a análise deses datos. O proxecto de código aberto *Bioconductor,* baseado na linguaxe de programación *R,* naceu para cubrir as necesidades deste novo eido de investigación con secuenciación de segunda xeración.

A bioinformática

O que hoxe se chama bioinformática, malia a ser xa definido no 1978 pero coma "o estudo de procesos de información en sistemas bióticos", xorde nos 1980s durante a "revolución xenómica" pola aplicación das ciencias computacionais e da información á análise de datos biolóxicos; particularmente dos datos froito da área da xenómica de secuenciación de DNA a gran escala. Nas últimas décadas o rápido desenvolvemento da xenómica e outras tecnoloxías de investigación molecular, en conxunto co das tecnoloxías da información, desataron a obtención dunha inxente cantidade de información no campo da bioloxía molecular. O concepto de *bioinformática* é, logo, o termo para definir as aplicacións computacionais e matemáticas que permiten elucidar os procesos biolóxicos que agochan, tales como os que agochan estes novos datos.

O eido da bioinformática ocúpase da análise de secuencias, a anotación xenómica, a bioloxía evolutiva computacional, a análise de expresión xénica, a regulación e expresión proteica, as mutacións, a xenómica comparativa, os sistemas de modelado biolóxico, a bioinformática estrutural para predición de estrutura proteica e interacción molecular e a análise de imaxes de alta resolución.

R e Bioconductor

R

R é unha linguaxe de código aberto para a programación estatística. Emprégase para xestionar datos, realizar análises estatísticas e representacións gráficas. Ao núcleo básico de R súmanselle paquetes adicionais que son desenvolvidos por unha ampla comunidade de voluntarios. Eses paquetes aportan funcionalidades específicas á instalación do núcleo.

R, nado no 1993, é arestora na análise estatística a lingua franca e é empregado nun amplo abano de áreas de investigación biotecnolóxica, como así mesmo tamén noutras áreas con aplicacións gobernamentais e industriais.

Bioconductor

Bioconductor é un proxecto de código aberto baseado na linguaxe de programación R no que se implementan ferramentas para a análise xenómica de datos de secuenciación de alto rendemento ou segunda xeración *(high-throughput genomics)*. O proxecto Bioconductor, que arrincou no 2001, xorde de primeiras da necesidade de estudar os datos obtidos de microarrays e posteriormente implementou o estudo de datos da secuenciación de alto rendemento. Dende o seu arrinque foise asentando como un dos principais proxectos neste campo grazas a súa credibilidade, gañada froito do seu rigor estatístico na análise de pre-procesamento de microarrays e a análise de deseños experimentais. O enfoque de integración e a capacidade de reprodución das análises son as características máis salientables de Bioconductor na bioinformática.

Arestora hai máis de 900 paquetes en Bioconductor que permiten, entre outras, o estudo de expresión xénica, de secuencias e aplicacións á citometría e a análise de imaxes.

R e Bioconductor na análise dos datos da secuenciación de alto rendemento

As nova variantes de protocolos e análises empregados na secuenciación de alto rendemento ou de segunda xeración permiten o estudo de expresión xénica, regulación e a codificación de variantes xenéticas. Iso na práctica supón que estes novos protocolos experimentais permiten obter unha enorme cantidade (millóns por mostra) de pequenas cadeas (de 35 a 100 pares de bases) de secuencias de nucleótidos. Aliñados a un xenoma, e en función da técnica aplicada, pódese inferir os niveis de expresión xénica (RNA-seq), a unión de elementos reguladores a puntos xenómicos concretos (Chip-Seq), ou a predominio de variacións estruturais (como os SNPs, indels curtos, reposicionamentos a escala xenómica...).

Téñase en conta ademais que a replicación das mostras varía dun par por tratamento a miles por individuo.

Sendo isto así, xa é inabordable o emprego de follas de cálculo para a análise dos datos obtidos destas novas técnicas de experimentación e precísase recorrer á programación con ferramentas como R e Bioconductor. As intricadas relacións entre os datos e a súa anotación e a variedade de cuestións a resolver precisan de flexibilidade, típica dunha linguaxe de programación. Na secuenciación de alto rendemento estes paquetes tanto permiten cubrir as necesidades da análise primaria, que abrangue os valores de expresión de microarrays, lecturas de secuencias e demais; como a anotación deses datos, que determinan a localización de xenes e características do tipo exóns e rexións de regulación, a súa implicación en rutas biolóxicas etc.

A meirande parte dos paquetes dispoñibles para Bioconductor abranguen os campos de análise que se precisan na secuenciación de segunda xeración. Para este tipo de análise Bioconductor dispón de paquetes para a representación de datos *(IRanges, GenomicRanges, GenomicFeatures),* de xestión de entrada e saída de datos (como *ShortRead* para ficheiros *fastq*), anotación *(GenomicFeatures, ChIPpeakAnno, VariantAnnotation),* aliñamento *(Rsubread, Biostrings),* visualización *(ggbio,Gviz)* e asesoramento da calidade *(htSeqTools,ShortRead)*. En función da técnica atopámonos con outros paquetes que abranguen a RNA-seq *(BitSeq,DESeq,DEXSeq,edgeR),* a ChIP-sep *(chipseq,ChIPseqR, mosaics, BayesPeak),* os motivos *(cosmo, rGADEM),* 3C *(HiTC)* e número de copias *(exomeCopy, segmentSeq)*. Así mesmo, dispón doutros específicos para xestión de traballo *(ArrayExpressHTS)* e bases de datos *(SRAdb)*.

Conclusións

No eido práctico, cabe destacar que R e Bioconductor facilitan a reproducibilidade na investigación. En R repetir a análise só supón a execución de novo dos scripts desenvolvidos e a modificación destes, a súa reciclaxe, permite o seu axuste a novos procesos. Así mesmo, R e Bioconductor facilitan coñecer as versións de cada funcionalidade empregada co cal a trazabilidade de problemas na análise está asegurada. Engadido a isto, sendo proxectos de código aberto, están suxeitos a avaliacións críticas que aseguran a súa eficacia.

Mais, xa en canto ao impacto destas ferramentas, os desenvolvemento de R e Bioconductor estiveron moi ligados a campos de investigación específicos nos que cabe destacar o forte vencellamento a institucións académicas e os seus grupos de traballo. Con Bioconductor pódese apreciar ben a súa ligazón á investigación con secuenciación de alto rendemento e no que a filosofía de código aberto permitiu que a achega de diferentes grupos de investigación cos mesmos obxectivos fora quen de dar froito a resultados neste emerxente e complexo campo da xenómica.

Bibliografía

[1] N. Le Meur and R. Gentleman, "Analyzing biological data using R: methods for graphs and networks," Methods Mol. Biol., vol. 804, pp. 343–373, 2012.

[2] "Bioconductor - Next generation sequencing data analysis with R/Bioconductor." [Online]. Dispoñible en: http://www.bioconductor.org/help/course-materials/2012/CSC2012/. [Último acceso: 1-Sep-2012].

[3] R. C. Gentleman, V. J. Carey, D. M. Bates, B. Bolstad, M. Dettling, S. Dudoit, B. Ellis, L. Gautier, Y. Ge, J. Gentry, K. Hornik, T. Hothorn, W. Huber, S. Iacus, R. Irizarry, F. Leisch, C. Li, M. Maechler, A. J. Rossini, G. Sawitzki, C. Smith, G. Smyth, L. Tierney, J. Y. H. Yang, and J. Zhang, "Bioconductor: open software development for computational biology and bioinformatics," Genome Biol., vol. 5, no. 10, p. R80, 2004.

[4] M. Kanehisa and P. Bork, "Bioinformatics in the post-sequence era," Nat. Genet., vol. 33 Suppl, pp. 305–310, Mar. 2003.

[5] R. Molidor, A. Sturn, M. Maurer, and Z. Trajanoski, "New trends in bioinformatics: from genome sequence to personalized medicine," Exp. Gerontol., vol. 38, no. 10, pp. 1031–1036, Oct. 2003.

[6] D. D'Elia, A. Gisel, N.-E. Eriksson, S. Kossida, K. Mattila, L. Klucar, and E. Bongcam-Rudloff, "The 20th anniversary of EMBnet: 20 years of bioinformatics for the Life Sciences community," BMC Bioinformatics, vol. 10 Suppl 6, p. S1, 2009.