Unified omics data containers for robust, transparent and flexible computational biology

Laurent Gatto[1,2] and Kathryn S. Lilley[1]

[1]Cambridge Centre for Proteomics, University of Cambridge, UK
[2]EMBL, European Bioinformatics Institute, Hinxton, Cambridge, UK
lg390@cam.ac.uk

Abstract

The analysis of data stemming from high throughput biology experiments offers multiple challenges to the computational biology and bioinformatics community, e.g. manipulation, annotation and visualisation of the data and transparent, traceable and reproducible analysis. These difficulties can be partly or completely lifted by appropriate data structures. A set of data representation for various omics data types, as implemented and successfully utilised in the frame of the Bioconductor project will be presented to illustrate how such unified interfaces to complex biological data and its associated meta-data enable scientists to focus on the processing, examination and interpretation of biological data.

Introduction

The pace at which modern biology generates data has often been coined as a deluge. The data analysis that transforms the raw, convoluted data into manageable and interpretable results has become the main bottleneck in high-throughput biology despite cross-disciplinary efforts and major advances in computational biology and bioinformatics.

The situation, in which the computational scientists spends a considerable amount of time transforming and managing the data to obtain a convenient format, prior to effective data exploration and analysis is, sadly, not an uncommon situation. In addition, the multivariate nature of biology often require substantial data annotation that needs to be collected and collated to the actual experimental data. Finally, the complexity of the biology and of the omics data itself, as well as the processing the latter needs to be subjected to, make it very difficult, even for experienced users, to track the computations and verify the results by merely looking at the input and the output. These challenges make a compelling case for robust and transparent infrastructure without compromising the flexibility needed to explore biological diversity.

The Bioconductor project [7] provides software for the analysis and comprehension of various high-throughput omics data. It distributes over 600 interoperable software and annotation packages developed for the `R` statistical programming language and environment [11], promotes collaborative and open scientific software development and is supported by a wide and multidisciplinary user and developer community.

Infrastructure for various omics data

The `eSet` data object type provides a template for the representation of high-throughput assays and experimental meta-data. Originally designed and implemented for the microarray technology (for instance the `AffyBatch` [6] and `ExpressionSet` [7] classes), implementation of the `eSet` structure have been proposed for high-throughput sequencing RNA-Seq count data (`CountDataSet`) [1] and proteomics data (`MSnExp` and `MSnSet`) [5].

All these data structures are build on the same principle: in addition to the main assay data, be it raw microarray probe intensities or transcript estimates, HTS read counts, raw or quantified mass-spectrometry based proteomics data, additional meta-data is provided for the experiment samples, termed `phenoData`, and meta-data describing the features under consideration, coined `featureData`. These respective meta-data structure can be updated when new information, for example biological annotation retrieved from databases or analysis metrics obtained in the frame of the data processing itself, is obtained. The validity of the data is guaranteed by coordinated and transparent data and meta-data handling.

Applications

Coordination of well defined data structure allows straightforward application of data processing algorithms across different omics domains. The *Variance Stabilization Normalization* [8, 9], initially developed for microarrays and applied to RNA-Seq data, had been described to be applicable to quantitative proteomics data [10] and is readily available through the adherence to common programming paradigms [5]. Similarly, application of compatible visualisation techniques to different data sets, like quantitative data of mRNA microarrays, RNA-Seq and proteomics data become straightforward.

Utilisation of the described framework is an important step in exploring the complexity of biology, making conceptually simple analysis trivial and more elaborated meta-omics analyses easier. Two use cases of such integrated analyses will be presented. The first one is work done by Vince J. Carey, a core member of Bioconductor project, integrating genetic data (using the `SnpMatrix` and `XSnpMatrix` infrastructure for genetic data [4]) and gene expression profiles [2]. I will also present some data from [3], combining expression data from Affymetrix GeneChips and iTRAQ 4-plex [12] quantitative proteomics data.

References

[1] S Anders and W Huber. Differential expression analysis for sequence count data. *Genome Biology*, 11:R106, 2010.

[2] Vincent J. Carey. *GGtools: Genetics of Gene Expression with Bioconductor*. R package version 4.7.0.

[3] J I Castrillo, L A Zeef, D C Hoyle, N Zhang, A Hayes, D C Gardner, M J Cornell, J Petty, L Hakes, L Wardleworth, B Rash, M Brown, W B Dunn, D Broadhurst, K O'Donoghue, S S Hester, T P Dunkley, S R Hart, N Swainston, P Li, S J Gaskell, N W Paton, K S Lilley, D B Kell, and S G Oliver. Growth control of the eukaryote cell: a systems biology study in yeast. *J Biol*, 6(2):4, 2007.

[4] David Clayton. *snpStats: SnpMatrix and XSnpMatrix classes and methods*, 2012. R package version 1.9.1.

[5] L Gatto and K S Lilley. MSnbase – an R/Bioconductor package for isobaric tagged mass spectrometry data visualization, processing and quantitation. *Bioinformatics*, 28(2):288–9, Jan 2012.

[6] L Gautier, L Cope, B M Bolstad, and R A Irizarry. affy – analysis of Affymetrix GeneChip data at the probe level. *Bioinformatics*, 20(3):307–315, 2004.

[7] 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 Yang, and J Zhang. Bioconductor: open

software development for computational biology and bioinformatics. *Genome Biol*, 5(10):R80, 2004.

[8] H Huber, A von Heydebreck, H Sueltmann, A Poustka, and M Vingron. Variance stabilization applied to microarray data calibration and to the quantification of differential expression. *Bioinformatics*, 18 Suppl. 1:S96–S104, 2002.

[9] W Huber, A von Heydebreck, H Sueltmann, A Poustka, and M Vingron. Parameter estimation for the calibration and variance stabilization of microarray data. *Stat Appl Genet Mol Biol*, 2:Article3, 2003.

[10] N A Karp, W Huber, P G Sadowski, P D Charles, S V Hester, and K S Lilley. Addressing accuracy and precision issues in itraq quantitation. *Mol Cell Proteomics*, 9(9):1885–97, Sep 2010.

[11] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2012. ISBN 3-900051-07-0.

[12] P L Ross, Y N Huang, J N Marchese, B Williamson, K Parker, S Hattan, N Khainovski, S Pillai, S Dey, S Daniels, S Purkayastha, P Juhasz, S Martin, M Bartlet-Jones, F He, A Jacobson, and D J Pappin. Multiplexed protein quantitation in saccharomyces cerevisiae using amine-reactive isobaric tagging reagents. *Mol Cell Proteomics*, 3(12):1154–1169, Dec 2004.