Data Set Descriptions

Big Data Biology (BIO00047I)
January 2021

Workshop material:
http://www-users.york.ac.uk/~dj757/BIO00047I/BIO00047I_index.html

Fungal metagenomic data

Overview of dataset: ITS_data_cleaned.csv

This data set looks at fungi that are associated with the soil near hedge roots, where the aim is to identify what fungi are associated with these roots. Instead of sequencing the whole genome of all the fungi in each sample, we sequence a single specific DNA region that can be used to barcode the species of fungus (in this case, Internal Transcribed Spacer 2 (ITS2)). Then, ITS2 sequences that are extremely similar to one another are grouped together: each group is referred to as an Operational Taxonomic Unit (OTU). Finally, we can count how often we find sequencing reads that map to each OTU in each environmental sample.

The rows of the table represent OTUs. See
https://en.wikipedia.org/wiki/Operational_taxonomic_unit

The columns of the table are as follows:

a. All columns whose names begin with the letter W represent different environmental samples. Each environmental sample will contain a mixture of many OTUs. The number of times each OTU is observed in the sample is listed in the table
b. 'sum': refers to the total number of each OTU observed across all environmental samples.
c. 'taxonomy': refers to the species that has the most similar ITS2 sequence to the OTU. It is formatted in the following way: k__[kingdom];p__[phylum];c__[class];o__[order];f__[family];g__[genus];s__[species]
d. 'p.value': refers to the significance of the match between the OTU and the reference ITS2 sequence from the 'taxonomy' column
e. 'something': who knows? We have a question addressing this in one of the workshops
f. 'sequence': the ITS2 sequence for this OTU.

This data set includes samples from the soil near hedges in four different fields, and there are 3 sampling points per hedge. The meta-data is available in the file 'GroupVars.csv'.

Overview of: GroupVars.csv

This table contains a description of where the environmental samples in ITS_data_cleaned.csv were collected.

Each row represents an environmental sample

The columns of the table are as follows:

a. 'Sample': the column name of ITS_data_cleaned
b. 'Field': the name of the field the sample was collected in. There are 4 unique fields: BSSE, BSSW, Copse, and Hillside
c. 'HedgeLocation': Each sampling location is given a unique ID. There were 12 unique sampling locations (3 in each field)
d. 'Root number': A unique ID for each root (not that useful for us)
e. 'Latitude': latitude of environmental sample
f. 'Longitude': longitude of environmental sample

Oilseed rape data

This is gene-centric data from plant *Brassica napus* that produces canola oil. Each row contains information about one gene.

The data in OSR101_sample.txt shows measurements of gene expression in a random sample of 101 oilseed rape cultivars. The gene expression measure used here is RPKM, or reads per kilobase per million aligned reads. The data in the file Glucosinolates.txt contains the quantitative measure of glucosinolates in the seed oil from each cultivar.

These data can be used to conduct an association study (comparing gene expression to glucosinolate levels). This is called associative transcriptomics (for example, see (*2*)).

Fission yeast data

This is gene-centric data from the fission yeast *Schizosaccharomyces pombe*. The data has been gathered from multiple different high-throughput studies, including some recent unpublished data (*1*). Much of it was downloaded from the Angeli website (http://bahlerweb.cs.ucl.ac.uk/cgi-bin/GLA/GLA_input).

Each row contains information about one gene. Both protein-coding genes and non-coding RNAs (ncRNAs) are included. Each column contains some categorical or quantitative information about each gene. The data columns are described in the table on the next page.

Fission yeast data table

Column name	Data	Type* (Q/C)	Notes	Reference
NumberIntrons	Number of introns in the gene	Q		(*3*)
NumResidues	Number of amino acid residues in the protein.	Q	Will be NA for ncRNAs.	(*3*)
protein_coding	Is the gene protein-coding?	C	1 = protein coding 0 = ncRNA	(*3*)
ncRNA	Is the gene a non-canonical ncRNA?	C	1 = ncRNA 0 = either protein-coding or a canonical ncRNA (rRNA, tRNA, snoRNA)	(*3*)
Rel_telomere	The relative distance to the telomere.	Q	0 = at the telomere, 1 = in the middle of the chromosome	(*3*)
mRNA_copies_per_cell	The number of mRNA copies per cell, from proliferating (growing) cells.	Q		(*4*)
protein_copies_per_cell	The number of protein copies per cell, from proliferating (growing) cells.	Q		(*4*)
mRNA.stabilities	The half-life of the mRNA in minutes.	Q		(*5*)
GeneticDiversity	The level of genetic diversity within *S. pombe* strains.	Q	The average pairwise similarity (π)	(*6*)
ProteinHalfLife	The half-life of the protein in minutes.	Q		(*7*)
Golgi, Mitochondrion, Nuclear_dots, Nuclear_envelope, Nucleolus, Nucleus, Vacuole	Where the protein is located.	C	1 = in this location 0 = not in this location	(*8*)
essential	Is this an 'essential' gene (required for cell survival)?	C		(*9*)

chromosome	Which chromosome the gene is on.	C	*S. pombe* has three chromosomes (I, II, III) and a mitochondria (MT).	(*3*)
start, end	The position of the gene in the chromosome	Q	Gene length = end - start +1	(*3*)
solid.media.KO.fitness	The colony size of a strain with this gene knocked out (on agar medium).	Q	The colony size is a proxy for knockout 'fitness'.	(*10*)
gene.expression.RPKM	The RNA expression level from RNA-seq, from proliferating (growing) cells.	Q		(*11*)
conservation.phyloP	The level of conservation in this gene.	Q	How fast the gene has changed over time.	(*1*)

* Quantitative or Categorical.

References

1. L. Grech *et al.*, Fitness Landscape of the Fission Yeast Genome. *bioRxiv*, 398024 (2018).
2. A. L. Harper *et al.*, Molecular markers for tolerance of European ash (Fraxinus excelsior) to dieback disease identified using Associative Transcriptomics. *Sci Rep*. **6**, 19335 (2016).
3. V. Wood *et al.*, The genome sequence of Schizosaccharomyces pombe. *Nature*. **415**, 871–880 (2002).
4. S. Marguerat *et al.*, Quantitative Analysis of Fission Yeast Transcriptomes and Proteomes in Proliferating and Quiescent Cells. *Cell*. **151**, 671–683 (2012).
5. A. Hasan, C. Cotobal, C. D. S. Duncan, J. Mata, Systematic Analysis of the Role of RNA-Binding Proteins in the Regulation of RNA Stability. *PLoS Genet*. **10**, e1004684 (2014).
6. D. C. Jeffares *et al.*, The genomic and phenotypic diversity of Schizosaccharomyces pombe. *Nature Genetics*. **47**, 235–241 (2015).
7. R. Christiano, N. Nagaraj, F. Fröhlich, T. C. Walther, Global proteome turnover analyses of the Yeasts S. cerevisiae and S. pombe. *Cell Rep*. **9**, 1959–1965 (2014).
8. A. Matsuyama *et al.*, ORFeome cloning and global analysis of protein localization in the fission yeast Schizosaccharomyces pombe. *Nat Biotechnol*. **24**, 841–847 (2006).
9. D.-U. Kim *et al.*, Analysis of a genome-wide set of gene deletions in the fission yeast Schizosaccharomyces pombe. *Nat Biotechnol*. **28**, 617–623 (2010).
10. M. Malecki *et al.*, Functional and regulatory profiling of energy metabolism in fission yeast. *Genome Biol*. **17**, 240 (2016).
11. S. R. Atkinson *et al.*, Long noncoding RNA repertoire and targeting by nuclear exosome, cytoplasmic exonuclease, and RNAi in fission yeast. *RNA*. **24**,

1195–1213 (2018).

Pseudosuchia macroevolutionary data

This is a macroevolutionary and environmental data set that explores the macroevolutionary drivers of a vertebrate clade called Pseudosuchia (crocodiles plus their extant & extinct relatives).

The data consist of:

1) A time-calibrated (dated) phylogenetic tree of Pseudosuchia (fossilCrocPhylogeny.tre).
2) Diversification rate data for Pseudosuchia (fossilCrocDiversificationData.txt).
3) Habitat data for Pseudosuchia (HabitatData.csv).
4) Environmental data:
    a) global temperature through time (temperatureTimeSeries.csv).
    b) global (eustatic) sea level through time (seaLevelTimeSeries.csv).

About the data:

1) This is a phylogenetic tree of Pseudosuchia. It results from a very recent piece of research and therefore is currently unpublished (Payne *et al.*, in prep) but the methods used to create this tree are the same as in Lloyd et al., (2016). The phylogeny contains 536 species and was built mostly using morphological data, as obtained from fossil pseudosuchians and also molecular sequence data used to place the extant taxa in the tree. The phylogeny has been calibrated to geological time. This means that the branch lengths are measured in units of millions of years and represent times in the geological past such as origination and extinctions of species.

2) This contains diversification dynamics data for the Pseudosuchia phylogeny. Put simply, diversification dynamics are composed of speciation rates and extinction rates. You will be looking at speciation rates in the workshops.

The diversification dynamics were modelled from the tree in a Bayesian framework using the fossilBAMM software (Rabosky, 2014; Mitchell *et al*., 2018). You will not be using this software in the workshops but you might find it interesting to read the papers and to take a look at the online documentation (BAMM documentation). It may also be useful to look at the documentation for the associated *R* package *BAMMtools* (Rabosky *et al*., 2014), which we will be using in the workshops.

3) This contains habitat data for most of the Pseudosuchia species, classified as either “terrestrial” or “marine”. You will be using the terrestrial partition in the workshops.

4a) This is a time series for global temperature through time. The first column is geological time, measured in millions of years, the second column is the temperature proxy. The temperature proxy is a ratio, giving us a relative measure of temperature through time, the numbers do not correspond to degrees of temperature. These data are from Veizer *et al*., (1999) and Zachos (2001).

4b) This is a time series for global sea level through time. The first column is geological time, measured in millions of years, the second column is sea level, measured in metres. These data are from Haq *et al*., 1987.

What can we do with these data?

Combined, these data can be used to analyse the environmental drivers of speciation and extinction of the species in the phylogeny through geological time. See the analysis

guidance document for tips and hints on how to analyse these data and some ideas for your independent analysis.

References

BAMM documentation: http://bamm-project.org/documentation.html

Haq BU, Hardenbol J & Vail PR. 1987. The chronology of fluctuating sea level since the Triassic: *Science* 235 (4793): 1156–1167. https://science.sciencemag.org/content/235/4793/1156

Lloyd GT, Bapst DW, Friedman M & Davis KE. (2016). Probabilistic divergence time estimation without branch lengths: dating the origins of dinosaurs, avian flight and crown birds. *Biology Letters* 12 (20160609). https://royalsocietypublishing.org/doi/pdf/10.1098/rsbl.2016.0609

Mitchell JS, Etienne RS & Rabosky DL. 2018. Inferring Diversification Rate Variation From Phylogenies With Fossils. *Systematic Biology* 68 (1): 1-18. https://academic.oup.com/sysbio/article/68/1/1/4999317

Payne ARD, Lloyd GT, Mannion PD & Davis KE. (In prep). Decoupling speciation and extinction reveals both abiotic and biotic drivers shaped 250 million years of diversity of crocodile-line archosaurs.

Rabosky DL. 2014. Automatic detection of key innovations, rate shifts, and diversity-dependence on phylogenetic trees. *PLoS ONE* 9 (e89543). https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0089543

Rabosky DL, Grundler M, Anderson C, Title P, Shi JJ, Brown JW, Huang H & Larson JG. 2014. BAMMtools: an R package for the analysis of evolutionary dynamics on phylogenetic trees. *Methods Ecol. Evol.* 5: 701–707. https://besjournals.onlinelibrary.wiley.com/doi/full/10.1111/2041-210X.12199

Veizer J, Ala D, Azmy K, Bruckschen P, Buhl D, Bruhn F, Carden GAF, Diener A, Ebneth S, Godderis Y, Jasper T, Korte C, Pawellek F, Podlaha OG & Strauss H. 1999. 87Sr/86Sr, δ13C and δ18O evolution of Phanerozoic seawater. *Chem. Geol.* 161: 59–88.

Zachos J. 2001. Trends, rhythms, and aberrations in global climate 65 ma to present. *Science* 292**:** 686-693.