Subject Section

This is a title

Corresponding Author [1,*], Co-Author [2] and Co-Author [2,*]

[1] Department, Institution, City, Post Code, Country and

[2] Department, Institution, City, Post Code, Country.

[*] To whom correspondence should be addressed.

Associate Editor: XXXXXXX

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Abstract

Motivation: Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text.

Results: Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text

Availability: Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text

Contact: name@bio.com

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

A common procedure in biological studies is the selection of representative sequences. Because the genome databases are commonly populated with species of interest for medical and economic research, one has to account for a taxonomic bias in the sample selection.

- Gains from the extended approach (breaking assumptions and filtering taxa, implications).

- use for low number of sequences/inputs

- stochastic procedure that still maximizes diversity

- allows for mixed taxonomic levels being used as input (important for NCBI taxonomy)

- impact over bias (?)

2 Approach

A sample is balanced if any two taxa with a common parent taxon receive the same number of observations in a sample, and diversified if it maximizes the number of taxa observed in each taxonomic level of a sample.

A random uniform sampling, in which each input sequence has the same probability of being sampled as any other sequence, is susceptible to any bias in the input. This can result in samples that aren't neither balanced nor diversified. This event happens when a few taxa have disproportionally more sequences than the others. The probability of obtaining any other taxon diminishes in a random uniform sampling, and may not appear at all in most cases.

The original algorithm is a recursive attribution of m observations from n inputs.

Given a non-leaf taxon node, the goal is to attribute a number of observations (m_i) for each of its children nodes (or sub-taxa) with at least one sequence, such that $sum(m_i) = m$. The algorithm is then recursively applied for each child node, in which the attributed m_i is passed as the child's m.

The method by which we attribute m_i to each individual node i determines how balanced or diversified the output will be.

The original algorithm attributes the mean for each node i, varying at most by 1 for when the mean isn't an integer, and achieves a balanced sample in terms of taxonomic representation. In other words, a given taxon won't have more sub-taxa sampled than another taxon in the same taxonomic level. With that, the algorithm limits bias from overrepresented or underrepresented taxa in the input, which can cause many taxa to have no samples, or a few ones being consistently sampled. This relies on the assumption that $n_i >= mean$ for every i, and requires repeating sequences to work if it is violated.

By changing the method to realocate excess attributions of m_i (when $n_i < mean(m_i)$) among other taxa, the algorithm can maximize diversity, as the likelihood

- assumption of $n_i >= m_i$ - balance vs diversity - randomized approach - cases when a taxon isn't a leaf node - filtering: requiring certain taxa (and their offspring) to be present or absent - n_i as sequences vs n_i as taxa (multiple sequences vs one sequence per taxon)

The algorithm requires previous knowledge of how many instances of each taxon exist in the input, where an instance is when a given input taxonomy ID is that taxon, or has it as an ancestor. This is obtained by searching the ancestor IDs of each input ID and keeping a table with the numeber occurences of each taxon, both input and ancestor.

Fig. 1. Caption, caption.

3 Methods

- for bulleted list, use itemize
- for bulleted list, use itemize
- for bulleted list, use itemize

3.1 This is subheading

3.1.1 This is subsubheading

3.2 Test1

4 Discussion

5 Conclusion

Acknowledgements

Funding

References

Bofelli,F., Name2, Name3 (2003) Article title, *Journal Name*, **199**, 133-154.

Bag,M., Name2, Name3 (2001) Article title, *Journal Name*, **99**, 33-54.

Yoo,M.S. *et al*. (2003) Oxidative stress regulated genes in nigral dopaminergic neur-nol cell: correlation with the known pathology in Parkinson's disease. *Brain Res. Mol. Brain Res.*, **110**(Suppl. 1), 76–84.

Lehmann,E.L. (1986) Chapter title. *Book Title*. Vol. 1, 2nd edn. Springer-Verlag, New York.

Crenshaw, B.,III, and Jones, W.B.,Jr (2003) The future of clinical cancer manage-ment: one tumor, one chip. *Bioinformatics*, doi:10.1093/bioinformatics/btn000.

Auhtor,A.B. *et al*. (2000) Chapter title. In Smith, A.C. (ed.), *Book Title*, 2nd edn. Publisher, Location, Vol. 1, pp. ???–???.

Bardet, G. (1920) Sur un syndrome d'obesite infantile avec polydactylie et retinite pigmentaire (contribution a l'etude des formes cliniques de l'obesite hypophysaire). PhD Thesis, name of institution, Paris, France.