Technical details for: A profile-based method for measuring the impact of genetic variation

Nicole E. Wheeler[1]*, Lars Barquist[2], Fatemeh Ashari Ghomi[1], Robert Kingsley[3], Paul P. Gardner[1,4]

Abstract
In the following we provide some mathematical justification for the Delta bitscore metric that we evaluate in the accompanying manuscript.

Keywords
genome variation — genotype — phenotype

[1] *School of Biological Sciences, University of Canterbury, Christchurch, New Zealand.*
[2] *Institute for Molecular Infection Biology, University of Wuerzburg, Wuerzburg, Germany.*
[3] *Institute of Food Research, Norwich Research Park, Norwich, Norfolk, United Kingdom.*
[4] *Biomolecular Interaction Centre and the Bio-Protection Research Centre, University of Canterbury, Christchurch, New Zealand.*
*__Corresponding author__: nicole.wheeler@pg.canterbury.ac.nz

Introduction

In this document we compare the mathematics between previously published profile HMM based methods for quantifying the likely phenotypic significance of genetic variation and our approach. Namely, the *logR.E-value* method [1], *FATHMM* [2, 3, 4] and *DBS* (this study).

1. Methods

1.1 logR.E-value

Clifford *et al.* (2004) suggest using the following measure to estimate the significance of a genetic variant:

$$\log R.E = \log_{10}\left(\frac{E-value_{var}}{E-value_{can}}\right) \quad (1)$$

Where $E-value_{var}$ and $E-value_{can}$ correspond to the expectation value derived from HMMER matches (to the same model) for a variant (*var*) and canonical (*can*) protein sequence.

$E-values$ are generally estimated by fitting an exponential distribution to an empirical (usually simulated) distribution. I.e.

$$E-value = \kappa MN e^{\lambda x} \quad (2)$$

Where x is the bit-score for a match between a profile HMM and a sequence, MN is the product of the database size and the model length and, finally κ and λ are parameters that ensure the intercept with the y-axis is correct and that the curve matches an empirical distribution .

In a breakthrough theoretical paper by Sean Eddy [5], he showed that the most computationally expensive parameter to estimate (λ) is a constant i.e. $\lambda = \ln(2)$.

Thus Equation 1 can be rewritten as:

$$\begin{aligned} \log R.E &= \log_{10}\left(e^{\lambda x_{var}}\right) - \log_{10}\left(e^{\lambda x_{can}}\right) \quad (3)\\ &= (x_{var} - x_{can}) * \log_{10}(e^{\lambda}) \\ &= DBS * constant \end{aligned}$$

If the base for the exponential and the logarithms had been equal, then *constant* the constant would equal λ. In either case, a constant multiplied by the difference between two bitscores is all that remains.

1.2 FATHMM

Shihab *et al* (2013) define the following unweighted measure for estimating the significance of a single non-synonymous SNP (the weighted version is trained to discriminate human disease from polymorphic variation, therefore is not directly comparable to our general approach). Their metric is a logit or log-odds value, comparing the emission probability of the wild-type variant (P_w) and a mutant variant (P_m) when the mutant is a single, non-synonymous point mutation (i.e. not a multiple point mutations or indels):

$$\begin{aligned} unweighted &= \ln\left(\frac{\frac{P_m}{1-P_m}}{\frac{P_w}{1-P_w}}\right) \quad (4)\\ &= \ln\left(\frac{P_m}{P_w}\right) + \ln\left(\frac{1-P_w}{1-P_m}\right) \quad (5)\\ &\approx DBS + \ln\left(\frac{1-P_w}{1-P_m}\right) \quad (6) \end{aligned}$$

The value $1-P_w$ and $1-P_m$ can be re-written as the following summation:

$$1 - P_w = \sum_{i \in amino-acids, i \neq w} P_i \quad (7)$$

$$1 - P_m = \sum_{j \in amino-acids, j \neq m} P_j \quad (8)$$

Equations 7&8 share 18 terms (for each of the 20 amino acids, less the ones corresponding to the wild-type (w) and mutant (m) variants. Therefore, $1 - P_w \approx 1 - P_m$ for most realistic biological results. As a consequence, the second term of Equation 6 is approximately zero (or at least, modest in comparison to the first term when there is a large difference between P_w and P_m). Therefore a difference between bitscores is the term that dominates Equation 4 (see the discussion below).

1.3 Delta bitscore (DBS)

Using the same nomenclature as above, we define delta bitscore (DBS) as:

$$DBS = (x_{var} - x_{can}) \quad (9)$$

The bitscore (x) for an HMM is defined as a product log of probability ratios [6]:

$$x = \log_2 \left(\frac{P(seq|M)}{P(seq|N)} \right) \quad (10)$$

Where M is a profile model derived from a sequence alignment. M generates and scores sequences based upon how likely they are to have generated by the same process as those in the sequence alignment. N is a null model, that generates and scores sequences based upon how likely they are to have generated by a random process.

Therefore, *DBS* can be re-written as:

$$DBS(seq_{var}, seq_{can}) = \log_2 \left(\frac{P(seq_{var}|M)}{P(seq_{var}|N)} \right) - \log_2 \left(\frac{P(seq_{can}|M)}{P(seq_{can}|N)} \right) \quad (11)$$

$$\approx \log_2 \left(\frac{P(seq_{var}|M)}{P(seq_{can}|M)} \right) \quad (12)$$

If we make the simplifying assumption that the null models for $P(seq_{var}|N)$ and $P(seq_{can}|N)$ are approximately equal (i.e. equal length and amino acid composition). Therefore, the first term of Equation 5 and Equation 12 are, in most situations, equivalent.

2. Discussion

As a consequence, the measures used by the log*R.E-value* (Equation 3) and the *FATHMM* (Equation 6) approach are approximations to the more direct estimation of significance, *DBS*. In the case of *FATHMM*, only single point mutations are considered, missing the wealth of variation due to insertions, deletions, multiple SNPs and and other larger-scale variants.

Consequently, *DBS* is a direct measure of the potential impact of genetic variation, that can be used on small as well as large and complex variants. We propose that this metric can be used to evaluate both population variation as well as variation between species. The mean of the distribution should be approximately zero, while the variance will increase with increasing phylogenetic distance (and different levels of selection).

One factor that may have an undue influence on *DBS* is in the rare cases where the optimal alignment between a the profile and the variant and the profile and the canonical sequence differ. For example, *HMMER3* currently only has a local mode (i.e. no "glocal" option). Therefore, splitting matches can happen, also slipped alignments also occur, particularly for repetitive sequences.

One way to mitigate these possibilities is to use *Forward Scores*, which rather than reporting just the value for an optimal alignment, reports instead the sum of all possible alignments between a query sequence and the profile model.

References

[1] R J Clifford, M N Edmonson, C Nguyen, and K H Buetow. Large-scale analysis of non-synonymous coding region single nucleotide polymorphisms. *Bioinformatics*, 20(7):1006–14, May 2004.

[2] H A Shihab, J Gough, D N Cooper, P D Stenson, G L Barker, K J Edwards, I N Day, and T R Gaunt. Predicting the functional, molecular, and phenotypic consequences of amino acid substitutions using hidden markov models. *Hum Mutat*, 34(1):57–65, Jan 2013.

[3] H A Shihab, J Gough, D N Cooper, I N Day, and T R Gaunt. Predicting the functional consequences of cancer-associated amino acid substitutions. *Bioinformatics*, 29(12):1504–10, Jun 2013.

[4] H A Shihab, J Gough, M Mort, D N Cooper, I N Day, and T R Gaunt. Ranking non-synonymous single nucleotide polymorphisms based on disease concepts. *Hum Genomics*, 8:11, 2014.

[5] S R Eddy. A probabilistic model of local sequence alignment that simplifies statistical significance estimation. *PLoS Comput Biol*, 4(5):e1000069, May 2008.

[6] R. Durbin, S. Eddy, A. Krogh, and G. Mitchison. *Biological sequence analysis*. Press, Cambridge U., 1998.