Enrich: software for analysis of protein function by enrichment and depletion of variants

Douglas M. Fowler1,∗, Carlos L. Araya1, Wayne Gerard1 and Stanley Fields1,2,3

1Department of Genome Sciences, 2Howard Hughes Medical Institute and 3Department of Medicine, University of Washington, Seattle, WA 98195, USA

Supplementary Information: Supplementary data is available at Bioinformatics online.

Received on June 8, 2011; revised on October 5, 2011; accepted on October 11, 2011

ABSTRACT

Summary: Measuring the consequences of mutation in proteins is critical to understanding their function. These measurements are essential in such applications as protein engineering, drug development, protein design and genome sequence analysis. Recently, high-throughput sequencing has been coupled to assays of protein activity, enabling the analysis of large numbers of mutations in parallel. We present Enrich, a tool for analyzing such deep mutational scanning data. Enrich identifies all unique variants (mutants) of a protein in high-throughput sequencing datasets and can correct for sequencing errors using overlapping paired-end reads. Enrich uses the frequency of each variant before and after selection to calculate an enrichment ratio, which is used to estimate fitness. Enrich provides an interactive interface to guide users. It generates user-accessible output for downstream analyses as well as several visualizations of the effects of mutation on function, thereby allowing the user to rapidly quantify and comprehend sequence–function relationships.

Availability and Implementation: Enrich is implemented in Python and is available under a FreeBSD license at http://depth.washington.edu/sfields/software/enrich/. Enrich includes detailed documentation as well as a small example dataset.

Contact: dfowler@uw.edu; fields@uw.edu

Supplementary Information: Supplementary data is available at Bioinformatics online.

1 INTRODUCTION

Understanding how variations in protein sequence relate to function is of fundamental importance. Measurement of protein activity is critical to engineer protein function, to understand how mutations relate to disease and to gain insight into catalytic mechanisms (Alper et al., 2006; Kato et al., 2003; Weiss et al., 2000). Efforts to parallelize measurement of protein activity rely on selection for a desired function present within a library of variants of a protein of interest using a display-based system that directly links a protein’s activity to its encoding DNA sequence (Levin and Weiss, 2006; Pal et al., 2006; Sidhu and Koide, 2007). Selection for function (e.g. ligand binding, catalytic activity or stability) alters the population of displayed proteins, and thus their associated DNA molecules. DNA sequences encoding highly functional variants are enriched, whereas DNA sequences encoding poorly functional variants are depleted.

Recent high-throughput sequencing has been used to significantly increase the number of variants assessed (Díaz-Neto et al., 2009; Ernst et al., 2010; Fowler et al., 2010; Hentpas et al., 2011; Hinkley et al., 2011; Ravn et al., 2010). Such ‘deep mutational scanning’ (Araya and Fowler, 2011) experiments engender significant analysis challenges.

Here, we present Enrich, a tool to address these challenges. Enrich identifies and enumerates unique protein sequences within high-throughput sequencing data. It calculates an enrichment ratio between unselected and selected libraries for each unique variant, and it creates a number of visualizations. Enrich is open-source, freely available and modular, creating easy-to-manipulate output files. Thus, users can customize Enrich and perform project-specific analysis.

2 APPROACH

Enrich is implemented in Python. Enrich requires ~2 h to run on a typical dataset on a desktop computer. To facilitate the analysis of multiple datasets in parallel, Enrich can function in a high-performance computing environment managed by the Oracle Grid Engine. Enrich uses the DRMAA distributed resource management API to facilitate extension to other environments (http://drmaa.org/).

Enrich supports command line execution and an interactive mode that guides users through the configuration and execution of Enrich runs.

Enrich takes as input FASTQ-formatted high-throughput sequencing data files acquired from an unselected and a selected library (Cock et al., 2010). Enrich can use reads from any sequencing platform, provided they are FASTQ-formatted. If overlapping paired-end reads have been acquired, Enrich corrects each read pair for sequencing error by examining agreement between the reads. At positions where the reads disagree, the nucleotide with the higher quality score is used. If both reads have identical quality scores at the position in question, the read pair is removed. More robust error models could improve error correction, particularly when overlapping paired-end reads are not available (e.g. ShoRAH) (Zagordi et al., 2011).

Variant sequences are identified and enumerated within the unselected and selected libraries. Variants containing insertions and deletions are removed. An enrichment ratio (selected/unselected)
frequency between two libraries is shown. (mutation combination. (within a library is illustrated by a heatmap of the frequency of each position–unobserved mutations. Red squares correspond to wild-type residues. Grey squares correspond to organized both by position and by amino acid (a single amino acid, serine, ratio for each position–mutation combination is plotted, individually providing these visualization options, Enrich produces easy to use enrichment and an all-residue enrichment ratio scan. In addition to starting point for further analyses (Fig. 1). The visualizations show (for a more detailed description, see the Supplementary Material).


Enrich visualizations. Enrich produces three visualizations; examples from the dataset included with Enrich are shown here. (a) The diversity within a library is illustrated by a heatmap of the frequency of each position–mutation combination. (b) The position-averaged change in mutational frequency between two libraries is shown. (c) The log2-scaled enrichment ratio for each position-mutation combination is plotted, individually organized both by position and by amino acid (a single amino acid, serine, is shown). Blue dots indicate the enrichment or depletion of substitutions. Red squares correspond to wild-type residues. Grey squares correspond to unobserved mutations.

is calculated for each variant. Enrichment ratios are evaluated using a two-sided Poisson exact test to calculate a P-value for the significance of enrichment or depletion for each variant. Multiple testing correction is performed using false discovery rates (Storey and Tibshirani, 2003). The resulting q-values enable the user to identify subsets of variants whose frequency is significantly altered by selection. To accomplish these tasks, the Enrich workflow is divided into seven modules that can run independently or all together (for a more detailed description, see the Supplementary Material).

Enrich uses matplotlib to produce any of three visualizations as a starting point for further analyses (Fig. 1). The visualizations show an estimation of library diversity, the position-averaged mutation enrichment and an all-residue enrichment ratio scan. In addition to providing these visualization options, Enrich produces easy to use output files that can be carried forward into project-specific analyses. Enrich can take advantage of high-performance computing to conduct many analyses in parallel. Enrich’s Python-based modular, extensible design enables users to customize the software. Enrich facilitates deep mutational scanning, which can be widely applied to the breadth of disciplines that depend on understanding protein sequence–function relationships.

ACKNOWLEDGEMENT

We thank Alan Rubin, Michael Hoffman and William Noble for helpful discussion and advice.

Funding: National Institutes of Health (P41RR11823 to S.F. and F32GM084699 to D.M.F.). S.F. is an investigator of the Howard Hughes Medical Institute.

Conflict of Interest: none declared.

REFERENCES


Zagordi, O. et al. (2011) ShoRAH: estimating the genetic diversity of a mixed sample from next-generation sequencing data. BMC Bioinformatics, 12, 119.