Bioinformatics Advance Access published April 6, 2015

Sequence analysis

PROVEAN web server: a tool to predict the functional effect of amino acid substitutions and indels

Yongwook Choi and Agnes P. Chan*
The J. Craig Venter Institute, Rockville, MD 20850, USA.

Associate Editor: Dr. John Hancock

ABSTRACT
Summary: We present a web server to predict the functional effect of single or multiple amino acid substitutions, insertions, and deletions using the prediction tool PROVEAN. The server provides rapid analysis of protein variants from any organisms, and also supports high-throughput analysis for human and mouse variants at both the genomic and protein levels.
Availability: The web server is freely available and open to all users with no login requirements at http://provean.jcvi.org.
Contact: achan@jcvi.org
Supplementary information: Supplementary data are available at Bioinformatics online.

1 INTRODUCTION

Large scale genotyping and genome re-sequencing projects have generated a large amount of sequence variant data for healthy and diseased individuals in human and model organisms. Sequence variants may reside in the protein-coding or non-coding regions of the genome. Whereas non-coding variants can potentially play a regulatory role and affect gene expression, coding variants can affect protein function and activity by altering the gene product such as substitution, insertion, or deletion of amino acids, frameshift, or truncation.

As the number of genome-wide variants is usually large, it is a challenging problem to identify the causal variant for a disease or specific phenotype of interest. For this purpose, computational tools such as PolyPhen (Adzhubei, et al., 2010; Sunyaev, et al., 1999) and SIFT (Kumar, et al., 2009; Ng and Henikoff, 2001) were developed to provide functional predictions for coding variants. Since then many other tools have been developed until recently (Cooper and Shendure, 2011; Ng and Henikoff, 2006). We have developed a new tool PROVEAN (Protein Variation Effect Analyzer), which uses an alignment-based score approach (Choi, 2012; Choi, et al., 2012). Unlike most existing tools, PROVEAN can generate predictions not only for single amino acid substitutions but also for multiple amino acid substitutions, insertions, and deletions using the same underlying scoring scheme. As the original PROVEAN web server (Choi, et al., 2012) was limited to supporting the analysis of single protein queries, we have expanded the PROVEAN web server to support high-throughput online analysis using the PROVEAN tool. The PROVEAN web server can now provide precomputed prediction for large sets of genome-wide nucleotide or amino acid variants for both human and mouse. For the original single protein query function, run time has been greatly reduced in the current new version by storing and reusing previously computed homologous protein sequences (supporting sequences) information for query proteins.

*To whom correspondence should be addressed.

2 RESULTS

2.1 Background

The PROVEAN method has been previously described and validated (Choi, et al., 2012). In order to build a large database of precomputed PROVEAN scores, an efficient method to compute the pairwise alignment scores between a protein sequence and a large number of single-locus variations of another protein has been previously developed (Choi, 2012).

2.2 Web server function

The web server currently supports three functions: (1) PROVEAN Protein for any organisms, (2) PROVEAN Protein Batch (human and mouse) and (3) PROVEAN Genome Variants (human and mouse). The web server is supported by an in-house MySQL database. Major data tables include Ensembl gene annotation, precomputed PROVEAN scores and precomputed supporting sequences for all protein sequences from human and mouse, and supporting sequence sets captured from first-time online user submissions of protein sequences that may originate from any organisms. The web server uses multiple queues for job submissions based on the expected job size so that small jobs are handled promptly on a designated queue to ensure a fast turnover. The database schema and data structure used in each of the three web server functions are described in the following subsections.

2.2.1. *PROVEAN Protein* The "PROVEAN Protein" tool provides online access to the function of the stand-alone PROVEAN software package distribution. Its primary function is to provide a prediction for a protein sequence originated from any organisms. The tool accepts a protein sequence and amino acid variations as input, performs a BLAST search to identify homologous sequences (supporting sequences), and generates PROVEAN scores. In general, it takes 10-20 minutes to generate prediction for a given protein query. To improve performance, we have implemented a caching approach to by-pass the most time-consuming steps of protein database search to collect homologous sequences, and clustering. In the caching approach, the supporting

sequences for all first-time protein query submissions were stored. A list of sequence identifiers for the supporting sequences and clustering information are stored in the database. The supporting sequence data are indexed based on the query protein sequence and reused in subsequent prediction requests. We showed that the implementation of the caching approach for intermediate results greatly reduces the overall run time from several minutes to a few seconds.

2.2.2. *PROVEAN Protein Batch* The "PROVEAN Protein Batch" function supports batch processing for a large number of protein variations based on precomputes. The input is a list of protein variations, with each variation consisting of a public protein identifier, amino acid position, reference amino acid, and variant amino acid. Public protein identifiers supported by this function include those from NCBI RefSeq (e.g. NP_000537.3), UniProt (e.g. P13569), and Ensembl (e.g. ENSP00000363868). Two approaches have been implemented to ensure a fast response time: precomputed scores and precomputed supporting sequences.

In the first approach, PROVEAN scores are directly retrieved from a database with precomputed scores. The precomputed scores have been generated for all protein sequences in our supported organisms (~90k human and ~46K mouse proteins) using the method described in (Choi, 2012). The precomputed scores cover 20 single amino acid substitutions and one amino acid deletion for every amino acid position.

In the second approach, PROVEAN prediction starts from precomputed supporting sequences. Due to a potentially infinitely large number of combinations for multiple amino acid substitutions, insertions, and deletions, the precomputed score approach is not applicable for all different types and lengths of amino acid changes. Thus, a supporting sequence set is precomputed and stored for every protein sequence in our supported organisms. The supporting sequence sets are then used for fast computation of prediction for amino acid variations that are not supported by the precomputed score approach. The scores computed in this approach are stored in database and reused for requests for the same variants.

2.2.3. *PROVEAN Genome Variants* The "PROVEAN Genome Variants" function supports batch processing for a large number of variations found across the entire genome. This tool accepts a list of genomic variations such as single nucleotide polymorphisms (SNPs), multiple nucleotide substitutions, insertions, or deletions. First, the genomic variations are classified as coding and non-coding variations based on the reference genome sequence and the Ensembl gene annotation. Second, the coding variants are classified at the protein sequence level as amino acid substitutions, insertions, deletions, nonsense mutations, or frameshifts. For amino acid substitutions, insertions, and deletions, PROVEAN scores and predictions are obtained in a similar approach as described above in the PROVEAN Protein Batch function. That is, the PROVEAN scores are either retrieved from the in-house database or computed using precomputed supporting sequence set information. The PROVEAN Genome Variants function also provides accessory information for the genomic variation input data including NCBI dbSNP reference accessions, and gene annotation obtained from Ensembl BioMart such as gene description, PFAM domain, and Gene Ontology. The steps for this function are shown in a flowchart in the Supplementary Material (Figure S1).

For efficient data storage, the PROVEAN scores are stored not at the genome level but at the protein sequence level. The coordinate conversion from genome variants to the corresponding protein variants is achieved instantaneously by real-time dynamic conversion. To support efficient coordinate conversion, the original Ensembl gene annotation in GFF/GTF format is re-organized so that each chromosomal position in the coding region is stored in a record along with the underlying nucleotide, codon triplet, reading frame, position of the codon in amino acid, gene ID, and gene orientation. Given a single nucleotide variant in a coding region, the record for the specific chromosomal position is retrieved. The amino acid change can be quickly determined based on the codon and reading frame information in the record. The conversion for other types of genomic variants to protein level alterations requires additional steps but can be achieved in a similarly straightforward way via the chromosomal position records.

3 DISCUSSION

To provide binary predictions, the cutoff for PROVEAN scores was set to -2.5 for high balanced accuracy. However, the users can reapply their own cutoffs for their analysis to achieve either higher sensitivity or higher specificity. The sensitivity and specificity obtained using different score cutoffs are shown in the Supplementary Material (Figure S2).

It was shown that the performance of PROVEAN for single amino acid substitutions is comparable with other tools (Choi, et al., 2012). Here we also examine the prediction consistency with other tools, SIFT and PolyPhen-2, using the UniProt human polymorphisms and disease mutations dataset. The three tools agree on a large portion of the variants, but there are still many variants for which three tools make different predictions. The results are summarized using Venn diagrams in the Supplementary Material (Figure S3).

Funding: This work was supported by the National Institutes of Health [5R01HG004701-04].

Conflict of interest: None declared.

REFERENCES

Adzhubei, I.A., *et al.* (2010) A method and server for predicting damaging missense mutations, *Nat Methods*, 7, 248-249.

Choi, Y. (2012) A fast computation of pairwise sequence alignment scores between a protein and a set of single-locus variants of another protein. *Proceedings of the ACM Conference on Bioinformatics, Computational Biology and Biomedicine*. ACM, Orlando, Florida, pp. 414-417.

Choi, Y., *et al.* (2012) Predicting the functional effect of amino acid substitutions and indels, *PloS one*, 7, e46688.

Cooper, G.M. and Shendure, J. (2011) Needles in stacks of needles: finding disease-causal variants in a wealth of genomic data, *Nature reviews. Genetics*, 12, 628-640.

Kumar, P., Henikoff, S. and Ng, P.C. (2009) Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm, *Nature protocols*, 4, 1073-1081.

Ng, P.C. and Henikoff, S. (2001) Predicting deleterious amino acid substitutions, *Genome research*, 11, 863-874.

Ng, P.C. and Henikoff, S. (2006) Predicting the effects of amino acid substitutions on protein function, *Annual review of genomics and human genetics*, 7, 61-80.

Sunyaev, S.R., *et al.* (1999) PSIC: profile extraction from sequence alignments with position-specific counts of independent observations, *Protein engineering*, 12, 387-394.