BIOINFORMATICS APPLICATIONS NOTE

Vol. 27 no. 15 2011, pages 2147–2148
doi:10.1093/bioinformatics/btr357

Genome analysis

Advance Access publication June 17, 2011

CHASM and SNVBox: toolkit for detecting biologically important single nucleotide mutations in cancer

Wing Chung Wong[1,†], Dewey Kim[1,†], Hannah Carter[1], Mark Diekhans[2], Michael C. Ryan[3] and Rachel Karchin[1,*]

[1]Department of Biomedical Engineering and Institute for Computational Medicine, Johns Hopkins University, Baltimore, MD, [2]Center for Biomolecular Science and Engineering, University of California, Santa Cruz, CA and [3]In Silico Solutions, Fairfax, VA, USA

Associate Editor: Alex Bateman

ABSTRACT

Summary: Thousands of cancer exomes are currently being sequenced, yielding millions of non-synonymous single nucleotide variants (SNVs) of possible relevance to disease etiology. Here, we provide a software toolkit to prioritize SNVs based on their predicted contribution to tumorigenesis. It includes a database of precomputed, predictive features covering all positions in the annotated human exome and can be used either stand-alone or as part of a larger variant discovery pipeline.

Availability and Implementation: MySQL database, source code and binaries freely available for academic/government use at http://wiki.chasmsoftware.org, Source in Python and C++. Requires 32 or 64-bit Linux system (tested on Fedora Core 8,10,11 and Ubuntu 10), 2.5∗≤ Python <3.0∗, MySQL server >5.0, 60 GB available hard disk space (50 MB for software and data files, 40 GB for MySQL database dump when uncompressed), 2 GB of RAM.

Contact: karchin@jhu.edu

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

Received on April 13, 2011; revised on June 1, 2011; accepted on June 8, 2011

1 INTRODUCTION

A fundamental goal of modern cancer genomics studies is to understand how alterations in DNA sequence contribute to disease susceptibility and prognosis. Targeted whole-exome deep sequencing is now affordable for many academic labs and the multitude of studies underway is yielding datasets of unprecedented magnitude. While researchers have previously developed methods to computationally predict the impact of single nucleotide variants (SNVs) (Kaminker *et al.*, 2007; Mooney *et al.*, 2010; Ng and Henikoff, 2003; Sunyaev *et al.*, 2001), to our knowledge there are no existing tools capable of fast classification of very large SNV datasets in cancer exomes.

We have previously developed a computational method Cancer-Specific High-throughput Annotation of Somatic Mutations (CHASM) (Carter *et al.*, 2009, 2010) that predicts whether tumor-derived somatic missense mutations are important contributors to cancer cell fitness. Here, we describe a software package that implements the CHASM method. The package includes a database of pre-computed predictive features called SNVBox that facilitates rapid feature retrieval and classification of very large SNV datasets. Furthermore, the features in SNVBox can be generally used to aid in the development of new classification algorithms that predict the impact of either germline or somatic SNVs.

2 METHODS AND IMPLEMENTATION

CHASM is an open-source collection of Python and C++ programs that takes a list of somatic missense mutations as input and ranks them according to their likely tumorigenic impact. It includes a curated set of driver mutations culled from the COSMIC database (Forbes *et al.*, 2008), which is used as a positive class for training a Random Forest classifier (Amit and Geman, 1997; Breiman, 2001). The negative class of mutations is generated *in silico* according to an estimated distribution of benign (passenger) variation, matched to the tumor type of interest. Users have the option to use their own estimates of passenger variant frequencies or to select from a library of pre-computed passenger frequency tables for several common cancers.

PyInstaller 1.4 was used to package Python source into dynamically linked, executable binaries. The `SnvGet`, `Build Classifier` and `RunChasm` executables are run by the user on the command line, while the others are called internally. The statically compiled C++ executable `waffles_learn` from the WAFFLES machine learning library is also called internally.

SNVBox is an MySQL database of 86 predictive features relevant to the biological impact of an SNV. The features have been pre-computed for each codon in all protein-coding exons of annotated human mRNA transcripts in the NCBI RefSeq, CCDS and EBI Ensembl databases (Birney *et al.*, 2004; Pruitt *et al.*, 2007, 2009). The `SnvGet` program enables fast retrieval of selected features from the database for classifier training and scoring of mutations input by the user.

3 WORKFLOW

(1) Prepare an input file of estimated passenger mutation rates in the cancer of interest. Optionally, select from one of several pre-computed passenger rate tables.

[*]To whom correspondence should be addressed.

[†]The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors.

(2) Prepare an input file of missense SNVs to be classified. Each row contains a protein accession identifier, codon number, and reference and variant amino acid residues.

(3) Run the `BuildClassifier` program.
- Produces a negative class of *in silico* passenger mutations by random nucleotide substitution in a library of expressed human mRNA transcripts from NCBI RefSeq, according to the distributions specified in the passenger mutation rate table (Supplementary Material).
- Retrieves a list of predictive features for each passenger (and driver) in the training set from SNVBox.
- Builds a Random Forest classifier, using `waffles_learn`.

(4) Run the `RunChasm` program.
- Retreives a feature list for all mutations supplied by the user.
- Applies the trained classifier to generate a CHASM score for each variant.
- Generates a second set of *in silico* passenger mutations, which (unlike the first set) is carefully filtered to eliminate mutations in any genes previously associated with cancer in either the Cancer Gene Census (Futreal *et al.*, 2004), the COSMIC cancer gene list and all cancer (C4 collection) genesets in MSigDB (Subramanian *et al.*, 2005).
- Filtered passengers are scored by the classifier to produce an empirical null distribution of variant scores.
- This null score distribution is used to compute a *P*-value for each variant supplied by the user (fraction of filtered passengers having CHASM scores less than or equal to the score of the variant).
- Benjamini–Hochberg multiple testing correction (Benjamini and Hochberg, 1995) is applied to the *P*-values.
- Outputs a list of the user-supplied mutations, with CHASM scores, *P*-values and Benjamini–Hochberg estimated false discovery rate (FDR).
- Outputs an ARFF formatted file of features for the submitted mutations.

4 DISCUSSION

The CHASM/SNVBox toolkit is the first distributable software package that specifically targets somatic missense mutations in cancer. The learning task of the Random Forest classifier is to discriminate between known drivers and a set of random passenger missense mutations that match the mutation spectrum in a cancer type of interest. CHASM results are sensitive to this definition of mutation spectrum and users are encouraged to use the somatic variant calls from their sequencing data to make the best possible estimates of the spectrum (Supplementary Material).

While many SNV classifiers are available through web interfaces [reviewed in Karchin (2009)], these are not currently capable of handling large size custom datasets (e.g. thousands to millions of SNVs discovered in sequencing projects). Some researchers have developed distributable packages that users can run on their local system to enable high-throughput SNV processing. These packages depend on third-party databases (sequences, alignments, protein structures, specialized protein annotations) and third-party software packages. The popular PolyPhen system, for example, requires installation of 10 third-party software packages, in addition to three Perl modules. To our knowledge, all available SNV classification tools base their inferences on predictive features computed when a custom dataset is input to the system (almost always using third-party databases and software). In contrast, the predictive features available in SNVBox (also calculated with many third-party tools) have been exhaustively pre-computed, allowing rapid retrieval for a custom dataset. In benchmark testing, retrieval of 86 features for one million SNVs took 11.39 h on a Dell R900 server with two AMD Opteron dual-core 64 bit CPUs and 16 GBs of RAM. CHASM score computation for these one million mutations took an additional 10 min and 33 s.

Finally, the predictive features available in SNVBox were designed to be useful for classification of both germline and somatic SNVs. We hope that SNVBox will enable design of new, improved machine learning algorithms to predict the impact of SNVs.

Funding: National Cancer Institute, National Institutes of Health grant (CA135877 to R.K., in part); National Science Foundation CAREER award (DBI 0845275 to R.K., in part); DOD NDSEG Fellowship 32 CFR 168a (to H.C., in part).

Conflict of Interest: none declared.

REFERENCES

Amit,Y. and Geman,D. (1997) Shape quantization and recognition with randomized trees. *Neural Comput.*, **9**, 1545–1588.

Benjamini,Y. and Hochberg,Y. (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc. Ser. B*, **57**, 289–300.

Birney,E. *et al.* (2004) An overview of ensembl. *Genome Res.*, **14**, 925–928.

Breiman,L. (2001) Random forest. *Mach. Learn.*, **45**, 5–32.

Carter,H. *et al.* (2009) Cancer-specific high-throughput annotation of somatic mutations: computational prediction of driver missense mutations. *Cancer Res.*, **69**, 6660–6667.

Carter,H. *et al.* (2010) Prioritization of driver mutations in pancreatic cancer using cancer-specific high-throughput annotation of somatic mutations (CHASM). *Cancer Biol. Ther.*, **10**, 582–587.

Forbes,S.A. *et al.* (2008) The Catalogue of Somatic Mutations in Cancer (COSMIC). *Curr. Protoc. Hum. Genet.*, **Chapter 10**, Unit 10.11.

Futreal,P.A. *et al.* (2004) A census of human cancer genes. *Nat. Rev. Cancer*, **4**, 177–183.

Kaminker,J.S. *et al.* (2007) CanPredict: a computational tool for predicting cancer-associated missense mutations. *Nucleic Acids Res.*, **35**, W595–W598.

Karchin,R. (2009) Next generation tools for the annotation of human SNPs. *Brief Bioinform.*, **10**, 35–52.

Mooney,S.D. *et al.* (2010) Bioinformatic tools for identifying disease gene and SNP candidates. *Methods Mol. Biol.*, **628**, 307–319.

Ng,P.C. and Henikoff,S. (2003) SIFT: predicting amino acid changes that affect protein function. *Nucleic Acids Res.*, **31**, 3812–3814.

Pruitt,K.D. *et al*. (2007) NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Res.*, **35**(Suppl. 1), D61–D65.

Pruitt,K.D. *et al.* (2009) The consensus coding sequence (CCDS) project: identifying a common protein-coding gene set for the human and mouse genomes. *Genome Res.*, **19**, 1316–1323.

Subramanian,A. *et al.* (2005) Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl Acad. Sci. USA*, **102**, 15545–15550.

Sunyaev,S. *et al.* (2001) Prediction of deleterious human alleles. *Hum. Mol. Genet.*, **10**, 591–597.