Critique

Sergei Lebedev

December 9, 2012

Cancer-Specific High-Throughput Annotation of Somatic Mutations: Computation Prediction of Driver Missense Mutations

Hannah Carter, Sining Chen, Leyla Isik, Svitlana Tyekucheva, Victor E. Velculescu, Kenneth W. Kinzler, Bert Vogelstein and Rachel Karchin

Cancer Research, 2009.

Abstract

The paper presents a Random Forest based approach for identifying and scoring somatic missense mutations, which are likely to generate changes that enhance cancer cell proliferation. The knowledge of these mutations is critical for understanding mechanisms involved in tumoronegensis.

Summary

Identifying DNA alterations, which are functionally related to tumor development, is one of the most significant challenges in cancer research. Based on their contribution to tumorogenesis, all DNA alterations can be classified as either *drivers*, mutations likely to enhance tumor proliferation and *passengers*, neutral with respect to cancer cell fitness.

It has been shown, that even though genes, mutated frequently in different tumors can almost certainly be classified as drivers, a significant fraction of genes is only mutated in some tumors. Which implies, that a good classification algorithm shouldn't make any assumptions about mutation frequency.

Carter *et al.* propose a novel machine learning method for determining which mutations are drivers and which are passengers, called CHASM, an abbreviation for the paper title. In essence, CHASM is just a Random Forest binary classifier, which is capable of assigning a score and a P-value ti each prediction output. How CHASM is different from previous works in the field?

Simulated Passenger Data Set Unlike most existing methods, CHASM doesn't use high MAF (minor allele frequency) nsSNPs as negative examples for the algorithm, based on the hypothesis that, passenger mutations are different from nsSNPS, since they operate in a different context and while high MAF nsSNP must be functionally neutral, passenger mutations may have a functional impact on the protein, as long as it's neutral with respect to cancer cell fitness. So, instead of high MAF nsSNP CHASM obtains passenger mutations from *in silico* simulations, performed, using mutation profiles that reflected tumor type and mutation context. Such an approach raises a natural question: do passenger mutations, obtained this way, make biological sense? results, given in the paper convince us the answer is positive.

Novel features CHASM uses over 50 features, some of which have not been previously used for missense mutation function prediction. These features include SNP density (the number of SNPs in the exon where the mutation occurs, normalized by exon length) and average nucleotide-level conservation of the exon in which a mutation occurs in vertebrates.

Sensitivity and Specificity CHASM have shown superior performance over existing methods (SIFT, PolyPhen, CanPredict, KinaseSVM) in terms of both sensitivity and specificity, predicting known mutations in P53 and EGFR genes.

Critique

Overall, CHASM looks like a well-engineered, practical method, which is confirmed[1] by the number of biological citations in cancer research journals, over the last few years. However, there are two things worth pointing out:

- The reliability of P-values, assigned to CHASM scores, is questionable, since the null distribution of passenger scores is modeled from the generated data set, which only makes sense under the assumption that mutations generated this way are actually *true* passenger mutations.
- Also, significance and reliability of some of the features, used for discriminating between drivers and passengers seems dubious: Predicted contribution to protein stability, Predicted secondary structure and Predicted residue solvent accessibility (see Supplementary Methods, pp. 16-17 for details). These features are based on the output of the predictive algorithms and may not have a clear biological interpretation, leading to meaningless classification results.

[1] `http://scholar.google.ru/scholar?cites=18219215566530706461`