Nucleic Acids Research, 2007, Vol. 35, Web Server issue ***W595–W598***
doi:10.1093/nar/gkm405

CanPredict: a computational tool for predicting cancer-associated missense mutations

Joshua S. Kaminker, Yan Zhang, Colin Watanabe and Zemin Zhang*

Department of Bioinformatics, Genentech, Inc., South San Francisco, CA 94080, USA

Received January 29, 2007; Revised April 17, 2007; Accepted May 3, 2007

ABSTRACT

Various cancer genome projects are underway to identify novel mutations that drive tumorigenesis. While these screens will generate large data sets, the majority of identified missense changes are likely to be innocuous passenger mutations or polymorphisms. As a result, it has become increasingly important to develop computational methods for distinguishing functionally relevant mutations from other variations. We previously developed an algorithm, and now present the web application, CanPredict (http://www.canpredict.org/ or http://www.cgl.ucsf.edu/Research/genentech/canpredict/), to allow users to determine if particular changes are likely to be cancer-associated. The impact of each change is measured using two known methods: Sorting Intolerant From Tolerant (SIFT) and the Pfam-based LogR.E-value metric. A third method, the Gene Ontology Similarity Score (GOSS), provides an indication of how closely the gene in which the variant resides resembles other known cancer-causing genes. Scores from these three algorithms are analyzed by a random forest classifier which then predicts whether a change is likely to be cancer-associated. CanPredict fills an important need in cancer biology and will enable a large audience of biologists to determine which mutations are the most relevant for further study.

INTRODUCTION

The study of mutations that drive tumorigenesis is a central focus of cancer biology. These mutations disrupt genes that regulate normal cellular processes, thereby providing growth advantages and metastatic capabilities to tumor cells. Understanding how such changes lead to an oncogenic phenotype can provide a deeper understanding of the molecular nature of different cancers while also revealing novel therapeutic targets. There are a number of well-known somatic mutations (1) and germline mutations (2,3) that have been implicated in cancer progression. However, there are likely many more mutations that have not yet been found (4). The identification and study of these additional mutations presents an important opportunity for further understanding of the biological processes and pathways underlying cancer.

Many large-scale screens have been initiated to identify novel cancer-causing mutations (4–7) (http://cancergenome.nih.gov). These efforts have relied on sequence analysis of a few hundred to several thousand genes across multiple tumor and cell line samples. While these screens are extremely important for further understanding of tumorigenesis, the results are difficult to interpret because the majority of identified changes are not cancer-causing. In fact, a recent large-scale survey of mutations in breast and colon cancers indicates that causal mutations likely account for less than 1% of all observed non-synonymous changes (4).

The high level of background signal can be attributed in part to single nucleotide polymorphisms (SNPs) and passenger mutations. SNPs can be distinguished from true cancer mutation data by a variety of methods including identifying the same change in a matched normal tissue sample, or identifying the same, change in a database of known SNPs such as dbSNP. However, such approaches can be complicated by many factors including a lack of matched normal samples for re-sequencing putative cancer mutations. Additionally, known SNP databases are largely incomplete (8) and can contain unreliable records, making it difficult to positively identify a particular change as an SNP.

It is even more difficult to distinguish passenger mutations from true cancer mutations as this usually requires laboratory experimentation. Recently, a method was developed by Sjoblom and colleagues (4) to identify passenger mutations by uncovering those changes that occur at a higher than expected frequency in a set of tumor samples. But, since this method is highly dependant on large numbers of representative tumor samples, well-known oncogenes such as BRAF were not identified due to their low observed frequency in the Sjoblom data.

*To whom correspondence should be addressed. Tel: 650-225-4293; Fax: 650-225-5389; Email: zemin@gene.com

Thus, without methods specifically designed to analyze the mutations generated from these genome-scale screens, it is likely that a large number of true causal mutations will be overlooked.

Different algorithms have been developed to measure the effect a particular mutation might have on protein function. These approaches include Sorting Intolerant From Tolerant (SIFT) (9), the Pfam-based LogR.E-value metric (10), Polyphen (11), LS-SNP (12), statistical geometry methods (13), support vector machine methods (14), decision trees (15) and random forest classifiers (16). Additionally, methods based on the gene ontology such as the Gene Ontology Similarity Score (GOSS) (17) can also provide a measure as to how similar a gene of interest is to other known cancer-causing genes. While these algorithms may provide some indication about the nature of a particular mutation, it remains unclear whether by themselves such methods could be directly applicable in cancer mutation analysis.

Recently, using algorithms described earlier, we found that relevant somatic missense mutations behave differently from SNPs, and based on this distinction we developed a computational method to predict whether a variant is likely to be cancer-causing or not (17). Our algorithm uses a random forest classifier to combine data from the SIFT, LogR.E-value and GOSS metrics to generate a prediction to distinguish relevant mutations from other missense changes. We demonstrated that this approach could be potentially useful in distinguishing causal from passenger mutations (17). While this method was described in detail, its implementation requires a thorough understanding of random forest algorithms and the R programming language, likely impeding a large number of experimental biologists from attempting to classify their mutations. Here, we present a web application, CanPredict, that provides a clean and straightforward interface to our algorithm. Changes identified on a RefSeq protein sequence can be submitted and a prediction is generated as to whether the changes are cancer-associated or not. This application provides the first public interface to an important algorithm that can provide insight into the large amount of mutation data being generated from cancer re-sequencing projects.

METHODS AND IMPLEMENTATION

The algorithm supporting the CanPredict application uses a random forest (RF) classifier to predict whether an amino acid change is likely to be cancer-causing or not. RF classifiers divide a large pool of data into smaller subsets based on characteristics of each datum (18). For the CanPredict application, the three characteristics used to describe each mutation are scores from SIFT, the Pfam-based LogR.E-value and the GOSS metrics. The SIFT algorithm uses similarity between closely related proteins to identify potentially deleterious changes (9). SIFT scores <0.05 are predicted to be deleterious (9) and only SIFT scores with a median information content score <3.25 are included for predictions since higher values likely indicate unreliable SIFT scores (9). Also, because the computation time to generate alignments used by the SIFT algorithm is lengthy, the alignments for all RefSeq protein sequences have been pre-computed and are stored on the server. The Pfam-based logR.E-value score predicts whether a change will alter protein function by determining the difference in fit of a wild-type version of the protein to a particular Pfam model (10). These scores were derived from values provided by the HMMER 2.3.2 software and the *ls* mode was used to search against the Pfam protein family database. The LogR.E-value score was calculated as: $\log_{10}(\text{E-value}_{\text{variant}}/\text{E-value}_{\text{canonical}})$. Lastly, the GOSS metric uses the gene ontology to measure the similarity of the submitted RefSeq gene to other known cancer-causing genes (17).

The training data set used to construct the classifier is composed of 200 randomly selected known somatic cancer mutations and 800 non-cancer, non-synonymous variants. The cancer mutations were downloaded from data stored in the COSMIC database (1) and the non-cancer variants were selected randomly from SNPs stored in dbSNP with a minor allele frequency >20%. For each mutation in the training data, a score from the SIFT, LogR.E-value, and GOSS algorithms was determined. These values were used to build the classifier using the package randomForest 4.5-16 (http://stat-www.berkeley.edu/users/breiman/RandomForests) for the R statistical environment (http://www.r-project.org). The out-of-bag error, an internal measure of the rate of misclassification of the classifier, was determined to be 3.19% suggesting that the classifier is very effective. The training data are freely available from http://share.gene.com/mutation_classification.

As shown previously (17), data from three different experiments suggest that the predictor can function very well to highlight putative cancer mutations. First, in a cross-validation experiment, the classifier consistently revealed a very low false-positive rate of 1.7% for distinguishing relevant mutations from common SNPs (17). Second, an experiment was performed to distinguish recurrently identified mutations from mutations occurring only one time; causal mutations are more likely than passenger changes to be seen in multiple different tumor samples because they are under positive selection in tumor samples. In this analysis, 58% of variants observed more than 10 times were predicted to be cancer-associated while only 43% of variants occurring only one time were predicted cancer-associated (*P*-value 0.018, two-tailed Fisher Exact test) (17). Third, the classifier was used to analyze recent data from a large-scale screen for cancer mutations performed by Sjoblom and colleagues (4). In the paper by Sjoblom, mutations were grouped into those genes likely to cause cancer and those genes unlikely to cause cancer, CAN genes and non-CAN genes, respectively. The CanPredict classifier revealed that mutations in CAN genes were more likely to be predicted as cancer-associated than mutations in non-CAN genes (26.3% to 13.3%, respectively; *P*-value 8.8e-6; two-tailed Fisher Exact test) (17).

The CanPredict user interface was designed using dynamic AJAX technology. The user-supplied mutations and protein sequence data are validated via a server

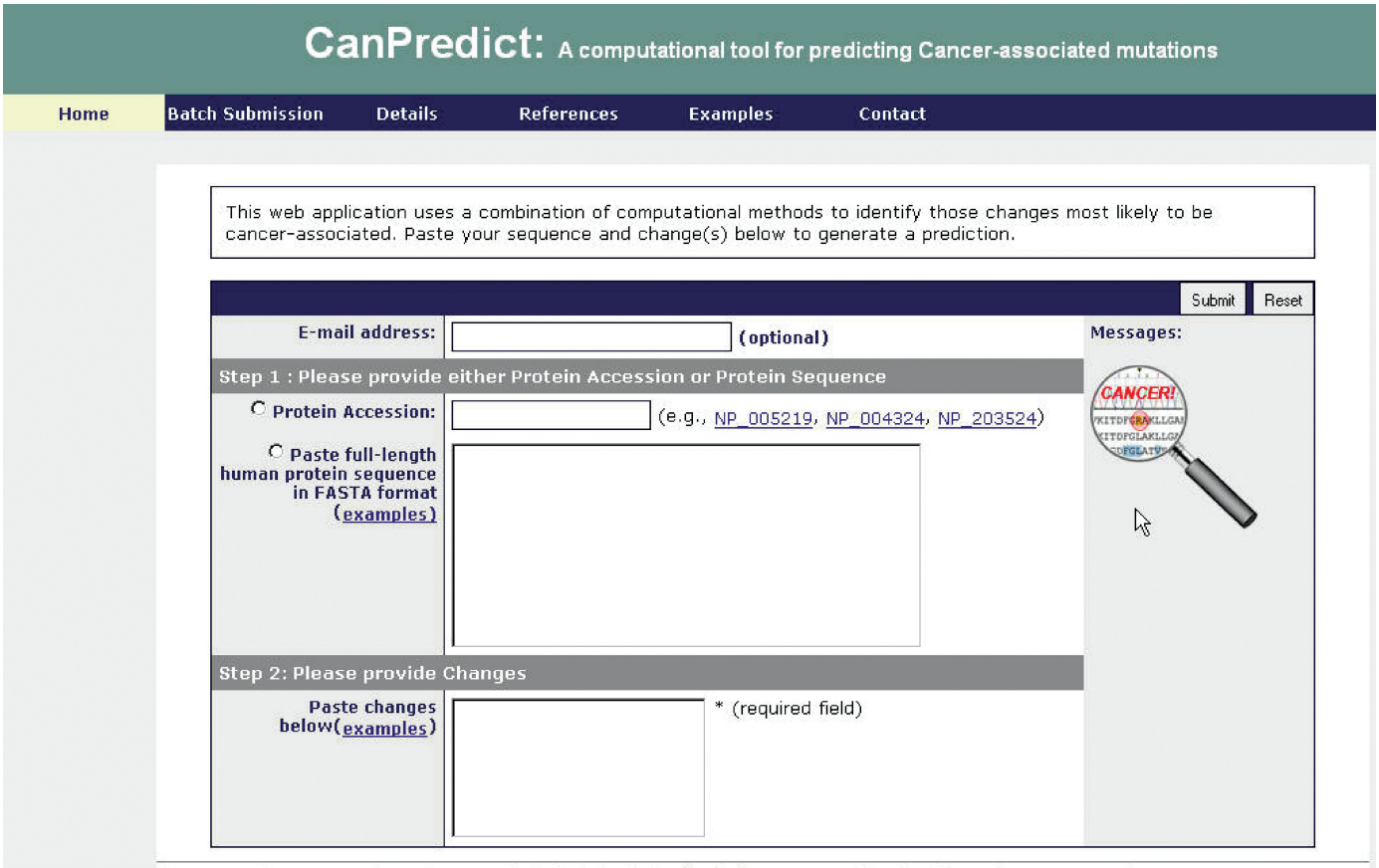

Figure 1. The home page of the CanPredict application.

process, and the analysis status is instantly updated without the user leaving the input page. The results summary page is automatically loaded when the AJAX call detects that the analysis is complete. The Dojo library (www.dojotoolkit.org) implements AJAX calls by providing support for the back and forward buttons, changing the URL in the address bar to allow for bookmarking, and gracefully degrading when AJAX or JavaScript are not fully supported on the client.

RESULTS AND DISCUSSION

The CanPredict application can be used to submit a single full-length RefSeq protein sequence or accession and multiple associated changes (Figure 1). Additionally, from the *Batch Submission* page, the application will accept multiple RefSeq protein accessions and associated changes. There is no limit to the number of changes that can be analyzed from the *Batch Submission* page. Changes are validated by the server to ensure that the amino acid specified in the change string occurs in the indicated sequence. For testing the application, users can either enter their own mutations or use the *test-it* link to submit example mutations. Included in these examples are known cancer-causing mutations in BRAF, KRAS and EGFR.

Results of the analysis are returned to the user in a summary page where they can also access all other submitted changes using links at the top of the summary (Figure 2). There is also a link directing users to a detailed description of the scores produced from each metric. Within the submission summary is a prediction from the classifier indicating *likely cancer*, *likely non-cancer* or *not determined*. The sequence flanking the change is included to allow the user to confirm the precise sequence used in the analysis. Below the submission summary are data from the SIFT, logR.E-value and GOSS analyses. As alignment files used by the SIFT algorithm are time-consuming to produce, they are available for download using the provided link. SIFT scores and median information content are also presented and only scores with a median information content of <3.25 are considered reliable (9) and will be used to generate a prediction from the classifier. The logR.E-value analysis indicates the domain altered by the submitted mutation. If there are multiple domains covering the same mutation, the domain with the most deleterious (largest) logR.E-value score will be selected for display and will be used by the classifier. The GOSS score is indicated last, and will be present only if the submitted change resides in a gene with a gene ontology description. The result pages can be bookmarked, and the associated data are saved in the server for a week. Finally, a link presented on the results summary page allows users to download their results in a tab-delimited format. Results from the batch submission page will be returned in a similar tab-delimited format.

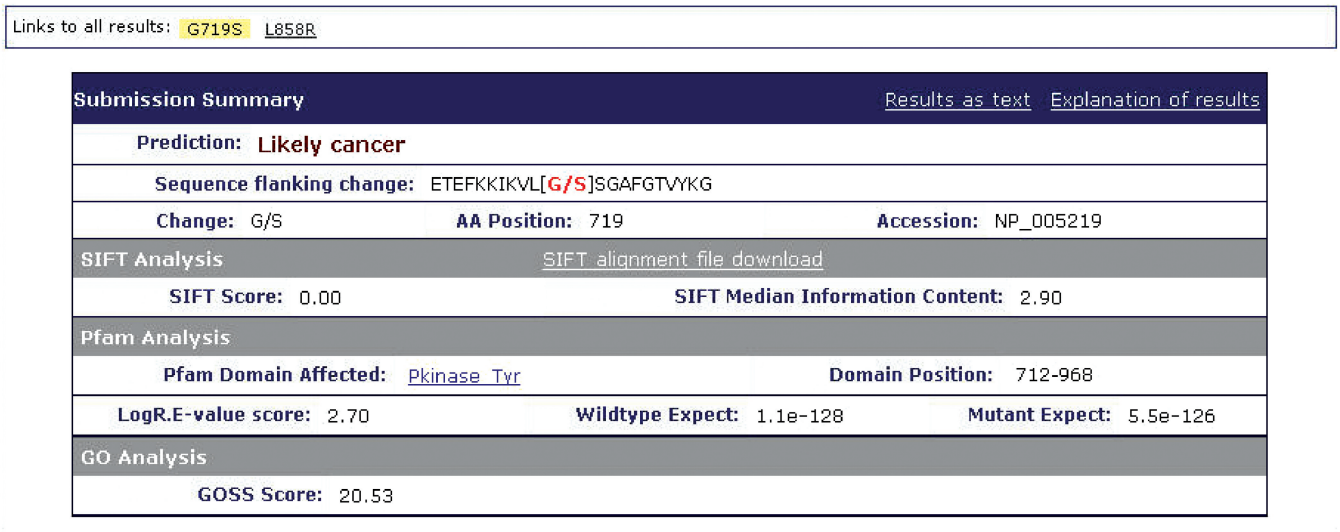

Figure 2. The results summary page of the CanPredict application.

The CanPredict application provides an easily accessible interface for users to determine if an amino acid change is likely to be cancer-causing. This application will likely be very useful for large-scale cancer genome projects.

ACKNOWLEDGEMENTS

We would like to thank Pete Haverty and Bill Forrest for discussions about the CanPredict algorithm, Shiuh-Ming Luoh, Lawrence Hon, Jerry Tang, Kiran Mukhyala and Reece Hart for helpful discussions, Sarah Kaminker for careful reading and editing of the manuscript and William Wood for guidance and support throughout the project. We would also like to thank the UCSF Computer Graphics Laboratory and Dr. Thomas Ferrin for hosting the CanPredict web application. Funding to pay the Open Access publication charges for this article was provided by Genentech, Inc.

Conflict of interest statement. None decalred.

REFERENCES

1. Forbes,S., Clements,J., Dawson,E., Bamford,S., Webb,T., Dogan,A., Flanagan,A., Teague,J., Wooster,R. *et al.* (2006) Cosmic 2005. *Br. J. Cancer*, **94**, 318–322.
2. Vierimaa,O., Georgitsi,M., Lehtonen,R., Vahteristo,P., Kokko,A., Raitila,A., Tuppurainen,K., Ebeling,T.M., Salmela,P.I. *et al.* (2006) Pituitary adenoma predisposition caused by germline mutations in the AIP gene. *Science*, **312**, 1228–1230.
3. Landi,M.T., Bauer,J., Pfeiffer,R.M., Elder,D.E., Hulley,B., Minghetti,P., Calista,D., Kanetsky,P.A., Pinkel,D. *et al.* (2006) MC1R germline variants confer risk for BRAF-mutant melanoma. *Science*, **313**, 521–522.
4. Sjoblom,T., Jones,S., Wood,L.D., Parsons,D.W., Lin,J., Barber,T.D., Mandelker,D., Leary,R.J., Ptak,J. *et al.* (2006) The consensus coding sequences of human breast and colorectal cancers. *Science*, **314**, 268–274.
5. Stephens,P., Edkins,S., Davies,H., Greenman,C., Cox,C., Hunter,C., Bignell,G., Teague,J., Smith,R. *et al.* (2005) A screen of the complete protein kinase gene family identifies diverse patterns of somatic mutations in human breast cancer. *Nat. Genet.*, **37**, 590–592.
6. Parsons,D.W., Wang,T.L., Samuels,Y., Bardelli,A., Cummins,J.M., DeLong,L., Silliman,N., Ptak,J., Szabo,S. *et al.* (2005) Colorectal cancer: mutations in a signalling pathway. *Nature*, **436**, 792.
7. Davies,H., Hunter,C., Smith,R., Stephens,P., Greenman,C., Bignell,G., Teague,J., Butler,A., Edkins,S. *et al.* (2005) Somatic mutations of the protein kinase gene family in human lung cancer. *Cancer Res.*, **65**, 7591–7595.
8. Kruglyak,L. and Nickerson,D.A. (2001) Variation is the spice of life. *Nat. Genet.*, **27**, 234–236.
9. Ng,P.C. and Henikoff,S. (2002) Accounting for human polymorphisms predicted to affect protein function. *Genome Res.*, **12**, 436–446.
10. Clifford,R.J., Edmonson,M.N., Nguyen,C. and Buetow,K.H. (2004) Large-scale analysis of non-synonymous coding region single nucleotide polymorphisms. *Bioinformatics*, **20**, 1006–1014.
11. Ramensky,V., Bork,P. and Sunyaev,S. (2002) Human non-synonymous SNPs: server and survey. *Nucleic Acids Res.*, **30**, 3894–3900.
12. Karchin,R., Diekhans,M., Kelly,L., Thomas,D.J., Pieper,U., Eswar,N., Haussler,D. and Sali,A. (2005) LS-SNP: large-scale annotation of coding non-synonymous SNPs based on multiple information sources. *Bioinformatics*, **21**, 2814–2820.
13. Barenboim,M., Jamison,D.C. and Vaisman,II. (2005) Statistical geometry approach to the study of functional effects of human nonsynonymous SNPs. *Hum. Mutat.*, **26**, 471–476.
14. Yue,P. and Moult,J. (2006) Identification and analysis of deleterious human SNPs. *J. Mol. Biol.*, **356**, 1263–1274.
15. Krishnan,V.G. and Westhead,D.R. (2003) A comparative study of machine-learning methods to predict the effects of single nucleotide polymorphisms on protein function. *Bioinformatics*, **19**, 2199–2209.
16. Bao,L. and Cui,Y. (2005) Prediction of the phenotypic effects of non-synonymous single nucleotide polymorphisms using structural and evolutionary information. *Bioinformatics*, **21**, 2185–2190.
17. Kaminker,J.S., Zhang,Y., Waugh,A., Haverty,P., Peters,B., Sebisanovic,D., Stinson,J., Forrest,W.F., Bazan,J.F. *et al.* (2007) Distinguishing cancer-associated missense mutations from common polymorphisms. *Cancer Res.*, **67**, 465–473.
18. Breiman,L. (2001) Random Forests. *Machine Learning*, **45**, 5–32.