SOFTWARE TOOL ARTICLE

iMutSig: a web application to identify the most similar mutational signature using shiny [version 1; peer review: awaiting peer review]

Zhi Yang, Priyatama Pandey, Paul Marjoram, Kimberly D. Siegmund

Department of Preventive Medicine, Keck School of Medicine of the University of Southern California, 2001 N.Soto Street, Los Angeles, CA, 91003, USA

Abstract

There are two frameworks for characterizing mutational signatures which are commonly used to describe the nucleotide patterns that arise from mutational processes. Estimated mutational signatures from fitting these two methods in human cancer can be found online, in the Catalogue Of Somatic Mutations In Cancer (COSMIC) website or a GitHub repository. The two frameworks make differing assumptions regarding independence of base pairs and for that reason may produce different results. Consequently, there is a need to compare and contrast the results of the two methods, but no such tool currently exists. In this paper, we provide a simple and intuitive interface that allows such comparisons to be easily performed. When using our software, the user may download published mutational signatures of either type. Mutational signatures from the pmsignature data source are expanded to probabilistic vectors of 96-possible mutation types, the same model specification used by COSMIC, and then compared to COSMIC signatures. Cosine similarity measures the extent of signature similarity. iMutSig provides a simple and user-friendly web application allowing researchers to compare signatures from COSMIC to those from pmsignature, and vice versa. Furthermore, iMutSig allows users to input a self-defined mutational signature and examine its similarity to published signatures from both data sources. iMutSig is accessible online and source code is available for download on GitHub.

Keywords

Mutational Signatures, pmsignature, COSMIC, Web interface, Shiny, R

This article is included in the RPackage gateway.
Corresponding author: Zhi Yang (zhiyang@usc.edu)

Author roles: Yang Z: Conceptualization, Formal Analysis, Funding Acquisition, Investigation, Methodology, Software, Visualization, Writing – Original Draft Preparation, Writing – Review & Editing; Pandey P: Validation, Writing – Review & Editing; Marjoram P: Funding Acquisition, Methodology, Project Administration, Resources, Supervision, Validation, Writing – Review & Editing; Siegmund KD: Funding Acquisition, Project Administration, Resources, Supervision, Validation, Writing – Review & Editing

Competing interests: No competing interests were disclosed.

Grant information: This work was supported by the National Cancer Institute [P01CA196569, P30CA014089 and R21 CA226106], and the National Institute of Environmental Health Sciences [P30ES07048]. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health (NIH).

The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Copyright: © 2020 Yang Z et al. This is an open access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

How to cite this article: Yang Z, Pandey P, Marjoram P and Siegmund KD. iMutSig: a web application to identify the most similar mutational signature using shiny [version 1; peer review: awaiting peer review] F1000Research 2020, 9:586 https://doi.org/10.12688/f1000research.24435.1

Introduction

Each human is subject to a variety of mutational processes throughout their lifetime, which result in the formation of a catalog of somatic mutations as his/her unique mutational profile1. A mutational signature captures the pattern of the mutations and contexts in which those mutations occur (i.e., the neighboring bases). Examples of important mutational processes with distinct mutational signatures include aging and ultraviolet (UV) radiation. Additionally, many research groups are performing analysis to discover de novo mutational signatures in cancer.7-11.

Currently, there are two frameworks used to characterize and visualize mutational signatures.7-9. The first, proposed by Alexandrov et al., used a vector of 96 probabilities to capture the possible composition of the six nucleotide substitutions (C>A, C>T, C>G, T>A, T>C, T>G) and the neighboring base immediately on each of the 5’ and 3’ side of the mutated base.4 A list of published mutational signatures can be downloaded from the Catalogue Of Somatic Mutations In Cancer (COSMIC) website3 (version 2, v2). Later, Alexandrov et al. published an expanded set of mutational signatures, which was referred to as version 3 (v3)4. The 67 COSMIC v3 Single Base Substitution (SBS) signatures include 30 v2 signatures. Based on the signature concept, but using different model assumptions, Shiraishi et al. proposed a mixed-membership model, pmsignature, which substantially reduced the number of parameters needed to characterize a signature. They achieved this by assuming independence across bases, thereby reducing the number of parameters from 6³ x 4² x 1 = 95 to (6 - 1) x (4 - 1) + (4 - 1) = 11². The reduction in the number of parameters is greater if more flanking bases are included. All Shiraishi’s signatures can be downloaded from their GitHub repository. In this paper, we will refer to signatures resulting from these two methods as “COSMIC signatures” with version numbers (for those resulting from Alexandrov et al.’s method) and “PM signatures” (for those resulting from Shiraishi et al.’s method).

A large number of researchers have published scientific findings resulting from the COSMIC signature-based method10-13, which was defined as the “gold standard” in the field by Baez-Ortega et al.8. Meanwhile, an increasing number of researchers are using the pmsignature-based method for samples with lower numbers of somatic variants due to it requiring fewer parameters.9,11,14. Given that both methods are widely used, investigators need the ability to compare results from their analysis with those reported in earlier databases, which may have been produced using the alternate method. For example, researchers have adopted both tools for gastric cancer and tried to compare and integrate the information from two data sources in a somewhat ad hoc manner15. No rigorous tool exists for this task. In this paper we present iMutSig, an easy-to-use tool that allows users to identify the most similar mutational signatures across methods and to compare the information characterizing those signatures using simple point-and-click navigation.

Methods

Implementation

In order to measure the similarity between mutational signatures across two databases, we need to represent PM signatures in a way that is comparable with those from COSMIC. To do this we convert the PM signature into a probabilistic vector with the same length as the COSMIC signature, i.e., 96. To calculate each of 96 resulting probabilities in the vector, we take the constituent components that make up the COSMIC signature - which refer to the nucleotide substitution and two flanking bases at the -1 and +1 position - calculate the probability of each component for the given PM signature, and then multiply those probabilities using pm-signature’s assumption of independence. For example, to calculate the probability of the COSMIC signature C[C >A]T we multiply three PM signature’s probabilities: P(C at pos -1), P(C >A), and P(T at pos +1). This example is shown in Table 1 and Equation 1.

\[
P(C[C >A]T) = P(C \text{ at pos } -1)P([C >A])P(T \text{ at pos } +1) \\
= 0.052 \times 0.012 \times 0.116 \\
= 7.24 \times 10^{-5}
\]

Table 1. An example of PM signatures.

<table>
<thead>
<tr>
<th>Nucleotide substitution</th>
<th>C&gt;A</th>
<th>C&gt;G</th>
<th>C&gt;T</th>
<th>T&gt;A</th>
<th>T&gt;C</th>
<th>T&gt;G</th>
</tr>
</thead>
<tbody>
<tr>
<td>C&gt; A</td>
<td>0.003</td>
<td>0.879</td>
<td>0.003</td>
<td>0.090</td>
<td>0.014</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Flanking bases</th>
<th>A</th>
<th>C</th>
<th>G</th>
<th>T</th>
</tr>
</thead>
<tbody>
<tr>
<td>-2</td>
<td>0.159</td>
<td>0.042</td>
<td>0.486</td>
<td>0.314</td>
</tr>
<tr>
<td>-1</td>
<td>0.044</td>
<td>0.052</td>
<td>0.870</td>
<td>0.034</td>
</tr>
<tr>
<td>+1</td>
<td>0.076</td>
<td>0.237</td>
<td>0.571</td>
<td>0.116</td>
</tr>
<tr>
<td>+2</td>
<td>0.245</td>
<td>0.247</td>
<td>0.256</td>
<td>0.252</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Transcription strand</th>
<th>Plus</th>
<th>Minus</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0.511</td>
<td>0.489</td>
</tr>
</tbody>
</table>

Now that we have represented both forms of signature using probabilistic vectors of length \(n = 96\), \(P\) and \(C\) say, we can directly compare the two signature types. In order to measure the similarity between them we use cosine similarity, \(CS\), defined as shown in Equation 2:

\[
CS(P, C) = \frac{P \cdot C}{||P|| \times ||C||} = \frac{\sum_{i=1}^{n} P Ci}{\sum_{i=1}^{n} Pi \times \sum_{i=1}^{n} Ci}
\]

Intuitively speaking, cosine similarity is the cosine of the angle between the two vectors. As such, cosine similarity
ranges from 0 to 1 (inclusive). In our context, if two mutational signatures have a cosine similarity of 1, they must be identical, i.e., the angle between them is 0°; in contrast, if two mutational signatures have a cosine similarity of 0, they are maximally dissimilar (i.e., orthogonal). Computing the cosine similarity between the input signature and each of the candidate signatures, and then sorting the similarities from highest to lowest value, we identify the candidate signature with the highest cosine similarity as the most similar mutational signature.

**Operation**

*iMutSig* is built in R with its key features depending on the R package, *pmsignature*. As shown in Figure 1, the Shiny app currently supports three possible workflows for users to choose from, depending on the type of signatures they have already obtained: 1) starting with a COSMIC signature; 2) starting with a PM signature; 3) starting with a self-defined signature that could follow either the COSMIC or PM format.

The first tab in the Shiny app window, “COSMIC to pmsignature”, allows users to select an input COSMIC signature via a drop-down list and returns the best-matched PM signature. The returned results are divided and organized separately in the top and the bottom portion of the page. The top half tab summarizes background information regarding the input signature by presenting: 1) visualized plots of the input signature and its membership among all cancer types, i.e., in which kind of cancers the mutational signatures has been found; 2) a table showing the cosine similarity between this signature and all PM signatures, sorted in decreasing order, along with a visualization of a similarity heatmap with color and intensity proportional to assessed similarity. The bottom half tab presents plots and descriptions of the input COSMIC signature, the most similar PM signature, and a second PM signature that the user can select. Thus, users can easily access all the vital information and results regarding these signatures rather than having to manually gather and organize information from publications. The top half of the tab will be automatically updated via a control panel in the middle section of the tab, which enables users to select a signature to start with and also highlights information about the currently selected signature, the most-similar signature from the alternate model framework, and the cosine similarity.

The second tab was designed in a similar manner to the first tab, but for the case in which we are starting with a PM signature and looking for the most similar COSMIC signature. For the first two tabs, users can choose which version of COSMIC signatures to input from the sub-menus, i.e., v2 or v3.

Unlike the first two tabs, the third tab enables users to enter a user-supplied signature, which can be in either PM or COSMIC format, and then identify the most similar signature from each online database. The user will be requested to enter a sub-menu based on the type of the input signature and to upload a comma-separated values (CSV) file containing a single signature. A sample CSV file is provided for download to give the user a better sense of the format of the input file. Then, the tab will be updated to display three tables, one from each data source (COSMIC v2, v3 and PM), listing the signatures from that data source and the cosine similarity of each signature with the user-uploaded signature. The tables are ordered from most similar to least similar signature. In addition, the user is able to view figures of the best-matched signatures (i.e., those with highest cosine similarity) from each data source, allowing users to observe any similarities and dissimilarities. Below, users will see a list of cancer types that contain the best-matched signature.

**Use cases**

We use *iMutSig* to identify the most similar signature for a given PM/COSMIC signature or a user-supplied signature. Figure 2 shows the input panel after inputting COSMIC v3 signature SBS1 and Figure 3 shows the input panel after inputting PM signature P1. If users provide a user-supplied signature of either COSMIC-kind or PM-kind, the results can be seen in Figure 4 and Figure 5. Consider the example shown in Figure 4,
Figure 2. Input a COSMIC v3 signature, SBS1.
Figure 3. Input a PM signature, P1.
Figure 4. Input a user-supplied COSMIC signature.
Figure 5. Input a user-supplied PM signature.
where we input COSMIC v2 signature C1. iMutSig returned the most similar signatures COSMIC v2 signature C1, COSMIC v3 signature SBS1, and PM signature P7 (similarity = 1.000, 0.947, and 0.948, respectively) along with the names of its associated cancer types. When providing PM signature P1, iMutSig returned COSMIC v2 signature C10, v3 signature C10a and PM signature P1 (similarity = 0.816, 0.957, 1.0, respectively).

Conclusions

iMutSig is a user-friendly interactive browser-based application that allows users who have a signature that they have discovered in an analysis of their own data to identify the best-matched existing mutational signature from the COSMIC and PM databases. It also allows users to directly compare signatures between the two databases. It does this in an interactive way, and also allows straightforward visualization of results. iMutSig enables researchers to easily identify the most similar mutational signature and to easily access characteristic information from both data sources without additional software installation and programming of their own.

Data availability

All data underlying the results are available as part of the article and no additional source data are required.

Software availability

Software available from: https://zhiyang.shinyapps.io/iMutSig/
Source code available from: http://www.github.com/USCbiostats/iMutSig
Archived source code at time of publication: https://doi.org/10.5281/zenodo.387388
License: MIT

References


Page 9 of 10
The benefits of publishing with F1000Research:

- Your article is published within days, with no editorial bias
- You can publish traditional articles, null/negative results, case reports, data notes and more
- The peer review process is transparent and collaborative
- Your article is indexed in PubMed after passing peer review
- Dedicated customer support at every stage

For pre-submission enquiries, contact research@f1000.com