

Claims

1. A method of predicting a mutation type of each of the plurality of variants in a single sample file comprising:

5 segregating 78 single-base substitutions (SBS) mutational signatures obtained from a Catalogue Of Somatic Mutations In Cancer (COSMIC) database containing valid and artifactual signatures into 96 different contexts to form a first data structure;

reconstructing a second data structure containing site features from the single sample file;

10 building a third data structure containing reference context features for each of the plurality of variants in the single sample file;

feeding the first, second and third data structures to a neural network architecture;

receiving from the neural network architecture a probability value for each of the plurality of variants being real or artifactual mutation; and

15 generating a report comprising a list of variants in the single sample file with corresponding one of predicted real and artifactual mutations.

2 The method according to claim 1 wherein the 78 SBS mutational signatures comprises 18 SBS mutational signatures from SBS27, SBS43, and SBS45-SBS6, and 60
20 SBS mutational signatures from valid mutational signatures that are found to be linked to mutational processes causing cancer.

3. The method according to claim 2 wherein the step of reconstructing the second data structure containing site features from the single sample file comprises:

25 extracting from an "INFO" column and a "SAMPLE" column from the single sample file to form the second data structure.

4. The method according to claim 3 wherein the step of reconstructing the second data structure containing site features from the single sample file further comprises:

applying one-hot encoding for discrete feature extracted from the single sample file,
and

5 applying standard scaling for continuous feature extracted from the single sample file.

5. The method according to claim 4 wherein the second data structure comprises columns determined by a number of categories in the single sample file and rows

10 determined by a number of rows in each category in the single sample file.

6. The method according to claim 5 wherein the step of building the third data structure containing reference context features for each of the plurality of variants in the single sample file comprises:

15 locating a plurality of mutation sites in a reference sequence file that correspond to each of the plurality of variants, the reference sequence file containing a full human genome sequence;

retrieving 10 bases at the left and right of each located mutation site in the reference sequence file to form a total of 21 bases for each of the plurality of variants; and

20 translating the 21 bases to form a 21 dimensional vector for each of the plurality of variants to form the third data structure.

7. The method according to claim 6 wherein the step of feeding the first, second and third data structures to the neural network architecture comprises:

25 passing the first data structure into a first embedding layer and converting into a point in a high-dimensional space, coordinates of which constitute an embedding vector for the type;

parsing the second data structure into a linear layer;

passing the third data structure into a second embedding layer and converting the 21-dimensional vector for each of the plurality of variants into a two-dimensional context matrix;

concatenating the outputs from the first embedding layer, linear layer and second embedding layer and feeding into a five-layer fully connected network, each layer in the five-layer fully connected network consists of dropout linear transformation, ELU (Exponential Linear Unit) activation function, and highway module with the following expressions,

$$\begin{aligned} \text{Linear}_i(X) &= W_i X + b_i \\ \text{ELU}(x) &= \begin{cases} x, & \text{if } x > 0 \\ \alpha * (\exp(x) - 1), & \text{if } x \leq 0 \end{cases} \\ H_{i+1} &= H_i + \text{ELU}(\text{Linear}(H_i)) \end{aligned}$$

where W and b are learnable parameters, the ELU hyperparameter α controls a value at which the activation function saturates for negative inputs, the highway module connects the input of each layer to the end, solving the gradient vanishing problem caused by overly deep networks, wherein a last layer of the five-layer fully connected network contains a sigmoid function that maps the scalar output from the previous layer to a value between 0 and 1 for each of the plurality of variants.

8. The method according to claim 7 further comprising:

determining a real mutation in response to the value being more than a predefined threshold; and

determine an artifactual mutation in response to the value being equal to or less than the predefined threshold.

9. The method according to claim 8 wherein the predefined threshold is 0.5.