Improvements in LymphGen 2.0

This version of the predictor loosens the requirements for feature inclusion when it is applied to incomplete data (for example, when copy-number information is not available). When the input data are complete, the results should be identical to those obtained in Version 1.0. When only incomplete data are available, this revised version should result in a prediction that more closely matches what one would expect from the complete model. The differences are described in detail below. To fully understand these differences, it is recommended that users refer to the STAR methods of the original paper [1] in which the algorithm was first presented.

In the original LymphGen algorithm, a gene would be identified for inclusion in a model of a given subtype only if there was a feature for that gene that was significantly enriched for that subtype in the training set with $p<0.001$. This stringent p-value was necessary to avoid including genes that were enriched purely by chance. Once a gene was identified, additional subfeatures would be considered for separate inclusion if they were found to be significantly enriched at a less stringent $p<0.05$ level. So, for example, the feature that combines truncating IRF4 mutations and IRF4 deletions is significantly overrepresented in MCD with a p-value of $7.3x10^{-4}$, so IRF4 was included as a gene in the model. Meanwhile, IRF4 truncations were found to be individually significant at $p=0.049$, and so they were included as a subfeature in the model.

In the original version of the algorithm, incomplete data were handled by refitting the model on the training data but excluding all features that were not available on the sample being predicted. So, if a sample lacked copy-number data, all features that required information on copy number would be excluded. In the IRF4 example given above, the combination truncation/deletion feature would be excluded since it involved a change in copy number. This would result in all features involving IRF4 being completely removed from the model (since there would no longer be an IRF4 feature significant at $p<0.001$). Therefore, even though they were available in the data, truncations of IRF4 were removed from the predictor.

In Version 2.0, we correct this by using the complete list of features on the training set when identifying a gene for possible inclusion. A gene is thus potentially included if any of its features were significant ($p<0.001$) on the training set, regardless of whether data for that feature was available for the sample being predicted. Once a gene is identified for possible inclusion, prediction is based on those features for that gene which are enriched in the training set with $p<0.05$ and which have data available on the predicted sample. So, in the example above, the presence or absence of truncations of IRF4 remain part of the predictor even when information about IRF4 deletions was not available.

1. Wright, G.W., et al., *A Probabilistic Classification Tool for Genetic Subtypes of Diffuse Large B Cell Lymphoma with Therapeutic Implications.* Cancer Cell, 2020. **37**(4): p. 551-+.