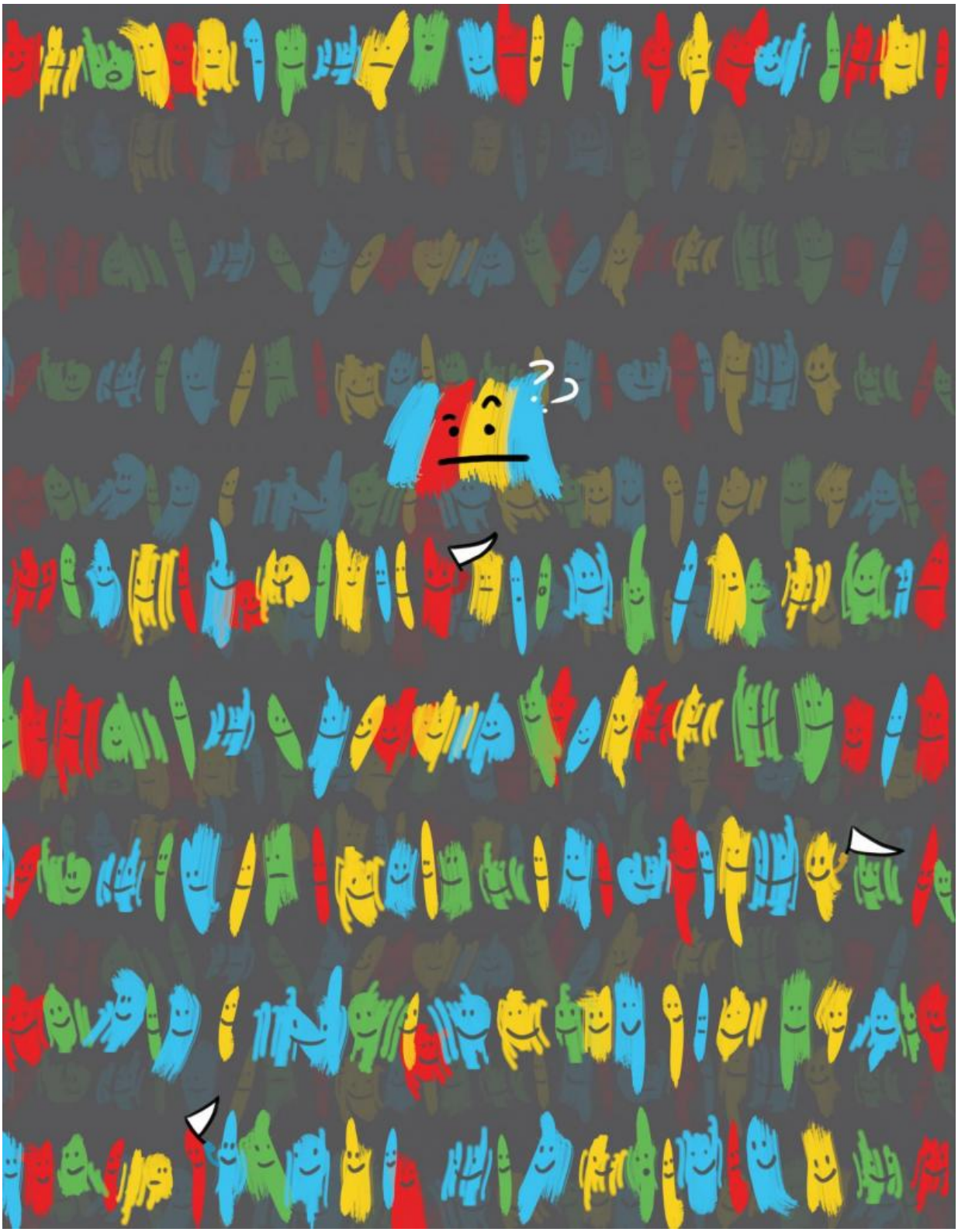
New model helps zero in on harmful genetic mutations

October 22 2015



A new University of Washington model and online tool help narrow down which

genetic mutations affect how genes splice and may contribute to disease. In this illustration, the cell's splicing machinery is trying to decide which of the splice sites (pictured as white flags) it should use. Credit: Jennifer Sunami

Between any two people, there are likely to be at least 10 million differences in the genetic sequence that makes up their DNA.

Most of these differences don't alter the way cells behave or cause health problems. But some genetic variations greatly increase the likelihood that a person will develop cancer, diabetes, colorblindness or a host of other diseases.

Despite rapid advances in our ability to map an individual's genome—the precise coding that makes up his or her genes—we know much less about which mutations or anomalies actually cause disease.

Now, a new model and publicly available Web tool developed by University of Washington researchers can more accurately and quantitatively predict which genetic mutations significantly change how genes splice and may warrant increased attention from disease researchers and drug developers.

The model—the first to train a machine learning algorithm on vast amounts of genetic data created with synthetic biology techniques—is outlined in a paper to be published in the Oct. 22 issue of *Cell*.

"Some people have variations in a particular gene, but what you really want to know is whether those matter or not," said lead author Alexander Rosenberg, a UW electrical engineering doctoral student. "This model can help you narrow down the universe—hugely—of the mutations that might be most likely to cause disease."

In particular, the model predicts how these genetic sequence variations affect alternative splicing—a critical process that enables a single gene to create many different forms of proteins by including or excluding snippets of RNA.

"This is an avenue that's unexplored to a large extent," said Rosenberg. "It's fairly easy to look at how mutations affect proteins directly, but people have not been able to look at how mutations affect proteins through splicing."

For example, a scientist studying the genetic underpinnings of lung cancer or depression or a particular birth defect could type the most commonly shared DNA sequence in a particular gene into the Web tool, as well as multiple variations. The model will tell the scientist which mutations cause outsized differences in how the gene splices—which could be a sign of trouble—and which have little or no effect.

The researcher would still need to investigate whether a particular genetic sequence causes harmful changes, but the online tool can help rule out the many variations that aren't likely to be of interest to health researchers. To validate the model's predictive powers, the UW team tested it on a handful of well-understood mutations such as those in the BRCA2 gene that have been linked to breast and ovarian cancer.

Compared to previously published models, the UW approach is roughly three times more accurate at predicting the extent to which a mutation will cause genetic material to be included or excluded in the protein-making process—which can change how those proteins function and cause biological processes to go awry.

That's because the UW team used a new approach that combines synthetic biology and machine learning techniques to create the model.

Machine learning algorithms—which enable computers to infer rules and "learn" from vast amounts of data—become more accurate the more data they're exposed to. But the human genome only has roughly 25,000 genes that create proteins.

Using common molecular biology techniques, the UW team created a library of over 2 million synthetic "mini-genes" by including random DNA sequences. Then they determined how each random sequence element affected where genes spliced and what types of RNA were produced—which ultimately determines which proteins get made.

That larger library of synthetic data essentially teaches the model to become smarter, said lead author Georg Seelig, a UW assistant professor of electrical engineering and of computer science & engineering.

"Our algorithm works super well because it was trained on these synthetic datasets. And the reason it works so well is because that synthetic dataset is orders of magnitude larger than the training set you get from the actual human genome," said Seelig.

"It is remarkable that a model trained entirely on synthetic data can outperform models trained directly on the human genome on the task of predicting the impact of mutations in people," he said.

Next research steps include expanding the approach beyond alternative splicing to other processes that determine how genes are expressed.

In the meantime, by making the Web tool free and publicly available, the team hopes other scientists will use their alternative splicing model—and ultimately make progress in narrowing down which natural genetic variations are most meaningful when it comes to health and disease.

"Other research groups and companies can use our model to rank the

areas of interest to them," Seelig said. "We hope other people will take this further to more clinical applications."

Provided by University of Washington