Machine-learning model provides detailed insight on proteins

12 March 2019

A novel machine-learning 'toolbox' that can read and analyse the sequences of proteins has been described today in the open-access journal *eLife*.

The study demonstrates that, when trained to read sequence data, artificial neural networks called Restricted Boltzmann Machines (RBM) can provide a wealth of information on protein structure, function and evolutionary features. It is believed to be the first method that can extract this level of detail from sequence data alone.

Proteins are formed of sequences of molecules called amino acids, which determine a given protein's structural and functional properties. But understanding which parts of the sequences are responsible for which properties is challenging. "Answering this question could have significant implications for pharmaceutical development," explains co-author Jérôme Tubiana, former Ph.D. student in the Physics Laboratory at l'École Normale Supérieure (ENS), Paris, France. "For example, it could help with the design of new proteins that have desired functions, or with predicting the future sequence evolution of proteins in living organisms, such as pathogens, and identifying appropriate drug targets."

To explore this question, Tubiana and his collaborators applied RBM to 20 protein 'families' - a group of proteins that share a common evolutionary origin. The researchers presented detailed results for four protein families, including two short protein domains called Kunitz and WW, one long chaperone protein called Hsp70, and synthetic lattice proteins for benchmarking.

They discovered that, after learning, the connections between the artificial neuronsin the RBM are interpretable and relate to the protein's structure, function (such asactivity) or phylogeny—the evolutionary relationships between protein sequences. Additionally, the team found that they could use RBM to design new protein sequences by composing and turning up or down the different artificial neural units at will.

"Our RBM model shows how machine-learning techniques can solve complex data recognition and draw conclusions from data in an interpretable way," says co-author Simona Cocco, CNRS Director of Research at the ENS Physics Laboratory. "This runs counter to the more complex, black-box models that are traditionally used in data science, as statistical analyses provided by these tools are largely uninterpretable. The interpretability of our method is a major benefit to scientists—it bears the promise of allowing them to generate proteins with desired functions in a controlled way."

"It will now be interesting to apply our model to proteins in pathogens," adds senior author Rémi Monasson, also CNRS Director of Research at the ENS Physics Laboratory, and Deputy Director of the Henri Poincaré Institute (CNRS/Sorbonne University), France. "Pathogens, particularly viruses, can often escape drugs through mutations that make treatments ineffective. Our method could be used to predict the mutational escape paths that are accessible to the functional protein from its current sequence, and help identify which combination of protein sites should be targeted by drugs to block all paths."

More information: Jérôme Tubiana et al, Learning protein constitutive motifs from sequence data, *eLife* (2019). DOI: 10.7554/eLife.39397

Provided by eLife

APA citation: Machine-learning model provides detailed insight on proteins (2019, March 12) retrieved 20 June 2019 from https://phys.org/news/2019-03-machine-learning-insight-proteins.html