Machine learning for new drug discovery

Finding suitable compounds for use in the development of new drugs is expensive and time-consuming, but researchers from STFC's Scientific Computing Department are working on new ways to improve both cost and time.

On average, drug development takes 12 years to get from the discovery stage to the marketplace. Although estimates will vary, the Association of the British Pharmaceutical Industry* puts the cost at more than £1 billion per drug. They estimate that the discovery stage, where researchers will need to identify a diseased protein cell (the target) and then find a molecule (or compound) to bind to the target and alter the disease, can cost £436 million.

Using conventional experimental methods, researchers will have to check thousands of molecules to predict how they will behave when exposed, for instance, to a cancerous cell. Will they bind to the cell to block it? Are they toxic enough to kill the cell?

Although fairly accurate theoretical and computational methods for prescreening exist, they require a huge amount of computing time and subsequent laboratory testing of the molecules. Up to 10,000 potential candidate molecules will be checked for each new drug.

The SOAP method of machine learning

A new approach known as the Smooth Overlap of Atomic Positions, or SOAP, can compare similarities in molecules at the atomic scale, giving near-perfect accuracy in identifying suitable new drug candidates for a fraction of the cost.

SOAP is a unique algorithm which trains computer programs to compare the 3D microscopic structures of molecules, including pharmaceutically active compounds – which are the building blocks of a material seen at the atomic scale – by mapping them together to identify similarities. Those molecules which do not have any similarities can immediately be eliminated from any further tests.

Using SOAP, which can be plugged into any machine that can run computer simulations, a database of 10,000 molecules that might be suitable to block a cancerous cell can be reduced in minutes to just a few of the most likely candidates.

This screening method could vastly accelerate identification of suitable molecules, saving huge amounts of time and reducing the need to run as many laboratory tests. It can predict with 99% accuracy whether or not a candidate molecule will bind to a target protein.

"SOAP could be a real game-changer. We found it is 99% accurate in predicting suitable candidate molecules for new drug development at just a fraction of the cost of conventional methods."
Albert Bartók-Pártay,
STFC Scientific Computing Department

Other applications

The SOAP algorithm provides a simple machine learning approach that can be widely used for a number of research applications. For instance, silicon is widely used in the semiconductor industry. By applying the SOAP method this industry will be able to minimise the computational effort required to model the structure, chemical stability and other complexities of silicon.

*Association of the British Pharmaceutical Industry
http://www.abpi.org.uk

SOAP has been developed by an international research team from the Science and Technology Facilities Council (STFC), the École polytechnique fédérale de Lausanne, the University of Cambridge, the University of Warwick and the US Naval Research Laboratory.

Published paper in the journal 'Science Advances: Machine learning unifies the modeling of materials and molecules; Albert P. Bartók et al http://advances.sciencemag.org/content/3/12/e1701816