BINF
sights.

ISSUE11
2021-T2-BREAK

Machine Learning in Bioinformatics

How computer science techniques for cybersecurity and music playlists are being used in rapid sequence analysis and novel drug development

P04 FEATURED ARTICLE
--
MACHINE LEARNING AND AI:
USING METHODS TO PREDICT MECHANISMS OF ACTION FOR NEW DRUGS USING PATTERN RECOGNITION IN CELL AND GENE PERTURBATION SIGNATURES

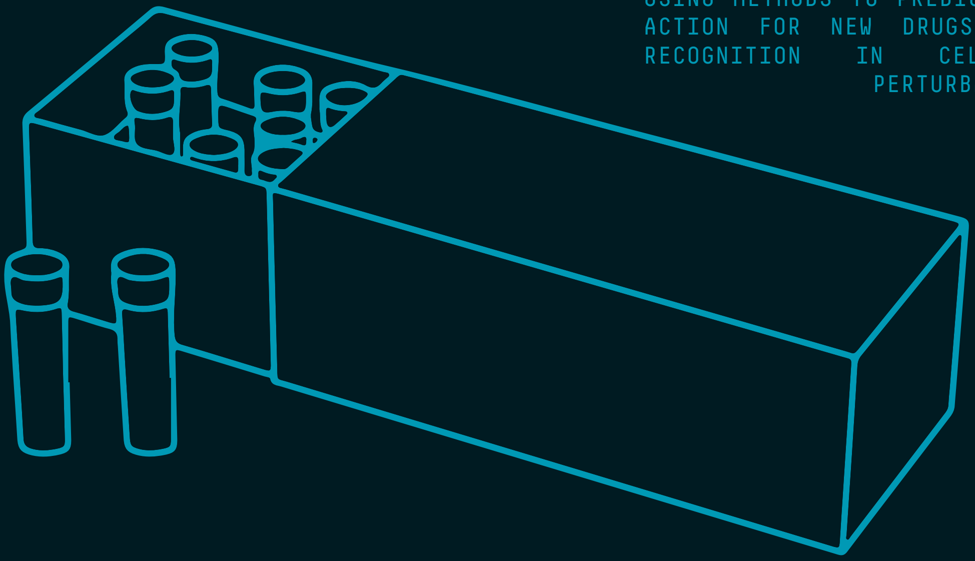


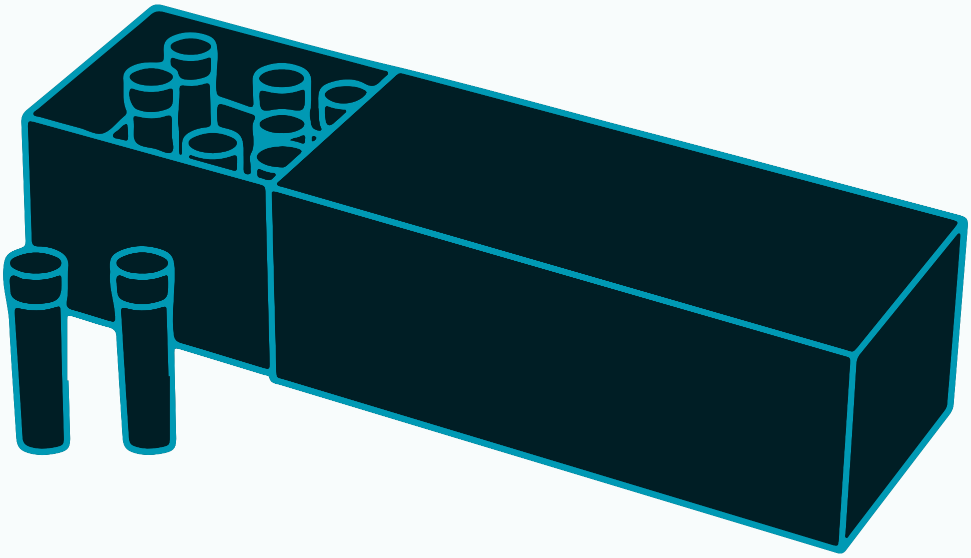
IMAGE SOURCE:
OXFORD NANOPORE TECHNOLOGIES LIMITED

ATTGCCTTTAGATUNSW
ACTBIOINFORMATICS
GCCGATCGASOCIETY_

“

I think the biggest innovations of the 21st century will be at the intersection of biology and technology. A new era is beginning.

— Steve Jobs



machine learning in bioinformatics

MACHINE LEARNING TECHNIQUES DEVELOPED IN OTHER AREAS OF COMPUTER SCIENCE OVER THE LAST FEW DECADES HAVE BEEN APPLIED TO THE VAST AMOUNTS OF BIOLOGICAL DATA, AND MAY OPEN THE DOOR TO NOVEL APPROACHES TO BIOMEDICAL DISCOVERIES.

Writer *Anthony Nguyen* ***Editor*** *Cam McMenamie*

Train, test, split...
It's not a gymnastics routine.

Have you ever needed to search up the definition of *bioinformatics* before? Perhaps it had been mentioned in one of your molecular biology classes, or maybe you needed to explain to somebody *what* you were studying. A simple online search tells us that bioinformatics is the 'intersection' of life science, computer science, mathematics, and statistics. But what exactly *is* this intersection? Let's try to answer this using a concrete example by exploring a branch of computer science currently experiencing tremendous attention.

Just a buzzword

Machine learning might seem like a buzzword, but for good reason. It is behind many of the online services offered by companies such as Netflix, Spotify, and Google (think about your film and music recommendations and the spam filters for your inbox), and are used to improve your experience on the platform by learning and making decisions so that you don't have to. In fact, this idea is at the core of machine learning. According to *IBM*, any algorithms that **use data to mimic human thinking** and **gradually improve its ability to perform tasks** based on that data can be thought of as machine learning.

Biological data and machine learning

We are no stranger to the explosion of rich biological and biomedical data that has occurred in recent years. Likewise, bioinformatics has become no stranger to machine learning approaches being developed to interpret this data. How can we use the results of proteomic, transcriptomic, and sequencing experiments? We learn from biological data in the same way we do other data.

Oracle describes the general life-cycle a machine learning program in a few simple steps:

1. Collect and prepare data
2. Choose your machine learning model(s)
3. Train and test your model(s)
4. Evaluate its performance

Let's see how this framework might be used to solve some real-life questions in biology.

Sequencing data

A fresh example is the base calling process involved in third-generation DNA sequencing pipelines such as in Oxford Nanopore's analysis system. The raw signal data from the nanopore sequencer is initially fed into a machine learning framework. The system learns to recognise patterns within the signal and assign the correct nucleotide sequences to the molecule, i.e. "calling" the bases. The base calling program does not necessarily ***know*** what an 'adenine' or 'cytosine' signal looks like; but it can certainly ***learn*** this during the training and testing phase. At the end of all of this, we are using a program that has already learnt how to decipher the raw signal like that of the molecule you are sequencing today, based on data it had seen and learnt from previously.

This is a far from trivial task, given that measurements taken at the molecular level are particularly prone to high levels of noise as well as the limited resolution of the molecular machinery involved. Nevertheless, algorithms which can learn these subtle patterns do exist and may even be tuned to detect individual nucleotide modifications, such as DNA methylation (a key process in regulating gene expression).

Transcriptomic data

The **Connectivity Map** (https://clue.io/cmap), a database resource managed by a number of institutes, houses gene expression and cellular signatures which can be used to “catalog transcriptional responses” to genetic or molecular perturbations. In other words, the transcriptomic data stored can be used to connect drugs, genes, and diseases to one another. This has enormous implications for novel treatment discovery and understanding gene pathways. In the same way, data from CMap can be used to train a chosen machine learning model to make predictions for us.

A closer look at drug discovery

Viewing and treating diseases at the protein level means that researchers will need to design a drug which can interact with specific proteins. Clearly, this involves knowing how drugs interact with a range of molecules, if at all. These interactions are called the drug’s **mechanisms of action**.

Connecting this back to what we were discussing earlier, the *Kaggle* competition ‘Mechanisms of Action (MoA) Prediction’ (more about Kaggle at https://kaggle.com/getting-started/44916) uses CMap data to apply machine learning to predict mechanisms of action. Here, participants cleaned up transcriptomic and cellular signals and fed them into machine learning models of their choice for training and testing. Models were fine-tuned and were ultimately able to weigh up how important each bit of data was in accurately predicting the biochemical activity of new, unseen drug compounds.

The ongoing development of such a tool is a promising avenue for researchers looking to derive hypotheses in the space of drug development.

At the intersection

We’ve explored two example applications at the intersection of biology, computer science, and statistics. However, there are countless other applications of machine learning ranging from genome-wide association analysis to protein structure and folding analysis. Bioinformatics tools, data, and analyses have grown in affinity with machine learning and will likely continue to incorporate techniques from the discipline to answer new and exciting biological questions.

BINFSOC

Merchandise Release.

BINFSOC HOODIES* available to order now!

Warm up and get toasty this lockdown season, whether you're just chilling at home, or practising BINF (also at home).

Our first merchandise item available, the COMPUTER-VIRUS hoodie design features a phylogenetic tree produced using R. The tree relates various viral DNA sequences together evolutionarily to one virus in particular (can you guess which?) -- you may have heard of it.

Each 'branch' of the tree is a unique viral strain, represented in braille for a visually stunning display of phylogeny and the wonders of bioinformatics -- ready for you to take home and wear indoors.

You can even flex it to your cat (or roommate) if they ever ask "hey, I wonder how COVID tracing works..?"
-- aesthetic, and educational!

COLOURS AVAILABLE:

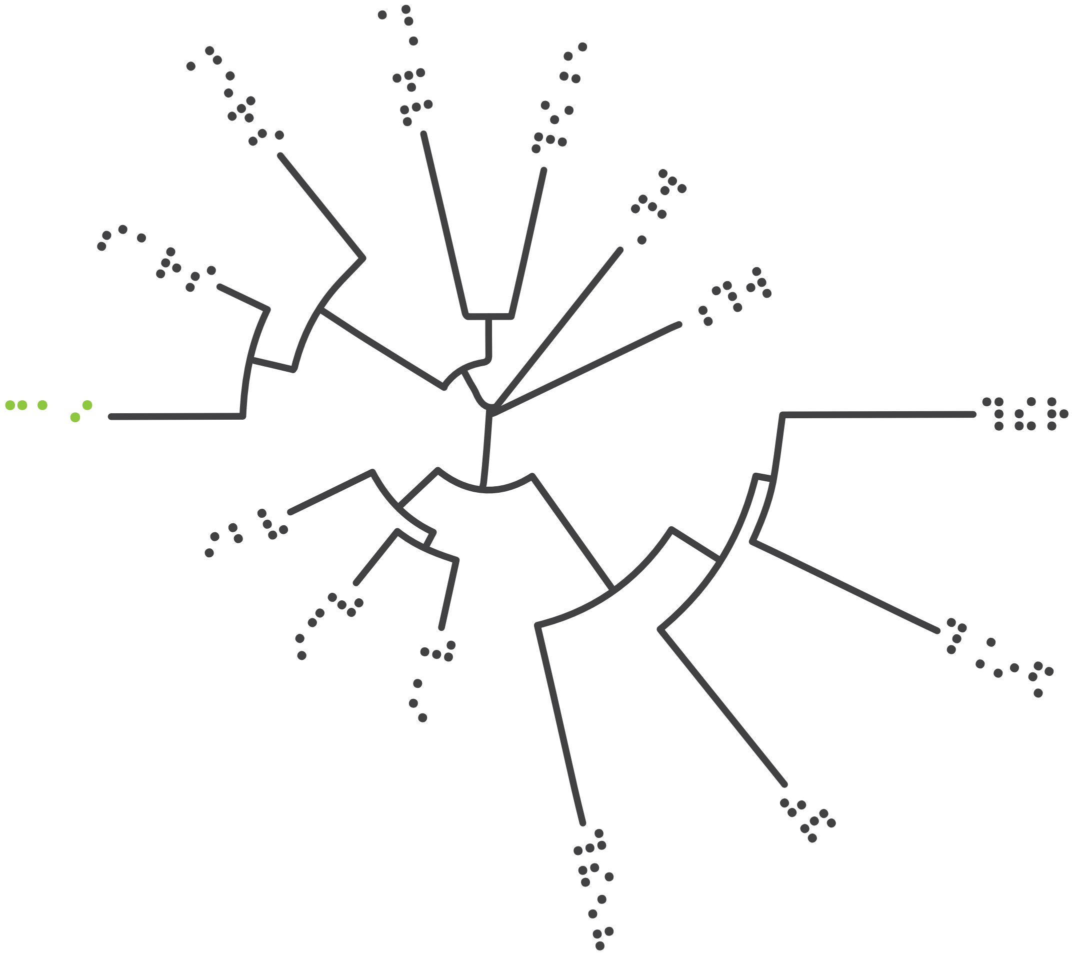
- BLACK
- CHARCOAL
- FOREST GREEN
- NAVY BLUE

ORDER NOW > https://unswbinfsoc.com/hoodie

*Viral DNA not included.

BINFSOC

ORDER NOW > https://unswbinfsoc.com/hoodie

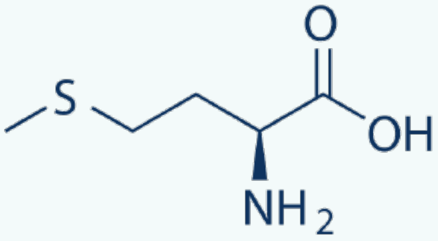


“Computer Virus”

AMINO ACID OF THE WEEK

[METHIONINE]

CHEMICAL STRUCTURE



METHIONINE

M

Met

149.21

RNA CODONS A U G

POLARITY:

NON-POLAR

DISCOVERY:

ISOLATED IN 1921 BY JOHN HOWARD MUELLER, AMERICAN BIOCHEMIST AND PATHOLOGIST

CANNOT BE SYNTHESISED BY HUMANS, BUT BIOSYNTHESIS IN PLANTS IS DERIVED FROM ASPARTIC ACID.

PROTEINOGENIC - BUILDING BLOCK OF PROTEINS. COMMON START CODON FOR PROTEIN SYNTHESIS. IMPORTANT IN GROWTH AND REPAIR OF BLOOD VESSELS.

METHIONINE IS AN IMPORTANT INTERMEDIATE SUBSTRATE IN BIOSYNTHESIS OF OTHER AMINO ACIDS. IMPROPER CONVERSION CAN LEAD TO ATHEROSCLEROSIS, WHERE WALLS OF ARTERIES DEVELOP LESIONS.

ONE OF TWO PROTEINOGENIC AMINO ACIDS CONTAINING SULFUR.

HIGH CONCENTRATIONS IN:

EGGS, FISH, MEAT, CHEESE, BRAZIL NUTS

LOW CONCENTRATIONS IN:

MOST FRUIT AND VEGETABLES

BINF sights.

Contact us

IF YOU HAVE ANY COMMENTS or feedback regarding BINFsights, please write to us at binfsights@unswbinfsoc.com

We also encourage anyone to share with us anything you'd like us to take a look at, be it a bioinformatics tool that you have made or find useful; or news in the bioinformatics world that you'd like to see written about in future issues.

TO VIEW PAST AND PRESENT issues of BINFsights, check out our website at unswbinfsoc.com/binfsights

Stay tuned on our Facebook page for updates regarding events and society news.

-- The BINFSOC Team

BINF
sights.