

BACKGROUND

DNA methylation age measures the cumulative effect of the epigenetic maintenance system.

A vast literature characterizes age-related hypermethylation or hypomethylation with specific genes or genomic regions that often are often associated with key developmental genes. By training a machine learning model on a set of over 7,000 DNA methylation samples across 50 datasets, we can predict your cell's DNA methylation age.

This novel epigenetic clock can be used to address a host of questions in developmental biology, cancer and aging research

What is DNA methylation?

DNA is composed of 4 building blocks called nucleotides: cytosine, guanine, adenosine, and thymidine. Cytosines in CpG dinucleotides (positions containing a cytosine and guanine connected by a phosphate) can be methylated to form 5-methylcytosine. In mammals, methylating the cytosine within a gene can affect the production of the protein from that gene. This regulation of production of important developmental proteins can lead to acceleration or deceleration of aging.

PROJECT

We implemented a penalized regression model (elastic net) in R on the cluster. Using CpG site coefficients for the informative CpG sites that have been documented in the literature [PMID: 24138928] the normalized model can predict a sample age (with associated confidence) along with other things like sample gender. Great care was given to ensure the final function call to predict age on a new sample is extremely optimized and fast for a web deployment, any long running tasks (normalization or continual model training) are saved for Celery distributed job queuing.

We aim to provide a service where anyone can upload a dataset they may have to predict age. Additionally, great potential for commercialization is in the testing process. For approximately \$30 (2015 processing cost) we could use a saliva, skin, blood, muscle, brain, or many other potential sample types to give a person's age within approximately 2 - 3 years. The applications in the criminal investigative industry would provide a revolution for crime scene analysis. When paired with other services (or given another few nights of hacking) a sample could inform the investigator an extremely accurate ethnicity. In approximately 72 hours, someone would know age, gender and ethnicity.

Additionally, delta age (chronological age minus predicted age) has been shown to predict all-cause mortality in later life [PMID: 25633388]. This provides recreational use of the service for individuals who would like personal health analysis of their biological age. We expect this application to be popular for individuals wanting to lead a healthier lifestyle.

All model normalization building (R), web application development (Django), model prediction (PANDAS, numPy) and design was completed since Saturday. As scientists and health industry bioinformaticians, we have been interested in developing a tool around this strong correlation finding for a few months now.





