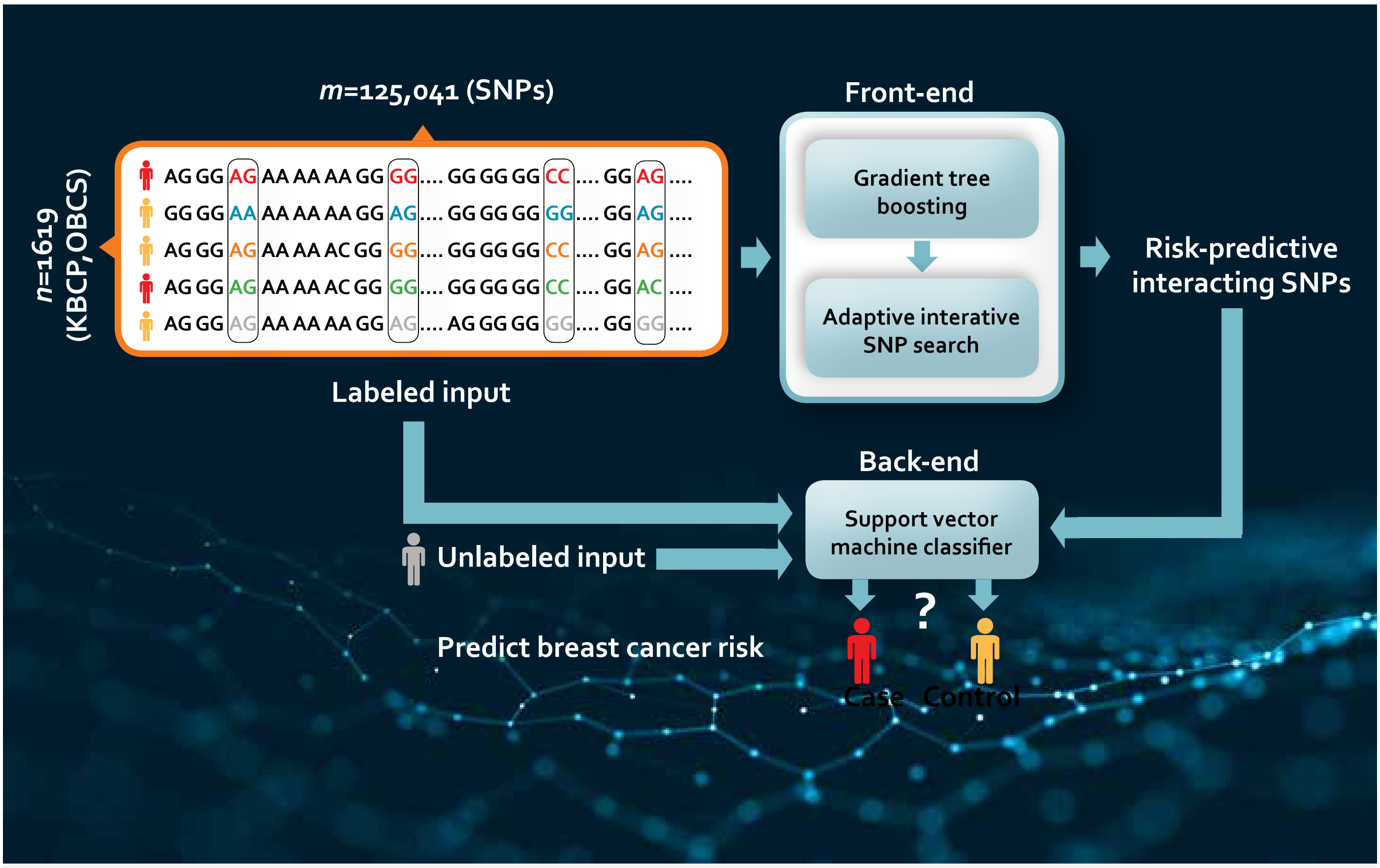
Searching markers for breast cancer by machine learning

In addition to gene variants there are also genomic variants in the locations of the single base pairs in the DNA stretch. The variations cause differences between individuals, but they can also help localise the disease-causing genes. These single nucleotide polymorphisms (SNP's) can act as markers indicating the disease. The artificial intelligence model developed at the University of Eastern Finland searches breast cancer interacting SNP's.

The huge amount of genomic data has made possible that researchers can now calculate what kind of gene variants are among the groups who have cancers. Hundreds or thousands of gene variants can have an impact to a single disease.

With statistical methods researchers can estimate how the gene variants of a single person can increase the disease risk. However, variations are also at the single base pair e.g. nucleotides in DNA, known as genetic variants or SNPs. DNA sequence variations occur when a single nucleotide (adenine, thymine, cytosine, or guanine) in the genome sequence is altered. Each SNP represents a difference in a single nucleotide. For example the nucleotide cytosine (C) can be replaced with the nucleotide thymine (T) in a certain stretch of DNA. It means that the base-pair cytosine-adenine can alter for thymine-adenine. Unlike gene mutations, SNP's are not necessarily located within genes. They can be also in the non-coding regions of the genes or regions between the genes. There are lots of SNP's in human genome. They occur almost once in every 1,000 nucleotides on average, which means there are approximately 4 to 5 million SNPs in a person's genome.

Genotyping of 125 000 SNPs was done by iCOGS chip in collaboration with Breast Cancer Association Consortium (BCAC). iCOGS is a genotyping array, designed to test genetic variants related to three hormone related cancers: breast, ovary and prostate. It has been genotyped on more than 250,000 subjects and SNPs across more than 50 regions known to harbour susceptibility variants for one of the target diseases.


m=125,041 (SNPs)
n=1619 (KBCP,OBCS)
Labeled input
Front-end
Gradient tree boosting
Adaptive interative SNP search
Risk-predictive interacting SNPs
Back-end
Support vector machine classifier
Unlabeled input
Predict breast cancer risk
?
Case
Control

AG GG AG AA AA AA GG GG.... GG GG GG CC.... GG AG....
GG GG AA AA AA AA GG AG.... GG GG GG GG.... GG AG....
AG GG AG AA AA AC GG GG.... GG GG GG CC.... GG AG....
AG GG AG AA AA AC GG GG.... GG GG GG CC.... GG AC....
AG GG AG AA AA AA GG AG.... AG GG GG GG.... GG GG....

Scientists have found more than 100 million SNPs (single nucleotide polymorphisms) in populations around the world. Most commonly, these variations are found in the DNA between genes. These variations may be unique or occur in many individuals. https://upload.wikimedia.org/wikipedia/commons/2/2e/Dna-SNP.svg

SNP's can be beneficial when searching the genetic risk factors for cancer. In biomedical research, SNP's are used for comparing regions of the genome between cohorts with and without a disease.

”When SNP's occur within a gene or in a regulatory region near a gene, they may play a direct role in disease by affecting the gene's function. We have a novel machine learning approach to identify group of interacting SNPs, which contribute most to the breast cancer risk,” says researcher **Hamid Behravan** from University of Eastern Finland. He works in Kuopio at the Institute of Clinical Medicine.

”We have published several findings about identifying the genetic component of the breast cancer risk that would reliably distinguish disease cases from healthy controls. Identifying the breast cancer-associated SNPs that reliably distinguish disease cases from healthy controls may be particularly useful in improving breast cancer risk prediction and developing individual treatment strategies”, says Behravan.

The standard hypothesis testing methods have measured only the association between a single SNP with a disease. However, the studies by University of Eastern Finland have demonstrated that risk factors for breast cancer can be predicted better when SNPs are examined as groups that actually interact with each other.

The idea of genome-wide association studies (GWAS) is to identify SNPs on the DNA, which explains the genetic component of the observed phenotype in genotyped people.

”Genome-wide association studies measure the association between an individual SNP's with a disease, but ignore the possible correlation among SNPs”, says Behravan.

”To date, population based genome wide association studies often use polygenic risk scoring (PRS), which aggregates the effects of risk alleles with the disease. However, PRS assumes that the disease-associated SNPs are independent of each other and the risk effects are linear and additive. We have shown that instead of evaluating the effect of single components (SNPs) one at a time, it would be particularly useful to improve breast cancer risk prediction by studying groups of interacting SNPs using an machine learning.”

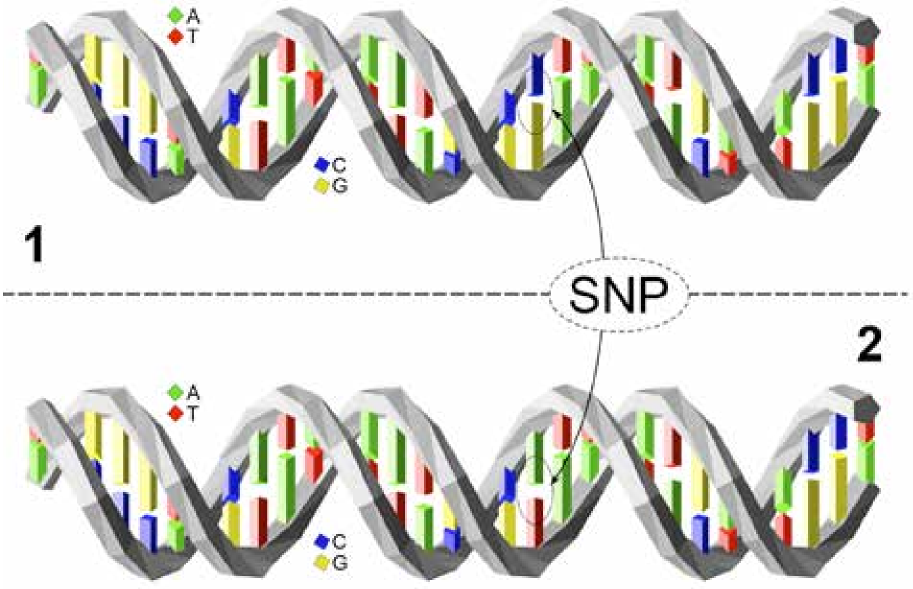
SNP's with true biological interpretation found by machine learning

The machine learning method developed in Eastern University of Finland has proven to be efficient.

”We found group of interacting SNPs that have true biological meaning. A biological analysis of the identified SNPs reveals genes related to important breast cancer-related mechanisms, such as Estrogen metabolism and apoptosis.”

Elevated endogenous estrogen levels are associated with increased postmenopausal breast cancer risk. There is also strong evidence that tumour growth is not just a result of uncontrolled proliferation but also of reduced apoptosis.

”So, we found genes behind those identified SNPs by our approach, and built gene interaction maps from those genes, and then we observed several separate networks related to breast cancer, such as Estrogen metabolism and apoptosis network. So not only our system found group of interacting SNPs with highest breast cancer risk predictive potential, but also those identified SNPs were behind a number of important biological entities in breast cancer. Therefore, interacting SNPs indicates both SNPs selected together, and SNPs involve in cancer related biological networks.”



Scientists have found more than 100 million SNPs (single nucleotide polymorphisms) in populations around the world. Most commonly, these variations are found in the DNA between genes. These variations may be unique or occur in many individuals. (SNP model by David Eccles).

Machine learns to search the genetic variants

The machine learning approach developed in Kuopio is based on a gradient tree boosting method followed by an adaptive iterative search algorithm. Boosting is the first module and searching the second module.

Boosting is an algorithm and method of converting weak learners into strong learners. Algorithm begins by training a decision tree. Weak classifiers are added sequentially to correct the errors made by existing classifiers towards building a strong classifier.

”The first module evaluates the accuracy of features, in this case the SNPs, on the breast cancer risk prediction. The first module provides an initial list of candidate SNPs with breast cancer-risk predictive features. ”

”The second module then uses the candidate SNPs in an adaptive iterative search to capture the interacting features. The best identified interacting SNPs are then used to predict the breast cancer risk for an unknown individual at the testing phase using a machine classifier. Classifier was trained to distinguish the breast cancer cases (positive samples) and healthy controls (negative samples).”

Since cancer is a multi-factorial disease caused by lifestyle, genetic, and environmental factors, individual analysis of the sources of genetic variants may not be enough to create a comprehensive view of the disease risk. According to Behravan other sources of data is needed.

“We are developing integrative machine learning approaches to combine different sources of data, such as demographic data.”

Ari Turunen

MORE INFORMATION:

School of Medicine, University of Eastern Finland
https://www.uef.fi/fi/web/laake

CSC – IT Center for Science
is a non-profit, state-owned company administered by the Ministry of Education and Culture. CSC maintains and develops the state-owned, centralised IT infrastructure.
http://www.csc.fi
https://research.csc.fi/cloud-computing

ELIXIR
builds infrastructure in support of the biological sector. It brings together the leading organisations of 21 European countries and the EMBL European Molecular Biology Laboratory to form a common infrastructure for biological information. CSC – IT Center for Science is the Finnish centre within this infrastructure.
http://www.elixir-finland.org
http://www.elixir-europe.org