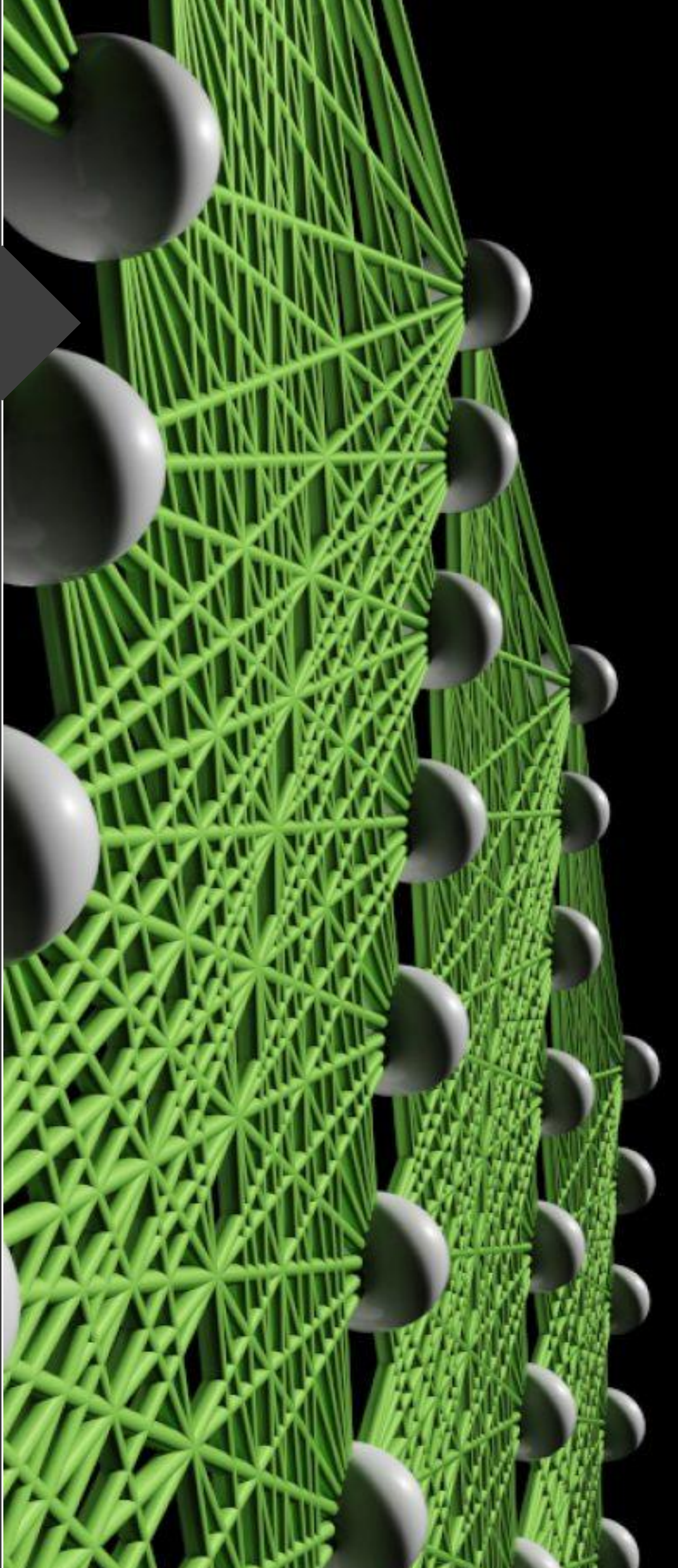
PATTERN
COMPUTER®

A MACHINE LEARNING APPROACH TO PATTERN DISCOVERY IN ONCOGENOMICS

Irshad Mohammed,[1] Nidhi Singh,[1,2] Meenakshi Venkatasubramanian,[1,2]

[1]Pattern Computer Inc., 38 Yew Lane, Friday Harbor, WA 98250.
[2]Molecular Ecosystems Biology Department, Biosciences Area, Lawrence Berkeley National Laboratory, Berkeley, CA 94720.



Pattern Computer Inc.

Pattern Computer Inc.
38 Yew Lane
Friday Harbor, WA 98250
USA

PATTERN COMPUTER MAKES NO WARRANTY OR REPRESENTATION, EITHER EXPRESS OR IMPLIED, WITH RESPECT TO THIS DOCUMENT, ITS QUALITY, ACCURACY, MERCHANTABILITY, OR FITNESS FOR A PARTICULAR PURPOSE. AS A RESULT, THIS DOCUMENT IS PROVIDED "AS IS," AND YOU, THE READER, ARE ASSUMING THE ENTIRE RISK AS TO ITS QUALITY AND ACCURACY.

IN NO EVENT WILL PATTERN COMPUTER BE LIABLE FOR DIRECT, INDIRECT, SPECIAL, INCIDENTAL, OR CONSEQUENTIAL DAMAGES RESULTING FROM ANY DEFECT, ERROR OR INACCURACY IN THIS DOCUMENT, even if advised of the possibility of such damages.

Some jurisdictions do not allow the exclusion of implied warranties or liability, in which case the above exclusion do not apply.

A Machine Learning Approach to Pattern Discovery in Oncogenomics

> *"This paper demonstrates the power of PCI's Pattern Discovery Engine that can sift through* ***sparse*** *and* ***unlabeled*** *data to uncover hidden patterns and trends of clinical relevance."*

Abstract

Owing to the advances in high-throughput technologies, the past decades have witnessed unprecedented growth in biological and medical data, such as genetic and epigenetic data, protein structures, as well as medical images. This data deluge requires computational tools that can effectively and efficiently transform it into valuable knowledge to allow for better understanding of human health and disease. In this regard, machine learning-based algorithmic frameworks show immense potential in extracting relevant features and to discover patterns from complex biomedical data that may provide clues to the underlying biological processes. In this paper, we present the development and validation of a novel algorithm for pattern discovery and its application in discovering biologically relevant genes and gene associations in unlabeled, and sparse genomic datasets. The strength of our method is not only in embracing the complexity of genetic architecture to identify distinct patterns embedded in a noisy background in a purely data-driven way but also, the ease with which it can be integrated into a comprehensive and flexible framework for data mining and actionable knowledge discovery that goes beyond human genetics and genomics.

(Video link: https://www.youtube.com/watch?v=LCdnCG8eUk4&feature=youtu.be)

Keywords: pattern discovery, drug repurposing, target identification, combination therapy, gene associations, oncogenomics , multidimensionality reduction, noisiness, heterogeneity, sparse, unlabeled data.

Introduction

Rapid innovation and technological advances have accelerated the mechanistic understanding of genome biology to an amazing level. Mapping genotype to phenotype, classifying different types of mutations, and predicting regulatory functional effects are all areas where exploiting the massive amount of genomic data obtained from large number of individuals can deliver new actionable insights. However, the extraordinary complexity and volume of information stored in human DNA presents a major hurdle to the "complete understanding" of genome with all its functions and interactions. Additionally, working with large volumes of heterogeneous data is challenging when conventional data-processing approaches are used. Novel and innovative methods are highly desirable in genome science to enhance our biological understanding of health and disease.

Application of machine learning, to genomics and biomedicine has major potential to revolutionize the state-of-the-art in genome analysis. Applying a targeted pattern discovery platform to systematically extract novel representations or features from input data, promises to leverage large datasets to uncover previously unknown structure residing within them, and enables accurate predictive outputs from the data. Implementation of various machine learning techniques has demonstrated breakthrough gains over conventional methods across a variety of domains such as medical imaging,[1-3] speech recognition,[4] computational chemistry,[5] high-energy physics[6] and robotics including self-driving cars,[7] and is particularly useful when large amounts of data are available. Since biology and medicine are data-intensive disciplines, novel algorithmic approaches are particularly well-suited to solve the problems of these fields. However, transferring the successes of these techniques from other domains into genomics is not as straight-forward given that biological data does not readily conform to the deterministic assumptions underlying many mainstream implementations. Thus, as these complex datasets present new opportunities, they also pose new challenges that need to be addressed.

Usually, high-dimensional data from biological experiments involving genomics or other high-throughput assays yield large and multifaceted datasets and are often represented as a data matrix. In these data matrices, each row corresponds to a gene or a feature, and each column corresponds to a sample or a patient. By analyzing these matrices computationally, it is possible to learn characteristic information within samples and to identify key features between samples to examine essential biological functions. However, the presence of high levels of noise and systematic biases in high-throughput experiments presents considerable challenge in genomic data analysis. Another fundamental challenge in biomedical data is their underlying representations which is often sparse. For instance, of an estimated 20,000-25,000 human protein-coding genes, alterations (or mutations) in only one gene can cause certain genetic diseases or disorders; a gene network is sparse as a gene either directly regulates or is regulated by a small number of genes relative to the total number of genes in the network; many biological signals are sparse or compressible as they have ‘concise representations’ when expressed in a transformed domain or basis. Furthermore, labeled data is often scarce and expensive to procure primarily due to the lack of knowledge of genetic process that makes discernment of an event difficult. Even when the data isn’t sparse and/or when clinical labels are obtainable, the best any standard machine learning algorithms, such as decision trees, gradient boosting machines, or neural networks, can do, is to build classifiers, or allow for importance sampling, or clustering. However, doing so wouldn't be of much use if the aim is to obtain actionable information. For that, one needs to find clear signals and quantify the inter-relationships of these genes i.e., how one gene is associating with another.

Complex diseases such as cancers, are caused by multiple interacting pathogenic genes. and so, the traditional ‘one target, one drug’ therapeutic mode has limited therapeutic potential. Recent studies have shown that targeting multiple disease-associated genes has greater therapeutic potential than single-target therapies,[8] and may not only bring synergistic or additive effects but also, reduced toxicity and delayed resistance for effective disease control. Inhibitors targeting synthetic lethal partners of genes mutated in tumors are already successfully utilized for effective and specific treatment in the clinic.[9,10] The search-space for possible signatures of interactions, however, is intractably large, and computational methods that limit the experimental effort of validating the interactions for therapy are highly desirable.

In this work, we address the very challenging yet fundamentally important problem of finding robust gene representations in sparse and unlabeled data. We have applied this strategy to the very sparse mutational data from lung cancer. The aim was to learn the standardized representations of all impacted genes that may allow the discovery of rare or non-obvious mutational patterns that can potentially illuminate new options for developing targeted therapies against cancer.

Method

In this section, we describe the general framework of our strategy to identify biologically and clinically relevant gene associations, outlined in Figure 1.

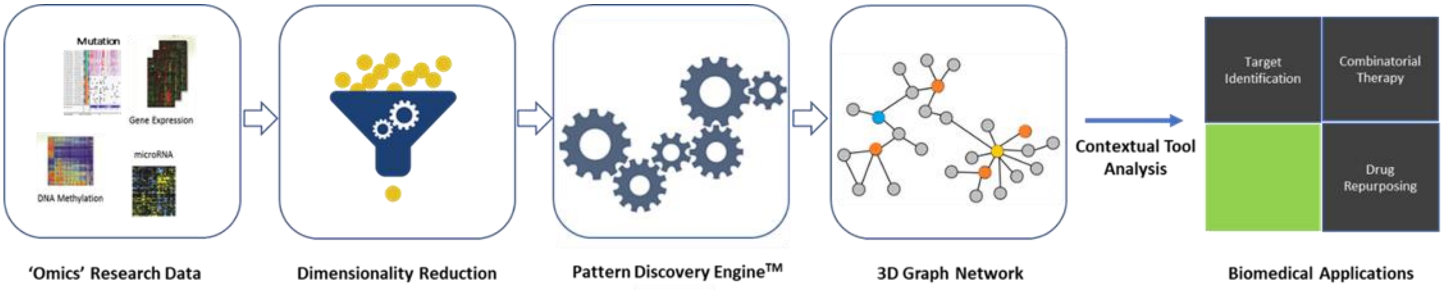


Figure 1. An outline of our approach to identify biologically relevant mutational patterns.

Accordingly, we developed a series of machine learning algorithms that among other things allows for mapping the input data to a very high-dimensional vector space to explore its topology, all as a means of locating embedding spaces for the standardized representations of mutated genes. We applied this tool to study the associations between the very sparse somatic mutation data obtained from 586 lung adenocarcinoma patients that was gathered by The Cancer Genome Atlas (TCGA). As a first step, a dimension reduction approach is applied to identify the most relevant features (i.e. genes) in the data. In general, this framework is compatible with multiple data sources including gene expression, mutation, methylation data and the criteria for the selection of genes varies based on the data considered (for e.g. frequency in mutational data, variability in gene expression data). This filtered data is then processed using our tool that finds gene representations as embeddings that can be visualized by constructing a 3D network graph to displays association between different genes; Figure 2. The resulting gene associations are then parsed through our contextual engine to identify, prioritize or validate clinically actionable gene associations that can be used to generate testable hypotheses and the final hand-selected candidates can be submitted for biological evaluation.

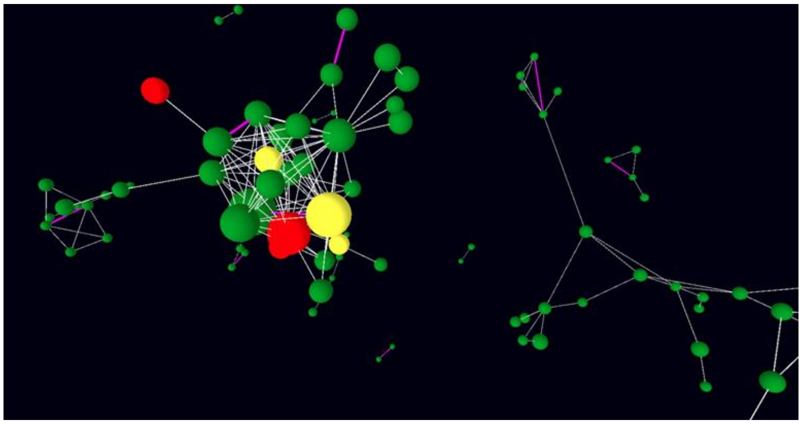


Figure 2. The figure shows a specific region of the three-dimensional gene association network graph. Each node in the network represents a mutated gene. The size of the node (or circle) represents the frequency of occurrence of that particular gene in lung cancer. Larger nodes, therefore, represent the most important genes. Different colors of the nodes represent different properties - Red: genes with an existing, approved drug for lung cancer treatment; Yellow: genes with an approved drug for other cancer treatments; Green: all other mutated genes in lung cancer. Please note not every gene is connected to every other gene, the connections are very specific. Finally, the connections in Magenta are stronger than the grey ones.

So, how did we know that the edges we drew were meaningful and not randomly constructed? To answer this question, we performed the following four tests:

(a) It stands to reason that the genes from the same family usually have similar functions and, therefore, should have stronger associations. To test this, we made a separate network graph and highlighted the genes belonging to one special family – UGT – that provides instructions for making particular enzymes. In the network shown in Figure 3a, we are showing all the members of UGT family genes shown in orange in Fig. 3a, and the rest of all the genes in green. It is worth noting that there are so many interconnections between the family members that it stands out. This is a complex network, containing thousands of genes and connections, yet we easily identified this family of genes, illustrating that even in the most complex networks, our method is able to find strong connections between functionally similar genes.

(b) While combining genomic datasets from different sources, the problem of gene aliases is quite common. When we apply our tool to integrated datasets, we find such genes are strongly associated with each other. As an example, Figure 3b shows strong connection between genes and their aliases (represented in purple). This is remarkable since no information regarding these aliases was provided in the dataset or in advance to the tool.

(c) Using statistical approaches, one can ascertain frequently co-occurring genes from the data itself. As expected, our

> “Extracting hidden patterns in biological and medical data is a highly non-trivial task.”

tool easily identified these pairs as shown in Figure 3c.

(d) There are some genes that are functionally similar and do not co-occur since only one of them is required to do the task. While co-occurrence is easily discernible, identifying functionally similar but mutually-exclusive pairs is not as straight-forward. Figure 3d shows that our tool can identified these mutually exclusive pairs that are connected by the cyan lines.

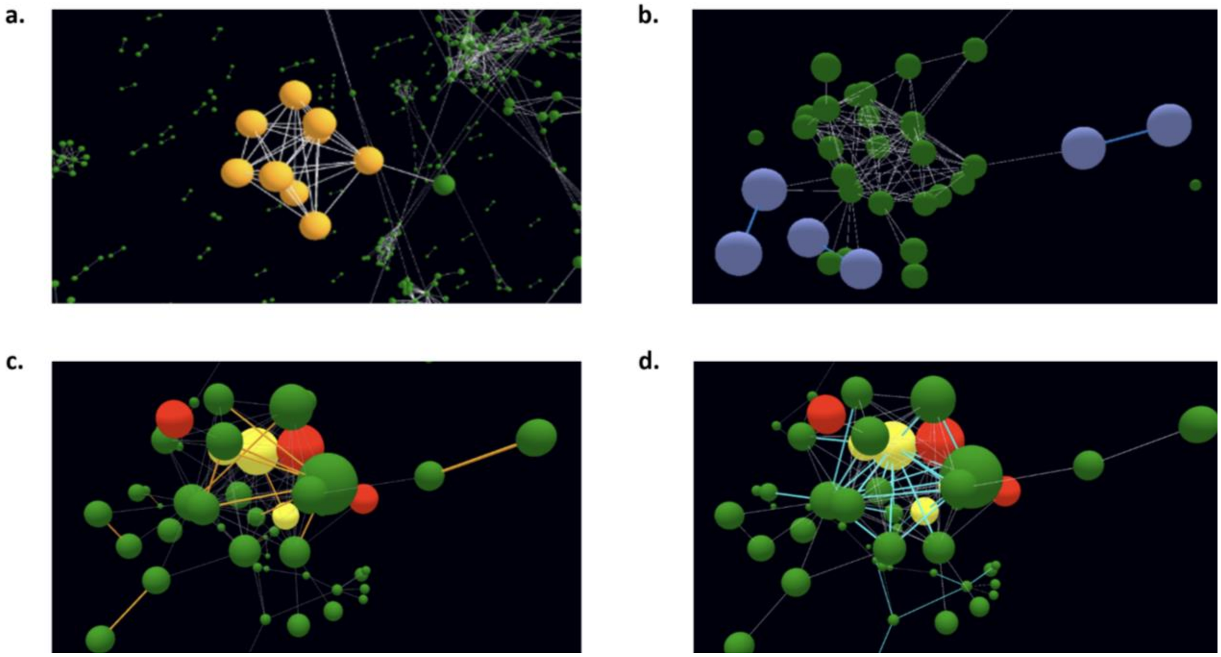


Figure 3a. A representation of genes that belong to the same family cluster together in the generated network graph. All the members of UGT family genes in orange, and the rest of all the genes in green. **b.** Network graph showing connected aliases in purple color. **c.** A graphical representations of co-occurring gene pairs connected by orange lines. **d.** A representation of functionally similar but mutually-exclusive gene pairs connected by the cyan lines.

Upon successful validation of our method, we can say with confidence that our standardized representation of genomic data advances the state-of-the-art.

Biomedical Applications

Traditional drug development is a long, tedious and expensive process, with an estimated cost of bringing a drug to market now exceeding $2.5 billion and taking between 10-15 years. Reducing this time frame, decreasing costs and improving success rates with more targeted approaches to discovery and streamlining process workflows in pharmaceutical R&D could significantly improve the profitability of these ventures. Accordingly, our approach can greatly speed up the traditional drug discovery process. Here are some specific application scenarios:

Identification of Novel Target(s)

Successful drug discovery hinges on choosing the right target or combination of targets with relevance to human disease and evidence that modulating them will be beneficial for patients. A vast majority of potential drug candidates fail because of poorly understood target-disease associations, pathway-target-drug-disease relationships and perhaps, adverse events profiling. Using our proprietary *in silico* approach to pattern discovery, we can effectively link genes and diseases to identify novel therapeutic targets with the potential to considerably decrease attrition rates in the drug discovery pipeline by significantly reducing the initial search space. The aim is to provide better evidence behind the role of the targets in diseases that might improve success rates and/or allow early termination of implausible drug development programs.

Repurposing

Drug repurposing, or drug repositioning, refers to the identification of new indications from existing drugs and the application of the newly identified drugs to the treatment of diseases other than the drugs' intended disease. One advantage of drug repositioning is that most of the repositioned drugs have already passed a series of tests, and so have lower risk of unexpected toxicity or side effects. With the application of our platform to real-world data, pharmaceutical companies can repurpose drugs faster and at lower costs than developing new drugs. This, in turn, may speed up the review of such drugs by the Food and Drug Administration (FDA) and, if approved, their integration into patient treatment.

Combination Therapies

One of the major issues with current treatment regimens has been that they ultimately stop working for patients resulting in recurrence of the disease. This is due to a phenomenon called drug resistance. This is particularly true for patients suffering from complex diseases, such as cancer, that are known to be caused by complex interactions between many genes. One of the solutions to overcome or delay the resistance has been to treat the patients with combinatorial target therapy rather than targeting only one gene at a time. However, it is difficult to select and experimentally evaluate effective combinations due to the large number of possible combinations. Using our techniques, it is now possible to identify the best synergistic gene combinations that can be used to develop more effective drug combination therapies.

References

1. Esteva A, Kuprel B, Novoa RA, Ko J, Swetter SM, Blau HM, Thrun S. Dermatologist-level classification of skin cancer with deep neural networks. Nature. 2017 Feb 2;542(7639):115-118. doi: 10.1038/nature21056. Epub 2017 Jan 25. Erratum in: Nature. 2017 Jun 28;546(7660):686.
2. Gulshan V, Peng L, Coram M, Stumpe MC, Wu D, Narayanaswamy A, Venugopalan S, Widner K, Madams T, Cuadros J, Kim R, Raman R, Nelson PC, Mega JL, Webster DR. Development and Validation of a Deep Learning Algorithm for Detection of Diabetic Retinopathy in Retinal Fundus Photographs. JAMA. 2016 Dec 13;316(22):2402-2410. doi:10.1001/jama.2016.17216.
3. Lakhani P, Sundaram B. Deep Learning at Chest Radiography: Automated Classification of Pulmonary Tuberculosis by Using Convolutional Neural Networks. Radiology. 2017 Aug;284(2):574-582. doi: 10.1148/radiol.2017162326. Epub 2017 Apr 24.
4. LeCun Y, Bengio Y, Hinton G. Deep learning. Nature. 2015 May 28;521(7553):436-44. doi: 10.1038/nature14539.
5. Goh GB, Hodas NO, Vishnu A. Deep learning for computational chemistry. J Comput Chem. 2017 Jun 15;38(16):1291-1307. doi: 10.1002/jcc.24764. Epub 2017 Mar 8.
6. Baldi P, Sadowski P, Whiteson D. Searching for exotic particles in high-energy physics with deep learning. Nat Commun. 2014 Jul 2;5:4308. doi: 10.1038/ncomms5308.
7. Massachusetts Institute of Technology. "Self-driving cars, robots: Identifying AI 'blind spots'." ScienceDaily. ScienceDaily, 25 January 2019.
8. Quan Y, Xiong L, Chen J, Zhang HY. Genetics-directed drug discovery for combating Mycobacterium tuberculosis infection. J Biomol Struct Dyn. 2017 Feb;35(3):616-621. doi: 10.1080/07391102.2016.1157037. Epub 2016 Jun 27.
9. Sessa C. Update on PARP1 inhibitors in ovarian cancer. Ann Oncol. 2011 Dec;22 Suppl 8:viii72-viii76. doi: 10.1093/annonc/mdr528.
10. Underhill C, Toulmonde M, Bonnefoi H. A review of PARP inhibitors: from bench to bedside. Ann Oncol. 2011 22, pp. 268-279.

To learn more about Pattern Computer and how to partner with us, e-mail us at inquiry@patterncomputer.com