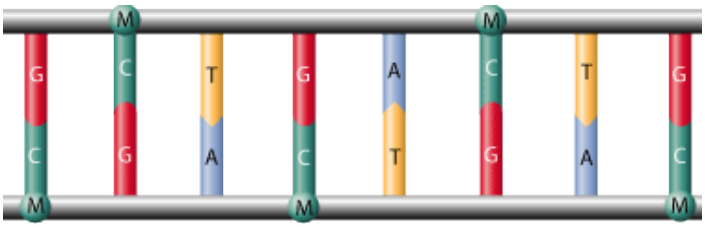
Computing the Cure to Cancer

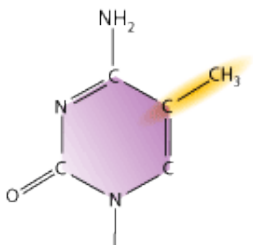
Kirk Mallett

The approaching future of health care is uniting humans and machines to tirelessly attack the most challenging diseases. Computers are essential to conquering recalcitrant diseases through extreme precision and awareness. Diseases such as cancer, that are predominantly controlled immunologically, genetically, and epigenetically necessitate individualized prognosis and treatment. During the next several years, we will see the coevolution of medical science and machine intelligence to provide such personalized care. In the following, we will see that to meet the instrumental and computational challenges of cancer and other diseases will require substantial progress beyond what presently occurs in oncology labs and clinics. By assessing cancer as if it were essentially physiological information, we see the importance of genetics and epigenetics during diagnosis and treatment. Then we look at the role of biomarkers in improving the state of practice in the lab and clinic, all the while emphasizing the imperative to work alongside machine intelligence.

"In epigenetics, there is an important asymmetry, that nearly your entire body is made of cells that are genetically identical, yet irreversibly differentiate into specialized roles."

Personalized cancer treatment requires large amounts of patient specific data. The central dogma of cancer progression is the buildup of mutations in the genetic sequence of certain genes. These oncogenes are especially critical in bestowing our cells with the tools and behaviors of cancer. Genetic analysis informs your doctor about what type of cancer you face, indicating the treatments more likely to defeat your tumor, which grows wildly in your organ. In some approaching year, the specific cells betraying your body will succumb to a treatment tailored uniquely to those cells. Few other cells will be harmed, dramatically minimizing side effects. Yet this specificity cannot be solely based on genetic information, which are the instructions on the construction of protein. Proteins are the nanomachines operating the complexities of both your healthy cells and your cancerous cells, and we know these cells are different by looking at their genetics. Genetics best informs us about what variants of proteins your cancer might express, and about how they differ from healthy cells. Genetics does not indicate at what levels proteins are expressed, if at all.





DNA methylation is the addition of a methyl group (M) to the DNA base cytosine (C).

"The larger a dataset is, and the more sophisticated the analysis becomes, the much greater the time required to process that data."

"Graphs, in the form of Hidden Markov Models (HMMs) underlie the symmetry involved in a lot of biomedical analysis."

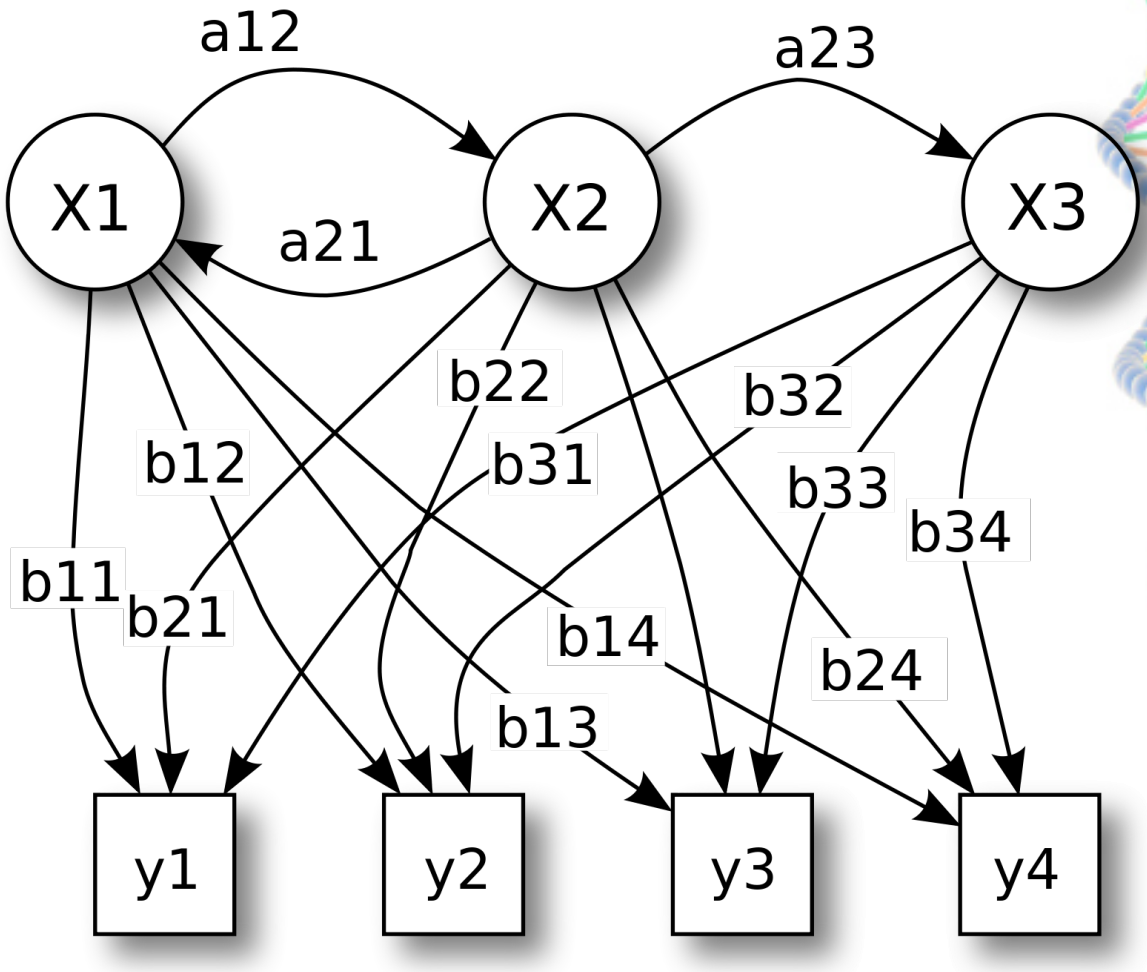
Transcriptomes, methylomes, and other epigenetic information, not genetics, tells us which proteins are being expressed at what levels and how they are being regulated (Dancey). Transcriptomes of RNA are difficult to access from cells, but will one day compliment genetics in deciphering what to target in your unique cancer. The pattern of DNA methylation and histone modification in tumorigenic cells could map out potential regulatory targets. In the future, there may be some drugs that stop the expression of critical proteins in your cancer. Epigenetics regards the regulation of genetic expression, when and which proteins are produced in a cell, and the variation of such regulation between different cell types and cells of the same type in different environments. In epigenetics, there is an important asymmetry, that nearly your entire body is made of cells that are genetically identical, yet irreversibly differentiate into specialized roles. Scanning through cells of every tissue we see a symmetry about their potential to produce any protein and perform any cellular role. This symmetry is broken across the axis of expression, the potential to be anything is suppressed in order to create specialized cells. Within tissues and between cells there is differential regulation of protein generation; we are a unified body of cloned cells that distinguish themselves solely through their varied expressions.

"...when available, a doctor today can sometimes, and to a limited extent, formulate a personalized treatment for a cancer patient."

Cancer also differentiates itself, but takes differentiation a step beyond expression; cancer is a polyclonal network of highly interdependent cells (Parsons). To conquer your body's particular brand of cancer, not only must changes in genetic sequences and expression levels be monitored and understood, this data must be isolated from different clonal populations in a highly heterogeneous tumor. To acquire this data requires a large assortment of tools and a robust biomedical industry. Sophisticated mathematics and continued growth in computational capacity will process this data until it presents insight and actionable information. If you are fortunate enough to outlive vascular diseases, you will someday be asking your physician about a lump or pain on your body somewhere. With the onset of cancer, you will be mollified by the quality of information your doctor has on your physical state, and by the variety of treatments that can be tailored to your specific tumor. You may or may not see yourself as a unique individual, but your doctor will know your cancer uniquely. However, this will require datasets of great size being analyzed at tremendous speeds. The larger a dataset is, and the more sophisticated the analysis becomes, the much greater the time required to process that data. Personalized treatment is therefore impossible without exploiting inherent symmetry through computationally efficient mathematics.

"So data parallelism and ILP both derive from the fact that complex genetic patterns can be based on simple premises."

Graphs, in the form of Hidden Markov Models (HMMs) underlie the symmetry involved in a lot of biomedical analysis. Interpreting sequences of DNA (Meng), DNA methylation patterns (Lee), or regulatory motifs in DNA (Wu) will become dominated by HMM methods. HMMs are graphs of observables, say a position in a DNA sequence that can be an A, C, G, or T nucleotide. A position in a particular genetic motif may have a 10% chance to be an A, a 33% chance to be a C, 42% chance to be a G, and a 15% chance to be a T. The motif itself may have a 30% chance of occurring after another motif 1a, a 24% chance of occurring after a motif 1b, and a 20% known chance of occurring before a motif 3. So we see that a particular position in a particular motif has a probability of being an A, T, C, or G, and this probability is dependent on where the nucleotide might be in which possible motif (Meng). The most probable description



Typical Hidden Markoz Model (HMM)

of the DNA sequence then becomes whichever arrangement of motifs is most likely to produce the observed sequence. This description is of a lower fidelity than in modern models, which should also include analysis between observables larger than a single nucleotide (Lee).

To train a model to recognize motifs correctly, the model must try many possibilities. This is computationally taxing, but there are inherent symmetries that speed the computations up. Though there are only four kinds of nucleotides to consider, they form very long combinations that can be as complex as they are long. These sequences can be nonrepeating, and potentially highly interdependent, in theory. HMMs avoid this problem by modeling each position in a sequence so that it has only one dependency, the nucleotide in the previous position (Lee). Since every one of the billions of nucleotide positions are symmetric in their limited dependencies, their computation can be distributed across many processing units (Meng). For the same reason, the storage and movement of nucleotides, and the instructions for analyzing them, can be efficiently managed (Meng). A second source of computational symmetry is called Instruction Level Parallelism (ILP), which HMMs elicit through their basic operations (Meng). While the properties of an HMM can be very difficult to prove, the model uses fundamentally simple operations, such as multiplying a small list of numbers together. Though these operations must occur many billions of times, one operation can be simultaneously performed on several positions in a sequence or across several steps in a chain for a single nucleotide. So data parallelism and ILP both derive from the fact that complex genetic patterns can be based on simple premises.

> "...despite very large datasets on genetics and expression, very little insight has emerged from analysis of that data, limiting the progress of oncology."

There are three stages of cancer treatment that are improved with information on genetic (or other) markers:

prognostic, predictive, and pharmacodynamic. Prognostic markers become relevant when deciding whether a patient needs further, more aggressive, therapy after excising the primary tumor (Majewski). Predictive markers then determine which such therapies are effective for a particular patient (Majewski). Pharmacodynamic markers indicate what dosages will be sufficient in fighting a specific patient's tumor, while not being too toxic (Majewski). Assembling the data from these markers, when available, a doctor today can sometimes, and to a limited extent, formulate a personalized treatment for a cancer patient. Personalized cancer therapy has had some reserved success, such as targeting HER2 in breast cancer, BCR–ABL translocations in chronic myelogenous leukemia (the less common, less aggressive leukemia), the EGF receptor in lung adenocarcinoma, and BRAF mutations in melanoma (Tyson). These targets tend to be fusion proteins arising from mistakes in the separation of chromosomes (Rodriguez-Antona). In horrifying irony, by targeting tumor cells that carry these mutations, selection for resistant cells occurs among close genetic variants of targeted cells (Tyson). These resistant cells perpetuate the tumor despite treatment. In other words, by killing only the most tumorigenic cells, closely related cells begin to dominate the tumor, and some of those will be or become resistant to the treatment. What we need are improved ways to identify and target families of tumorigenic cells.

"To make mouse models suitable for personalized medicine, mouse avatars are being developed from patient derived tumor xenografts."

"The growth of this data implies the need for Natural Language Processing (NLP) agents that inform doctors about potential surgical outcomes and responses to drug and radiological treatments."

Only half of recently approved drugs have known biomarkers associated with them to indicate whether a patient might respond to the drug, what extent that response might be, or when dosage becomes toxic (Rodriguez-Antona). This means that despite very large datasets on genetics and expression, very little insight has emerged from analysis of that data, limiting the progress of oncology. Inadequate biochemical tools and techniques in the laboratory is a part of this retardation. Insufficient biomarkers, their assays, and the means for their bioconjugation limits information on molecular pathways (signal cascades) in cancers. Costs and risks have continually increased, and drug development slowed, for the discovery and refinement of new drug targets (de Castro). The quest to make personalized cancer treatments robust and routine has consequently faltered. Moreover, without critical information on drug pathways, side effects cannot be predicted. Long term concerns of genetic toxicology, how acute chemo- and radio- treatment affects genetic stability, remains unknowable in healthy tissue. Nevertheless, hope is high that progress will triumph over these challenges. In addressing these shortcomings and revolutionizing treatments, very large datasets are expected to be assembled and must be processed.

One source of this data will come from mice, which, besides being genetically similar to humans, are relatively easy to genetically alter. Mice also have a short gestation period, and are cheap to house. However, mouse models have lacked the clonal and signaling heterogeneity of human tumors; they are simply too simple to model our diseases. To make mouse models suitable for personalized medicine, mouse avatars are being developed from patient derived tumor xenografts. Immunodeficient mice are transplanted with a biopsy sample from a patient's tumor. From this, a mouse line is raised and used to test various cancer therapies, looking for ideal agents and dosages for an individual patient (Malaney). This data can be combined with whole-exome sequencing that can also be performed on biopsy samples (Rodriguez-Antona). This combined approach has recently been trialed, successfully treating thirteen patients. Six patients saw partial remission and the other seven experienced disease stabilization, having no progression of the tumor (Garralda). In eleven cases the

avatar model mimicked the patient response (Garralda). Over ten million codons of selected genetic regions, from each patient's tumorous and healthy samples, were analyzed to find only an average of 45 mutations (Garralda). Apparently, observing all the significant mutations of a cancer requires a high degree of fidelity. Yet even this level of scrutiny is not considered sufficient for general application of personalized cancer treatment (Parsons). We see the need for detailed yet efficient analysis of the millions of biopsies per year.

The most common and most important source of information in the clinic and in the medical science laboratory, today, is text based documentation and communication (Jensen). The growth of this data implies the need for Natural Language Processing (NLP) agents that inform doctors about potential surgical outcomes and responses to drug and radiological treatments (Jensen). NLP agents interpret and make sense of normal human language, such as English or Chinese, usually as blocks of text rather than spoken phrases. IBM is developing such an expert agent, a medical assistant based on their Watson project. IBM's product relies on numerous techniques to analyze text for patterns meaningful to doctors. In Watson, text is turned into nodes and lines on graphs. Phrases that emphasize nouns are designated as branch points, and phrases emphasizing verbs become the connections between those branch points (Kalyanpur). These phrases are created by Watson from simpler linguistic features, like parts of speech, and from dictionaries (Kalyanpur). An extension to Watson, called WatsonPaths, breaks a question into a set of smaller questions after Watson provides its top rated responses to the original question (Lally). WatsonPaths also asks the doctor questions, and utilizes the doctor's response to improve Watson's answer (Lally).

Microenvironment, developmental state, cell type, and other factors modify the expression and activity levels of hundreds of relevant molecular components in each tumor (Chin). The varieties of cancer with their diversity of genetic mutations compound the number of assays and molecular tools needed in both the laboratory and clinic (Chin). Great concentration of resources will compel progress in these areas, and to interpret and create value from the amassing data demands proportionate computational advances. There is a glimpse of the future today in the development of IBM's Watson and its successor. Beyond the symmetries in genetics and epigenetics, there is more symmetry to exploit in immunology, and endocrinology (Kolch; Brock; Melero). These symmetries become apparent in the theory and computation of modeling the complex interactions ubiquitous in those domains. By exploiting symmetries of physical and combinatorial structures that are medically relevant, several critical problems have become tractable. Simplifying the identification of disease relevant genes in an individual tumor is one instance. Symmetry is integral to simplifying and solving other principle challenges, like mapping the nearly inscrutably attenuated regulatory pathways of disease progression. These eventual triumphs await long tribulations of discovery, invention, and investment that will require greater integration of health industries and consumers (Fagnan).

References

Brock, A., Krause, S., & Ingber, D. E. (2015). Control of cancer formation by intrinsic genetic noise and microenvironmental cues. Nature Reviews Cancer, 15(8), 499-509.

Gonzalez de Castro, D., Clarke, P. A., Al Lazikani, B., & Workman, P. (2013). Personalized cancer medicine: molecular diagnostics, predictive biomarkers, and drug resistance. Clinical Pharmacology & Therapeutics, 93(3), 252-259.

Chin, L., Andersen, J. N., & Futreal, P. A. (2011). Cancer genomics: from discovery science to personalized medicine. Nature medicine, 17(3), 297-303.

Dancey, J. E., Bedard, P. L., Onetto, N., & Hudson, T. J. (2012). The genetic basis for cancer treatment decisions. Cell, 148(3), 409-420.

Fagnan, D. E., Fernandez, J. M., Lo, A. W., & Stein, R. M. (2013). Can financial engineering cure cancer?. The American Economic Review, 103(3), 406-411.

Garralda, E., Paz, K., López-Casas, P. P., Jones, S., Katz, A., Kann, L. M., ... & Hidalgo, M. (2014). Integrated next-generation sequencing and avatar mouse models for personalized cancer treatment. Clinical Cancer Research, 20(9), 2476-2484.

Jensen, P. B., Jensen, L. J., & Brunak, S. (2012). Mining electronic health records: towards better research applications and clinical care. Nature Reviews Genetics, 13(6), 395-405.

Kalyanpur, A., & Murdock, J. W. (2015). Unsupervised Entity-Relation Analysis in IBM Watson. In Proceedings of the Third Annual Conference on Advances in Cognitive Systems ACS (p. 12).

Kolch, W., Halasz, M., Granovskaya, M., & Kholodenko, B. N. (2015). The dynamic control of signal transduction networks in cancer cells. Nature Reviews Cancer, 15(9), 515-527.

Lally, A., Bachi, S., Barborak, M. A., Buchanan, D. W., Chu-Carroll, J., Ferrucci, D. A., ... & Welty, C. A. (2014). WatsonPaths: Scenario-based Question Answering and Inference over Unstructured Information. Technical Report Research Report RC25489, IBM Research.

Lee, K.-E., & Park, H.-S. (2014). A Review of Three Different Studies on Hidden Markov Models for Epigenetic Problems: A Computational Perspective. Genomics & Informatics, 12(4), 145–150. http://doi.org/10.5808/GI.2014.12.4.145

Majewski, I. J., & Bernards, R. (2011). Taming the dragon: genomic biomarkers to individualize the treatment of cancer. Nature medicine, 304-312.

Malaney, P., Nicosia, S. V., & Davé, V. (2014). One mouse, one patient paradigm: new avatars of personalized cancer therapy. Cancer letters, 344(1), 1-12.

Melero, I., Berman, D. M., Aznar, M. A., Korman, A. J., Gracia, J. L. P., & Haanen, J. (2015). Evolving synergistic combinations of targeted immunotherapies to combat cancer. Nature Reviews Cancer, 15(8), 457-472.

Meng, X., & Ji, Y. (2013). Modern computational techniques for the HMMER sequence analysis. ISRN bioinformatics, 2013.

Parsons, B. L. (2008). Many different tumor types have polyclonal tumor origin: evidence and implications. Mutation Research/Reviews in Mutation Research, 659(3), 232-247.

Rodríguez Antona, C., & Taron, M. (2015). Pharmacogenomic biomarkers for personalized cancer treatment. Journal of internal medicine, 277(2), 201-217.

Tyson, D. R., & Quaranta, V. (2013). Beyond genetics in personalized cancer treatment: assessing dynamics and heterogeneity of tumor responses. Personalized medicine, 10(3), 221.

Wu, J., & Xie, J. (2010). Hidden Markov model and its applications in motif findings. In Statistical Methods in Molecular Biology (pp. 405-416). Humana Press.

Image Sources

https://upload.wikimedia.org/wikipedia/commons/thumb/8/8a/HiddenMarkovModel.svg/2000px-HiddenMarkovModel.svg.png

http://www.abemployersolutions.com/Images/Slide%20Pics/DNA/DNA%20strand%20istock.jpg

http://www.scq.ubc.ca/wp-content/uploads/2006/08/methylation%5B1%5D-GIF.gif

Layout by Kara Turner