Research & Innovation

Computational Medicine – Breaking the Code

The human genome is made up of two long chains each with three billion letters, a living opus written in a four-letter alphabet. These letters make up a person's DNA, and can be thought of as an enormous book of instructions for building many molecules, not all of which are yet known.

Traditionally, we thought these letter combinations only convey instructions that tell the body how to build proteins, which are molecules that participate in virtually every process that happens in our bodies. Thanks to the ground-breaking work of the Human Genome Project, we now know that less than two percent of a person's DNA codes for proteins. We know very little about what the other 98 percent does.

Discoveries by my team at Jefferson have revealed an outsize role for many thousands of sequences that are located in the "uncharted" 98 percent. We focus on two categories of molecules, microRNAs (miRNAs) and transfer-RNA-derived fragments (tRFs). Together, the two categories comprise tens of thousands of different molecules, each about two-dozen-letters long.

For more than 10 years, fewer than 2,000 sequences in the human genome were known to contain instructions for making miRNAs. In 2015, my team reported findings that tripled that number. Our projections indicate that there could be tens of thousands additional such regions within our genome awaiting discovery.

Whether previously known or newly discovered, the scientific community quickly noticed that the miRNA-making instructions are used to build multiple molecules at the same time, the "isomiRs." My team showed that the specific combination of isomiRs made from each such sequence, and their abundances, depend on variables such as gender, population origin, ethnicity, specific diseases, etc.

For tRFs, molecules key to producing proteins, the story proved to be analogous. The community knew that tRFs arose from the regions of the genome containing the instructions for making transfer RNAs (tRNAs), helper molecules used in protein production. Again, my team showed that the specific combination of tRFs made from each such region, and their abundances, depend on variables such as gender, population origin, ethnicity, specific diseases, etc.

The number of events that could be regulated by isomiRs and tRFs is potentially massive and increases a great deal every time a new isomiR or tRF is identified. These molecules have a cascading effect that can have considerable downstream consequences on the expression of a particular trait, cancer, or other phenomena we have yet to discover. And because the specifics of these molecules depend on a patient's attributes, their regulatory consequences are central to precision medicine.

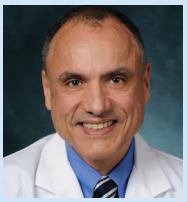
My team's research approach is "data-driven." We use high-performance computing techniques to look at these regulatory

Continued on back page

relationships in detail and begin to unravel the underlying complexity. To do this, we start with data, typically collected from many hundreds, or thousands, of individuals, which we then mine.

Data mining helps us quickly discard uninteresting information and zoom in on potentially actionable nuggets. With information from many people, computers can point out chunks of information that recur consistently in different groups of individuals. And it is typically the case that we do not know what these recurring nuggets do, if anything. Again, we resort to computing to prioritize them, and to limit the number of possible functions for each nugget. Once we have distilled this information, we transition to the experimental phase of our work that takes place in the wet laboratory.

Not surprisingly, these unconstrained, data-driven approaches at times suggest areas of inquiry that are not part of the current body of knowledge. In fact, during the last several years, using this approach we made many observations that we expect will impact the field of precision medicine. Some hold potential for designing diagnostics that can offer more accurate and earlier detection for diseases, while others could lead us to better treatments for cancer and other maladies.



Isidore Rigoutsos, PhD, was originally trained in physics and later in computer science. After graduation, he joined IBM's Research Division where he co-founded the Computational Biology Center more than 25 years ago. Soon thereafter, he moved deeply into studying genomic architecture and computational molecular genetics. At Jefferson since 2010, he is the founding director of the Computational Medicine Center and was recently selected to receive the Richard W. Hevner Professorship.