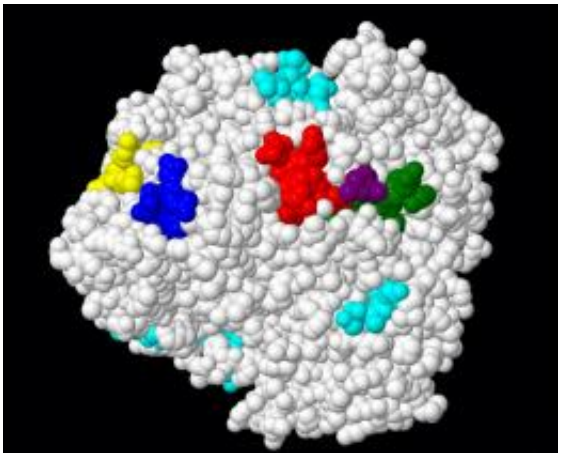
Locating patterns in human proteins

April 26 2011, By: Colin Poitras



A genetic motif. Credit: Sanguthevar Rajasekaran

A national research team, led by University of Connecticut engineering professor Sanguthevar Rajasekaran, is developing a new generation of exact algorithms that will help biologists locate patterns in human proteins and DNA. The work could eventually lead to new medicines to help fight disease.

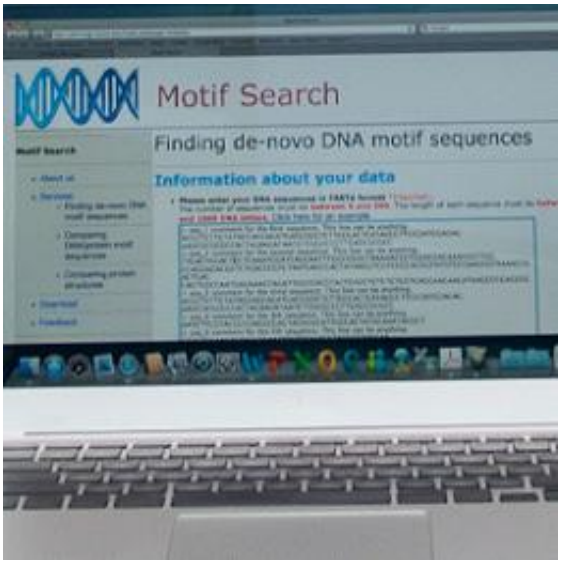
The research is supported by a $1.5 million, four-year grant from the National Institutes of Health that will allow the scientists to develop novel algorithms that can be used to analyze genomes for complex, biologically-relevant patterns called “motifs.”

“Genetic analysis and other approaches have identified many mutations, often found in protein coding regions, that are associated with inherited

human disease," says Rajasekaran, principal investigator for the grant. "If we can identify a drug that interferes with the protein containing the mutation, we can devise effective treatments. Analysis of protein and DNA sequences is an important approach for predicting protein function, and therefore an important part of the pipeline in drug discovery."

Rajasekaran, UTC Chair Professor in the Department of Computer Science and Engineering, is joined by Reda A. Ammar, professor and head of the computer science and engineering department, on the research team. Others involved in the initiative are Sartaj Sahni, a professor at the University of Florida, and Martin Schiller, an associate professor at the University of Nevada-Las Vegas.

Rajasekaran, who received his Ph.D. from Harvard, is considered an international expert in the field of applied algorithms. He holds nine U.S. patents, and in 2010 was inducted as a Fellow of the American Association for the Advancement of Science, an international non-profit organization dedicated to advancing science around the world.



An online motif search.

Supercomputers and efficient algorithms have become crucial tools for biologists trying to sort through the vast amount of data generated by the Human Genome Project, in which researchers set out to identify the approximately 20,000 to 25,000 genes of the human genome and determine the sequences of the three billion chemical base pairs that make up human DNA.

Finding patterns in genomes that are repeated over many sequences – and possibly over many species – is one way of identifying potentially useful information. For instance, if a particular motif is found in a protein that is believed to repress a certain disease trait, and researchers go on to discover a mutation of that motif in individuals with the disease, then drugs can be developed that may be able to repress the disease in those individuals and help them lead healthier lives.

But existing algorithms used in this kind of research tend to be complicated and take up large amounts of computing time and memory, which can be problematic for research teams with limited resources. Using the currently best known algorithms, identifying motifs of length 27 can take more than a month on a regular PC, Rajasekaran says. Identifying motifs of length 31 or more can take more than 5 years. Biologists would benefit greatly from algorithms that can find these long and complex motifs quickly and reliably. The longer and more complex the identified motif, the greater its usefulness and the less likely that comparative matching will lead to "false positives."

"Our role is to make things faster and more efficient while running in real time," says Ammar. "We are building tools that need to be friendly and easy to use for a non-technical person. In the past, biologists relied

solely on experiments in the lab; now they can turn to the computer, which is faster and can give those results in just minutes."

As part of their earlier work in this area, members of the research team created a web tool called the Minimotif Miner to search for motifs. The tool is now used by biologists worldwide.

With the new grant, the team will develop a web-based system incorporating three variations of the problem: Planted Motif Search, Edit-distance Motif Search, and Simple Motif Search. Rajasekaran says the new algorithms will help biologists find highly reliable short strains of genomic sequences among the huge number of possible strains available. It is akin to directing scientists to key shelves in a library full of millions of books.

"One can look at the entire genomic sequences of healthy individuals and compare them to those with cancer. There could be millions of differences, because those genomes are so huge," Rajasekaran says. "That's why we target mutations in motifs, because they are very fundamental and instrumental in protein-protein interactions."

Sahni, distinguished professor and chair of the computer and information science and engineering department at the University of Florida, is excited about developing sequential and parallel algorithms for motif search as a member of the team. He hopes the research will improve the well-being of society at large.

Schiller, a biologist and bioinformatician who was a co-developer of the Minimotif Miner system when he was an associate professor at the UConn Health Center from 2000 to 2009, likens the algorithm and motif research to scientists trying to understand hieroglyphics.

"We have all this information that appears to us as symbols but we don't

have the cipher key, the Rosetta Stone," says Schiller. "We're trying to put things in some order to extract meaningful information. By doing a pattern search, we are pulling the rules of life out of the genome and finding out what it means."

Provided by University of Connecticut