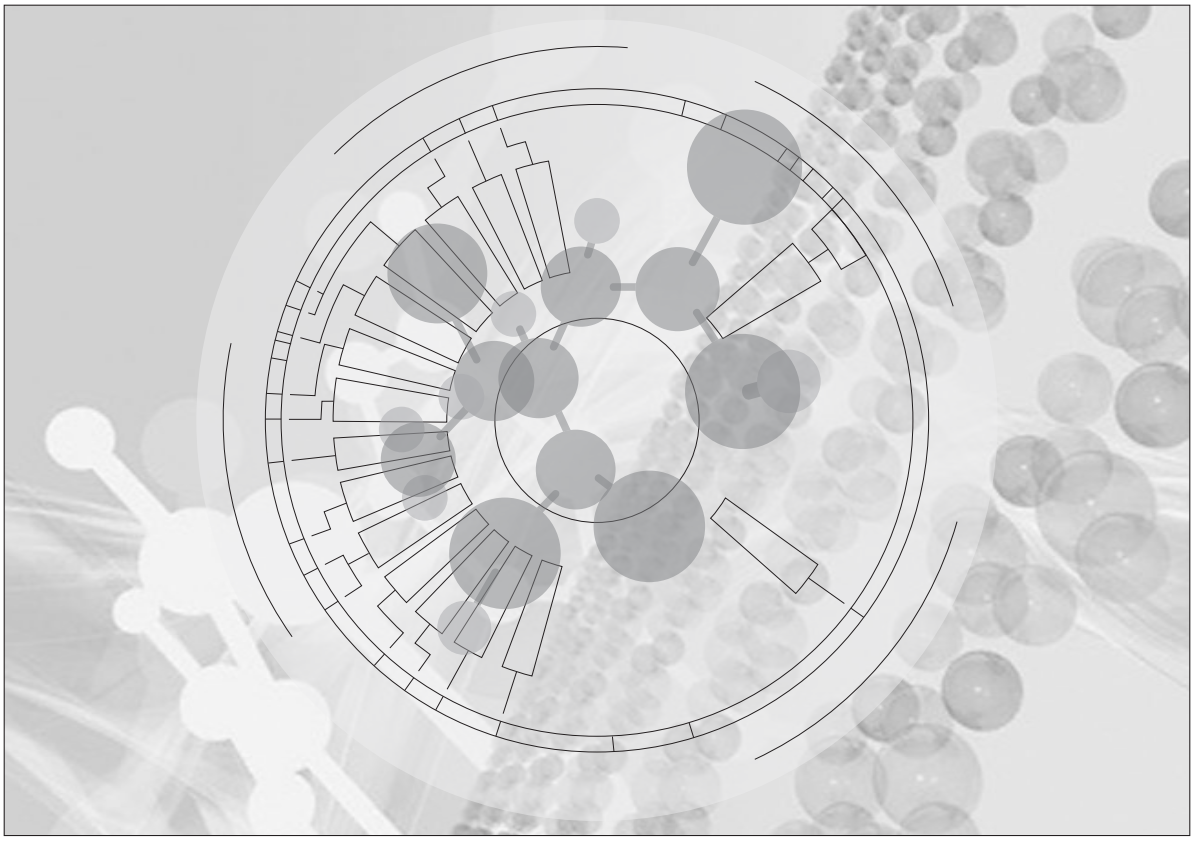

Trends in Molecular Sciences

ISSN 1994-5469

science alert

ANSInet
an open access publisher
http://ansinet.com

New Annotated Database Sifts Through Mountains of Sequencing Data to Find Gene Promoters

Researchers at The Wistar Institute announce the release of an online tool that will help scientists find "gene promoters" -- regions along a DNA strand that tell a cell's transcription machinery where to start reading in order to create a particular protein. The Mammalian Promoter Database (MPromDb) integrates the genome sequencing data generated at Wistar with publicly available data on human and mouse genomics. MPromDb pinpoints known promoters and predicts where new ones are likely to be found, the researchers say.

"Several complete genome sequences are available, including highly accurate assembled sequences from more than 1,000 individuals from the '1000 Genome Project,' with the goal of providing a comprehensive resource on human genetic variation and guiding us into the personal genomics era," said Ramana V. Davuluri, Ph.D., Associate Professor in Wistar's Molecular and Cellular Oncogenesis Program and Associate Director of The Wistar Institute Center for Systems and Computational Biology. "With this information, researchers can design personalized diagnostics and therapeutics or delve deeper into the study of gene regulation than previously thought possible."

Davuluri and his colleagues published details of how they built MPromDb in the journal Nucleic Acids Research.

Contrary to what was once the textbook view of genetics, one gene may not encode just one protein. In fact, scientists now know that a single gene may encode multiple versions of a given protein -- called a protein's isoforms -- which allows cells to make almost 100,000 distinct proteins even though our DNA only encodes about 20,000 protein-coding genes. As the body grows in the womb, cells may use different isoforms at different stages of development. Likewise, different adult cells may also use different isoforms of a protein depending on what type of cell it is, such as a neuron versus a skin cell.

"We have evolved this beautiful system where our DNA creates tremendous diversity from a limited set of genetic instructions," Davuluri said. "Recent evidence shows that at least half of all of our genes have alternative promoters that allow cells to make transcript variants and protein isoforms."

Earlier reports from the Davuluri laboratory showed that nearly 40 percent of genes use alternative promoters to create protein isoforms. According to Davuluri, integrating this information with data from other studies would surely find significantly more of these alternative promoters.

"Much of the genetic variations occur outside protein coding regions, such as gene regulatory regions," Davuluri said. "MPromDb provides context for data in the form of gene promoter annotations that can tell you where and when our bodies make a particular protein variant."

MPromDb mines its information from huge databases maintained by national and international consortiums of researchers, such as Gene Expression Omnibus (GEO) maintained by the National Center for Biotechnology Information and ENCyclopedia of DNA Elements run by the National Human Genome Research Institute. Essentially, MPromDb looks for key DNA sequences that could be potential binding sites for Polymerase II, an enzyme that creates the RNA transcript that the cell later translates into protein. The current database contains information on over 42,000 human promoters found in six different cell types and over 48,000 mouse gene promoters found in 10 different cell types.

"In fact, scientists are so good at generating this sort of information using next generation sequencing methods, that they collect information far in excess of what they might

need for a given experiment or project," Davuluri said. "This information all ends up in places like GEO, waiting to be discovered by groups like ours."

According to Davuluri, the Wistar Center for Systems and Computational Biology plans to expand MPromDb to include epigenetic data -- information on modifications to DNA that affect gene regulation; protein-DNA interaction data; and genetic variation data for both humans and mice.

Funding for this project comes from the National Human Genome Research Institute of the National Institutes of Health, the American Cancer Society, and the Philadelphia Healthcare Trust. Davuluri holds the Philadelphia Healthcare Trust Professorship at The Wistar Institute.

Co-authors of this study include Wistar Center for Systems and Computational Biology researchers Ravi Gupta, Ph.D.; Anirban Bhattacharyya, Ph.D.; Francisco J. Agosto-Perez; and Priyankara Wickramasinghe, Ph.D.

H. Sun, J. Wu, P. Wickramasinghe, S. Pal, R. Gupta, A. Bhattacharyya, F. J. Agosto-Perez, L. C. Showe, T. H.- M. Huang, R. V. Davuluri. Genome-wide mapping of RNA Pol-II promoter usage in mouse tissues by ChIP-seq. Nucleic Acids Research, 2010; DOI: 10.1093/nar/gkq775