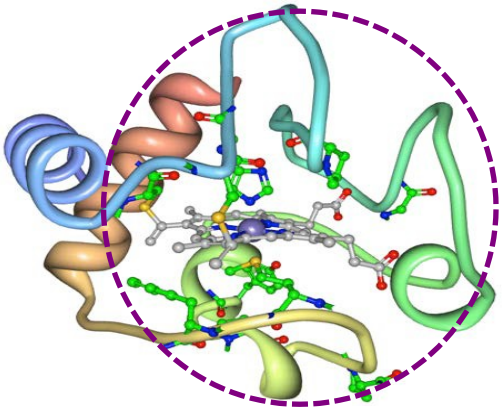
UNIVERSITY OF WATERLOO

TECHNOLOGY SUMMARY

Pattern Alignment Clustering to Reveal Functionality in Biosequences - WeMine™



Structure/Function Revealed

Reference
8810-7360

Patent status
U.S. Patent issued
Patent Pending in Canada

Stage of development
Working prototype and validating application data available

Contact
Scott Inwood
Director of Commercialization
Waterloo Commercialization Office
519-888-4567, ext. 33728
sinwood@uwaterloo.ca
uwaterloo.ca/research

Background
With the increased accessibility of lower cost high-throughput genome sequencing hardware, individuals will soon be able to secure personalized genome information for under $100. Similarly, this increase in sequencing efficiency will significantly enhance the productivity of basic medical and drug research efforts. Thus there is a pressing need for software that can unlock actionable information contained in this explosion of big data. Currently available biosequence software tends to be highly targeted to specialized narrow aspects of analysis and can be highly variable and inaccurate depending on the application.

Description of the invention
Waterloo researchers have developed a novel software tool that discovers, locates, aligns and clusters non-redundant statistically significant sequence patterns from a diverse range of biosequence data (e.g. protein, DNA and amino acid sequences). When these patterns are integrated into a tool called WeMine, researchers will be able to discover and identify conserved sequence patterns, protein binding sites, partition subgroups and uncover co-occurring interacting regions of biomolecules. Identifying highly conserved and non-redundant sequence patterns can lead to the discovery of genetic and/or protein markers (biomarkers), which can aid in drug development, disease treatments, and uncovering crucial information contained within big data.

Advantages
- Does not require any prior knowledge to generate and analyse the data thereby reducing wetlab biology and bioinformatics costs
- Comprehensive representation of the sequence residue associations and distant structural and functional associations
- Unaffected by mislabelled or unbalanced class labelling, which is a problem for other traditional algorithms, such as Hidden Markov Model and Support Vector Machine
- The software is flexible; it can process DNA sequences, Protein sequences, and discovery patterns of DNA-Protein binding regions
- Compared to existing methods, this software is 616x faster, more accurate, and not restricted by input data parameters, than traditional motif-finding software.

Potential applications
- Drug development
- Genomic, transcriptomic and protemic data analysis from such things as Microarray, RNA-Seq, and ChIP sequencing.
- Genetic testing\diagnostics – personalized medicine
- Biologics