Candidate Gene Screening and Novel Splice Form Detection from EST and Full-Length Sequence Databases Through Accurate EST assembly and CRAW analysis.

John Burke(1) and Hui Wang(2)

(1) Pangea Systems, jburke@panbio.com, (2) Affymetrix Inc.

ABSTRACT

A method (CRAW) is presented that 1) automatically screens for candidate genes connected to disease and development, 2) discovers previously hidden levels of information concerning differential expression of alternate gene forms (for example, regulated alternative splicing), and 3) corrects for artifacts (such as chimerism) in EST and "full-length" databases. CRAW is rapid and can process millions of EST sequences in a short amount of time. CRAW extends, corrects, and enhances the information obtained from sequence clusters through rigorous sequence assembly and through partitioning sequence clusters into sub-clusters based upon alternate exon usage, alternative gene termination, and novel domains.

It is proven that the organization imposed by effective assembly and partitioning can greatly increase the sensitivity/specificity of gene expression studies by accounting for the existence and tissue or pathology specific expression of gene isoforms and polymorphisms. Also, it is shown that this method can ameliorate the damage that artifacts, such as chimerism, inflict on EST cluster integrity.

To demonstrate CRAW effectiveness in modeling alternative splicing events, a broad-band study on regulated gene splicing in libraries associated with various stages of tumor development and different forms of cancer is performed.

Reference:

J. Burke, et al. 1998. Alternative Gene Form Discovery and Candidate Gene Selection from Gene Indexing Projects. Genome Research vol.8 no.3 pp.276-290. (March 1998).

J. Burke and A. Chou. 1999. CRAWview: for viewing splicing variation, gene families, and polymorphism in clusters of ESTs and full-length sequences. Bioinformatics 15(2) pp.376-381. (May 19990).