Bioinformatics 2005 – Paper 8 Answer 9 (PL)

(*a*) Present the aim of Blast software.

(*i*) Describe in words how the algorithm works. [5 marks]

BLAST (Basic Local Alignment Search Tool) is created to do similarity searches of a query against a database. The standard BLAST algorithm parameters are word length w, word score threshold T and segment score threshold S. The similarity search begins with identifying short words of length w in the query sequence that either match or satisfy some positive-valued threshold T when aligned with a word of same length in a database sequence. This is done by building an automaton of all the neighbors of the words. These hits act as seeds for initiating searches to find longer segments pair containing them. The word hits are extended in both directions along each sequence for as far as the cumulative alignment score can be increased. Extension of the word hits are halted when: the cumulative alignment score falls off by the quantity S from its maximum achieved value or the end of either sequence is reached. PSI-BLAST (Position Specific Iterated BLAST) is the state of the art Blast software. It uses significant alignments to construct position specific score matrix. This matrix is used in the next round of database searching until no new significant alignments are found. Widely used in genome studies is MegaBlast. Other software, such as PatternHunter, uses multiple non-consecutive position words that do not overlap heavily.

(*ii*) Describe the output of a Blast search [5 marks]

The fundamental unit of BLAST output is two sequence fragments of arbitrary but equal length whose alignment score meets or exceeds a threshold score. Each alignment shows a bit score (S) which is normalized and higher values mean more significant, and an E value, which is function of the number of hits of score that are expected by chance. It is based on a random database of similar size and lower value means more significant - zero means the sequences are identical.

(*b*) Describe the aim of microarray data analysis

(*i*) Describe the format of microarray data [2 marks]

Physically, microarrays (also known as gene chips, DNA arrays, or gene arrays) are small chips (often made of glass) with thousands of small embedded wells or spots. There are two basic methodologies, using large RNA fragments (full-length or near full-length genes) or oligonucleotides, 20mers to 70mers. Each spot represents one gene, and all of the genes in the genome may be represented on the same chip. Each spot contains fluorescent dye chemically bonded (red/green) to the DNA or RNA at that spot. The relative levels of the red and green dye can be optically analyzed, stored as a Tiff image and used to infer the relative levels of expression of each gene. For example red represents the control mRNA and green represents the experimental mRNA, the colors on each spot will reflect how the expression of each gene was changed by the manipulation of the experimental condition. Image analysis is performed to obtain the raw signal data for every spot. Poor quality data are filtered out and the remaining high quality data are normalized. Any spot with intensity lower than the background plus two standard deviations is excluded. The intensity ratios is log-transformed so that up-regulated and down-regulated values are of the same scale and comparable. Finally raw microarray data is in a tab-delimited text file with genes in rows and experiments or time measurements (individual arrays) in columns.

(*ii*) Describe in words how a cluster algorithm works [8 marks]

The purpose of clustering is to group genes with significant changes in expression levels that behave similarly under different conditions. Clustering methods can be roughly divided into two groups: partitioning and hierarchical methods. A partitioning algorithm describes a method that divides the data set into k clusters, where the integer k needs to specified by the user. Typically, the user runs the algorithm for a range of k-values. For each k, the algorithm carries out the clustering and also yields a quality index, which allows the user to select the best value of k afterwards. Hierarchical Algorithms. A hierarchical algorithm describes a method yielding an entire hierarchy of clusterings for the given data set. The Hierarchical clustering algorithm computes a distance matrix, after which genes are assigned to the nearest neighbor gene or cluster. After all grouping has been accomplished, the clusters are linked in a dendrogram (tree relationship), where each branch representing a group of genes with similar behavior One of the most well-known partitioning methods is k-means. In the k-means algorithm the observations are classified as belonging to one of k groups. Group membership is determined by calculating the centroid for each group (the multidimensional version of the mean) and assigning each observation to the group with the closest centroid. Euclidean and Manhattan distance measures are often used. The k-means algorithm alternates between calculating the centroids based on the current group memberships, and reassigning observations to groups based on the new centroids. Centroids are calculated using least-squares, and observations are assigned to the closest centroid based on least-squares. This use of a least-squares criterion makes k-means less resistant to outliers than the medoid based methods.