BIOINFORMATICS APPLICATIONS NOTE Vol. 18 no. 3 2002
Pages 488–489

ProbeMatchDB—a web database for finding equivalent probes across microarray platforms and species

Pinglang Wang, Fei Ding, Hsienyuan Chiang, Robert C. Thompson, Stanley J. Watson and Fan Meng

Department of Psychiatry and Mental Health Research Institute, University of Michigan Medical School, 205 Zina Pitcher Place, Ann Arbor, MI 48109, USA

Received on July 21, 2001; revised on October 25, 2001; accepted on October 29, 2001

ABSTRACT

Summary: ProbeMatchDB is a web-based database designed to facilitate the search of EST/cDNA sequences or STS markers that can be used to represent the same gene across different microarray platforms and species. It can be used for finding equivalent EST clones in the Research Genetics sequence verified clone set based on results from Affymetirx GeneChip®s. It will also help to identify probes representing orthologous genes across human, mouse and rat on different microarray platforms.

Availability: The database is accessible at http://brainarray.mhri.med.umich.edu/MARRAY/BC_ASP/brainarray.htm by clicking the 'Query ProbeMatchDB' link.

Contact: mengf@umich.edu

Microarray-based expression assay provides an efficient way to determine the expression level of thousands or even tens of thousands of genes in parallel. The fast growth in microarray usage creates the need for comparing microarray data across different array platforms and different species. While there are many issues involved in cross-platform and cross-species comparisons, the foremost problem is to identify probes that represent the same gene across different platforms and species. Furthermore, due to the relatively high variability inherent in microarray data, it is usually necessary to independently confirm the results with non-array methods such as Northern blot, *in situ* hybridization, RT-PCR, Taqman®, etc. Beyond such gene discovery studies, it is often desirable to verify the function of the identified genes through experiments in other model species, such as creating gene knockouts in mice. All such follow up studies require the identification of equivalent gene probes across platform or across species.

Since different platforms may use different sequences to represent the same gene, finding equivalent probes across platforms or species is usually very tedious and time-consuming for a list of genes, since there is no public domain batch probe match database available. In order to provide a user-friendly tool for finding equivalent probes across different platforms and species, we constructed the ProbeMatchDB by integrating NCBI UniGene (Schuler, 1997), HomoloGene (Zhang *et al.*, 2000) and UniSTS (NCBI, 2001) databases as well as the probe/clone information provided by Affymetrix (2000); Research Genetics (2000) and Operon (2001).

Cross platform probe match

Currently, there are two major platforms for expression microarray, the oligonucleotide-based Affymetrix GeneChip®s and the EST-based spotted arrays. Although both platforms use EST/cDNA sequences to represent unique genes, sequences selected by Affymetrix for GeneChip®s are usually different from those included in the popular sequence-verified Research Genetics EST clone sets used for spotted arrays. Since GeneChip®s are relatively easy to use and usually offer higher density than spotted arrays, they are often used in first round microarray experiments. Interesting genes identified by GeneChip®s are then used for additional studies with other methods. At this stage, it is usually advantageous to find EST clones that represent the differentially expressed genes detected on GeneChip®s, since sequence-verified clone sets offer a convenient library for selecting specific probes for many different applications, including the making of custom arrays.

Such a probe matching problem is very simple conceptually. For example, GeneChip®s utilize gene-specific nucleotide sequences for their oligonucleotide probe design. Although the exact sequence information for those oligos are not available, the accession numbers used for oligo design are readily available from Affymetrix (2000), which can be used to identify the UniGene clusters represented by oligonucleotides on GeneChip®s by searching the UniGene database for a particular species. Since UniGene databases already contain detailed cluster

member information (i.e. EST/cDNA sequences included in a cluster), one can then use a UniGene cluster ID to find all the accession number(s) belonging to that cluster. Such a cluster-specific accession number list can then be compared with the accession number list for the Research Genetics sequence-verified clone set from the same species. Whenever there is a unique accession number match, it means there is an EST clone in the Research Genetics clone set that represents the same cluster as the corresponding oligonucleotide probes on a GeneChip®.

Although this process is conceptually simple, it is very tedious to implement, particularly for a large probe list. Our approach is to establish an integrated Oracle database system called ProbeMatchDB that provide a one-step solution for this problem. ProbeMatchDB stores the Affymetrix probe accession number list, the clone information provided by Research Genetics as well as the periodically updated UniGene clustering information generated by NCBI. We also built a web-based interface that allows batch accession number queries, which is essential for microarray experiments.

Most recently, we incorporated identity information for the 70mer oligonucleotide probe set generated by Operon for spotted microarrays (Operon, 2001) as well as the UniSTS database. The ability to use Clone ID to search for probes is also added. As a result, ProbeMatchDB enables cross platform searches for equivalent probes among Affymetrix GeneChip®s, EST/cDNA arrays, STS arrays, and Operon oligonucleotide arrays using accession number, clone ID or STS names.

Cross species probe match

There are many situations where results obtained in one species need to be verified or studied in greater detail in another species. For example, interesting genes identified by microarray experiments in human disease tissue samples are usually only the starting point of a comprehensive study. These genes are commonly studied in rat or mouse disease models, as rat and mouse are usually more amenable to a variety of physiological, pharmacological or genetic manipulations. Frequently, the expression levels of these genes require monitoring by equivalent rat or mouse nucleic acid probes. Similarly, there are also situations where candidate genes revealed by rat or mouse models need to be investigated in human. These experiments demand the ability to readily identify equivalent probes across different species as well as across different array platforms.

In order to implement the cross species probe search function, we incorporated data from the HomoloGene database into ProbeMatchDB. The HomoloGene database is generated by NCBI for cross-referencing similar genes across several species. It is a very useful database for finding homologous genes in several species using UniGene cluster ID, LocusLink Locus ID, gene name and/or keyword in queries. Nonetheless, although accession numbers representing sequences that are used for gene similarity calculations by HomoloGene may be used in searches, the HomoloGene database itself does not contain sequence cluster member information. Consequently, the accession numbers for EST or cDNA sequences that are not included in gene homology calculations cannot be used to query the HomoloGene database. This is a serious problem for microarray applications since most of the sequences used in different microarray platforms, particularly EST clone and STS sequences, are not used in the gene similarity calculations by the HomoloGene database.

To solve this problem, the HomoloGene data set is internally linked to UniGene databases for human, mouse and rat in ProbeMatchDB. As a result, any accession number included in UniGene databases can be used for cross-species searches. Furthermore, since probe information from various platforms is already integrated in our database, the ProbeMatchDB allows the cross-species searches for every possible combination of platforms, such as Affymetrix-human versus EST-rat or EST-rat versus EST mouse, etc.

In summary, ProbeMatchDB provides a one-step solution for cross-species and cross-platform probe matching. It should be helpful for the design and validation of microarray experiments. The interface for ProbeMatchDB is intuitive although our website also provides more detailed information about sample input/output screens as well as dataflow in ProbeMatchDB.

ACKNOWLEDGEMENTS

We want to thank Dr Huda Akil for her critical comments on this manuscript. This work was supported by the University of Michigan Microarray Network funding, the Nancy Pritzker Depression Research Network and NIMH program project grant L99 MH60398 to S.J.W. The Department of Psychiatry pilot study grant to F.M. and the National Institute on Drug Abuse R21 DA13754-01 to F.M.

REFERENCES

Affymetrix (2000) The GeneChip expression analysis sequence information database. http://www.affymetrix.com.

NCBI (2001) UniSTS. http://www.ncbi.nlm.nih.gov/genome/sts/.

Operon (2001) Human genome oligo set. http://www.operon.com/arrays/arraysets.php.

Research Genetics (2000) Research genetics sequence verified EST clones. http://www.resgen.com/include/menus/cdna_menu.php3.

Schuler,G.D. (1997) Pieces of the puzzle: expressed sequence tags and the catalog of human genes. *J. Mol. Med.*, **75**, 694–698.

Zhang,Z., Schwartz,S., Wagner,L. and Miller,W. (2000) A greedy algorithm for aligning DNA sequences. *J. Comput. Biol.*, **7**, 203–214.