DOI: 10.1093/nar/gkh452

ProbeLynx: a tool for updating the association of microarray probes to genes

Fiona M. Roche, Karsten Hokamp, Michael Acab, Lorne A. Babiuk[1], Robert E. W. Hancock[2] and Fiona S. L. Brinkman*

Department of Molecular Biology and Biochemistry, Simon Fraser University, Burnaby, BC, V5A 1S6, Canada, [1]VIDO, Saskatoon, SK, Canada and [2]Department of Microbiology and Immunology, University of British Columbia, Vancouver, BC, Canada

Received February 15, 2004; Revised and Accepted April 21, 2004

ABSTRACT

As genome sequence data and gene prediction improve, probes developed for a given microarray experiment should be continuously re-evaluated for their specificity for given genes. ProbeLynx (www.pathogenomics.ca/probelynx) is a new web servicewhichusescurrentgenomicsequenceinformation to re-examine microarray probe specificity and provide annotation updates relevant to determining which gene(s) and transcript(s) are associated with a given probe. Probe sequences (either oligonucleotide- or cDNA-based) are uploaded in FASTA format and the results returned as a tab-delimited flat file for insertion into a spreadsheet application or database management system for further analysis. ProbeLynx has been initially developed to focus on arrays derived from human, mouse, chicken and bovine genomes, but may be expanded to handle other genomic datasets. ProbeLynx offers microarray users the important ability to continuously assess the potential of a probe to cross-hybridize to paralogous genes and the suitability of a given probe to investigate a transcript of interest. By also including the latest gene function annotation information in the output, ProbeLynx provides the critical first step in updating microarray data annotation.

INTRODUCTION

Microarray technology has become increasingly more common in biological research, shifting the focus from the study of single genes to studying entire transcriptomes in one experiment (1). In the case of non-commercial arrays, a microarray slide is typically spotted with thousands of short stretches of DNA sequences, each specific to a single gene in the genome of interest. In this paper, we will refer to the DNA spotted on the array as the 'probe' and to the labeled cDNA, which hybridizes to the array, as the 'target'. The specificity of a probe sequence to a single gene is important to permit extrapolation of meaningful biological information from signal intensity measurements. Non-specific probe sequences lead to cross-reactivity with non-target cDNAs and result in a mixture of signals for a given probe spot that are difficult to interpret. Many tools currently exist to ensure optimal probe design (2–4) and reduce the potential for cross-hybridization. However, most arrays are built from model systems in which the genome is still incomplete, with many being derived from expressed sequence tags (EST) sequence data until the full genome draft is available. Confidence in probe specificity is therefore limited as new sequence updates may make users aware of additional probe targets that can complicate data interpretation. In addition, eukaryotic gene prediction is widely known to be one of the biggest challenges in bioinformatics today (5), and such predictions are changing as gene prediction algorithms are improved. Therefore, a given probe developed to be specific to a particular gene in a genome may be found later to have a different level of specificity or utility than was originally envisaged.

To date there is no publicly available web service that allows microarray users of cDNA spotted arrays to upload their probe sequences and receive an automatic update of gene annotations. The GeneAnnot system (6) provides pre-processed up-to-date annotations on Affymetrix probe sets by comparing probe sequences against mRNA sequence databases. Other annotation tools such as ASAP (7) and SOURCE (8) facilitate the annotation of an array only if gene information (such as external identifiers) is available. We have therefore developed a novel web-based tool called ProbeLynx which compares probe sequences directly against current genomic sequence and gene prediction data to assess the issue of cross-hybridization and provide updated annotations on identified targets. An interface has been designed

*To whom correspondence should be addressed. Tel: +1 604 291 5646; Fax: +1 604 291 5583; Email: brinkman@sfu.ca which allows uploading of a list of probe sequences to query against several eukaryotic genomes found in either the Ensembl or the TIGR database. User-defined annotations are returned in tab-delimited format and form the basis for further linkage to other biological data resources.

METHODS

Sequence similarity searching

Direct sequence comparison of array probe sequences against the latest version of a particular genome is carried out using the BLAT algorithm, which offers improved speed over BLAST for larger genomes and sufficient sensitivity for detection of cross-hybridization targets (9). The program is run using the -oneOff parameter set to 1 to allow one base pair mismatch per tile during sequence alignment. For oligonucleotide probes, this allows BLAT to reliably detect cross-hybridization candidates which cause >15% of the overall target signal based on cutoffs from previous specificity studies (10). Since microarrays frequently suffer from noisy signal intensities, the aim of this tool was to focus on detecting major targets that would significantly contribute to the overall signal intensity measurement of a probe. BLAT was chosen as best suited for this purpose. To efficiently compute the required sequence similarity searches, a wrapper script utilizes a 40-node Linux cluster to parallelize BLAT runs.

Cross-hybridization analysis

As a web service, ProbeLynx currently accepts upload of oligonucleotide probe sequences (50–80 bp) from four genome types (human, mouse, chicken and bovine). For cDNA based probes (>100 bp), web-based querying is only available against mRNA databases, such as for bovine and chicken. Querying of cDNA probes against large genome databases such as human and mouse is restricted to contacting the corresponding author directly by email to control lengthy processing times.

Following BLAT analysis, the detected target sequences are passed on to a filtering step to detect cross-hybridization candidates. Sequence similarity cutoffs are based on previous cross-hybridization studies for oligonucleotide or cDNA probes that were performed experimentally (10–13). For oligonucleotide probes, sequences showing >80% sequence similarity over >95% of the oligonucleotide probe sequence or which contain >25 bp of identical match to the probe sequence will contribute to the hybridization signal and therefore are assigned as a probe target. For cDNA probes, sequences showing >70% sequence similarity over at least 200 bp of the cDNA probe sequence or which contain >100 bp of identical match to the probe sequence will also be reported as a target for that probe. In addition to detailed hit-quality data, flags are added to probe annotations to distinguish between different levels of cross-reactive targets, thereby allowing users to filter out which probes they consider to be problematic.

Integration of annotation

Initially developed for a large microarray project (Genome Canada Pathogenomics Project; www.pathogenomics.ca), ProbeLynx has focused on arrays derived from human, mouse, chicken and bovine genomes but could be expanded to incorporate other species. Once probe targets are detected by sequence similarity analysis and cross-hybridization filtering, users can select from a series of annotation types to be linked to their probe set. Human and mouse annotations are retrieved from Ensembl (http://www.ensembl.org), which offers a comprehensive source of up-to-date functional annotations and gene prediction data. Since whole genome data is unavailable for bovine and chicken, annotations are sourced from the TIGR gene indices database (http://www.tigr.org/tdb/tgi/), which uses clustered EST data to predict and annotate tentative gene transcripts.

Ensembl annotations offered to the user for each probe include (i) location of the target, including whether a target lies within the exon of a predicted coding region, or within an intron or other intergenic region, (ii) Ensembl transcript and gene identifiers, provided that the probe lies within a coding region, and (iii) gene functional information such as functional descriptions, HUGO names, gene ontologies, protein domain and family information, external database identifiers and chromosomal locations. Knowledge of a target's mapping location within a gene structure (i.e. intron, exon, intergenic) can help users filter out non-relevant cross-hybridization candidates from their results. Furthermore, a broad range of functional annotations can help users better predict a target gene's biological role where no known function has yet been assigned. The gene functional information provided thus forms the basis for annotation of an array and can be linked via external identifiers to other biological databases with known functional information.

For arrays derived from bovine and chicken genomes, for which only EST sequence information is available, ProbeLynx exploits the TIGR gene index database and maps probes to Tentative Consensus (TC) sequences (14). TIGR-generated annotations such as functional descriptions and gene ontology information are available for linkage to probe targets. Access to the Eukaryotic Gene Orthologs (EGO) database allows interspecies information to be used to link putative orthologs to uncharacterized TC sequences for inference of gene function. Since both bovine and chicken genomes are incomplete and the EGO database was compiled using the reciprocal best BLAST hit approach, many putative orthologs inferred from this analysis may in fact be paralogs, as the true ortholog may not yet be sequenced. Taking this into account, all ortholog annotations derived from EGO are highlighted in ProbeLynx as being putative. A new tool designed to scan through pairs of sequences tentatively described as being orthologous and flag those pairs which are likely paralogous will be integrated into the next version of ProbeLynx to provide higher confidence in ortholog predictions.

Probe-to-gene annotations are available as tab-delimited files and can be integrated into microarray analysis tools for downstream mining, as was accomplished in-house with ArrayPipe (www.pathogenomics.ca/arraypipe).

Program implementation

ProbeLynx has been written in Perl and makes use of the MySQL DBMS for querying and storing results. It can be run as a standalone package or it can be used from a web server. For the web server, the user must specify an email address. Results are posted on a website and the URL is mailed to the user. Contact the authors to obtain the freely available standalone package, which currently runs on Linux Redhat 9.0.

RESULTS

ProbeLynx was used to update probe annotations from the Qiagen human array-ready oligo sets version 1.0 and version 2.0 to highlight the importance of performing such analysis (see ProbeLynx website for downloads). For each oligo set, Qiagen provided one gene annotation per probe. In contrast, using the Ensembl human build v34, ProbeLynx found that 1409 and 842 of the probes from each array, respectively, were found to map to multiple gene targets, some mapping to exon regions in >30 different genes, such as probe id H004395_01 from the Qiagen version 1.0 array. Moreover, ProbeLynx offers specific transcript information linked to each gene target. As expected, analysis of array expression data demonstrated that probes mapping to an increased number of transcripts showed a correspondingly increased signal intensity measurement compared with probes mapping to a single transcript. Flagging of non-specific probes is vital for accurate interpretation of microarray data. This allows users to critically assess these probes and determine whether they should be removed from downstream analysis.

Despite being originally derived from human mRNA sequences, a large proportion of probes (~30%) from both oligo sets were predicted to lie within intergenic regions, often mapping close to the proximal regions of genes. This is not surprising, since most microarray probe sets focus on the 3′ end of genes, in part because of the efficiency of reverse transcription to increase signal intensities and also because sequence divergence is typically greater in these regions. Since human EST and cDNA sequence data are renowned for sequencing errors, Ensembl requires at least three forms of transcript evidence before extending the coordinates of a gene. Methods used for gene prediction in eukaryotes are still quite poor (15), and coordinates for genes represent a continuously moving target as additional experimental evidence arises. For this reason, ProbeLynx flags probes which map to intergenic regions but offers annotation for those derived from mRNA sequence which map to within 5 kb of the proximal region of a single transcript. This increases the probe annotation coverage while still conservatively flagging probes as putatively intergenic based on currently available gene prediction data. To take advantage of this additional feature, accession numbers such as those available in the Genbank, RefSeq and EMBL databases are required for upload along with probe sequence information.

CONCLUSIONS

ProbeLynx is the first freely available tool that specifically offers microarray users of spotted arrays the ability to continuously assess the specificity and annotation of array probes based on the latest releases of genomic sequence data and gene predictions. ProbeLynx utilizes probe sequence information to first identify the targets before assigning annotation information to them, ensuring a clear and accurate link between the actual probe sequence and any gene annotation information. Cross-hybridization is an important issue that should be continuously addressed when using arrays built from incomplete genomic sequence data. ProbeLynx gives users increased confidence that when an increased microarray signal intensity is observed, they have the best data currently available for assessment of the probable transcript(s) involved.

SUPPLEMENTARY MATERIAL

Supplementary Material is available at NAR Online.

ACKNOWLEDGEMENTS

K.H. and F.S.L.B. hold postdoctoral and scholar career awards, respectively, from the Michael Smith Foundation for Health Research. Funding was provided by the Functional Pathogenomics of Mucosal Immunity (FPMI) Project of Genome Canada/Genome Prairie/Genome BC, with support from Inimex Pharmaceuticals Inc. (BC, Canada). R.E.W.H. is supported by a Canada Research Chair.

REFERENCES

1. Gerhold,D., Rushmore,T. and Caskey,C.T. (1999) DNA chips: promising toys have become powerful tools. *Trends Biochem Sci.*, **24**, 168–173.
2. Chang,P.C. and Peck,K. (2003) Design and assessment of a fast algorithm for identifying specific probes for human and mouse genes. *Bioinformatics*, **19**, 1311–1317.
3. Rouillard,J.M., Zuker,M. and Gulari,E. (2003) OligoArray 2.0: design of oligonucleotide probes for DNA microarrays using a thermodynamic approach. *Nucleic Acids Res.*, **31**, 3057–3062.
4. Reymond,N., Charles,H., Duret,L., Calevro,F., Beslon,G. and Fayard, J.M. (2004) ROSO: optimizing oligonucleotide probes for microarrays. *Bioinformatics*, **20**, 271–273.
5. Burge,C., Birney,E. and Fickett,J. (2002) Top 10 future challenges for bioinformatics. *Genome Technol.*, **17**, 1–3.
6. Chalifa-Caspi,V., Shmueli,O., Benjamin-Rodrig,H., Rosen,N., Shmoish,M.,Yanai,I., Ophir,R., Kats,P., Safran,M. and Lancet,D. (2003) GeneAnnot: interfacing GeneCards with high-throughput gene expression compendia. *Brief Bioinform.*, **4**, 349–360.
7. Kossenkov,A., Manion,F.J., Korotkov,E., Moloshok,T.D. and Ochs,M.F. (2003) ASAP: automated sequence annotation pipeline for web-based updating of sequence information with a local dynamic database. *Bioinformatics*, **19**, 675–676.
8. Diehn,M., Sherlock,G., Binkley,G., Jin,H., Matese,J.C., Hernandez-Boussard,T., Rees,C.A., Cherry,J.M., Botstein,D., Brown,P.O. and Alizadeh,A.A. (2003) SOURCE: a unified genomic resource of functional annotations, ontologies, and gene expression data. *Nucleic Acids Res.*, **31**, 219–223.
9. Kent,W.J. (2002) BLAT—the BLAST-like alignment tool. *Genome Res.*, **12**, 656–664.
10. Kane,M.D., Jatkoe,T.A., Stumpf,C.R., Lu,J., Thomas,J.D. and Madore,S.J. (2000) Assessment of the sensitivity and specificity of oligonucleotide (50mer) microarrays. *Nucleic Acids Res.*, **28**, 4552–4557.
11. Evertsz,E.M., Au-Young,J., Ruvolo,M.V., Lim,A.C. and Reynolds,M.A. (2001) Hybridization cross-reactivity within homologous gene families on glass cDNA microarrays. *Biotechniques*, **31**, 1182–1192.
12. Hughes,T.R., Mao,M., Jones,A.R., Burchard,J., Marton,M.J., Shannon,K.W., Lefkowitz,S.M., Ziman,M., Schelter,J.M., Meyer,M.R., Kobayashi,S., Davis,C., Dai,H., He,Y.D., Stephaniants,S.B., Cavet,G., Walker,W.L., West,A., Coffey,E., Shoemaker,D.D., Stoughton,R.,

Blanchard,A.P., Friend,S.H. and Linsley,P.S. (2001) Expression profiling using microarrays fabricated by an ink-jet oligonucleotide synthesizer. *Nature Biotechnol*., **19**, 342–347.

13. Lyne,R., Burns,G., Mata,J., Penkett,C.J., Rustici,G., Chen,D., Langford,C., Vetrie,D. and Bahler,J. (2003) Whole-genome microarrays of fission yeast: characteristics, accuracy, reproducibility, and processing of array data. *BMC Genomics*, **4**, 27–41.
14. Quackenbush,J., Cho,J., Lee,D., Liang,F., Holt,I., Karamycheva,S., Parvizi,B., Pertea,G., Sultana,R. and White,J. (2001) The TIGR Gene Indices: analysis of gene transcript sequences in highly sampled eukaryotic species. *Nucleic Acids Res*., **29**, 159–164.
15. Mathe,C., Sagot,M.F., Schiex,T. and Rouze,P. (2002) Current methods of gene prediction, their strengths and weaknesses. *Nucleic Acids Res*., **30**, 4103–4117.