D880–D884 *Nucleic Acids Research, 2009, Vol. 37, Database issue*
doi:10.1093/nar/gkn878

CleanEx: new data extraction and merging tools based on MeSH term annotation

Viviane Praz[1,*] and Philipp Bucher[2]

[1]ISREC, Swiss Institute of Bioinformatics, Boveresses 155, Epalinges, VD 1066 and [2]ISREC/EPFL, Swiss Institute of Bioinformatics, Epalinges, VD, Switzerland

Received September 16, 2008; Revised October 16, 2008; Accepted October 17, 2008

ABSTRACT

The CleanEx expression database (http://www.cleanex.isb-sib.ch) provides access to public gene expression data via unique gene names as well as via experiments biomedical characteristics. To reach this, a dual annotation of both sequences and experiments has been generated. First, the system links official gene symbols to any kind of sequences used for gene expression measurements (cDNA, Affymetrix, oligonucleotide arrays, SAGE or MPSS tags, Expressed Sequence Tags or other mRNA sequences, etc.). For the biomedical annotation, we re-annotate each experiment from the CleanEx database with the MeSH (Medical Subject Headings) terms, primarily used by NLM (National Library of Medicine) for indexing articles for the MEDLINE/PubMED database. This annotation allows a fast and easy retrieval of expression data with common biological or medical features. The numerical data can then be exported as matrix-like tab-delimited text files. Data can be extracted from either one dataset or from heterogeneous datasets.

INTRODUCTION

Given the increasing size of public databases storing high-throughput gene expression experiments, namely GEO (1) and ArrayExpress (2), the problem of specific expression data retrieval is becoming a serious challenge, mainly related to two major annotation issues: a precise and up-to-date link between the numerical results of the expression measurements and the name of the corresponding transcripts, the tag-to-gene mapping, and second, a correct and universal biomedical annotation of the samples and experiments, or experiments' annotation. Due to the rapid evolution of genomes' annotations, the tag-to-gene mapping needs to be refreshed continuously to maintain its usefulness. On the other hand, the experiments' annotation, often based on a free-text variable description, needs to be standardized once via a common biological ontology.

To meet these annotation standards for public expression data, we developed the CleanEx (3) tags-to-gene re-annotation and biomedical integration system, which allows to keep an up-to-date gene annotation for expression data, and which provides a standard biomedical annotation of Human and Mouse public expression datasets, based on the Medical Subjects Headings (MeSH http://www.nlm.nih.gov/mesh/) terms. The CleanEx database now contains nearly one thousand of Human and Mouse datasets, most of them were uploaded from the GEO database and biologically re-annotated by CleanEx. The upload rhythm depends on the post-processing needed for each dataset. During last year, a mean of 15 datasets per month have been integrated in the CleanEx re-annotated data. The first choice for data upload goes to cancer data, in particular breast and lung cancer data, as these are the main research topics at the Swiss Institute for Experimental Research. It is estimated that at least 60% of all publicly available data meeting these criteria are currently available via CleanEx.

Regarding tags-to-gene mapping, CleanEx newly provides a mapping for SAGE (4) or MPSS (5) tags. The CleanEx Affymetrix (http://www.affymetrix.com) mapping files are now also accessible by an external Affymetrix probes viewer called Splicy (http://host10.bioinfo3.ifom-ieo-campus.it/splicy/) (6).

Major improvement in the web interfaces has lead to an efficient mechanism for both data retrieval and analysis. The expression measurements can be searched via gene names, as well as via the biomedical MeSH terms. Specific data can then be analyzed on the CleanEx server or downloaded as a matrix-formatted text file. The CleanEx data analysis system is now coupled with the Signal Search Analysis (SSA) server (http://www.isrec.isb-sib.ch/ssa/) (7), and a promoter sequence retrieval system.

*To whom correspondence should be addressed. Tel/Fax: 4121 692 5956; Email: viviane.praz@unil.ch

GEO, ArrayExpress AND CleanEx

Amongst the three official gene expression repositories selected by the MIAME consortium, two are heavily populated and offer user-friendly data access interfaces, namely ArrayExpress (http://www.ebi.ac.uk/microarray-as/ae/) and GEO (http://www.ncbi.nlm.nih.gov/geo/). ArrayExpress is the second in terms of data quantity, while GEO has become the major repository and contains now nearly 10 000 experiments, or series, generated with very heterogeneous experimental procedures. Though the first aim of these two databases was to serve as public expression data archives and distribution centers, they both evolved considerably beyond this scope and offer a number of tools to query an integrated database of expression profiles.

GEO provides, via the NCBI Entrez (http://www.ncbi.nlm.nih.gov/) system or directly via dataset identifiers, some interesting data retrieval tools, going from the single-gene expression retrieval with expression graphs to complex queries allowing data clustering. However, clustering is possible only for a subset of annotated GEO series (GSE), called datasets (GDS), which have been manually annotated and classified according to specific experimental criteria. The data retrieval and analysis is limited to a single series/dataset at a time.

ArrayExpress offers a search system which retrieves all the annotated experiments for one gene at once. The ArrayExpress team has also set up its own ontology (Ontology—EFO: http://www.ebi.ac.uk/ontology-lookup/browse.do?ontName=EFO) (8), based on the MGED (http://www.mged.org/) Ontology Work Group (OWG http://mged.sourceforge.net/ontologies/index.php) proposals, as a trial to classify the data via a list of standard terms. This ontology contains biomedical terms, as well as terms associated with gene expression profiling protocols. The expression data query interface, however, is not yet linked to the tree-like structure of this ontology.

Despite the emergence of biomedical ontologies (9), the expression experiments' annotation is still quite inconsistent, often using different synonyms for the same concept. This heterogeneity problem has already been faced for the numerical expression values. While the numerical values, though not standardized across different platforms, are now stored in well-defined formats that make retrieval and downstream analysis easy the usual format for experimental annotation is stil based on a free text description. The annotation standardization is mainly done after data integration, by the repository maintainers. An example is the previously mentioned 'datasets' (GDS) of GEO, which are pools of manually annotated experiments.

In the following part we will describe the methods developed in the CleanEx database to handle the two annotation issues mentioned above, namely the gene mapping and the experiments' annotation standardization.

CleanEx TAG-TO-GENE MAPPING

The CleanEx 'tag-to-gene' mapping, which provides all expression measurements for one gene under one single name (usually the official gene symbol), has already been described in the former paper (3). In brief, all the sequences which appear in a CleanEx dataset (Affymetrix probe set sequences, SAGE or MPSS tags, RefSeq entries (http://www.ncbi.nlm.nih.gov/RefSeq/), Expressed Sequence Tags (ESTs) or oligonucleotides) called 'targets' are remapped to known RNA sequences (RefSeq, ESTs, mRNAs, etc.), or 'features'. In the second step, these sequences, through their accession numbers, are mapped onto gene names via the Unigene (http://www.ncbi.nlm.nih.gov/sites/entrez?db=unigene) database. Note that the tag-to-gene mapping may initially associate one target with multiple genes. For instance, individual sequences of an Affymetrix probeset may match different mRNAs. The way such ambiguities are handled has been described in detail in the previous paper (3). Note further that the accuracy of the annotation depends both on the genome's annotation accuracy, and on the depth of the arrays' or platforms' annotation. If the authors of the experiments give access to the raw sequences for their platform (Affymetrix, SAGE and MPSS), we regularly re-map these sequences onto the continuously growing RNA sequence repositories to increase the mapping precision. For other experiments, the platform annotation often refers to accession numbers of published sequences which matched the spotted sequence at the moment the experiment was carried out. Then, only the mapping of the sequence accession numbers onto Unigene clusters is possible, and there is a risk of information loss with time as sequence accession numbers may disappear from Unigene. The regular remapping of accession numbers on the Unigene database can anyway improve the accuracy of the gene annotation, as sequences considered as 'unknown' when entered in the databases later become members of Unigene clusters of known genes. To follow the continuous evolution of the organisms' specific gene catalogs, this mapping procedure is done on a monthly basis.

In addition to the mapping of dual channel and Affymetrix platforms, CleanEx now also provides an accurate mapping for SAGE, LONGSAGE, MPSS and LONGMPSS tags via a procedure which resembles the one used for Affymetrix, except that we reduce the sequence search space by using only RNA sequences which have already been associated with SAGE and MPSS tags. This filter is based on an in-house built Human and Mouse transcriptome called the Trome database (http://www.isrec.isb-sib.ch/tromer/) (10). Using this filtered sequences database, the procedure maps tags to sequences via the 'tagger' program (11), and then links sequence identifiers to gene names via Unigene clusters. Each SAGE or MPSS tag is flagged with a quality tag which indicates its specificity (Figure 1). This quality tag is also provided for other targets. For Affymetrix probesets, the quality tag is based on the mapping of the individual probes sequences. The CleanEx mapping can be retrieved either on a single-gene basis, or via a batch web query which can handle multiple identifier types, such as Unigene, RefSeq, GenBank/EMBL (http://www.ncbi.nlm.nih.gov/Genbank/) gene symbols or Entrez GeneID (http://www.ncbi.nlm.nih.gov/sites/entrez?db=gene). The Affymetrix probes mapping on ESTs,

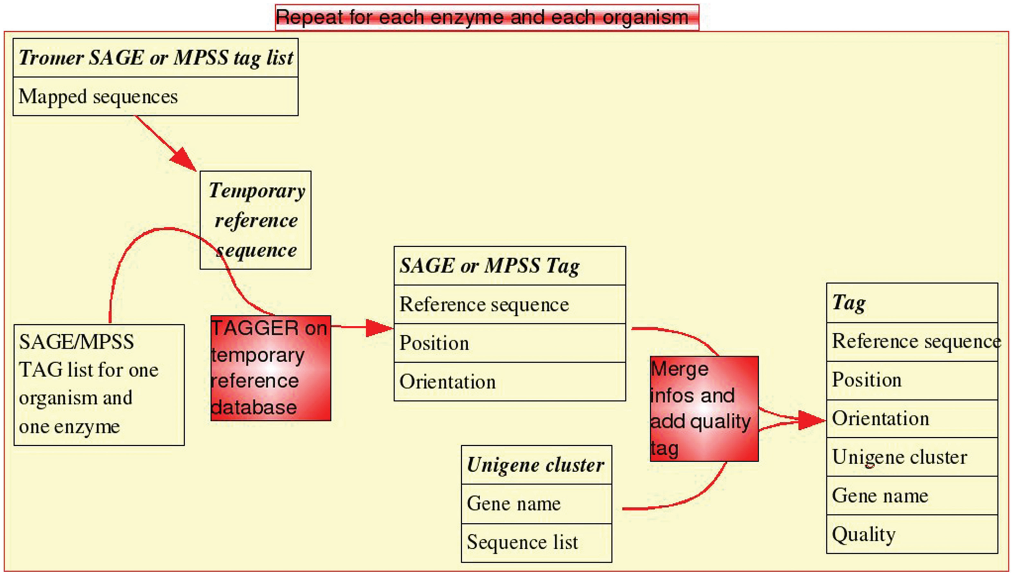

Figure 1. Mapping process for SAGE and MPSS tags. Sequences belonging to Trome clusters are stored in a tempory database The SAGE or MPSS tags are then mapped onto these sequences via the "tager" program. The corresponding gene name for the stored sequence identifires is extracted from Unigene. The quality is given according to the following criteria, for each SAGE or MPSS tag : if all the targeted sequences belong to one or two Unigene clusters, the quality is set as "High", If all the targeted sequences belong to more than two but less than five clusters, the quality is set as "Medium". For more than clusters, is "low". Otherwise, the quality is considered as "Unknown". The process is done for human and mouse.

mRNAs from GenBank, RefSeq mRNAs and HTCs for each chip and each organism is provided as separated text files on the ftp server (ftp://ftp.epd.unil.ch/pub/databases/CleanEx/Affy_mapping).

CleanEx EXPRESSION EXPERIMENT ANNOTATION

In order to get an accurate experiment annotation from the usual free-text descriptions provided in the data repositories, we have re-annotated all the CleanEx datasets with the terms belonging to the MeSH controlled vocabulary dictionary. MeSH is the National Library of Medicine's controlled vocabulary thesaurus, which is constantly improved and developed. It consists of sets of terms naming descriptors in a hierarchical structure that allows searching at various levels of specificity. The MeSH thesaurus is used to index articles from biomedical journals for the MEDLINE/PubMED database.

The hierarchical structure of this catalog, as well as the fact that it is already used to retrieve biomedical data, makes it a good start to annotate gene expression experiments. We generated a semi-automatic procedure which first associates the corresponding MeSH terms to all its stored synonyms, then searches these terms in the free-text expression experiment annotation and stores the official terms list for each experiment. In a second phase, the accuracy of the association between the MeSH terms and the experiments is manually verified. The automatically selected MeSH terms are either accepted or rejected, according to their relevance for the given experiment. This two-step procedure ensures that:

- Only MeSH terms are kept in the experiments' annotations, and all the recorded synonyms for each MeSH term are checked.
- The terms retrieved are consistent with the experiment original free-text description.

CleanEx WEB-BASED TOOLS

To ease data retrieval, the final annotation is stored in an indexed file for both the whole datasets and the individual experiments. To keep the MeSH tree in a reasonable range, only the part of the tree which is related to CleanEx datasets or experiments is reproduced on the CleanEx query web server. The biomedical annotations of the datasets can be searched in different ways via the proposed web-based tools. All the tools are listed on the database main page (http://www.cleanex.isb-sib.ch). A list of output examples is also available from the CleanEx server (http://www.cleanex.isb-sib.ch/examples.html).

First, the MeSH-oriented retrieval tool is linked to the gene-oriented search tools. This allows users to retrieve specific expression data together with a specific gene description with very high annotation accuracy. As an example, one can retrieve the list of all the CleanEx datasets related to Breast Cancer AND associated with the fibronectin one (FN1) gene via this dual gene/experiment query form.

Second, the MeSH-based annotation can be used to retrieve and extract numerical expression data from datasets with specific biomedical criteria (see examples 7 and 8 at http://www.cleanex.isb-sib.ch/examples.html). All the checked terms are then searched in the indexed file, and the corresponding experiments are extracted from the CleanEx datasets. The search along the MeSH tree is fast and accurate and can be done either at the datasets level or at the experiments level. Data can thus be extracted either from one single dataset or from multiple, possibly platform-heterogeneous datasets. In the former case, the user can select the numerical field to extract from the original data. For example, for a dual channel experiment, one could extract only the 'red' channel, or the ratio, or directly the normalized data, if these fields are provided by the authors of the experiment. For heterogeneous data retrieval, a rank-based scaling is applied on the numerical data, where all the numerical values for all individual samples or arrays are scaled between 0 and 1000, in order to make the comparison between data coming from different datasets possible. The procedure extracts these numerical data for all the common genes across the selected experiments.

All the search tools using the MeSH based datasets and experiments annotation can be queried by either using the javascript 'walk down the MeSH tree' page to select the appropriate terms, or by directly listing the specific MeSH terms to search in the database. In both cases, the terms to search can be linked by different operators (AND, OR or BUT NOT), allowing the generation of complex queries for the indexed annotation files. An example of such a query could be: 'Colon OR Pancreas BUT NOT Cell Line AND Neoplasms BUT NOT Neoplasm Metastasis'.

After data selection and extraction, the numerical values are provided as a matrix file in which the columns represent the experiments, or samples, and the lines represent either targets, for single dataset extraction, or genes, for heterogeneous data extraction. The targets' annotation and the original experiments' description are provided as additional files. These files can then be used as inputs for specific expression data analysis programs, such as R (http://www.r-project.org/) and Bioconductor (http://www.bioconductor.org/), or on-line expression analysis tools, such as EPClust (http://www.bioinf.ebc.ee/EP/EP/EPCLUST/) or the SotArray server (http://www.transcriptome.ens.fr/cgi-bin/gepas/sotarray) (12).

CleanEx also provides a comparison tool, in which you can generate two experiments pools based on different biomedical criteria, and extract the scaled numerical values of the corresponding experiments. Numerical data are provided as two different matrices, one for each pool. The experiments' description, as well as the gene list, is given in two separated files. One can then compare the expression level in these two pools via the CleanEx analysis system. This tool is based on a basic 'mean difference ranking' system, which is described in more details in the CleanEx online documentation (http://www.cleanex.isb-sib.ch/tutorial/CleanEx_tutorial.html#analysis_step). For both pools, the expression mean of each gene is calculated. The gene list is sorted according to this value for the two biomedical conditions. The difference of the gene position in each pool gives the gene expression bias between the two experiments pools. From the given result (a list of genes ordered according to their expression mean ranking difference), one can extract the promoter regions of a selected subset of genes. The sequences provided in FASTA format can then be uploaded to web-based motif discovery tools such as Signal Search Analysis (http://www.isrec.isb-sib.ch/ssa) (6) or 'MEME' (http://meme.nbcr.net) (13). This sequence download system is unique in the sense that the sequence list can be generated from a comparison across heterogeneous expression experiments. As an example, Figure 2 shows the graph obtained by using the SSA server with sequences of genes differentially expressed in cancer tissues. The two genes' lists and their corresponding sequences have been obtained via the CleanEx comparison tool. All the CleanEx online tools are described in details in the CleanEx tutorial (http://www.cleanex.isb-sib.ch/tutorial/). The description of each tool is illustrated with specific examples.

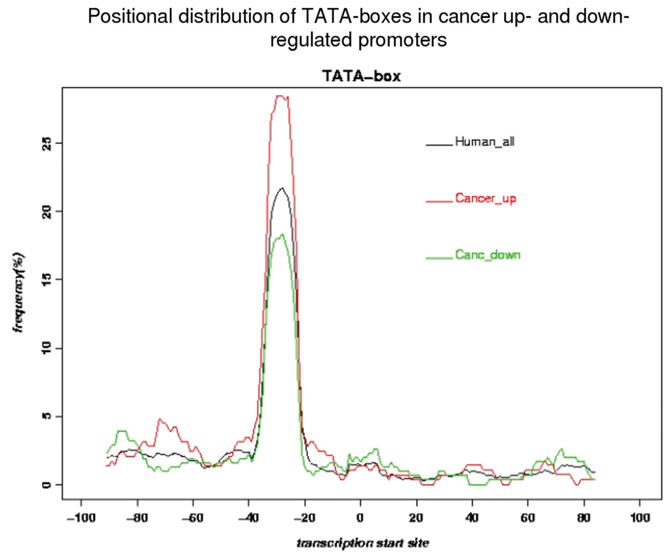

Figure 2. Over-representation of TATA-box-occurrences in human genes which show up-regulation in cancer tissues. Promoter sequences for up- and down-regulated sets of genes havae been extracted via the cleanEX two pools comparision tool and-box freaquency has been analysed via the SSA server with default options.

DISCUSSION

Adding extraction and analysis tools to transcriptomics repositories has greatly facilitated the use of public data by external users. The use of a tree-based biomedical ontology system to retrieve specific data is probably well adapted for these repositories which contain such a vast amount of data, as it reduces the search space while going down the ontology tree. This type of ontologies exists already. As an example, the EFO ArrayExpress ontology is a member of the Open Biomedical Ontologies (OBO) (http://www.obofoundry.org/). All the ontology databases in OBO can be browsed via the Ontology lookup service (OLS), a centralized query interface for ontology and controlled vocabulary lookup. There are now more than forty ontologies integrated in OBO. Though very complete, the

OBO system is also very complex and stores data coming from a wide range of concepts, amongst which only a few are relevant for expression data annotation. For that reason, we decided not to use the complete OBO ontology, but to focus on biomedical standardized terms used in MeSH. This catalog has already proven to be very useful for retrieval of biomedical scientific publications, and adapting it to gene expression experiments was indeed straightforward. The CleanEx database and its collection of MeSH- and gene-oriented query systems can be considered as a 'proof of concept' that using MeSH terms, or any other well structured biomedical ontology, for transcriptomics data retrieval is doable and effective.

FUNDING

Funding for open access charge: The Swiss Government.

Conflict of interest statement. None declared.

REFERENCES

1. Barrett,T., Troup,D.B., Wilhite,S.E., Ledoux,P., Rudnev,D., Evangelista,C., Kim,I.F., Soboleva,A., Tomashevsky,M., Edgar,R. *et al*. (2006) NCBI GEO: mining tens of millions of expression profiles–database and tools update. *Nucleic Acids Res*., **35**, D760–D765.
2. Parkinson *et al*. (2007) ArrayExpress – a public database of microarray experiments and gene expression profiles. *Nucleic Acids Res*., **35**, D747–D750.
3. Praz,V., Jagannathan,V. and Bucher,P. (2005) CleanEx: a database of heterogeneous gene expression data based on a consistent gene nomenclature. *Nucleic Acids Res*., **32**, D542–D547.
4. Velculescu,V.E., Zhang,L., Vogelstein,B. and Kinzler,K.W. (1995) Serial analysis of gene expression. *Science*, **270**, 484–487.
5. Brenner,S., Johnson,M., Bridgham,J., Golda,G., Lloyd,D.H., Johnson,D., Luo,S., McCurdy,S., Foy,M., Ewan,M. *et al*. (2000) Gene expression analysis by massively parallel signature sequencing (MPSS) on microbead arrays. *Nat. Biotechnol*., **18**, 630–634.
6. Rambaldi,D., Felice,B., Prav,V., Bucher,P., Cittaro,D. and Gufanti,A. (2007) Splicy: a web-based tool for the prediction of possible alternative splicing events from Affymetrix probeset data. *BMC Bioinformatics*, **8**, S17.
7. Ambrosini,G., Praz,V., Jagannathan,V. and Bucher,P. (2003) Signal search analysis server. *Nucleic Acids Res*., **31**, 3618–3620.
8. Malone,J., Rayner,T.F., Bradley,X.Z. and Parkinson,H. (2008) Developing an application focused experimental factor ontology: embracing the OBO Community. In Proceedings of ISMB 2008 SIG meeting on Bio-ontologies.
9. Rubin,D.L., Shah,N.H. and Noy,N.F. (2008) Biomedical ontologies: a functional perspective. *Brief. Bioinform*., **9**, 75–90.
10. Sperisen,P., Iseli,C., Pagni,M., Stevenson,B.J., Bucher,P. and Jongeneel,C.V. (2004) trome, trEST and trGEN: databases of predicted protein sequences. *Nucleic Acids Res*., **32**, D509–D511.
11. Iseli,C., Ambrosini,G., Bucher,P. and Jongeneel,C.V. (2007) Indexing strategies for rapid searches of short words in genome sequences. *PLoS One*, **2**, e579.
12. Herrero,J., Al-Shahrour,F., Díaz-Uriarte,R., Mateos,Á., Vaquerizas,J.M., Santoyo,J. and Dopazo,J. (2003) GEPAS, a web-based resource for microarray gene expression data analysis. *Nucleic Acids Res*., **31**, 3461–3467.
13. Bailey,T.L., Williams,N., Misleh,C. and Li,W.W. (2006) MEME: discovering and analyzing DNA and protein sequence motifs. *Nucleic Acids Res*., **34**, W369–W373.