

# Expression Array Annotation Using the BioMediator Biological Data Integration System and the BioConductor Analytic Platform

H. Mei<sup>1</sup>, P. Tarczy-Hornoch, M.D.<sup>1,4</sup>, P. Mork, M.S.<sup>1,3</sup>,

A.J. Rossini, Ph.D.<sup>2</sup>, R. Shaker<sup>4</sup>, L. Donelson<sup>1</sup>

Biomedical & Health Informatics<sup>1</sup>, Biostatistics<sup>2</sup>, Computer Science & Eng.<sup>3</sup>, Pediatrics<sup>4</sup>

University of Washington, Seattle, WA

*This paper presents the implementation of a model for expression array annotation (EAA) using the BioMediator biological data integration system along with BioConductor, an analytic tools platform. The model presented addresses the need for annotation sources identified during BioConductor's development. Annotation provides us with well-curated genomic background knowledge for expression array analysis and interpretation. Annotation requests are constructed and posted to the query interface of the EAA package (the EAA model implemented as a component of BioConductor). The software enumerates all possible annotation paths for queries. These are then transformed to PQL queries and processed by BioMediator. Annotation entities returned from the EAA package answer the annotation request.*

## INTRODUCTION

Expression array technologies report expression levels of thousands of genes in one experiment, and produce large amounts of data. These huge datasets are more valuable when analytic tools are available and annotated information on the datasets can be provided. For example, a supervised learning algorithm<sup>1</sup> combined with background knowledge from Gene Ontology<sup>2</sup> explores large scale expression datasets for gene function prediction<sup>3,4</sup>. In this paper, we introduce the EAA model (and its implementation) to provide a bridge between the BioConductor<sup>5</sup> analytic platform for expression data and the BioMediator<sup>6</sup> data integration system for annotation of a gene-based expression array.

## BACKGROUND

**BioConductor** is a collaborative open source project to develop a modular general framework for the analysis of expression array<sup>5</sup> data. This project has developed many innovative software packages for solving problems in Bioinformatics. All software packages are implemented in the R statistical programming language<sup>7</sup>. The goal of BioConductor is to provide access to a wide range of numerical and visual statistical methods for the analysis of genomic data. In this paper, we introduce an expression array annotation (EAA) package that annotates (BioConductor) expression arrays using results from a biological data integration system (BioMediator).

**Expression Array Data** from laboratories contain raw expression data along with many important properties (metadata) like array characteristics, descriptions of labeling, hybridization and washing conditions. To efficiently utilize the increasing volume of expression data and metadata, public databases allowing submission, storing, sharing and analysis of microarray data are desired. There exist many public and private expression array database repositories including Gene Expression Omnibus (GEO)<sup>8</sup>, ArrayExpress<sup>9</sup> and SMD<sup>10</sup>. For the purpose of this paper, GEO will be considered, though the other two could easily be incorporated into this model. GEO serves as a complementary tertiary resource for the storage and retrieval of public high-throughput gene expression and genomic hybridization data<sup>11</sup>. GEO attempts to cover the broadest spectrum of high-throughput experimental methods by loose requirements and standards for entry<sup>11</sup>, which make possible storing flexible basic information for many kinds of expression array data. However, GEO does not provide extended information on the gene-based expression data (e.g., metadata such as annotation data). To support efficient data analysis and successful interpretation, external well-curated data sources for annotation should be provided. We interfaced GEO to BioMediator as a key data source for expression array data. GEO and other well-curated data sources in BioMediator form the initial basis for EAA.

**BioMediator**<sup>12</sup> is a general-purpose biological information system, which permits integration and analysis of diverse types of biological data to help answer biological questions.

In BioMediator, interoperability of multiple data sources is achieved by constructing a mediated schema designed for the domain of discourse (i.e., the universe of anticipated queries). To incorporate a data source into the data integration system, BioMediator uses semantic mappings to relate the concepts and terms of the data source to the mediated schema; queries are posed using the terms of the mediated schema. The important advantage of this architecture is that the data stays at the source (and hence can be managed and updated there), and at query time, appropriate accesses are made to the data sources to obtain the most up-to-date data. BioMediator currently supports twelve data sources including Entrez<sup>13</sup>,

SwissProt<sup>14</sup>, HUGO<sup>15</sup>, LocusLink<sup>16</sup>, OMIM<sup>17</sup>, IMAGE<sup>18</sup>, GEO<sup>8</sup>, GO<sup>2</sup>, et al.

The current mediated schema includes biologically meaningful entities such as Gene, Protein, Phenotype and Nucleotide Sequence. The interface is PQL<sup>19</sup>, a path-based declarative query language.

**Vision:** BioConductor provides efficient data structures to store expression array data and a rich set of analytics (provided as R packages) for data analysis. We developed the EAA model and implemented the EAA package to bridge BioMediator and BioConductor, so that external knowledge for dynamic expression array annotation flows seamlessly from BioMediator to BioConductor, which can be used directly by existing analytic tools running under BioConductor.

## RELATED WORK

With the broad application of the expression array technique, a multitude of expression array systems now store the high-throughput gene expression data. Most of these systems (e.g., SMD<sup>10</sup> and ArrayExpress<sup>9</sup>) adopt the Minimum Information About a Microarray Experiment (MIAME<sup>20</sup>) standard to store experiment information. To facilitate analysis of expression arrays, some systems are integrated with analytic packages. E.g., the Stanford Microarray Database (SMD<sup>10</sup>) uses the XCluster software package<sup>21</sup> to provide hierarchical and K-means clustering. Though some systems are not associated with an analytic package, expression data can be exported in a form suitable for a specific analytic tool (e.g., AMAD<sup>22</sup>). To help analyze expression data, some systems also provide extended annotation (e.g., GeneX<sup>23</sup> can store BLAST hits and map positions associated with a sequence).

To our knowledge, nearly all current expression array systems store expression data in a data warehouse. Though some systems integrate an analytic package, most of these packages are focused on a specific type of data analysis (e.g., clustering) and are not open source. A review of the leading expression array systems reveals no single system that can provide powerful extended annotation and analysis for an expression array dataset. Therefore, we developed the EAA model, to bridge the BioMediator data integration system and the BioConductor analytic platform to establish a powerful integrated expression array annotation and analytic system.

## PRELIMINARY WORK

To support our goal of expression array annotation, we first integrated the NCBI Gene Expression Omnibus (GEO) into BioMediator. This involved extend-

ing the mediated schema to include entities and relationships (nodes and edges) relevant to expression arrays. We chose GEO because it is a public domain expression array database for data deposition and query with broad support for a variety of different expression array types. To achieve an open and flexible design, GEO uses a semi-structured data model implemented as tab-delimited free text. This semi-structured data is transformed by the BioMediator wrapper and metawrapper layers into a data stream with mediated schema semantics represented in XML syntax.

At times, the annotation of expression arrays requires mapping from nucleotide sequences to cDNA clones. Expression arrays are designed and synthesized based on nucleotide sequence information. To extract nucleotide sequence from a cDNA clone array, the I.M.A.G.E database<sup>18, 24</sup>, the world's largest public collection of genes, was included in the BioMediator mediated schema and appropriate wrappers were developed. I.M.A.G.E integration provides two-way annotation between cDNA clones and nucleotide sequences. Furthermore, there is on-going two-way annotation from I.M.A.G.E cDNA clones to two types of gene clusters, Known Gene Clusters based on NCBI's Reference Genes and Candidate Gene Clusters having no known gene association.

## METHODS

The EAA model is an intermediate layer between BioMediator and BioConductor. The information flows bi-directionally through the EAA package. Entity annotation queries from BioConductor flow into EAA and are translated to PQL queries for BioMediator. Well-curated information retrieved from source databases by BioMediator enters the EAA tool and leave as an annotated entity in a data structure supported by BioConductor. The EAA model and software package can be divided into the following components: 1) BioConductor-EAA interface, 2) Graph Searching Schema, 3) Metadata Translation 4) Entity Coalesce and Classification 5) EAA-BioMediator interface. See Figure 1 for the graphical representation of the model underlying the EAA tool and the information flow. Each of these components is described in detail below.

### 1. BioConductor-EAA interface

At the top-level, analytic tools interact directly with the EAA interface (Fig 1A), which processes the annotation query request. The EAA model relies on querying several biologically meaningful entities, e.g., Gene, Clone, Nucleotide, Protein, and Phenotype. Queryable entities and their annotation relationships are defined in the EAA Searching Schema (Fig 1B).

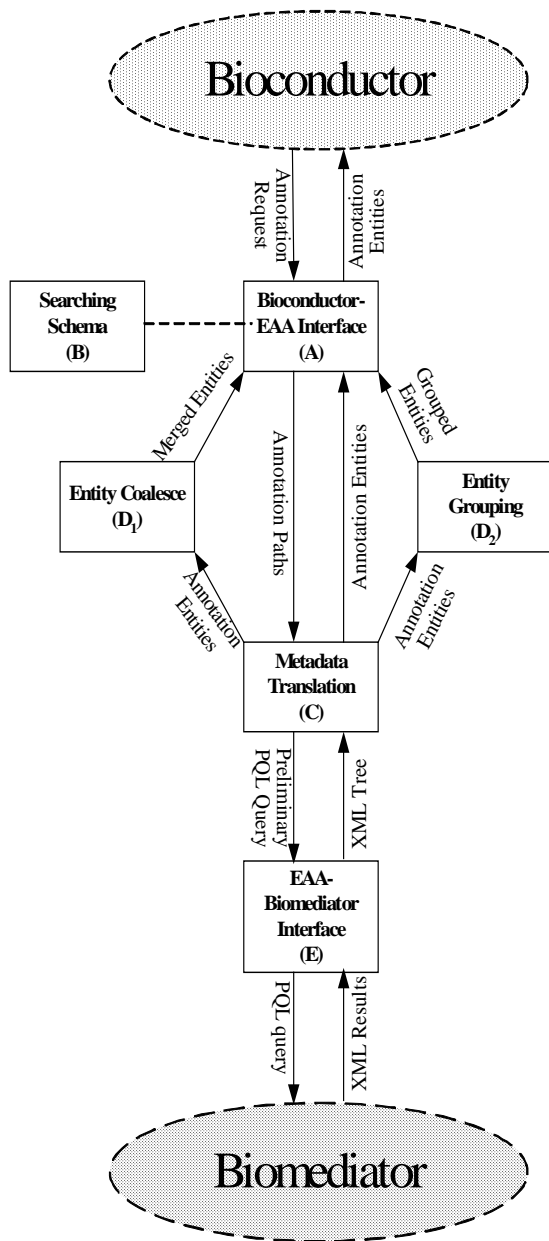


Figure 1: Expression Array Annotation (EAA) Model Bridge Bioconductor and Biomediator

The BioConductor-EAA interface uses a graph-searching algorithm to find possible annotation paths in the Searching Schema. The paths then are distributed to the Translation Layer (Fig 1C) for further processing. Responding to an annotation query request, the interface (Fig 1A) returns entities with well-curated annotations for directly analysis in BioConductor.

## 2. Graph Searching Schema

We designed the Graph Searching Schema (Fig 1B) in EAA to assist with the automatic construction of PQL queries. PQL is a graph-based declarative query

language: The USING clause defines which relationships are well-curated (based on metadata in a source knowledge base). The WHERE clause declares the entities of interest. The CREATE and LINK clauses define the nodes and edges to be returned (see an earlier PQL<sup>19</sup> paper for a more detailed description). In BioMediator, users manually construct PQL queries based on the mediated schema. The final PQL query is posed to PQL reformulator to create query plans over the actual sources. In EAA, we need a mechanism to automatically construct and pose PQL queries to BioMediator.

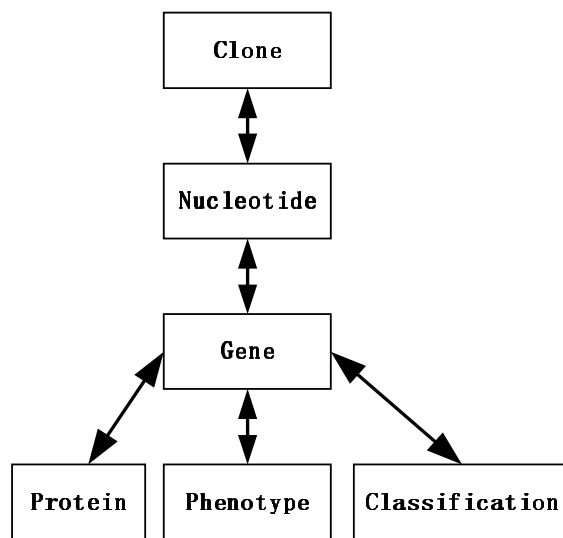


Figure 2: Searching Schema

Because EAA uses nested data structures and BioMediator output is flat (i.e., relational) EAA must iteratively construct nested results. The Graph Searching Schema (Fig 1B) is used for graph traversal. Given the searching schema of Figure 2, we can find only one annotation path from Clone to Gene, namely: Clone → Nucleotide → Gene. Starting from Gene, the annotation path to Phenotype will be Gene → Phenotype. Based on the annotation path, a preliminary PQL query is provided by the Metadata Translation (Fig 1C) component.

The Graph Searching Schema (Fig 1B) is a dynamic schema stored as a directed graph. It is dynamic because the search schema can be defined in BioConductor and passed to the BioConductor-EAA interface or a default searching schema can also be loaded when no schema is defined.

## 3. Metadata Translation

This is a bidirectional translation layer (Fig 1C). In the forward translation, the input stream contains the entities and annotation paths. These are transformed

to a preliminary query with four fragments, which correspond to the USING, WHERE, CREATE and LINK<sup>19</sup> clauses in PQL.

In the backward translation, the input stream contains annotated results from the BioMediator system. This component traverses the XML results and populates the attributes of the corresponding BioConductor entities, which are returned to the upstream process (Fig 1C, Fig 1D<sub>1</sub>, or Fig 1D<sub>2</sub>).

#### 4. Entity Coalescing and Grouping

Before passing annotated entities back to the BioConductor-EAA interface (Fig 1A), entities are pre-processed for merging (Fig 1D<sub>1</sub>) and grouping (Fig 1D<sub>2</sub>). Both merging and grouping are based on the entities' attributes. If more than one attribute is provided, a logical 'AND' relationship is applied. By merging (Fig 1D<sub>1</sub>), redundant entities can be removed and clean entities returned. For example, gene entities can be merged on their gene names, so gene entities with the same name from different external data sources will be merged to a single gene entity. Entity grouping (Fig 1D<sub>2</sub>) provides a clustering mechanism based on an entity's annotated attributes. Thus expression array probes corresponding to entities with common annotated attributes can be clustered into the same group (for example, different probes of the same gene could be clustered into a single group). Entities can be returned to BioConductor with or without merging/grouping.

#### 5. EAA-BioMediator interface

This layer (Fig 1E) communicates with BioMediator. It accepts preliminary queries and constructs the final PQL queries to be processed by BioMediator. Preliminary queries are represented using BioConductor data structures. The EAA-BioMediator component maps these data structures to PQL statements, which can be directly processed by BioMediator. When preliminary queries are sent to the EAA tool, the interface creates an active connection to the remote BioMediator PQL query processor and submits the final PQL query. This interface parses the XML and converts it to a tree representation, which is returned to the Metadata Translation (Fig 1C) component.

### IMPLEMENTATION & RESULTS

The EAA model is implemented as a software package in R. This facilitates the integration of the two systems and permits EAA to be distributed as a BioConductor package. Each component of the EAA package was designed and tested independently. Because R is an interpreted functional language for statistical analysis, interfacing layers in the EAA im-

plementation communicate with each other via input parameters and output results. Analytic tools in BioConductor can post annotation requests to the BioConductor-EAA interface (Fig 1A), as described above, or communicate directly with EAA-BioMediator interface (Fig 1E) by manually constructing preliminary queries.

Under BioConductor, user posts annotation request to EAA-BioMediator interface for retrieval of related annotation entities. Annotation requests consist of two entities, known entity and annotation entity. For example (refer to EAA example online<sup>25</sup> and EAA thesis<sup>26</sup> for detailed explanation and output results), in the request "Find protein information related to probe identifier, cloneID=6133969", the known entity is "Clone" with known id attribute 6133969 and the annotation entity is "Protein". When the BioConductor-EAA interface gets this annotation request, it explores searching schema and constructs annotation path: Clone -> Nucleotide Sequence -> Gene -> Protein. Metadata translation uses the annotation path and clone id attribute, 6133969, to form a preliminary PQL query. The preliminary query is organized to form a final PQL query by EAA-BioMediator interface, which is forwarded to BioMediator for processing. The EAA-BioMediator interface converts XML results from BioMediator into an XML tree. Metadata translation traverses the XML tree and extracts seven protein entities related to the clone with id attribute 613359. In this annotation example, Entity Coalesce merges seven proteins to three proteins based on their name attribute (e.g., all proteins with name "myc-associated zinc finger protein" are merged to one protein). Entity Group classifies those proteins to three groups based on their organism attribute (e.g., proteins from "human" are classified in same group). By seamless information flowing and parsing within EAA, annotation entities with well-curated information wrapped in attributes are finally returned for easy and efficient control under BioConductor.

### DISCUSSION & CONCLUSION

**Successes:** The implementation of the EAA model as a software package is able to link the BioConductor analytic platform to the BioMediator data integration system. Users of BioConductor can easily retrieve well-curated biological knowledge without having to explicitly consider heterogeneous data sources with different data storage formats and different semantics. From the annotation example above, we can see that this process can be easily realized by posting simple annotation requests to the query interface. The EAA system for annotation is efficient. It removes onerous work to access independent data source and integrate

heterogeneous data manually. The uniform and well-structured data format makes automated analysis of huge datasets possible. The EAA model uses its own searching schema to define the annotation strategy. This allows the biologist to focus only on the biologically meaningful annotations. For example, we can provide constraints that make Nucleotide → Gene → Protein a valid annotation path, but not Nucleotide → Protein → Gene. Furthermore, as described above, each layer in the EAA tool is implemented as an independent module, which makes it easily modified and expanded without changing other components.

**Current Development:** We currently (March, 2003) have implemented all the basic modules of the EAA model. These modules process the basic biological entities' annotation. In the next step of development, we expect to define new entities to represent extended expression array annotation (e.g., homologous genes related to interesting probes). Current EAA model has an important limitation. Entities in the EAA model are derived from entities defined in the BioMediator knowledge base, but BioMediator is a highly expandable (and evolving) system<sup>27</sup> and thus the structure and characteristics of entities in the mediated schema may be modified for various applications. We do not yet have automated mechanisms to keep the entities in the EAA model (Fig 1B) consistent with the evolving BioMediator mediated schema. One solution could be to have the EAA model directly access the Protégé knowledgebase stored on the BioMediator server (an approach used by many components of BioMediator).

**Future challenges:** In face of increasingly broad applications of expression arrays, more sophisticated annotation is desired. Therefore, more heterogeneous data sources will be added to BioMediator for expanded annotation. For example, to find a potential SNP for a given sequence, dbSNP in NCBI could be included as a source. To cluster EST sequence, the NCBI sequence cluster database, UniGene, needs to be supported. To explore genetic regulatory networks, data sources of metabolic and signaling pathways need to be supported. These sources could be called *data analysis resources*<sup>28</sup>.

Incorporation of these resources into BioMediator requires that additional wrappers be written and the mediated schema modified. Moreover, the current system architecture is based on standard data integration techniques. We are exploring ways to extend these ideas to handle such analytical tools.

In light of these anticipated extensions and modifications, the main challenge for EAA will be to ensure that the BioConductor and BioMediator data models are kept congruent.

## ACKNOWLEDGEMENTS

We would like to thank Sara Thiebaud for assistance in query testing and paper writing. Funding was provided by NHGRI (R01HG02288, Tarczy-Hornoch, PI), NLM training grant (T15LM07442, trainees: Mork & Donelson), and the University of Washington UIF fund (Mei).

## REFERENCES

1. Brown MPS, Grundy WN, Lin D, et al. Knowledge-based analysis of microarray gene expression data using support vector machines. *Proceedings of the National Academy of Sciences* 1999;97(1):262-267.
2. Gene Ontology Consortium (GO). Gene ontology (go). <http://www.geneontology.org/>
3. Hvidsten TR, Komorowski J, Sandvik AK, Laegreid A. Predicting gene function from gene expressions and ontologies. In: *Pacific Symposium on Biocomputing*; 2001; 2001. p. 299-310.
4. Dettling M, Buhlmann P. Supervised clustering of genes. *Genome Biology* 2002;3(12):1-15.
5. Bioconductor. <http://www.bioconductor.org/>
6. Biomediator.
7. The R project for statistical computing. <http://www.r-project.org/>
8. Gene expression omnibus. <http://www.ncbi.nlm.nih.gov/geo/>
9. Arrayexpress. <http://www.ebi.ac.uk/arrayexpress/>
10. Stanford microarray database. <http://genome-www5.stanford.edu/MicroArray/SMD/>
11. Edgar R, Domrachev M, Lash AE. Gene expression omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Research* 2002;30(1):207-210.
12. Mork P, Halevy A, Tarczy-Hornoch P. A model for data integration systems of biomedical data applied to online genetic databases. In: *American Medical Informatics Association Fall Symposium*; 2001 November 3-7; Washington, DC, USA: AMIA; 2001. p. 473-477.
13. Entrez search and retrieval system. <http://www.ncbi.nlm.nih.gov/Entrez/>
14. Swiss-prot protein knowledgebase. <http://us.expasy.org/sprot/>
15. The human genome organisation (hugo). <http://www.gene.ucl.ac.uk/hugo/>
16. Locuslink. <http://www.ncbi.nlm.nih.gov/LocusLink/>
17. OMIM online mendelian inheritance in man. <http://www.ncbi.nlm.nih.gov/omim/>
18. The I.M.A.G.E. Consortium. <http://image.lnl.gov>
19. Mork P, Shaker R, Halevy A, Tarczy-Hornoch P. PQL: A declarative query language over dynamic biological schemata. In: *American Medical Informatics Association Fall Symposium*; 2002 Nov; San Antonio, TX: AMIA; 2002.
20. Brazma A, Hingamp P, Quackenbush J, et al. Minimum information about a microarray experiment (MIAME)-toward standards for microarray data. *Nature Genetics* 2001;29(4):365-371.
21. Xcluster. <http://genetics.stanford.edu/~sherlock/cluster.html>
22. Arraymaker. <http://www.microarrays.org/software.html>
23. Genex? <http://www.ncgr.org/genex/>
24. Lennon G, C A, Polymeropoulos M, Soares MB. The I.M.A.G.E. Consortium: An integrated molecular analysis of genomes and their expression. *Genomics* 1996;33:1512.
25. Eaa example online. [www.biomediator.org/EAA/EAA\\_example.html](http://www.biomediator.org/EAA/EAA_example.html)
26. Mei H. Expression array annotation using the biomediator biological data integration system and the bioconductor analytic platform. 2003.
27. Shaker R, Mork P, Barclay M, Tarczy-Hornoch P. A rule driven bi-directional translation system remapping queries and result sets between a mediated schema and heterogeneous data sources. In: *American Medical Informatics Association Fall Symposium*; 2002 November 9-13; San Antonio, Texas: AMIA; 2002. p. 692-696.
28. Wheeler DL, Church DM, Federhen S, et al. Database resources of the national center for biotechnology. *Nucleic Acids Research* 2002;31(1):28-33.