Proceedings of the 3rd International Workshop on Distributed Statistical Computing (DSC 2003)
March 20–22, Vienna, Austria *ISSN 1609-395X*
Kurt Hornik, Friedrich Leisch & Achim Zeileis (eds.)
http://www.ci.tuwien.ac.at/Conferences/DSC-2003/

RDBMS in Bioinformatics: The Bioconductor Experience

Vincent J. Carey

Abstract

Bioconductor (http://www.bioconductor.org/) is an open source collection of resources aimed at transparently advancing the theory and practice of bioinformatics, with a focus on expression arrays and the R statistical computing environment. I will sketch the key data structures and data flow processes addressed in Bioconductor thus far. I will review the role played by RDBMS in the development and curation of packaged annotation networks and in the analysis of Serial Analysis of Gene Expression (SAGE) libraries. Non-relational database technologies such as BerkeleyDB and HDF5 have also played a role in tools for archiving and navigating expression array data. At present the role of RDBMS in Bioconductor is less pronounced than had been anticipated. This will change as requirements for query optimization, data structure standardization, and greater volumes of data and metadata emerge.

1 Introduction

The recent explosion in volume of biological data and metadata impacts scientific computing in many ways. Figure 1 is a common depiction of the problem, schematizing the exponential growth of quantitative data on sequence, diversity, signaling and expression. Not shown are metadata resources required to allow accurate and efficient use of the burgeoning experimental and observational data. These are schematized in Figure 2, and empirical growth of data in these databases is sketched in Figure 3.

Bioconductor (http://www.bioconductor.org/) is an open source collection of resources aimed at transparently advancing the theory and practice of bioinformatics, with a focus on expression arrays and the R statistical computing environment. The primary objective of Bioconductor is to provide a basis for getting the best statistical and computational methods of bioinformatics into the hands of practicing biologists. This has necessitated considerable work on the infrastructure of R, and the success of this has allowed us to limit the role of RDBMS technology in

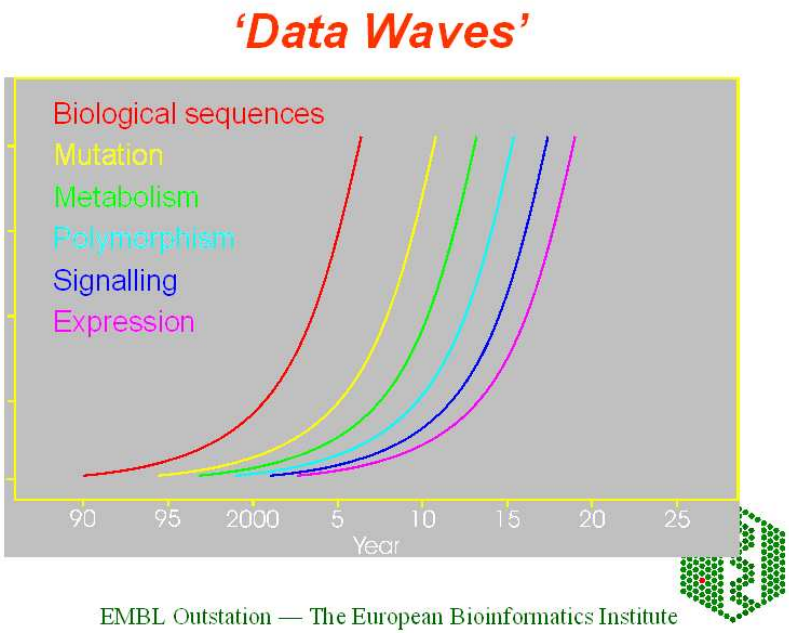


Figure 1: Data explosion. ©2002 EBI.

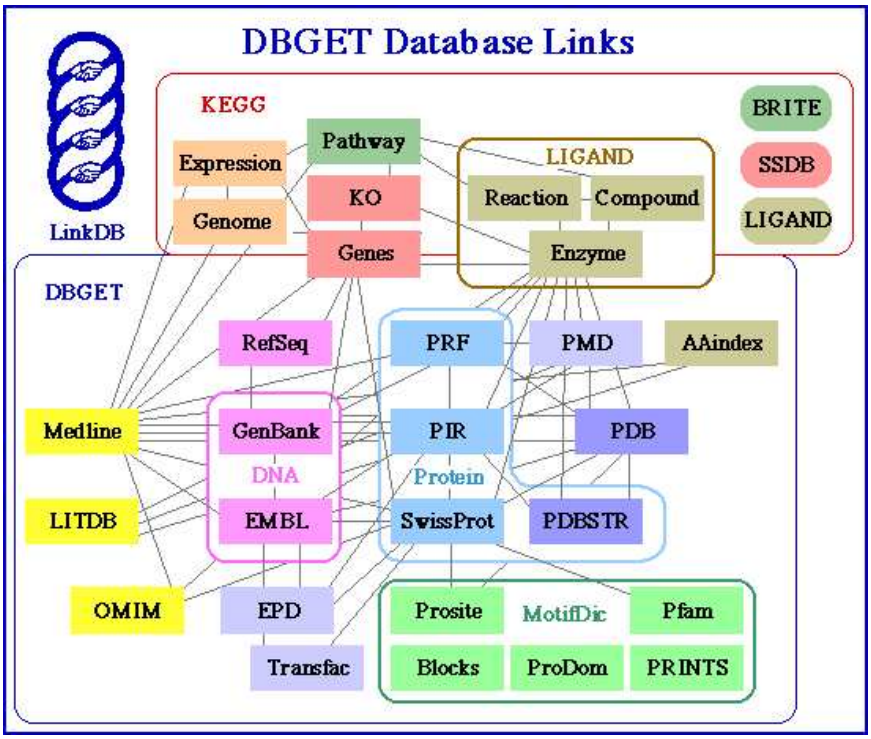


Figure 2: Data/metadata network. ©2002 GenomeNet Japan.

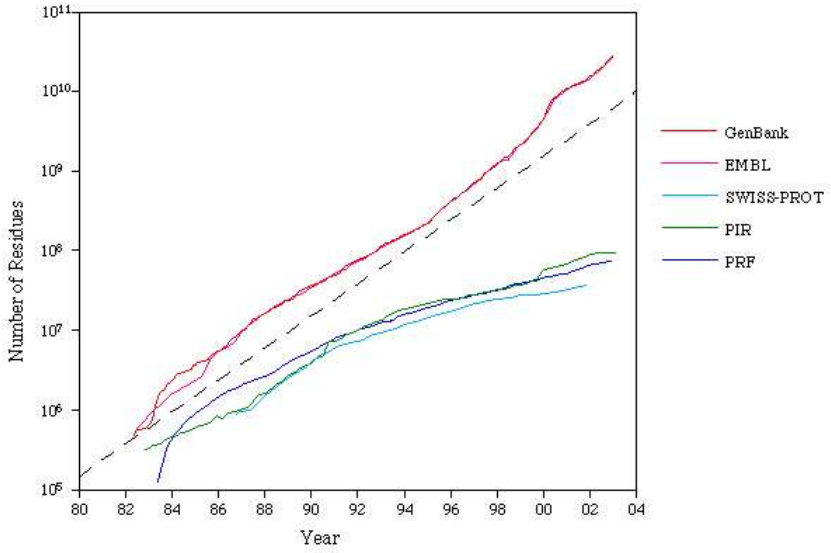


Figure 3: Residue load in major bioinformatics databases. ©2002 GenomeNet Japan.

Bioconductor to date. In this report I will sketch various aspects of bioinformatic data structure and data flow processes supported by Bioconductor. Issues in the creation and curation of packaged metadata and annotation networks will be reviewed in some detail, and the explicit role of Postgres in new tools for the analysis of SAGE libraries will be described. Longer term involvements with DBMS technologies related to volume, standards, and query optimization are discussed in the conclusion.

2 Data structures and flow related to microarrays

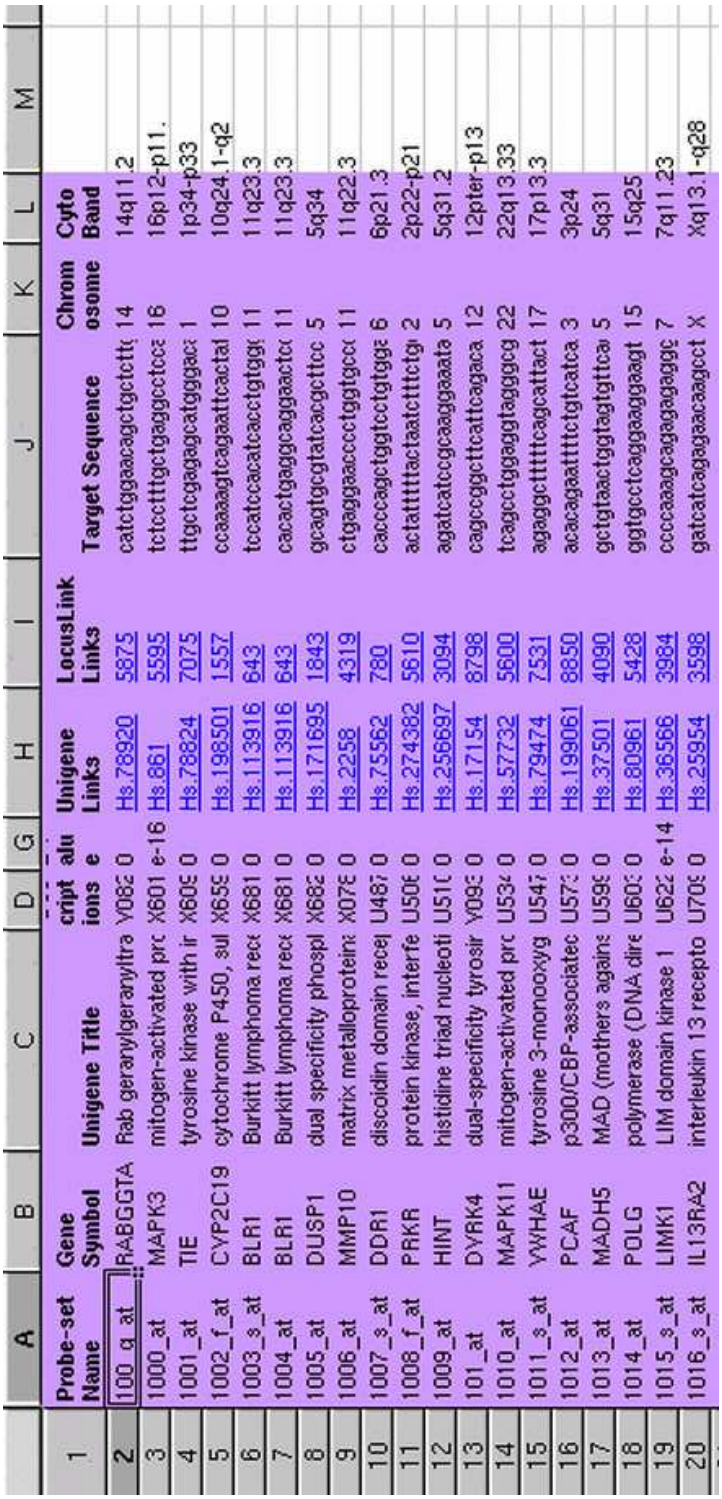
2.1 Genomic metadata

Microarrays are rectangular layouts of DNA sequences. I will focus on the AffymetrixTM Hu6800 or U95Av2 chips to describe annotation issues arising in Bioconductor. The location of sequence on the chip is recorded in `.sif` or `probe_tab` files supplied by the manufacturer.

```
Probe Set Name
               Serial Order
                       Probe X Probe Y Probe Interrogation Position
                                                 Probe Sequence

A28102_at      1       5       1       1084    TACCTAAAGTGGCATATGCGACGGC
A28102_at      2       6       1       1174    TCAACTATTTCACCAAGCGGAGTTG
A28102_at      3       7       1       1216    AGGTGCCAGAGGCCCTGGAGATGAA
...
```

This particular annotation table (for Hu6800) has 131542 lines.



	A	B	C	D	G	H	I	J	K	L	M
1	Probe-set Name	Gene Symbol	Unigene Title	cript ions	alu e	Unigene Links	LocusLink Links	Target Sequence	Chrom osome	Cyto Band	
2	100_g_at	RABGGTA	Rab geranylgeranyltra	Y08(	0	Hs.78920	5875	catctggaacagctgctctt	14	14q11.2	
3	1000_at	MAPK3	mitogen-activated prc	X601	e-16	Hs.861	5595	tctcctttgctgaggcctcc	16	16p12-p11.	
4	1001_at	TIE	tyrosine kinase with ir	X60(	0	Hs.78824	7075	ttgctcgagagcatgggac	1	1p34-p33	
5	1002_f_at	CYP2C19	cytochrome P450, sul	X65(	0	Hs.198501	1557	ccaaaagtcagaattcactat	10	10q24.1-q2	
6	1003_s_at	BLR1	Burkitt lymphoma rec	X681	0	Hs.113916	643	tccatccacatcacctgtgg	11	11q23.3	
7	1004_at	BLR1	Burkitt lymphoma rec	X681	0	Hs.113916	643	cacactgaggcaggaactc	11	11q23.3	
8	1005_at	DUSP1	dual specificity phospl	X68(	0	Hs.171695	1843	gcagtgcgtatcacgcttcc	5	5q34	
9	1006_at	MMP10	matrix metalloproteins	X07(	0	Hs.2258	4319	ctgaggaacccctggtgccc	11	11q22.3	
10	1007_s_at	DDR1	discoidin domain rece	U48(	0	Hs.75562	780	cacccagctggtcctgtggc	6	6p21.3	
11	1008_f_at	PRKR	protein kinase, interfe	U50(	0	Hs.274382	5610	actatttttactaatctttctg	2	2p22-p21	
12	1009_at	HINT	histidine triad nucleoti	U51(	0	Hs.256697	3094	agatcatccgcaaggaaata	5	5q31.2	
13	101_at	DYRK4	dual-specificity tyrosir	Y09(	0	Hs.17154	8798	cagccggcttcattcagaca	12	12pter-p13	
14	1010_at	MAPK11	mitogen-activated prc	U53(	0	Hs.57732	5600	tcagcctggaggtaggggcg	22	22q13.33	
15	1011_s_at	YWHAE	tyrosine 3-monooxyg	U54(	0	Hs.79474	7531	agaggctttttcagcattact	17	17p13.3	
16	1012_at	PCAF	p300/CBP-associatec	U57(	0	Hs.199061	8850	acacagaattttctgtcatca	3	3p24	
17	1013_at	MADH5	MAD (mothers agains	U59(	0	Hs.37501	4090	gctgtaactggtagtgttcaa	5	5q31	
18	1014_at	POLG	polymerase (DNA dire	U60(	0	Hs.80961	5428	ggtgcctcaggaaggaagt	15	15q25	
19	1015_s_at	LIMK1	LIM domain kinase 1	U62(	e-14	Hs.36566	3984	ccccaaagcagagagaggc	7	7q11.23	
20	1016_s_at	IL13RA2	interleukin 13 recepto	U70(	0	Hs.25954	3598	gatcatcagagaacaagcct	X	Xq13.1-q28	

Figure 4: U Michigan mapping from Affy probe sequences to the annotated genome; see http://dot.ped.med.umich.edu:2000/ourimage/micro-arrays/Affy_annot/Unigene/index.html.

The relationship from probe sequence to the mapped human genome is characterized by a group at U Michigan. Figure 4 is an excerpt from a spreadsheet used to disseminate the coordinated mapping, which is accomplished by blasting the probe sequences against the evolving Unigene representation of the genome.

Once the connection (or connections, as the biological role of the sequence fragment may depend upon context) between probe sequence and genome is made, a number of annotation directions can be pursued. Based on the `hgu95av2` and `KEGG` data packages, a probe ID can be mapped to various nomenclatures or characterizations of the gene involved.

```
> annoSumm2("1005_at",30)
                    gbAcc                                    chr
                "X68277"                                    "5"
                   chrloc                                 chrori
             "172896942"                                    "-"
                   enzyme                                   name
              "3.1.3.16"    "dual specificity phosphatase 1"
                     grif                                     ll
              "12080474"                                 "1843"
                      map                                   path
                  "5q34" "04070 (Phosphatidylinositol si..."
                     pmid                                   summ
              "12080474" "Non-receptor protein-tyrosine ..."
                      sym                                  unig
                 "DUSP1"                          "Hs.171695"
```

With the handwritten function `annoSumm2` idiosyncratic abbreviations are used to obtain a concise sketch of the available information. The `gbAcc` field is for the GenBank accession number in which the probe sequence can be found. The `unig` field identifies the UniGene 'gene oriented transcript cluster' in which the sequence can be found. The gene DUSP1 is found on chromosome 5 in the antisense direction; more precisely the sequence is in band 5q34 at basepair 172986942. The `pmid` and `grif` fields are for PubMed publication accession numbers in which key information about this sequence has been published (a 2002 paper in *Oncogene*, in this case).

2.2 Experimental data

The structure and flow of experimental data is generally institution-specific and hard to predict. Many analytic and managment utilities function by requiring the user to load a directory with files, or to interactively browse and select files that are to be analyzed. With the `affy` package, collections of raw array outputs (Affymetrix CEL files) can be collected into `AffyBatch` objects. These are then processed to background-corrected, co-normalized expression data and organized for the user, in conjunction with *phenotype* or experimental treatment data, as an `exprSet`. The external presentation of an `exprSet` is mediated by the formal `show` method. For the Leukemia study (Golub et al., 1999), we have

```
> golubTrain
Expression Set (exprSet) with
    7129 genes
    38 samples
         phenoData object with 11 variables and 38 cases
           varLabels:
```

```
Samples: Sample index
ALL.AML: Factor, indicating ALL or AML
BM.PB: Factor, sample from marrow or peripheral blood
T.B.cell: Factor, T cell or B cell leuk.
FAB: Factor, FAB classification
Date: Date sample obtained
Gender: Factor, gender of patient
pctBlasts: pct of cells that are blasts
Treatment: response to treatment
PS: Prediction strength
Source: Source of sample
```

Not shown in this display is a collection of fields responding to an emerging standard for microarray metadata: the MIAME (Minimum Information About a Microarray Experiment) protocol (Brazma et al., 2001).

```
> slotNames("MIAME")
 [1] "name"           "lab"            "contact"        "title"
 [5] "abstract"       "url"            "samples"        "hybridizations"
 [9] "normControls"   "preprocessing"  "other"
```

The fields in this structure must be filled in a regimented manner in order for the results of the experiments to be publishable in any of the high-visibility scientific journals that subscribe to the MIAME standard.

The numerical data are excerpted by:

```
> exprs(golubTrain)[500:504,1:5]
           [,1] [,2] [,3] [,4] [,5]
D50930_at   512  666 1161 1025  785
D50931_at -477  -88 -850  185  -96
D55638_at -152 -197 -434 -139  -55
D55640_at  793  525  848  817  255
D55654_at 1094 2133 2858 2107 3510
```

Here rows are genes (ESTs putatively associated with genes) and columns are samples. When these data are subjected to statistical analysis, some ESTs will be typically be identified as playing a role in the biological process of interest. The analyst then uses the genomic metadata to rationalize the selection of these ESTs.

Clearly the conormalized array outputs and the "phenodata" are rectangular data structures that could straightforwardly be managed in an RDBMS. At present Bioconductor has not emphasized this possibility because the capacity to efficiently perform arbitrary statistical computation on RDBMS contents is not sufficiently developed.

2.3 Summary of data structure and flow

The basic information streams reviewed thus far are

- *Disseminated genomic metadata.* This includes the probe sequence data, its location on manufactured chips, its mapping to the genomic sequence, and the functional characterization of the genomic sequence. It flows from manufacturer to mapping specialists (e.g., U Michigan or Bioconductor) and from institutional bioinformatics centers (e.g., NCBI, EBI) into packaged annotation resources.

- *Local experimental data.* This includes the raw microarray scanner outputs, the experimental design metadata and phenotype-related information. This will flow from a lab or institutional warehouse to the analysis team/platform.
- *Results of local statistical analysis.* Here the disseminated genomic metadata are queried in a focused manner to elucidate the expression patterns discovered in analysis.

Bioconductor has contributed to the structure and curation of disseminated genomic metadata to facilitate the efficient interpretation of statistical analyses of microarray experiments. In the next section I provide some details of how this occurs.

3 Curating high-availability metadata structures for use in R

WWW distribution of high-resolution genomic metadata is well-established, and indeed most of the annotation materials used in Bioconductor are obtained by download from central bioinformatics repositories. However, network latencies and restrictions on frequent queries make it infeasible to employ the WWW as a high-availability annotation query resolver at this time.

The `Data packages` node of http://www.bioconductor.org/ includes 16 R packages (regimented collections of functions, data and documentation with integrated testing and quality control tools) embodying metadata on the hgu133a-b, hgu95a-e, mgu74a-c, rgu34a-c microarray platforms, and on KEGG and GO biological metadata resources. The `hgu95av2` package is a collection of environments

```
> objects("package:hgu95av2")
 [1] "hgu95av2"              "hgu95av2ACCNUM"       "hgu95av2AFFYCOUNTS"
 [4] "hgu95av2CHR"           "hgu95av2CHRLOC"       "hgu95av2ENZYME"
 [7] "hgu95av2ENZYME2PROBE"  "hgu95av2GENENAME"     "hgu95av2GO"
[10] "hgu95v2aGOPROBE"       "hgu95av2GO2ALLPROBES" "hgu95av2GRIF"
[13] "hgu95av2LOCUSID"       "hgu95av2MAP"          "hgu95av2PATH"
[16] "hgu95av2PATH2PROBE"    "hgu95av2PMID"         "hgu95aPMID2PROBE"
[19] "hgu95av2QC"            "hgu95av2SUMFUNC"      "hgu95av2SYMBOL"
[22] "hgu95aUNIGENE"
```

constituting a variety of mappings between terms in different nomenclatures. These mappings are used manually as follows:

```
> get("1005_at", env=hgu95av2GO)
[1] "GO:0004726" "GO:0006979"
> get("1005_at", env=hgu95av2PATH)
[1] "04070"
> get("1005_at", env=hgu95av2ENZYME)
[1] "3.1.3.16" "3.1.3.48"
> get("1005_at", env=hgu95av2MAP)
[1] "5q34"
```

An important mapping for reasoning prospectively from gene function to sets of probes is provided in the `*GO2ALLPROBES` environments. We see that Gene Ontology term `GO:0006979` was assigned to `1005_at`. The `GOBPID2TERM` environment of the `GO` package allows us to decipher this:

```
> get("GO:0006979", env=GOBPID2TERM)
[1] "oxidative stress response"
```

Now to find all probes connected with this oxidative stress response process, we use

```
> get("GO:0006979", env=hgu95av2GO2ALLPROBES)
 [1] "41776_at"   "1005_at"    "34715_at"   "41323_at"   "41324_g_at"
 [6] "38386_r_at" "41631_f_at" "33284_at"   "35723_at"   "1403_s_at"
[11] "1404_r_at"  "1405_i_at"  "33789_at"   "34363_at"   "36620_at"
[16] "34666_at"   "39729_at"   "36937_s_at" "41432_at"   "39136_at"
[21] "35605_at"   "399_at"     "40104_at"   "770_at"
```

Clearly the mappings must be composed in complex ways in order to be fully useful. For example, it is not particularly meaningful to find out that a given probe is involved in pathway 04070. In order to appreciate this we need to be able to decode the pathway tag and also to examine the behavior of other probes implicated in the same pathway (if such exist). The design of efficient tools to permit programmatic composition and navigation of annotation mappings is an ongoing project. At this time, we focus on issues related to the construction of the various mapping components.

Several broad issues of database design and management arise in dealing with the annotation mapping problem.

- *Resource and output scope.* We have decided that the experimental platform defines a useful level of focus for mapping tools. Instead of providing a general mapping between, e.g., all LocusLink identifiers and GO terms, we first scale down the problem to the set of LocusLink identifiers associated with a specific chip, and then carry out the map construction. This leads to some redundancies among the various platform-specific packages, but brings the map resource components to a manageable and focused scale.

- *Multiplicity of sources.* As noted above a variety of parties are involved in annotation. The manufacturers and various institutions (e.g., EBI, Gene Expression Omnibus) may provide overlapping and conflicting annotation data with various levels of trustworthiness.

- *Evolving versions.* The central annotation resources on gene location, structure and function are constantly changing. Bioconductor resources need to be updated, but concepts of backward compatibility also play a role as ongoing analyses may be adversely affected by real-time modifications to annotation.

A programmatic approach to annotation resource curation is developed in the `AnnBuilder` package (Zhang et al., 2003). The basic components of this package address problems of

- *Central file access:* identifying and collecting annotation information from the primary bioinformatics servers responsible for its basic curation.

- *Parsing base files:* LocusLink and UniGene data are distributed as weakly marked up flat files. Perl parsers are provided to collect needed fields for import to R environments; because Gene Ontology data is distributed as XML, XML parser functions based on Duncan Temple Lang's `XML` library are provided.

```
ANNBUILDER:ANNOTATE
 |_(annbuilder:attr+,
 |    |_(annbuilder:target+,
 |    |    |_EMPTY
 |    |
 |    |__annbuilder:datemade+,
 |    |    |_EMPTY
 |    |
 |    |__annbuilder:version+,
 |    |    |_EMPTY
 |    |
 |    |__annbuilder:sourcefile*,
 |    |    |_EMPTY
 |    |
 |    |__annbuilder:element*)
 |         |_EMPTY
 |
 |
 |__annbuilder:data+)
      |_(annbuilder:entry*)
          |_(annbuilder:item*)
              |_EMPTY
```

Figure 5: `annotate.dtd` tree structure.

- *Unifying multiple incomplete maps with variable trust.*
- *Navigating and harvesting the GO DAG.* This phase employs Postgres to store the DAG and to extract paths from root to given nodes.
- *Exporting XML.* Data packages are formatted as XML according to the `annotate.dtd` maintained at the Bioconductor site. See Figure 5 for the structure.

Improved performance of R and inevitable complications of requiring a properly configured Postgres at the user's end have led to a reformulation of the curation software in a postgres-free package now known as `pubRepo`.

In summary, the development of microarray-oriented metadata via R environments appears to be a successful approach to supporting curation of high-availability genomic annotation networks. Once the data packages are installed, the annotation is continuously available (WWW queries not required), and the metadata can be programmatically linked to the analysis, facilitating general computation on annotation data for biological and clinical inference. It is likely that graphical structures relating translator environments will be used to support navigation of annotation networks in forthcoming versions of Bioconductor.

4 SAGElyzer

Serial Analysis of Gene Expression (SAGE) is a comprehensive tool for discovering genes and quantifying gene or transcript expression (Velculescu et al., 1995). The

basic organizing element is the *tag*, a 14 or 21-base pair segment of DNA from a defined location in each expressed gene as a unique identifier for that gene.

The frequency distribution of tags in mRNA obtained from cells or tissues of interest is standardized and organized in a *SAGE library*. SAGE permits measurement of transcript expression without restriction to a predetermined set of genes, thus avoiding a basic limitation of microarray analysis.

In 199 SAGE libraries collected for use at Dana Farber Cancer Center, over 500000 unique tags (with associated frequencies irregularly available from library to library) were present. A postgres database was set up to archive the libraries. `RODBC` (Windows) and `Rdbi/RPgSQL` (Unix) interfaces are used to perform a 'chunked' ranking of tag profiles in terms of similarity to the profile associated with a user-selected tag. Specifically, a tag is selected and its frequency profile across libraries is computed. The user also specifies how large a set N of 'neighbors' is desired. Chunks of records are extracted from the database, exported to R, the distance from the target profile to each chunk element is computed, and the tags and distances for the N nearest neighbors are extracted across all chunks. This vector is then annotated and returned to the user.

5 Alternative DBMS technologies

HDF5 (Hierarchical Data Format version 5) has been interfaced to R for the purpose of archiving and supporting high-performance navigation of microarray image data. See the package `rhdf5`. An alternate implementation of `exprSet`s with BerkeleyDB C structures has been experimentally deployed as the `exprDB` package.

6 Future directions

The role of MAGE-ML import/export will undoubtedly grow as more organizations adopt this standard. Support for the MAGE-OM object model and schemas is planned. It is likely that the volume of metadata will necessitate an approach in which comprehensive annotation archives are maintained in RDBMS in contrast to R environments/packages. This will introduce novel problems of configuration (to adapt to institution- or lab-level idiosyncrasies in RDBMS environments) and more focused work on the R/RDBMS interfaces required to permit seamless binding of metadata to components of the data analysis information flow.

References

A Brazma, P Hingamp, and J Quackenbush. Minimum information about a microarray experiment: towards standards for microarray data. *Nature Genetics*, 29: 365–371, 2001.

T. R. Golub, D. K. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J. P. Mesirov, H. Coller, M.L. Loh, J. R. Downing, M. A. Caligiuri, C. D. Bloomfield, and E. S. Lander. Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring. *Science*, 286:531–537, 1999.

V. E. Velculescu, L. Zhang, B. Vogelstein, and K.W. Kinzler. Serial analysis of gene expression. *Science*, 270:484–487, 1995.

J. Zhang, V. Carey, and R. Gentleman. An extensible application for assembling annotation for genomic data. *Bioinformatics*, 19(1):155–56, 2003.

Affiliation

Vincent J. Carey
Harvard Medical School
181 Longwood Ave Boston MA USA 02115
E-mail: stvjc@channing.harvard.edu