entity-attribute-value model,
relational database management system,
DNA microarray

Tomasz WALLER[1], Damian ZAPART[1], Magdalena TKACZ[2], Zygmunt WRÓBEL[1]

ANALYSIS OF ENTITY-ATTRIBUTE-VALUE MODEL APPLICATIONS IN FREELY AVAILABLE DATABASE MANAGEMENT SYSTEMS FOR DNA MICROARRAY DATA PROCESSING

Large volumes of data are generated during DNA microarrays experiments. Database management systems (DBMS) are increasingly applied to these data, providing optimum processing and management from multiple microarray experiments. In this study, freely accessible DBMS software versions were compared (Microsoft SQL Server 2008 Express Edition, Oracle Database 10g Express Edition, DB2 Express-C 9.7.2, MySQL 5.1, and PostgreSQL 9.0). We examined them in the context of possible Entity-Attribute-Value (EAV) application as an optimal organization method for microarray data.

It was confirmed in the comparative analysis of component data processing methods, consistent with the EAV model, that efficient methods for microarray data analysis are available in Microsoft SQL Server 2008 Express Edition and PostgreSQL 9.0 systems. Also, DNA microarray data processing was confirmed to be more efficient with Microsoft SQL Server 2008 Express Edition as compared with PostgreSQL 9.0.

The EAV method was also shown to be suitable for use with open-source versions of DBMS software as an optimum storage model for DNA microarray data. In terms of data processing methods and performance, the Microsoft SQL Server 2008 Express Edition proved to be the best among compared database systems.

1. BACKGROUND

DNA microarray is a commonly used technology to measure the transcription activity (expression) for tens of thousands of genes. Microarrays are widely used in genotyping, medical diagnosis, and pharmacy [8], and they can be constructed by the direct synthesis of oligonucleotide or cDNA spotting.

Microarray data were developed and specified based on the properties of data obtained with the oligonucleotide method developed by Affymetrix (HG-U133A microarray model).

DNA microarray data are standard numerical data obtained by processing the microarray images used in the experiment. Each image reflects a gene activity profile for a genetic material tested with a specific microarray. Processed image data are converted into a table of numerical data, where each row corresponds to a specific gene and each column to a tested sample. Each column presents a gene expression profile for a specific sample, and each row presents an expression level for each gene for all analyzed genetic materials. The table size depends on the microarray model and the number of tested samples. The type of microarray determines the number of rows, which are constant for the applied model (for example, for HG-U133A by Affymetrix, is 22283), and the number of used microarrays corresponds to the number of columns. The data are the basis for all tests and analyses.

A well-designed microarray database can provide valuable information on gene expression levels [1, 2, 4, 5]. The migration of data from a microarray experiment to a relational database is a simple process, generally involving the design of a suitable table within the DBMS. However if data from several experiments are to be stored, a problem will occur with in determining a suitable number of columns for the relational database. For each experiment, a different number of microarrays can be used, and the general number of columns cannot be determined in advance. Admittedly, one can assume that the experiment will include, for example 256 samples. However this fixed allocation of samples, is not a perfect solution, since it limits the possibility of using the database in experiments where the number of

[1] Institute of Computer Science, Division of Biomedical Computer Systems, University of Silesia, Katowice, PL
[2] Institute of Computer Science, Division of Information Systems, University of Silesia, Katowice, PL

samples exceeds the number assumed for the database design. Additionally, for experiments with smaller numbers of microarrays, the table will include unused columns and NULL values, which represent a superfluous load at the data processing stage.

	GENE ID	MM1	MM2	MM3	MM4	MM5	MM6	MM7	MM8	MM9	MM10
1	1007_s_at	9,37064075469971	8,76850986480713	9,25258922576904	8,60418319702148	8,746158599853 52	8,66250228881836	8,68928050994873	7,92920970916748	8,90814590454102	9,1297388...
2	1053_at	4,47357654571533	4,38682031631 47	4,49260234832764	4,66426610946655	4,32363891601563	4,42414379119873	4,48229265213013	4,51816320419312	4,54742527008057	4,1949443...
3	117_at	6,13452911376953	6,04070711135864	6,51620006561279	6,07580280303955	6,19125604629517	5,79320001602173	5,8508129119873	6,09781169891357	6,16555452346802	6,0399031...
4	121_at	8,55980682373047	8,35822582244873	8,46381282806396	8,54459762573242	8,41231250762939	8,49077224731445	8,23797607421875	8,64408874511719	8,45146369934082	8,3434219...
5	1255_g_at	3,84340405464172	3,78949451446533	3,65429615974426	3,63046836853027	3,65746855735779	3,50215697288513	3,62728118896484	3,68150925636292	3,71160626411438	3,7044839...
6	1294_at	7,20534229278564	7,51769495010376	7,53055429458618	7,97169923782349	7,40089893341064	8,0644998550415	8,03535270690918	7,28476428985596	7,96531581878662	7,6804437...
7	1316_at	4,52201223373413	4,51401853561401	4,75651359558105	4,49362087249756	4,59989833831787	4,40654754638672	4,56029415130615	4,80376815795898	4,54615497589111	4,3895716...
8	1320_at	4,07777214050293	4,1824803352356	4,22394609451294	4,46391344070435	4,30546712875366	4,39015531539917	4,21411609649658	4,45786237716675	4,4286937713623	4,1741542...
9	1405_i_at	6,0865044593811	6,25917530059814	6,09619808197021	6,54993581771851	5,83881282806396	7,11537265777588	5,91771268844604	5,79089689254761	5,76026344299316	5,0135374...
10	1431_at	4,18942213058472	3,75942778587341	3,69134736061096	3,66687726974487	3,52786612510681	3,62515163421631	3,63023567199707	3,73535227775574	3,73457670211792	3,7210195...
11	1438_at	6,31258153915405	6,0294885635376	6,36651086807251	6,00778770446777	5,71528339385986	5,69585180282593	5,8590259552002	6,52168273925781	5,95231485366821	5,9801826...
12	1487_at	6,79423809051514	6,94645643234253	6,80649089813232	6,84117889404297	6,77983951568604	6,81056785583496	6,97372579574585	6,75760221481323	6,8597240447998	7,2602572...
13	1494_f_at	6,6067943572998	6,5712833404541	6,918903350830 08	6,59816122055054	6,47270965576172	6,30407094955444	6,35133028030396	6,63644123077393	6,46989583969116	6,5157737...
14	1598_g_at	8,59897994995117	9,7918815612793	8,86520004272461	9,64746284484863	8,48329830169678	9,70395183563232	9,66138648986816	9,40286540985107	9,85118770599365	8,9336395...
15	160020_at	8,83163356781006	7,91780471801758	7,99596357345581	7,33910703659058	7,59892654418945	7,43334627151489	7,84055519104004	7,88918304443359	7,60577011108398	7,5832681...
16	1729_at	7,39481639862061	7,3694863319397	7,86437368392944	7,45801401138306	7,0842719078064	7,27780818939209	7,2202262878418	6,88254499435425	7,16400051116943	7,0482087...
17	177_at	4,51230049133301	4,67804002761841	4,30085802078247	4,44540929794312	4,18501853942871	4,31222915649414	4,35478782653809	4,27660655975342	4,2929835319519	4,6313910...
18	1773_at	5,07591533660889	4,88596677780151	5,15866184234619	5,31757640838623	5,04552173614502	5,18654441833496	5,15675401687622	5,23807191848755	5,21436643600464	4,8422889...
19	179_at	7,95317220687866	8,50562858581543	8,46837902069092	8,52435874938965	7,6840181350708	7,85075187683105	8,55474090576172	8,54386043548584	8,29907131195068	7,6492094...
20	1861_at	4,47176647186279	4,72104549407959	4,65391826629639	5,37926816940308	4,11285877227783	4,89171314239502	4,89813852310181	4,83189296722412	4,93066787719727	4,2937450...
21	200000_s_at	7,71340274810791	8,28806591033936	7,85552835464478	8,96768283843994	7,57816076278687	8,74876117706299	8,60437965393066	7,56769227981567	8,59657955169678	8,7023010...
22	200001_at	8,79023456573486	9,42018222808838	9,00786113739014	10,0877122879028	8,74620151519775	10,50758934021	9,84454822540283	9,96816921234131	10,486686706543	9,5679540...
23	200002_at	10,0439348220825	10,0369787216187	9,67755699157715	9,997125625610 35	9,73553848266602	10,1909303665161	10,092212677002	9,58203125	9,90908813476563	10,282584...
24	200003_s_at	10,7595491409302	10,5324144363403	10,7752017974854	10,7745237350464	9,79375457763672	10,4330539703369	10,8360004425049	10,3128614425659	10,3913946151733	10,823941...
25	200004_at	9,86493015289307	9,88998889923096	9,39234066009521	9,87339687347412	10,0218982696533	10,2710742950439	9,82971382141113	9,09655094146729	9,59739303588867	10,119415...
26	200005_at	8,62274265289307	8,19677543640137	8,09800720214844	8,89187526702881	8,33071899414063	8,90583515167236	8,40023803710938	8,08819484710693	8,71073150634766	9,4119415...
27	200006_at	9,37872409820557	9,30368995666504	9,492995262146	9,86097049713135	9,76126289367676	9,76937770843506	9,73130416870117	9,09993934631348	9,49797534942627	10,407783...
28	200007_at	8,79712772369385	9,48137092590332	9,07268714904785	9,63889122009277	9,7044038772583	9,89698028564453	9,64221477508545	9,25639724731445	9,45306491851807	9,8333749...
29	200008_s_at	8,12747478485107	8,34486675262451	7,80654335021973	8,62778091430664	7,14637660980225	8,48642349243164	8,4940128326416	7,69132375717163	8,39269733428955	7,4823117...
30	200009_at	9,71415138244629	9,36833381652832	9,41126155853271	9,49438285827637	9,29961490631104	9,25418472290039	9,70801639556885	8,89087581634521	9,15568828582764	9,4586944...
31	200010_at	10,559775352478	10,8501291275024	10,5792179107666	10,6138200759888	10,3851442337036	11,0986299514771	10,8362112045288	10,2884702682495	10,8100490570068	11,150650...
32	200011_s_at	8,3623628616333	7,95209455490112	8,92352104187012	8,82637500762939	8,31218910217285	8,95365333557129	8,35516452789307	8,24428462982178	8,75062274932861	8,7724657...
33	200012_x_at	11,3398380279541	11,5184316635132	11,2614307403564	11,458869934082	11,3013668060303	11,5927362442017	11,5992069244385	10,9908199310303	11,4342565536499	11,515610...
34	200013_at	9,84353637695313	9,91970157623291	9,83184146881104	10,2002744674683	9,6886043548584	9,94840049743652	9,983229637146	9,42029094696045	9,75630378723145	10,304139...

Query executed successfully. | MIV\SQLEXPRESS (10.0 SP1) | MIV\Tomek (54) | MGR | 00:00:05 | 22283 rows

Fig. 1. An example of microarray data. Table section with experimental data - activity of 22,283 genes in 10 samples. 10 HG-U133A microarrays by Affymetrix were used. The data were pre-processed. Author's own data from DNA Microarray Laboratory of the Department of Molecular Biology of the Medical University of Silesia.

An interesting alternative to a relational model is the EAV model promoted by the authors [2, 7]. Similar to the relational model, the data are also stored in tables. However, with the EAV model, the tables include 3 columns (entity ID, attribute name and attribute value) [6, 7]. With specific reference to microarray data, EAV model tables [15, 16, 17] will include the following columns: sample/microarray ID, gene ID and expression value (Fig. 2).

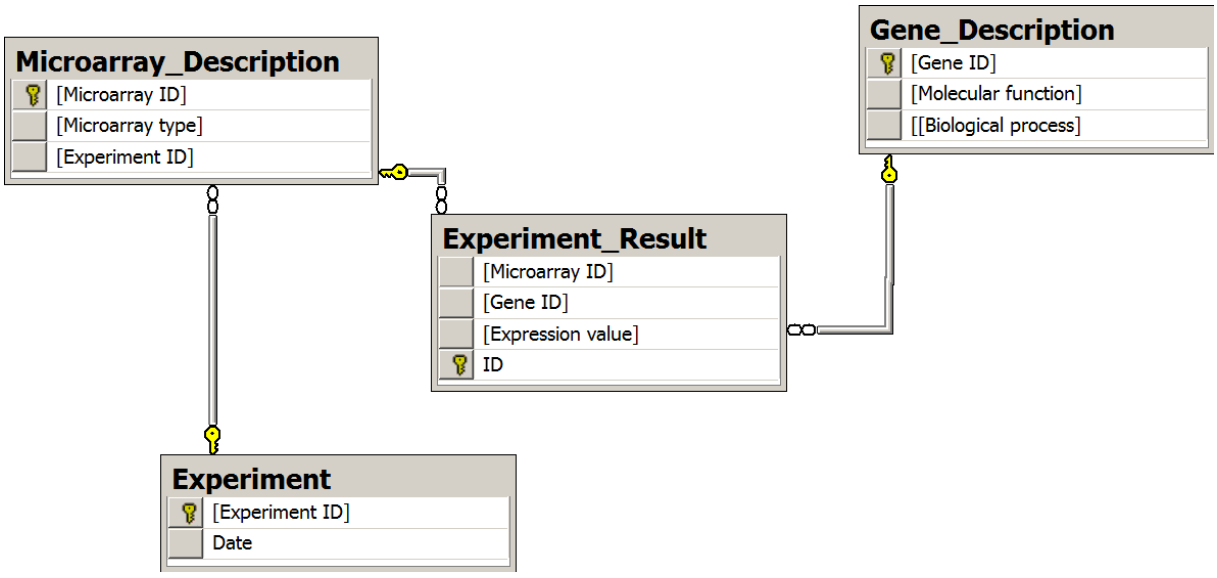

Fig. 2. Entity-relationship diagram. Simple database design based on EAV model. Author's own data.

EAV model application in database design allows storage of data from several microarray experiments in a single table, where the data are diversified by the microarray type and the number of samples. An advantage of this approach is that it does not limit the number of attributes, and it is not burdened with NULL values.

At the same time, the EAV data structure leads to an undesirable increase in the number of rows in the table [6]. Numerous rows are a significant impediment in creating complex SQL queries, since applied microarray data analysis methods (e.g., Significance Analysis of Microarrays [14] or cluster analysis [14]) require a standard data layout, as illustrated in Fig. 1. Therefore, a main criterion for usefulness of EAV model in a specific database system is the availability of a cross table feature. With this feature, data columns and rows can be selected and rotated to change the structure of data, creating a typical microarray experiment table, having the structure compatible with the EAV model.

2. METHODS

Two comparative analyses were conducted to investigate whether the EAV model can be used with an open-source database system for efficient DNA microarray data storage. The first analysis compared the most popular database systems in terms of accessibility of cross table development functionality. Systems offering no such feature were excluded from further analysis. Another criterion subject in conduced analysis concerns the processing of performance of DNA microarray data (consistent with the EAV model in selected DMBS systems). These analyses were conducted on a PC (CPU: Intel Core 2 Duo CPU, RAM 4 GB) with Windows 7 Professional 32 bit.

3. RESULTS

3.1. COMPARATIVE ANALYSIS OF AVAILABLE CROSS TABLE FEATURE IN DBMS

Several database management systems are currently available. The most popular versions, based on Google search engine ratings (see Figure 3) were included in the study. The comparative analysis covers only the free DBMS versions.

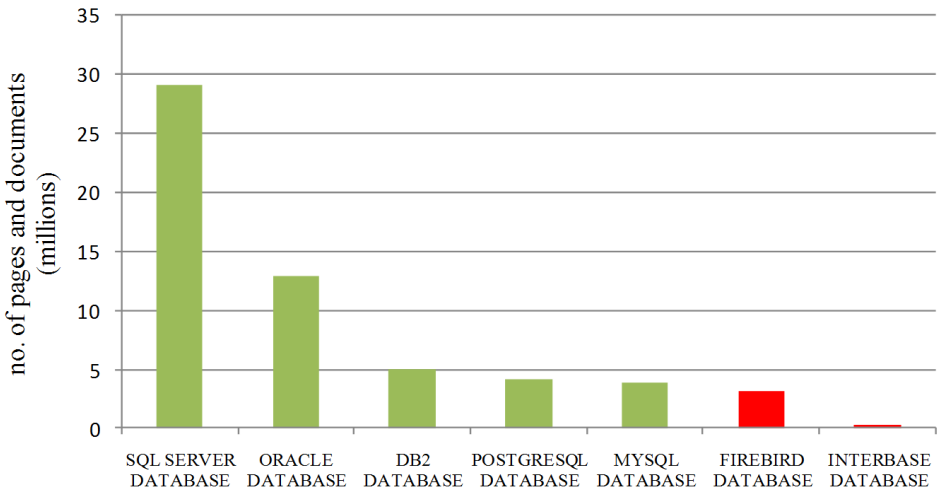

Fig. 3. Google search engine ratings of the DBMS. DBMS marked green were included in the comparative analysis. Author's own data based on [http://www.google.pl].

The following systems were included: Microsoft SQL Server 2008 Express Edition, Oracle Database 10g Express Edition, DB2 Express-C 9.7.2, MySQL 5.1 and PostgreSQL 9.0. The analysis demonstrates that a dedicated and convenient method of creating cross tables is only provided by Microsoft SQL Server 2008 Express Edition and PostgreSQL 9.0 (see Table 1).

Table 1. Benchmarks. List of selected database management systems regarding cross table feature [9, 10, 11, 12, 13]. Author's own data.

Database management system	Cross table feature
Microsoft SQL Server 2008 Express Edition	YES, PIVOT relational operator
Oracle Database 10g Express Edition	NO
DB2 Express-C 9.7.2	NO
MySQL 5.1	NO
PostgreSQL 9.0	YES, CROSSTAB function

Thus, Oracle Database 10g Express Edition, DB2 Express-C 9.7.2, and MySQL 5.1 systems were eliminated from further study, since no cross table feature is available.

3.2. COMPARATIVE ANALYSIS OF AVAILABLE CROSS TABLE FEATURE IN DBMS

The DBMS usability analysis included a comparison of the time required to create a cross table based on the EAV table (see Fig 1) for 25, 50, 100 and 200 DNA microarrays. Fig. 4 provides the test results. Significantly better results were obtained using Microsoft SQL Server 2008 Express Edition, where the cross table creating method is based on an embedded PIVOT relational operator.

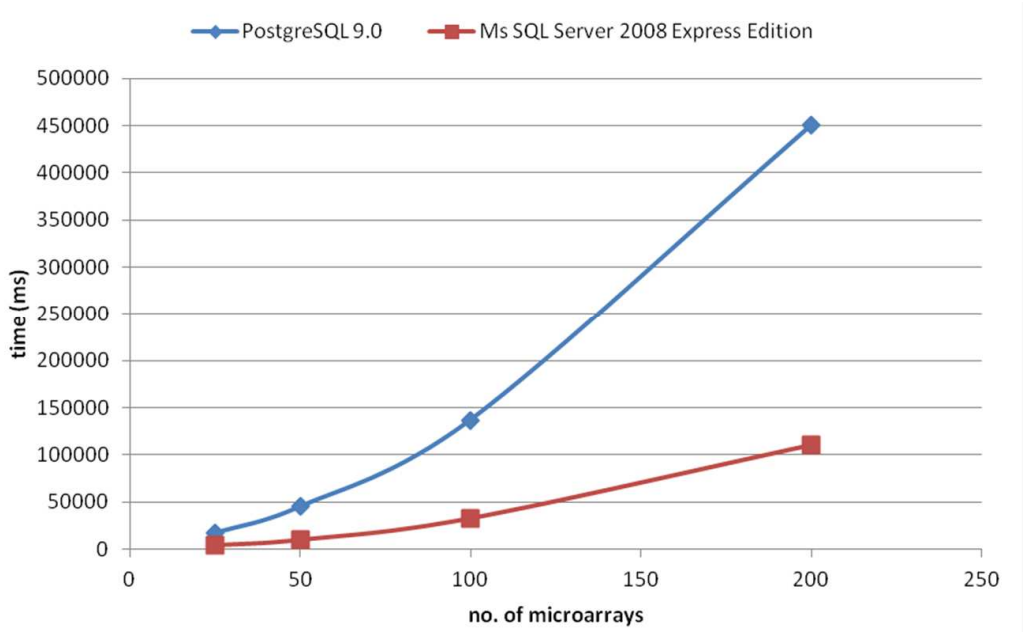

Fig. 4. Comparison of DBMS efficiency. Comparison of efficiency of cross table creation depending on the data quantity. Author's own data from DNA Microarray Laboratory of the Department of Molecular Biology of the Medical University of Silesia.

In the case of PostgreSQL 9.0, the cross table is generated with a standard function written in C programming language, which is much less effective.

4. CONCLUSIONS

The EAV method was shown to be eligible for use in freely available versions of DBMS software as an optimum storage model for DNA microarray data. In view of the data processing methods offered by developers (and in particular, the availability of a simple method for the creation of cross tables), Microsoft SQL Server 2008 Express Edition and PostgreSQL 9.0 are recommended as the preferred methods. Microsoft SQL Server 2008 Express Edition proved to be the most efficient in the tests; however, it cannot be considered as the best choice due to its limited maximum database capacity of 4 GB (10 GB in the R2 version).

5. ACKNOWLEDGEMENTS

We thank the Department of Molecular Biology of the Medical University of Silesia for the DNA microarray data. We also thank the reviewers for their critical comments which helped us improve our article.

Tomasz Waller and Damian Zapart received a scholarship under the project "DoktoRIS - Scholarship Program for Innovative Silesia" co-financed by the European Union under the European Social Fund.

BIBLIOGRAPHY

[1] TSOI L.C. ZHENG W.J.: A Method of Microarray Data Storage Using Array Data Type, 2007, Comput Biol Chem, Vol. 31, No. 2, pp. 143-147.
[2] MASYS D.: Database designs for microarray data, 2001, The Pharmacogenomics Journal, Vol. 1, pp. 232–233
[3] SPLENDIANI A., BRANDIZI M., EVEN G., OTTAVIO B., Pavelka N., PELIZZOLA M., MAYHAUS M., FOTI M., MAURI G., RICCIARDI-CASTAGNOLI P.: The Genopolis Microarray Database. BMC Bioinformatics 2007, 8(Suppl 1):S21.
[4] KILLION P., SHERLOCK G., IYER V.: The Longhorn Array Database (LAD): An Freely-accessible, MIAME compliant implementation of the Stanford Microarray Database (SMD), 2003, BMC Bioinformatics, pp. 4-32.
[5] SHAH S. P., HUANG Y., XU T., YUEN M., LING J., OUELLETTE F.: Atlas – a data warehouse for integrative bioinformatics, 2005, BMC Bioinformatics, pp. 6-34.
[6] GRAMACKI A., GRAMACKI J.: Modelowanie typu Entity-Attribute-Value w bazach danych, 2007, Pomiary Automatyka Kontrola, No. 5, pp. 54-56.
[7] HONG-HAI D., TORALF K., ERHARD R.: Comparative Evaluation of Microarray-based Gene Expression Databases, BTW 2003, In Proc. 10. Fachtagung Datenbanksysteme für Business, Technologie und Web, Leipzig.
[8] LENOIR T., GIANNELLA E.: The emergence and diffusion of DNA microarray technology, 2006, Journal of Biomedical Discovery and Collaboration, 1:1.
[9] PostgreSQL 9.0.5 Documentation - http://www.postgresql.org/
[10] Microsoft SQL Server 2008 Books Online (October 2009) - http://technet.microsoft.com/
[11] MySQL 5.1 Reference Manual - http://dev.mysql.com/doc/
[12] Oracle Database 10g Documentation Library - http://docs.oracle.com/
[13] DB2 database product documentation - https://www-304.ibm.com/support/
[14] RUSSELL S., MEADOWS L., RUSSELL R.: Microarray Technology in Practice, 2009, Elsevier Inc.
[15] THALLINGER G.G. et al., TAMEE: data management and analysis for tissue microarrays, 2007, BMC Bioinformatics, pp. 8-81.
[16] JOHNSON S.B.: Generic data modeling for clinical repositories, 1996, J Am Med Inform Assoc, No. 3, pp. 328-339.
[17] NADKARNI P.M. et al.: Organization of heterogeneous scientific data using the EAV/CR representation, 1999, J Am Med Inform Assoc, Vol. 6, pp. 478-493.