THE ETUMOUR DATABASE: A TOOL FOR ANNOTATION AND CURATION OF MULTIDIMENSIONAL HUMAN BRAIN TUMOR DATA

M. Julià-Sapé[1,2], M. Mier[3,4], M. Lurgi[2,3], F. Estanyol[3], X. Rafael[3,4], T. Delgado-Goñi[1,2], M. Camisón[2], M. Martinez-Bisbal[5,6], B. Celda[5,6], C. Arús[1,2], and C. eTUMOUR[7]

[1]CIBER-BBN, UAB, Cerdanyola del Vallès, Barcelona, Spain, [2]Bioquimica i Biologia Molecular, Universitat Autonoma de Barcelona, Cerdanyola del Valles, Barcelona, Spain, [3]Microart SL, Barcelona, Spain, [4]School of Informatics, The University of Edinbourgh, Edinbourgh, United Kingdom, [5]CIBER-BBN, Universitat de Valencia, Valencia, Valencia, Spain, [6]Quimica Fisica, Universitat de Valencia, Valencia, Valencia, Spain, [7]http://www.etumour.net

Introduction: Biological databases are becoming more common nowadays, with more than 1000 listed in 2008 by the Molecular Biology Database Collection [1]. In contrast, to our knowledge there are only 2 specialized medical databases for human brain tumor MRS data [2, 3], and they store a limited dimension of biological data (i.e. clinical information, MRI and MRS data).

Purpose: To develop and annotate a multicentre, web-based and curated database of human brain tumor patient data: the eTUMOUR project (eT) database (eTDB). To develop strategies and database-associated tools to allow curation [4, 5]. To annotate the eTDB with clinical, MRI, MRS (both *in-vivo* and *in-vitro:* HRMAS) and microarray data.

Methods: Design and user-interface: The eTDB was designed based on specifications obtained from users (clinicians, spectroscopists and molecular biologists) according to what would be an ideal way of organizing and storing the information collected following the acquisition protocols set up for the eT project [6] (Figure 1). The eTDB was implemented on an Oracle 10g Database Management System (DBMS) server running on a linux box. It is based on the Three-Tier Application model, which is composed by three main components: a client, a server, and a database. In the particular case of the eTDB, the client would be the browser application running on the side of the system's user, the server would be the web server (Apache Tomcat 5.5.17), whereas the database functionality is provided by the Oracle server mentioned above. The Graphical User Interface (GUI) developed as the front-end for the eTDB is based on Java Server Pages (JSP), which are dynamically-generated web pages that can receive information from the JavaBeans data structures running on the server side, which in turn maintain the application's information coming from the database. Additional functionalities provided by the system, such as file upload or the interface between the GUI and the JavaBeans, among others; are implemented using Java Servlets, a common and efficient way of doing this on Java-based web applications. All incoming patient data (*in-vivo* MRS and MRI) from the different manufacturers were anonymized via dedicated applets. ***Curation strategies:*** Curation was divided in two activities: Data entry (DE) and data validation (DV). **DE**: for each patient different data dimensions were divided into different sections, as follows: clinical, *in-vivo* SV MRS, *in-vivo* MV MRS, *in-vivo* MRI, *in-vitro* HR-MAS and microarrays. **DV**: two validation levels were set: 1st) "Quality control" (QC), in which the purpose was to establish for each section, if the data entered were in accordance with established acquisition conditions [6], for example, absence of acquisition artifacts in MRI, MRS and microarray data, or whether the diagnostic biopsy had been acquired from the same location where the in-vivo MRS had been acquired, among other checks. 2nd) "Quality assessment" (QA), in which the purpose was to establish whether the eTDB entries were the same as source records, i.e. hospital files. ***Curators' roles:*** Users' roles (writer, reader, and downloader) were set according to the tasks to be performed. In general, each person performing DE could edit and download data from his/her centre only, while both the "data manager" and the "database administrator" could edit all data. For QC, "senior curator" (SC) roles were defined so that the SC could edit defined fields for recording the quality of the data type of the centers assigned, but could not edit the raw data from the centre. In general, a three-SC's system was established for the following dimensions: histopathology, *in-vivo* SV and HRMAS, where two SCs had to agree on a judgment and in case of disagreement, a third one acted as tie-breaker, like in [3]. SC's did not judge cases from their own centers. For microarrays, an automated system was established, in which certain values were automatically read by the eTDB and fed into the DB as values (Figure 1), and depending on values and according to a pre-set algorithm, each microarray dataset classified as valid or not valid. A protocol for QA was established, with periodic, pre-defined samplings as in [3] in which each centre checked for the consistency of their own data. QA was performed on a section basis for each case, taking into account that there can be different levels of completeness depending on data type for each case. Previously to any QA check, the user had to close the section. ***Metrics***: Among others, metrics that account for data usability and completeness were devised. The concepts of "indispensable dataset", "ideal dataset" were defined being respectively the minimum amount of data that makes a case usable for its analysis and the ideal amount of data (per case) aimed for the project. ***Validated eTDB***: The subset of cases that had at least the indispensable dataset, that fulfilled QC and QA for the sections available were considered suitable for their inclusion into the validated-eTDB, which will serve as reference database for the eTUMOUR project.

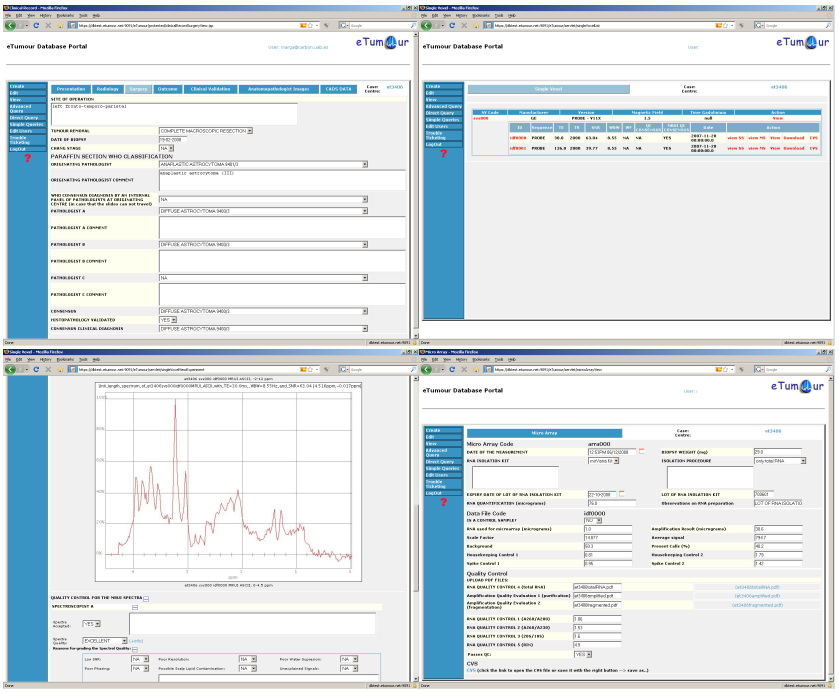

Figure 1*: Four representative screens from the eTDB. Top left: Clinical information. Top right: Single voxel MR data. Bottom left: Spectrum viewer, a spectrum automatically processed from the raw data file by the database. Bottom right: Microarrays information.*

Results: ***Design and user interface***: A relational database with more than 60 tables of metainformation and a file system dedicated to the storage of the different types of data coming from these experiments has been implemented. Automatic processing and visualization tools for SV and HR-MAS data have been implemented in order to allow QC and QA. ***Data stored***: Currently (Novembre 5th, 2008), there are 1929 cases (1625 prospective eT and 304 from [3]) with clinical information, of which 656 eT prospective cases fulfill the "indispensable dataset" requirements. ***QC and QA***: These are continuous activities throughout the time-span of the project and final results will be presented at the end of the project (Jan 31st, 2009). Presently, 879 prospective eT cases have consensus histopathology diagnosis and 1220 SV spectra have been analyzed by the SC spectroscopists.

Discussion/Conclusion: eTUMOUR has developed not only a database with multidimensional brain tumor patient data (*in-vivo*, transcriptomic, metabolomic) but also a set of strategies to curate and to ensure the usability of the entries. The eTDB complies with the 95/46/EU directive regarding data protection [7].

References: [1] Nucl. Acids Res. (2008) 36 (Supplement 1): D2. [2] Magn Reson Med (2002) 48:411–418. [3] Magn Reson Mater Phy MAGMA (2006) 19: 22–33. [4] PLoS Comput Biol (2006) 2(10): e142. [5] PLoS Comput Biol (2006) 2(10): e125 [6] http://www.eTUMOUR.net/ [7] European Parliament and Council. Directive 95/46/EC (1995) Official Journal 281:0031–0050.