300–302 *Nucleic Acids Research, 2001, Vol. 29, No. 1*

OrCGDB: a database of genes involved in oral cancer

Alan E. Levine* and David L. Steffen[1]

Department of Basic Sciences, The University of Texas Health Science Center Houston, Dental Branch, Houston, TX 77030, USA and [1]Biomedical Computing, Houston, TX 77005, USA

Received September 1, 2000; Revised and Accepted October 31, 2000

ABSTRACT

The Oral Cancer Gene Database (OrCGDB; http://www.tumor-gene.org/Oral/oral.html) was developed to provide the biomedical community with easy access to the latest information on the genes involved in oral cancer. The information is stored in a relational database and accessed through a WWW interface. The OrCGDB is organized by gene name, which is linked to information describing properties of the gene. This information is stored as a collection of findings ('facts') that are entered by the database curator in a semi-structured format from information in primary publications using a WWW interface. These facts include causes of oncogenic activation, chromosomal localization of the gene, mutations associated with the gene, the biochemical identity and activity of the gene product, synonyms for the gene name and a variety of clinical information. Each fact is associated with a MEDLINE citation. The user can search the OrCGDB by gene name or by entering a textword. The OrCGDB is part of a larger WWW-based tumor gene database and represents a new approach to catalog and display the research literature.

INTRODUCTION

Oral cancers comprise 5% of all cancers in the United States, with ~60 000 new cases and 15 000 deaths each year (1). Surviving patients suffer many forms of cosmetic and/or functional compromise due to the extensive surgical resection necessary to treat the tumors. It is clear that genetic information about oral cancer would be useful in the diagnosis and treatment of this disease. To meet the need for a readily accessible source of information about the genes involved in oral cancer, an Oral Cancer Gene Database (OrCGDB) has been developed (http://www.tumor-gene.org/Oral/oral.html).

The amount of biomedical research reported in the literature continues to increase at a dramatic rate. It is nearly impossible for a researcher or clinician to keep up with his or her own field let alone allied fields. In response to this dilemma, more efficient methods of information retrieval have been developed relying on ever improving computer technology. Bibliographic (e.g. MEDLINE) and factual [e.g. GenBank (2) or Online Mendelian Inheritance in Man, OMIM (3)] databases have been developed to meet this need. In the case of bibliographic databases, the references returned in response to a query must be read in order to obtain the desired fact (i.e. protein size, activating mutation, etc.). Even with the availability of online full text journals, this can be a laborious process. Factual databases can return specific facts that can be linked to a bibliographic database such as MEDLINE.

The long-term goal of the present work was the development of model systems for storing scientific information. Such systems would enable access to information on demand. As a short term step towards the fully structured database required by that long term goal, our strategy was to begin by constructing partially structured databases and incrementally increase the fraction of data therein which is structured. Previously, a Tumor Gene (4) and a Breast Cancer Gene (5) Database were developed. The OrCGDB extends the model of data representation found in the Tumor Gene Database to the genes associated with oral cancer. As of August 2000, the OrCGDB contained 15 genes known to be involved in oral cancer and 1367 facts associated with these genes.

DATABASE SOFTWARE

The data structure, WWW front end and other features are implemented in the database schema and approximately 5000 lines of database scripting language. A detailed description of this software has been published (4).

IDENTIFICATION OF REFERENCES

Facts in the OrCGDB are derived from research papers identified by searching MEDLINE. The search strategy used was based on that developed earlier for the Breast Cancer Gene Database (5), and is: (((((“oncogenes”[MeSH Terms] OR oncogene[Text Word]) OR “genes, suppressor, tumor”[MeSH Terms]) OR “proto-oncogene proteins”[MeSH Terms]) OR “protein-tyrosine kinase”[MeSH Terms]) AND ((((“Mouth neoplasms”[MeSH Terms] OR “Esophageal neoplasms”[MeSH Terms]) OR “Head & Neck neoplasms”[MeSH Terms]) OR “Laryngeal neoplasms”[MeSH Terms]) OR “Otorhinolaryngolic neoplasms”[MeSH Terms])).

Development and analysis of the OrCGDB search strategy is in progress, but work to date indicates that the above strategy retrieves about half the relevant papers, and that about half the papers retrieved are relevant. Although it is inconvenient that half the papers retrieved are irrelevant, missing half the relevant literature is the more serious problem. To counter that, a second set of searches are done which combine the names of the genes identified in the first search with the organ site specific search terms. Preliminary analysis suggests that this

*To whom correspondence should be addressed. Tel: +1 713 500 4497; Fax: +1 713 500 4500; Email: alevine@mail.db.uth.tmc.edu

increases the percentage of relevant papers retrieved to ~90%. An example of such a search for the gene MYC is: ((“genes, myc”[MeSH Terms] OR MYC[Text Word]) AND (((((“Mouth neoplasms”[MeSH Terms] OR “Esophageal neoplasms”[MeSH Terms]) OR “Head & Neck neoplasms”[MeSH Terms]) OR “Laryngeal neoplasms”[MeSH Terms]) OR “Otorhinolaryngolic neoplasms”[MeSH Terms]))

DATABASE CONTENT

The OrCGDB consists of a collection of facts for a given gene that are organized by topic. To date we have identified 15 genes as being associated with oral cancer. Table 1 lists these genes and the number of facts currently associated with each gene. The number of genes and facts should be considered as a current ‘in progress snapshot’ of the OrCGDB. Clearly, more genes will be identified as being important to oral cancer and this list will be continually updated.

Table 1. Genes associated with oral cancer

Gene name (symbol)	No. of facts in OrCGDB
BAX	18
BCL-2	125
Cyclin D1 (CCND1)	76
Cyclin-dependent kinase 4 (CDK4)	16
Cyclin-dependent kinase inhibitor 1A (CDKN1A)	91
Cyclin-dependent kinase inhibitor 1B (CDKN1B)	27
Epidermal growth factor receptor (EGFR)	68
ERB-B2	193
Heat shock protein 70 (HSP70)	5
MYC	170
Proliferating cell nuclear antigen (PCNA)	15
Transforming growth factor-α (TGFA)	20
Transforming growth factor–β (TGFB)	34
FAS antigen (TNFRSF6)	8
P53 (TP53)	501

DATA ENTRY

Voluntary curators entered all data into the OrCGDB. Literature references identified by the working search strategy were imported into the database and appear in a WWW interface data entry screen (see Supplementary Material). The curator reads the abstract, enters the gene locus and selects a topic from a dropdown menu. The topics include oncogenic activation, chromosomal localization, biochemical type, regulation and clinical information. The complete list of 31 topics and their definitions can be found in the Supplementary Material. The fact (up to 80 characters) is entered in a semi-structured format. A comment field (of unlimited length, not structured) is available to further elaborate on the fact entered. The entry screen also contains information necessary for interaction with the database software. The information, entered by the curator, is reviewed by an editor and revised if necessary. The fact is then published and available to the public via the WWW interface. Any authorized curator anywhere in the world can enter data through the password protected data entry interface.

DATA RETRIEVAL

Users can access the OrCGDB freely via a WWW interface and obtain a list of all the genes associated with oral cancer. The database can be searched either by gene name or by keyword. One complication of the tumor gene literature is the number of different names associated with many genes in the literature. To minimize this problem, the database search engine incorporates a synonym list as part of the database query. A gene name search therefore returns all relevant facts for a given gene when any one of the gene name synonyms is entered in the search field.

The result of a gene name query is returned to the user as a page organized by topic. Under each topic is a list of facts with the reference linked by hypertext to PubMed. A link to the gene in OMIM is also presented on this data page.

A search by keyword returns a list of genes that contain the word in at least one fact associated with that gene. By clicking on the gene name the user is returned the same information as if gene name was originally searched for.

To give one concrete example, the fact ‘Alters bcl-2 expression in oral keratinocyte cell lines’ was entered by a curator for the gene TP53 under the topic ‘Function’. If the user searched for the gene names ‘TP53’ or ‘P53’ or for the text words ‘keratinocyte’ or ‘expression’, the page for the gene P53 would be returned with this associated fact.

DISCUSSION

A database containing facts about genes involved in oral cancer has been created (http://www.tumor-gene.org/Oral/oral.html). As of August 2000, the OrCGDB contains 15 gene names and 1367 associated facts. We do not believe that these are the only genes involved in oral cancer. As work on building the OrCGDB continues, more genes will be identified and many more facts entered for the new and existing genes. The Cancer Genome Anatomy Project (CGAP) has to date identified 1445 genes, of which 195 have only been identified in oral cancer or adjacent normal oral tissue (6). By combining the results of all approaches, including CGAP, microarray technology, laser capture microdissection (7) and the general literature, a total of 1891 genes have been associated with oral cancer. We suspect that only a fraction of these will ultimately be identified as genes that are targets for cancer causing mutations, the definition of tumor gene used in the OrCGDB. The final list of genes in the OrCGDB will almost certainly become larger.

The OrCGDB is part of a larger Tumor Gene Database that is structured to allow sharing of information between organ-specific databases. Therefore, when a user searches for facts about a gene involved in oral cancer, the results of that query include all information in the database associated with that gene. For example, searching for information about the involvement of p53 in oral cancer will yield all the facts in the database about p53. This is valuable because findings about a gene could be of general importance and not just important to the gene’s involvement in a specific cancer type.

The information returned by the current version of the OrCGDB includes hyperlinks to the MEDLINE reference

allowing ready access to the source of the fact. Links to other sources such as OMIM allow the user to obtain more information from other sources. The linking of factual and bibliographic databases can be bi-directional. Including links to the OrCGDB through the 'LinkOut' feature of PubMed is an example of this. The linking of these databases in a bi-directional manner would allow access to the maximum amount of information independent of the source of entry.

SUPPLEMENTARY MATERIAL

A table of topics used in the Oral Cancer Gene Database and a sample data entry screen are available at NAR Online.

REFERENCES

1. Slavkin,H.C. (1996) The war on oral cavity and pharyngeal cancer. *J. Am. Dent. Assoc.*, **127**, 517–520.
2. Benson,D.A., Boguski,M.S., Lipman,D.J., Ostell,J., Ouellette,B.F., Rapp,B.A. and Wheeler,D.L. (1999) GenBank. *Nucleic Acids Res.*, **27**, 12–17.
3. McKusick,V.A. (1998) *Mendelian Inheritance in Man. Catalogs of Human Genes and Genetic Disorders.* Johns Hopkins University Press, Baltimore, MD.
4. Steffen,D.L., Levine,A.E., Yarus,S., Baasiri,R.A. and Wheeler,D.A. (2000) Digital reviews in molecular biology: approaches to structured digital publication. *Bioinformatics*, **16**, 639–649.
5. Baasiri,R.A., Steffen,D.L., Glasser,S.R. and Wheeler,D.A. (1999) The breast cancer gene database: A collaborative information resource. *Oncogene*, **18**, 7958–7965.
6. Shillitoe,E.J., May,M., Patel,V., Lethanakul,C., Ensley,J.F., Strausberg,R.L. and Gutkind,J.S. (2000) Genome-wide analysis of oral cancer—early results from the Cancer Genome Anatomy Project. *Oral Oncol.*, **36**, 8–16.
7. Leethanakul,C., Patel,V., Gillespie,J., Pallente,M., Ensley,J.F., Koontongkaew,S., Liotta,L.A., Emmert-Buck,M. and Gutkind,J.S. (2000) Distinct pattern of expression of differentiation and growth-related genes in squamous cell carcinomas of the head and neck revealed by the use of laser capture microdissection and cDNA arrays. *Oncogene*, **19**, 3220–3224.