LATEST PAPERS SEARCH for PAPERS Printer-friendly PDF Comment(s)

>Introduction >Database structure >Database functionalities >Database curation >Perspectives >References >Contact author

Cancer Immunity, Vol. 1, p. 3 (30 March 2001)

Towards a Cancer Immunome Database

Victor Jongeneel

Ludwig Institute for Cancer Research, Office of Information Technology, Epalinges, Switzerland

Keywords: human, tumor antigens, factual databases

Introduction

The reality of the human immune response to cancer is now beyond question. The accumulated experimental evidence is overwhelming, documenting both humoral and cellular responses to an increasing number of specific antigens. Because the amount of available information is increasing at such a dizzying rate, it is clear that a well-organized and professionally curated repository has become indispensable for scientists in the field.

The Academy of Cancer Immunology, with support from the Ludwig Institute for Cancer Research (LICR), has taken up the task to setup a database that will present through a single point of access information about all of the gene products against which an immune response has been documented in cancer patients. This project is a continuation, in a more comprehensive and organized form, of the SEREX database maintained by the LICR since the fall of 1997 (1). It will be linked to the online Journal of the Academy, Cancer Immunity, as well as to continuing efforts to characterize in more detail the antigens discovered by the SEREX program. The project has also received support from the European Union, through the funding of the European Cancer Immunome Program (EUCIP). It is my aim in this short document to set out our plans for the structures and functionalities that will be incorporated into the Cancer Immunome Database.

Database structure

To reflect the complexity of the knowledge about cancer antigens, the database will be organized in multiple tables, which will be linked to each other using common keys (a standard relational database structure). A central role will be played by a table describing the **genes** that encode the antigens. For each of these, we will document the sequence of the gene itself, of the mRNA(s) it encodes, and of the protein(s) translated from these mRNAs. Whenever possible, these sequences will be taken from the RefSeq database (2) maintained at the NCBI (for nucleic acids) and from the SWISS-PROT database (for proteins) (3). If the gene sequences have not been fully characterized, we will try to reconstitute them from partial expressed sequence tags (EST) and genome data, index them based on their poly(A) proximal sequences, and link them to the UniGene database (4) when applicable. Ancillary information, such as chromosome map position and genome coordinates, will also be included. Additionally, we hope to document somatic events affecting the gene in tumors, such as loss of heterozygosity (LOH), amplification, homozygous deletion, methylation status and point mutations.

One gene can be the source of multiple antigenic **epitopes**, which can themselves be either **peptides** presented by **class I** or **class II** MHC molecules and recognized by T cells, and **subsequences** (linear or conformational epitopes) within proteins being recognized by antibodies. For antigenic peptides, we will document the **cells** ($CD4^+$ or $CD8^+$ **clones**) that recognize them. For epitopes on intact proteins, it will be **patient sera** and/or **monoclonal antibodies**. We will also document the frequency with which epitopes are recognized by sera or cells from normal and diseased individuals with various types of pathologies.

Much of this information has to be linked to anonymized data about **patients**. At the least, this should include the type of cancer from which they were suffering, and when available some indication about the stage of the disease at the time samples were obtained. Patient data should of course be linked to all of the reagents that were derived from the patient, such as serum, cell lines and clones, cDNA libraries, etc. The **SEREX** methodology, as documented in the current SEREX database, has generated its own specific information: cDNA clones (linked to genes), with length and end-points of inserts and sequences, cDNA libraries, and sera.

Finally, where appropriate, we would like to give access to some of the experimental evidence documenting the facts stored in the database. These could be microscope images, serological results, cytotoxic assays, or any of the many types of data collected in a laboratory setting.

One could go on forever, or almost, enumerating all of the individual types of information that should be available in the database. To a large extent, these will be determined by the contributing scientists. The central principle is that we want to create a single resource that will integrate information from many different sources into a coherent and interlinked environment.

Database functionalities

As with any database, we want to enable visitors to find the information they are looking for. This will be implemented using a few forms allowing them to specify their queries in a simple and intuitive manner. Reports will then present the most pertinent information, and links to additional data.

But the Web offers ways to add many additional functionalities, and hosts a large number of public information resources that are relevant to our project. Those that come to mind immediately are bibliographic databases [PubMed (5)], sequence databases [EMBL (6)/GenBank (7), SWISS-PROT (3) and RefSeq (2)], genetic mapping databases [LocusLink (8), GDB (9)], and gene information databases [OMIM (10), GeneCards (11)]. Records in the Cancer Immunome Database will be linked to the appropriate external source(s), so as to make this additional information immediately available to the user. It may also become evident at some point that more specialized databases (e.g. mutation databases, gene-specific databases, and the like) contain information relevant to our project. Every effort will be made to identify such sources of information and to link to them where appropriate.

Similarly to the SEREX database, we plan to make a set of sequence analysis tools available to database users, so that they can detect features of individual nucleic acid or protein sequences that may not be explicitly documented in the database. These tools can also be very useful to scientists contributing new information, as they allow a first level of analysis of the data and its documentation in the database. While the tools will be very similar to those already incorporated into the SEREX environment, they will reach further by making full use of the draft human genome and transcriptome data, and in particular of the mapping of partial transcripts to the genome that is being performed at the LICR and the National Cancer Institute.

Database curation

The Cancer Immunome Database will have multiple partners, each with different expectations. The Academy of Cancer Immunology and its Journal, Cancer Immunity, expect a stable and authoritative source of documentation about the human immune response to cancer. The EUCIP and SEREX programs will want a living resource documenting ongoing progress in understanding the molecular events driving this response. The participants in the LICR's SEREX program may also expect to see a continuity in the way their data are organized and presented. While these expectations may seem contradictory, I believe that they can be reconciled by presenting views of the data that reflect the needs of this varied community of users. This can be implemented if each user is assigned a "role" when s/he logs in or accesses the database through a public portal, and if this role then determines which data the user will see, and in which form.

An essential element in the success of any database is the quality of the information that it contains. This quality can be attained, and maintained, only if a team of dedicated and knowledgeable scientists is committed to keeping the information up to date, and to verify it using their own experience and the scientific literature. We will be relying on the knowledge available within the LICR and its Affiliates, and from the EUCIP partners, to ensure the accuracy of the database. In particular, Prof. Boon and his colleagues at the Brussels Branch of the LICR will maintain the database of peptides presented by MHC class I and II molecules, while Drs Chen and Scanlan at the New York Branch will curate the information about epitopes recognized by antibodies. In addition, we will be collaborating with the group of Prof. Kolchanov in Novosibirsk, where several scientists will be checking the consistency and quality of the data, and assist in data entry. This group will also participate in the design and maintenance of the database.

Perspectives

The Cancer Immunome Database will be the first publicly available resource aiming to document all aspects of the immune response to cancer. It will have strong institutional support from the Ludwig Institute for Cancer Research, the Cancer Research Institute, the Academy of Cancer Immunology, and the European Union. We fully expect that the combined expertise of its designers and curators, who are among the most knowledgeable scientists in the field, will make this database an indispensable resource. We also hope that it will survive the vicissitudes of time, and be able to maintain itself as an up-to-date, authoritative and reliable institution.

Abbreviations

LICR, Ludwig Institute for Cancer Research; SEREX, serological analysis of recombinant expression cloning.

References

1. SEREX database. URL: http://www.licr.org/SEREX.html
2. RefSeq database. URL: http://www.ncbi.nlm.nih.gov/LocusLink/refseq.html
3. SWISS-PROT database. URL: http://www.expasy.ch/sprot/
4. UniGene database. URL: http://www.ncbi.nlm.nih.gov/UniGene/
5. PubMed database. URL: http://www.ncbi.nlm.nih.gov:80/entrez/query/static/overview.html
6. EMBL database. URL: http://www.ebi.ac.uk/embl/
7. GenBank database. URL: http://www.ncbi.nlm.nih.gov/Genbank/GenbankOverview.html
8. LocusLink database. URL: http://www.ncbi.nlm.nih.gov/LocusLink/
9. GDB database. URL: http://gdbwww.gdb.org/
10. OMIM database. URL: http://www3.ncbi.nlm.nih.gov/omim/
11. GeneCards database. URL: http://bioinformatics.weizmann.ac.il/cards/

Contact

Address correspondence to:

Victor Jongeneel
Ludwig Institute for Cancer Research
Office of Information Technology
Chemin des Boveresses 155
CH-1066 Epalinges
Switzerland
Tel: + 41 21 692 59 94
Fax: + 41 21 692 59 45
E-mail: Victor.Jongeneel@licr.org