SHORT COMMUNICATION

Data standards for minimum information collection for antibody therapy experiments

M. Yong1, B. Tolner, S. Nagl, R.B. Pedley, K. Chester, A.J. Green, A. Mayer, S. Sharma and R. Begent

UCL Cancer Institute, University College London, Paul O’Gorman Building, London WC1E 6BT, UK

1To whom correspondence should be addressed. E-mail: m.yong@ucl.ac.uk

Research groups developing antibody therapies generate diverse data sets; the value of these sets would be compounded when shared or amalgamated. A complete amalgamation of diverse data sets requires data standards for information collection during experiments. We propose to define elements of the data standards in the form of common data elements (CDEs) in order to clarify each experiment’s targets and data values. We have created a set of core information elements which we suggest should be collected from antibody therapy experiments. We propose these as a basis for community consultation with a view to defining a set of data standards which can be developed under the auspices of the Antibody Society.

Keywords: antibody therapy/common data elements/controlled vocabulary/data standards

Introduction

A data standard which gains the widespread support of its target research community can bring huge economic benefits (Ball et al., 2004) in addition to fulfilling a fundamental requirement to enable data sharing, as implemented by the US National Institutes of Health (NIH) and partners in the UK National Cancer Research Institute (NCRI). A discussion on the readiness of biological research for data standards (Anonymous, 2006) generated some positive responses (Jonker et al., 2007; Klipp et al., 2007; Reddington et al., 2007) from the community.

Data standards for minimum information collection have been defined for diverse fields from genome sequencing (minimum information about a genome sequence, MIGS) (Field et al., 2008), to the recording of molecular interaction experiments (minimum information required for reporting a molecular interaction experiment, MIMIx) (Orchard et al., 2007). Experiences from further initiatives such as standards for functional genomics (Microarray Gene Expression Data, MGED) (Ball and Brazma, 2006) and proteomics (Human Proteome Organisation Proteomics Standardisation Initiative, HUPO-PSI) (Hamacher et al., 2008) suggest that a venture to create data standards for collecting information from antibody therapy experiments could be similarly rewarding.

All too often, resources have been wasted when experiments have to be repeated in order to record key elements that could have easily been recorded the first time. When data sharing among groups of researchers fails, it is often because each group has developed its own culture regarding fields of information describing the research aim, assessment methods and results to record. It is therefore difficult to make valid comparisons between experiments from different groups. Support for a set of data standards from researchers means an agreement between researchers to collect a set of minimum information from each experiment, in order that the rest of the community may compare experiments and avoid unnecessary repetition.

In addition, as the biological sciences move away from the reductionist era to approach the age of integrative translational medicine, the need for meaningful data sharing between disciplines and disease areas becomes ever more pressing. For example, statistical and machine learning methods are effective at detecting complex behaviours. This is potentially very useful in the field of antibody therapy research where multiple parameters interact in a complex manner. However, these methods require large data sets in order to produce robust algorithms and statistically significant results. Although a single antibody therapy experiment is unlikely to produce enough data, an amalgamation of data from different experiments will improve statistical power and may be sufficient for a machine learning method to yield previously unseen and statistically valid interactions which introduce fresh viewpoints. This can only work if a large number of researchers are willing to record defined minimum information about their experiments.

We advocate the approach of early stage community involvement in the development of a data standard for collecting minimum information from antibody therapy experiments. Efforts of similar initiatives recommend this as the most effective way to ensure that the output generated is both pragmatic and useful, and that such standards are accepted by the community (Lindon et al., 2005; Patel et al., 2006; Mohanty et al., 2008).

Methods

Members of the Antibody Society collaborated to identify key elements which should be recorded for antibody therapy experiments. These elements form the backbone of the data standard and consist of properties to describe target proteins and therapeutic agents as well as cellular, animal and clinical models.

The elements are defined in the form of a data structure called common data elements (CDEs) in compliance with International Standards Organisation (ISO 11179). The purpose of using CDEs is the clarity gained in terms of the meaning and content of the information to describe element. As each element is independently created as a CDE, an object-oriented structure is given to the data standard. New pieces of information can be added and outdated fields...
removed without affecting the rest of the minimum information data set fields.

**Guidelines for information about antibody therapy experiment tree**

The principal components of antibody therapy are identified and organised, in order to generate guidelines for information about antibody therapy experiments (GIAATE). The GIAATE is presented as a tree structure, with branches to broadly denote fields of information and leaves denoting individual properties to describe those fields. Further detail can be added in additional levels of the hierarchy.

GIAATE incorporates information from established community databases where possible. The information from those databases are referred to using the accession numbers, or in the case of International ImMunoGeneTics Information system (IMGT) (Lefranc, 2008), the sequence of the protein. Other databases included are Protein Data Bank (PDB) (Berman et al., 2002), Universal Protein Resource (Uniprot) (Bairoch et al., 2005), Systematized Nomenclature of Medicine Clinical Terms (SNOMED CT) (Wasserman and Wang, 2003) and Pharmacogenetics and Pharmacogenomics Knowledge Base (pharmGKB) (Owen et al., 2008).

Using information from established databases has the benefits of both additional clarity and simplifying the data collection process. For example, researchers who want to record the structure of the target protein of an experiment would provide the Protein Data Bank ID of the protein. This removes any ambiguity about the target protein, whereas simply naming a protein may create confusion as researchers use different naming conventions and notations. This also provides an instantly accessible source of information to other researchers. In addition, users who provide the protein ID for a database such as the UNIPROT simultaneously provide a link to other pieces of information managed by the database regarding the protein (e.g. protein sequence, interactions and genetic variations).

**Common data elements**

A CDE is a data structure that defines a datum in terms of a data element concept and a value domain. A data element concept is made of a definition of an object and a definition of a property of the object. The value domain defines the property’s range of possible values. Defining the GIAATE tree with CDEs removes any ambiguity about the pieces of information to be recorded, both in terms of meaning and range of permissible data values.

The data elements required for GIAATE may already exist as ISO 11179—compliant CDEs or may need to be defined de novo. When data elements are defined and shared, they can be used in the protocols for experiments relating to antibody therapy and a body of data which can validly be compared will begin to build. When these data are shared and re-used, the development of effective and safe therapeutics will be accelerated and made more cost-effective. Shared data are still subject to protection of intellectual property and confidentiality (Contreras, 2008).

We have used two approaches to developing the CDEs for GIAATE: we have found and reused some CDEs created by other institutes, and we have attempted to create the rest. One of the caBIG resources maintained by the National Cancer Institute is a metadata registry called caDSR (Warzel et al., 2003). caDSR allows users to upload new CDEs; it examines the CDEs for ISO compliance and ensures that no similar CDEs have been previously stored by caDSR. caDSR also provides a search engine which allows users to search for and download CDEs which have been accepted by the service.

Reusing caDSR-sourced CDEs in the GIAATE means that our data standards are defined using terms which have been similarly used by other institutions, and which carry the same meaning and data value range. The sharing of CDEs is strongly encouraged for reasons of flexible semantic data integration. The set of CDEs that are sourced from caDSR consists mostly of fields that overlap with other research areas.

We have also created some new CDEs which have not yet been defined in the caDSR. Our own CDEs are populated using definitions from the National Cancer Institute Thesaurus (Fragoso et al., 2004). The thesaurus contains a set of definitions which are updated regularly and many research groups use these definitions in order to provide a shared meaning to their terms. For our purpose of creating data standards for antibody therapies, the NCI Thesaurus provides a controlled vocabulary which can be openly accessed by researchers wishing to propose new elements or more accurate definitions to any of the existing elements of the GIAATE tree. In some cases, we were also required to define new relationships for the new CDEs; these relationships are created as an extension to the ones provided by the NCI Thesaurus. Our CDEs were created using the Protégé software.

However, there are fields in the GIAATE which utilise terms not defined in the NCI Thesaurus. Overwhelmingly, these terms describe properties which are specific to antibody therapy, for example, the expression system used for antibody production. Members of the Antibody Society are working to define a set of definitions for these terms. These new terms will be submitted to the NCI Thesaurus after being vetted by the Antibody Society, but are currently kept as our own extension to the thesaurus.

**Results**

GIAATE is included in the Supplementary Materials section available at PEDS online, or can be viewed through the Antibody Society website (http://www.antibodysociety.org/data/giaate.php).

Every field on the GIAATE tree has been recorded in the form of a data structure called CDEs; this is done with the purpose of clarifying every piece of information that is to be recorded. A similar approach in defining data standards for integrating. The set of CDEs that are sourced from caDSR consists mostly of fields that overlap with other research areas. The sharing of CDEs is strongly encouraged for reasons of flexible semantic data integration. The set of CDEs that are sourced from caDSR consists mostly of fields that overlap with other research areas.

We have also created some new CDEs which have not yet been defined in the caDSR. Our own CDEs are populated using definitions from the National Cancer Institute Thesaurus (Fragoso et al., 2004). The thesaurus contains a set of definitions which are updated regularly and many research groups use these definitions in order to provide a shared meaning to their terms. For our purpose of creating data standards for antibody therapies, the NCI Thesaurus provides a controlled vocabulary which can be openly accessed by researchers wishing to propose new elements or more accurate definitions to any of the existing elements of the GIAATE tree. In some cases, we were also required to define new relationships for the new CDEs; these relationships are created as an extension to the ones provided by the NCI Thesaurus. Our CDEs were created using the Protégé software.

However, there are fields in the GIAATE which utilise terms not defined in the NCI Thesaurus. Overwhelmingly, these terms describe properties which are specific to antibody therapy, for example, the expression system used for antibody production. Members of the Antibody Society are working to define a set of definitions for these terms. These new terms will be submitted to the NCI Thesaurus after being vetted by the Antibody Society, but are currently kept as our own extension to the thesaurus.

| Table 1. Proportion of CDEs which were reused from caDSR, defined using Protégé and not yet defined, according to each branch of the GIAATE tree |
|-----------------|----------|---------|------------|
| GIAATE tree branch | Reused | Defined | Not yet defined |
| Target           | 6       | 5       | 1           |
| Therapeutic      | 6       | 2       | 5           |
| Molecular model  | 8       | 5       | 0           |
| Animal model     | 14      | 6       | 6           |
| Patient          | 13      | 6       | 8           |
GIAATE tree. Terms which are not yet defined are mostly terms that are specialised to the antibody therapy field. The category to which element belongs is shown on the GIAATE tree in the Supplementary Materials section available at PEDS online.

To illustrate, Fig. 1 shows a portion of the GIAATE which relates to the target protein of an antibody therapy experiment. The annotation on each piece of information on the figure denotes its progress on the CDE-creation front. For example, ‘Protein Data Bank ID/PDB ID’ is a CDE which has been successfully downloaded from caDSR. A CDE for ‘Elimination pathway’ has been created using Protégé. Finally, ‘Rate of Replacement’ has not yet been formulated as a CDE because there are no appropriate definitions that describe this term in the NCI Thesaurus.

**Discussion**

Minimum information collection guidelines have been successfully developed and widely adopted in other fields. The novelty of creating guidelines to antibody therapy experiments lies in the fact that these guidelines will encompass diverse parameters reflecting the many aspects of antibody therapy; as a result, there are fields of information which may be critical to one set of experiments but irrelevant to another.

We have identified a set of core parameters whose recording we advocate in antibody therapy experiments; this is described in GIAATE. We have also created a set of CDEs in accordance with GIAATE to clarify the meanings and permissible values to these parameters.

Researchers who collect the minimum set of information as presented in GIAATE will be able to share their data and make valid comparisons between experiments. An example of this can be illustrated using two pivotal publications (Cutsem et al., 2007; Jonker et al., 2007) of well-known therapeutics as case studies. Panitumumab (Cutsem et al., 2007) and Cetuximab (Jonker et al., 2007) are both monoclonal antibody therapies directed against epidermal growth factor receptor (EGFR) activity in colorectal cancers. Few GIAATE elements were extracted from the publications, but it was found that both publications contained information detailing model clinical features and particularly the drug administration route, dose and time course. The Panitumumab (Cutsem et al., 2007) study included additional information on the concentration of the target molecule in the tumours by providing the percentage of cells with EGFR membrane staining as well as staining intensity. The publications are comparable as clinical trials as they both follow the data standard set by Consolidated Standards of Reporting Trials (CONSORT). However, in terms of experimental detail provided, the publications are not comparable as they do not contain the same information.

The idea of introducing long-term guidelines which are widely adopted by the antibody therapy community and are flexible enough to include novel techniques and data fields as they arise is certainly ambitious. It becomes more feasible if the implementation of the data standard has services in place...
to allow researchers to easily record feedback and add data elements which they feel are necessary components of their experiments.

Users should be the main source for garnering feedback regarding the feasibility of collecting information contained in GIAATE. The first stage has already begun, we have created software containing forms which allow users to view the meanings of each CDE, and fill in the appropriate values according to an experiment they have conducted. In the near future, these forms will be made publicly available on the Antibody Society website. Users may completely or partially fill the forms; any feedback would be useful in assessing the clarity and feasibility of this data standard.

Feedback so far on the selection of key elements to represent the minimum information has generated the viewpoint that these data standards should not be too strict as to constrain the scientist’s approach to a problem. GIAATE can help in clarifying the aims of an experiment but as the antibody therapy field is so broad, it should be viewed as a general guideline and not a constrictive data standard for every experiment. Researchers who require new elements necessary for describing their experiments are encouraged to build new CDEs and share them with the rest of the community.

Funding
This work was supported by the UCL Experimental Cancer Medicine Centre, Cancer Research UK and the UCL Cancer Institute Research Trust.

References

Received October 1, 2008; revised December 19, 2008; accepted January 5, 2009

Edited by Dario Neri