Letter to the Editor

Data storage: bringing us a step closer to data sharing?

The work by Saito et al. (2005) presented in a recent issue of this journal is laudable and timely for two reasons. First of all, it raises awareness about the importance and the challenges of data storage and management, an issue the field of nutrition cannot ignore. Second, the article reminds us that, beyond data storage, data sharing is fundamental to establishing such resources. As pointed out by the authors, out of 250 publications currently stored in their tool, only a handful are actually available for comprehensive review (i.e. include raw data). Surveying PubMed with keywords such as ‘adipose tissue’, ‘microarray’, ‘obesity’ and ‘diabetes type 2’ identifies over thirty array-based publications (Cao et al. 2001; Atzmon et al. 2002; Barta et al. 2002; Guan et al. 2003; Almind & Kahn, 2004; Sreekumar et al. 2002; Castro-Chavez et al. 2003; Fujiwara et al. 2003; Gabrielson et al. 2003; Tanaka et al. 2003; Lopez et al. 2003; Moraes et al. 2003; Fujiwara et al. 2003; Nakazato et al. 2003; Almind & Kahn, 2004; Becker et al. 2004; Crott et al. 2004; Deng et al. 2004; Dhabhi et al. 2004; Much et al. 2004; Recinos et al. 2004; Tseng et al. 2005; Tsuda et al. 2005; van Breda et al. 2005; van Schothorst et al. 2005; Xiao et al. 2005; Yagil et al. 2005), accounting for roughly 600 hybridisations worth of data. Assuming 8500 genes per array and twenty hybridisations on average per experiment, this corresponds to over 5 million data points. With little doubt, these datasets could be used as seeding material for establishing a nutrient–gene interaction knowledge base. Sadly for the nutrition community, only one out of these thirty datasets has been deposited in a public repository (GSE1392) (Much et al. 2004) and, for that one, the raw data are not available.

Is so the nutrition community really ready for data sharing? Is it that effective data sharing is just too complex to be feasible yet? Well, one can arguably say no to the latter, considering the vast amount of effort spent in the field of microarrays to enhance data exchange and access. It is almost 4 years since the publication of the ‘Minimum Information About a Microarray Experiment’ (MIAME) paper by the Microarray Gene Expression Society (MGED) (Brazma et al. 2001). The same group of individuals has come up with an object model for database implementation (MAGE-OM) that would enable data persistence and an XML format (MAGE-ML) to enhance data exchange between institutions. Last, MGED has made an attempt to provide the community with a common set of descriptors (controlled vocabularies) arranged in an ontology referred to as ‘MO’, standing for MGED Ontology. Why generate such an ontology? If MIAME defines the amount of information to describe, it does not formulate any recommendation about which terminology should be used to provide annotation. Hence, one can be MIAME compliant with simple free text. This is where the sting is. Making sense of free text is computationally expensive and currently available text-mining techniques clearly under-perform. A way around this hurdle is to promote the use of community-vetted annotation standards, such as controlled vocabularies and ontologies. Their use eventually ensures that terms and descriptors are employed consistently throughout a community, which is made resource aware.

This is one drawback of the work by Saito et al. (2005), as the resource they have set up makes little use of controlled vocabularies. Still, their work should be viewed as an essential reminder that more attention should be paid to consistent data annotation and at establishing curated resources.

But why make so much fuss about annotation and does missing annotation really matter? Again, let us take a simple example. How does the presence or absence of information impact the understanding of a study? What if one does not report the sex of the animal or the strain used or the developmental stage or the target organ? Can the information be reconstructed simply by reading the articles? The latter task is certainly manageable when dealing with, say, four or ten datasets, but what about 1000 datasets?

This is where journals and funding agencies can play a central role by insisting on having raw data and metadata (i.e. all the necessary descriptive ancillary data enabling use of raw data files) deposited in public databases (for example, Gene Expression Omnibus (GEO) at the National Center for Biotechnology Information (NCBI) (Barrett et al. 2005) or ArrayExpress at the European Bioinformatics Institute (EBI) (Parkinson et al. 2005)). Furthermore, making sure that data can be peer-reviewed for publications or grant appraisal is a critical step of scientific assessment (Ball et al. 2002). Therefore, data sharing is vital for ensuring that knowledge is not lost and work is not unnecessarily duplicated.

Last, depositing data is a long-term investment: facilitated data access allows bioinformaticists to develop new algorithms and tools while enabling students and trainees to get acquainted with complex datasets combining data from omics technologies with classical phenotypic anchoring.

Nor should we stop at microarray data. Mass spectrometers and NMR instruments are becoming more broadly available for protein characterisation and metabolic studies. These techniques are adding new dimensions to the space of metrics nutritionists can use to explore biological systems.

Structured management of these data is simply becoming a necessity. Reassuringly, both the proteomics and metabolomics communities have organised themselves and are striving to develop data exchange standards in their own fields, namely the Human Proteome Organisation Proteomics Standards Institute (HUPO-PSI) and the Standard Metabolic Reporting Structure (SMRS) group (http://smrs-group.sourceforge.net/mrn_report.html) respectively (Taylor et al. 2003; Lindon et al. 2002; Crott et al. 2004; Mutoh et al. 2004; Sreekumar et al. 2005; Oana et al. 2005; Rota et al. 2005; Tseng et al. 2005; Tsuda et al. 2005; van Breda et al. 2005; van Schothorst et al. 2005; Xiao et al. 2005; Yagil et al. 2005).
HUPO-PSI has delivered the mzData format for describing MS data, which has been accepted by major hardware companies and a data repository, ‘Proteomics IDENTifications database’ (PRIDE), relying on the format that has been set up at EBI (http://www.ebi.ac.uk/pride/) (Martens et al. 2005). The good thing is that these efforts are not being undertaken in isolation. Rather, synergies and integration capabilities are highly encouraged. To this end, all these groups are working collaboratively alongside the Reporting Structure for Biological Investigation Working Group (RSBI-WG), an offshoot from the MGED society, which brings together representatives from Nutritional Genomics, Toxigenomics and Environmental Genomics communities (http://www.mged.org/Workgroups/rsbi/rsbi.html). The RSBI working group has a key liaising role and the group reports about its activities to the Functional Genomics (FuGE) and Functional Genomics Ontology (FuGO) groups in order to bring forward use cases and specific need from the nutrition research arena, a body working to provide tools, exchange formats and annotation standard for describing the complexity of functional genomics experiments (http://fuge.sourceforge.net and http://fuge.sf.net/fugo).

All this will be ultimately beneficial to the field of nutritional science but on one condition, that of data sharing.

The authors are members of The European Nutrigenomics Organisation (NuGO). The European Nutrigenomics Organisation, linking genomics, nutrition and health research (NuGO, CT-2004-505944), is a Network of Excellence funded by the European Commission’s Research Directorate General under Priority Thematic Area 5 Food Quality and Safety Priority of the Sixth Framework Programme for Research and Technological Development.

Philippe Rocca-Serra1 and Ruan M. Elliott2

1EMBL-EBI
Wellcome Trust Genome Campus
Hinxton
Cambridge CB10 1SD
UK

2Institute of Food Research
Norwich NR4 7UA
UK
Fax: 44 1603 507723
Email: ruan.elliott@bsbrc.ac.uk

References


