Data Integration in the Life Sciences: Fun, Findings and Frustrations

Norman W. Paton

School of Computer Science, University of Manchester
Oxford Road, Manchester M13 9PL, UK
npaton@manchester.ac.uk

Abstract. This paper concerns the research topic of data integration in the life sciences. The paper presents no technical results, but rather provides a classification of research activities in terms of the contributions they seek to make to the life sciences, bioinformatics or computer science.

1 Introduction

Research involving data integration in the life sciences is diverse in nature, being conducted by researchers with different backgrounds and objectives. Research can be classified into the five areas represented by the overlapping circles in Figure 1, which in turn can be characterised (left-to-right) as follows:

Life Science for its own sake: The use of informatics to obtain biological insights. Typically, where the aim is to obtain insight into some biological system or experimental method, existing informatics techniques are deployed. Results are published in the life sciences literature (e.g. [2]).

Bioinformatics for Life Science: The use of novel bioinformatics to learn specific biological lessons. Such an activity requires the development of a novel result in bioinformatics to enable a specific biological system or technique to be better understood. Results are typically published in the life sciences or computational biology literature (e.g. [7]).

Bioinformatics for its own sake: The development of novel generic (organism independent) bioinformatics techniques. Typically, the new technique is not widely applicable outside the life sciences, and results are not necessarily accompanied by new insights into biological systems. Results are typically published in the biotechnology or bioinformatics literature (e.g. [6]).

Bioinformatics for Computing: The use of the life sciences as a source of challenging computing problems. Results are typically published in the bioinformatics or computing literature (e.g. [3]).

Computing for its own sake: Computing research motivated by or illustrated using biological problems. Results are typically published in the computing literature (e.g. [1]).

The diverse range of types of result (from discoveries in the life sciences to generic techniques in computer science) from research under the heading of "data integration in the life sciences" has a number of implications for researchers working in the area, as discussed in the next section.

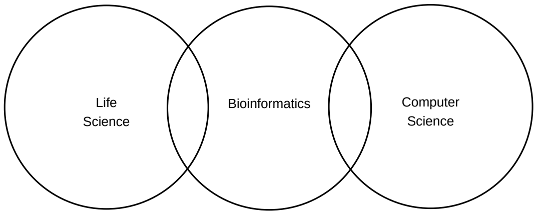

Fig. 1. Research areas of relevance to data integration in the life sciences

2 Observations

The following observations can be made about research on data integration in the life sciences:

Individuals: Few researchers are successful across the full range of areas in Figure 1, and those who are generally play a supporting role at one or both edges. This is neither surprising nor problematic, as it enables interdisciplinary teams to publish across the full spectrum.

Projects: Few projects are successful across the full range of areas in Figure 1, and those that are rarely apply the most novel computing when obtaining biological results. Indeed, individual projects typically occupy one or a few (adjacent) segments in Figure 1. This is not surprising, as deploying emerging computational techniques on applications that require dependable outcomes is a risky strategy. It may be considered problematic, however, as an interdisciplinary team that includes both computer and life scientists may be unlikely to generate research of direct interest to all its participants. Furthermore, the effectiveness of new computing techniques may not be subject to much practical evaluation in relevant applications.

The “Data Integration in the Life Sciences” (DILS) Workshop series is probably most naturally located in the *Bioinformatics for Computing* segment described in Secton 1. If so, then the community is principally seeking to refine computational techniques for data integration in the light of challenges identified in life science applications. In common with other research in computer science, techniques under current investigation are not the finished article, and implementations are typically early prototypes or proofs of concept; as a consequence, results generated by this community are often not ready for large-scale deployment.

Overall, reflecting the focus on novel data integration techniques, there is little evidence of technical consolidation. The diversity of research reflects both differences in requirements in different areas of the life sciences, and the fact that various aspects of data integration in the life sciences are difficult in ways that are hard to address systematically. For example, independently developed and autonomously maintained data resources often provide integrators with rapidly changing models and interfaces, inconsistent descriptions of common concepts, incompatible identification schemes, etc. Such features make high-quality data integration solutions (e.g. through warehouses or distributed query processing) costly to develop and maintain.

As a result, there is increasing interest in approaches with reduced up-front costs (e.g. [4]), which in turn often leads to more loosely coupled models. In the life science, a particular focus has been on workflow technologies, in which services interoperate, but data need not be "integrated" in any meaningful sense. Such platforms provide consistent access to data and computational resources, and may yet provide a framework within which different data integration technologies can be brought together, accommodating as they do both pay-as-you-go [8] and plan-ahead [9] integration. However, understanding the relative costs and benefits of different data integration techniques continues to be a challenging undertaking [5], and no less so in the life sciences than elsewhere.

As such, data integration in the life sciences potentially involves both *fun* and *frustrations* while trying to produce *findings*: *fun* in that the area is a source of worthwhile problems involving diverse collaborators; and *frustrations* in that the domain continues to manifest problems that elude elegant solutions. The latter in turn means that individual projects rarely generate findings of value across the range depicted in Figure 1.

Acknowledgement: Research on data integration in the life sciences at Manchester is supported by the BBSRC and the EPSRC, whose support we are pleased to acknowledge.

References

1. Khalid Belhajjame et al. Automatic annotation of web services based on workflow definitions. In *International Semantic Web Conference*, pages 116–129, 2006.
2. M.J. Cornell et al. Comparative genome analysis across a kingdom of eukaryotic organisms: specialization and diversification in the fungi. *Genome Research*, 17(12):1809–1822, 2007.
3. C. A. Goble et al. Transparent access to multiple bioinformatics information sources. *IBM Systems Journal*, 40(2):532–551, 2001.
4. A. Y. Halevy et al. Principles of dataspace systems. In *PODS*, pages 1–9, 2006.
5. B. Howe et al. Smoothing the roi curve for scientific data management applications. In *CIDR*, pages 185–195, 2007.
6. A.R. Jones et al. The Functional Genomics Experiment model (FuGE): an extensible framework for standards in functional genomics. *Nature Biotech.*, 17(12):1809–1822, 2007.
7. R.D. King et al. Functional genomic hypothesis generation and experimentation by a robot scientist. *Nature*, 427(6971):247–252, 2004.
8. T.M. Oinn et al. Taverna: lessons in creating a workflow environment for the life sciences. *Concurrency and Computation: Practice and Experience*, 18(10):1067–1100, 2006.
9. J. Syed et al. Supporting scientific discovery processes in discovery net. *Concurrency and Computation: Practice and Experience*, 19(2):167–179, 2007.