EMBO reports

correspondence

Informatics and hypothesis-driven research

The amassing of enormous data sets in genomics, proteomics and imaging has led a number of scientists to envision a future in which automated data-mining techniques, or 'data-driven discovery', will eventually rival the traditional hypothesis-driven research that has dominated biomedical science for at least the past century. It is no surprise that prominent scientists have expressed their scepticism—to say the least—about this point of view (Allen, 2001). However, I believe that framing the debate in terms of hypotheses versus informatics, with the subtext of man versus machines, misses an important point: currently available informatics techniques can greatly assist traditional hypothesis-driven research, but only if investigators slightly alter their practice to take advantage of this opportunity.

For example, informatics tools exist that can assist investigators in formulating, assessing and prioritising their hypotheses. Many hypotheses are, in fact, straightforward extrapolations from current findings: for example, knowing that apolipoprotein E4 is a risk factor for Alzheimer's disease, it is almost an automatic process to ask whether E4 may also be a risk factor for other neurological diseases or whether it interacts with other known risk factors; if one knows that RNA interference occurs in plants and lower organisms, it is logical to wonder whether it may occur in mammals as well. Publicly available tools, such as Arrowsmith (http://arrowsmith.psych.uic.edu), do not attempt to bypass scientists, but rather help them to integrate knowledge that is retrievable from the scientific literature in order to formulate hypotheses quickly, systematically and comprehensively (Swanson and Smalheiser, 1997; Smalheiser and Swanson, 1998). These tools can be thought of as analogous to word processors: they do not write manuscripts, and they do not do anything that people cannot do by themselves, but they do promise a new standard of efficiency and productivity.

Likewise, data mining of research databases need not be thought of as bypassing the traditional hypothesis-driven analysis of data, but rather as providing significant 'added value'. Consider a commercial database consisting of credit-card transactions: its purpose is to keep track of individual accounts, and most of the queries to the database are specific, focused and initiated individually. In contrast, automated data-mining techniques permit the same database to be characterised in terms of significant large-scale correlations that provide a rich array of market research data. More importantly, one can search on an ongoing basis for anomalous patterns of activity that raise the possibility of fraud; in fact, a commercial database that does not carry out such automated 'data-driven discovery' might even be considered negligent. I suggest that research databases that are populated and analysed according to specific hypotheses (Valencia, 2002) should also benefit from being monitored by computer programs that search for unanticipated correlations and anomalous patterns.

One of the basic concepts of informatics is the 'future value of primary data'. It is envisioned that the primary data—and, if possible, the actual samples—collected by one investigator will be archived and made available to other investigators, who may re-analyse the data from a different point of view, employ part of the data set not relevant to the first investigator, pool data with other studies or conduct new measurements on the original samples (Koslow, 2000). This is entirely compatible with hypothesis-driven research. Indeed, a good hypothesis is not one that is likely to be correct, but one that opens up a new arena of investigation. Since this arena cannot be fully perceived in advance, one must be prepared to carry out new analyses not included in the original hypothesis. Yet, most current experimental design simply ignores this fact: the investigator collects only those data that are deemed relevant to the original hypothesis, and when new information causes the original hypothesis to change, the investigator must plan a new experiment from scratch.

Ultimately, informatics should be viewed neither as a bag of tools and programmes nor as inextricably linked to the idea of artificial intelligence, but rather as pointing to a new approach to experimental design that takes into account the future use of primary data. If investigators and funding agencies simply included archiving of samples and data into research projects together with the metadata needed to understand how the data were collected, the increased efficiency and productivity that would accrue via data recycling should allow them to recoup their investments many-fold. Admittedly, most fields within biomedical science still lack an effective infrastructure for data archiving, sharing and collaboration. But this only means that investigators need to become actively involved to make this a reality and not retreat in the belief that informatics represents a threat to hypothesis-driven research.

References

Allen, J.F. (2001) *In silico veritas*. Data-mining and automated discovery: the truth is in there. *EMBO rep.*, **2**, 542–544.

Koslow, S.H. (2000) Should the neuroscience community make a paradigm shift to sharing primary data? *Nat. Neurosci.*, **3**, 863–865.

Smalheiser, N.R. and Swanson, D.R. (1998) Using Arrowsmith: a computer-assisted approach to formulating and assessing scientific hypotheses. *Comput. Methods Programs Biomed.*, **57**, 149–153.

Swanson, D.R. and Smalheiser, N.R. (1997) An interactive system for finding complementary literatures: a stimulus to scientific discovery. *Artif. Intell.*, **91**, 183–203.

Valencia, A. (2002) Search and retrieve. Large-scale data generation is becoming increasingly important in biological research. But how good are the tools to make sense of the data? *EMBO rep.*, **3**, 396–400.

Neil R. Smalheiser

Neil R. Smalheiser is at the UIC Psychiatric Institute in Chicago, IL.
E-mail: smalheiser@psych.uic.edu

DOI: 10.1093/embo-reports/kvf164