L&PS – Logic and Philosophy of Science
Vol. IX, No. 1, 2011, pp. 397-403

Saving the Phenomena in Molecular Biology

Cecilia Nardini
University of Milan and European Institute of Oncology (IEO)
Campus IFOM-IEO, Milan (Italy)
e-mail: cecilia.nardini@ifom-ieo-campus.it

ABSTRACT. The current understanding of the articulation between the-ory construction and experimental observation is still strongly influenced by the way it is instantiated in Physics. In the last decades, however, the upsurge of molecular Biology has provided a different framework and a source of novel insight. A discipline with openly explanatory and pre-dictive goals, molecular Biology relies to a large extent on quantitative experiments. In this paper I explore the merit of the distinction between "data" and "phenomena" originally proposed by Bogen and Woodward (1988) as applied to molecular Biology, and I argue that it provides a useful tool for understanding the epistemology of modern experimental Biology.

KEYWORDS: Epistemology, experiment, theory-ladenness, data-phenomena, biology

1. Introduction

In a seminal paper of 1988, Bogen and Woodward initiated the field of philosophy of data analysis. They proposed a distinction between data, the immediate result of observation, and phenomena, underlying structures that me-

Published by L&PS – Logic and Philosophy of Science
http://www2.units.it/ ~episteme — ISSN: 1826-1043

diate between empirical observations and scientific theories. Their proposal promoted a reflection in the philosophical community that eventually resulted in a deeper understanding and articulation of the relation between theory and experiment.[1]

Molecular Biology is fast becoming a strongly quantitative discipline, increasingly involving the analysis of large datasets from biological experiments in a fashion fairly similar to Physics -even though with potentially important differences as will be pointed out later. In this paper, I want to argue that the distinction proposed by Bogen and Woodward (hence BW) provides a useful key to understand the role and interpretation of experiments in modern biology; in addition, I claim that current developments in experimental biology, namely high-throughput experiments, provide an interesting case that goes further in support of BW's point.

The paper will proceed as follows: I will at first briefly recapitulate the point made by BW about the data–phenomena distinction, together with some critiques to their position. I will then discuss the relevance of the data–phenomena distinction to modern Molecular Biology, drawing on a recent paper by Napoletani et al. (2011). In closing, I will present a case study, an experiment in quantitative proteomics, in support of the view that I put forward.

2. The Data-Phenomena Distinction and Its Critics

Bogen and Woodward claim that there exists a level of analysis of experimental results that is theory-free in a relevant sense. The phenomenon is, in BW's idea, derived from the results of observation by making use of a set of substantive empirical assumptions. However, and importantly, the assumptions involved in deriving phenomena from data do not depend on the theoretical body that the experiment is meant to put to the test. Statistical analysis of a body of data, to extract its most salient features, is a typical example of the operation in point. By separating phenomena from data, BW achieve the isolation of phenomena as the stable candidate for explanation and prediction that can be derived or constructed out of intrinsically unstable observational results. Furthermore, the introduction of phenomena as mediators between observation and theory results in a partial weakening of the problem of theory-ladenness of observation. The positivist position that BW were challenging

[1] A state-of-the-art overview is provided by Morgan and Morrison (1999)

viewed observation as a single entity, in direct relation to theory that is supposed to predict and explain what is observed. BW's distinction separates observation into two very distinct stages, of which only one, the level of the phenomenon, is in direct relation to theory. Since, according to BW, the step of construction of phenomena from data is theory-free in the sense discussed above, theory-ladenness is a problem that does not affect the whole process of observation, but only part of it: namely, the step of putting phenomena in relation with theoretical entities.

Bogen and Woodward's picture has been challenged on the grounds that no extraction of a pattern from experimental data is as free of reference to the theo-retical framework as BW like to think. From one side, McAllister (1997, 2010) has issued a constructivist critique by showing that what counts as "pattern" in a set of data is unavoidably determined by preconceptions and interests of the measuring scientist. Along a different line, Harris (2003) has argued that what BW call "phenomenon" seems to stand for a model of data in the sense of Suppes (1962), a construct that contains the most salient features of a body of data. If this is the case, then the kind of manipulation that data are subjected to responds to theoretical objectives and it is influenced by theory-laden assump-tions, such as for example specification of the mathematical function that has to be used to fit a population of data.

The two critiques to BW data–phenomena distinction outlined above appear to deny that the distinction is a fruitful one. If phenomena are interest-dependent *qua* patterns, or they are theoretically laden *qua* models, the issue of theory-ladenness of observation is neither dispelled nor alleviated by the introduction of the distinction. In the following my aim is to to bring an example from molecular biology to show that the space for theory-free construction of phenomena is actually wider than what is acknowledged by both Harris and McAllister, and this is in fact quite in line with the picture delineated by Bogen and Woodward.

3. Data and Phenomena in Molecular Biology

Molecular biology has become in recent years a data-heavy scientific discipline. Along with more traditional experiments aimed at dissecting individual mechanisms, so called high-throughput experiments are designed and performed. These are extensive assays aimed at collecting large amounts of quantitative in-formation on the system under study. Statistical and computational techniques are developed to deal with the large datasets and to detect patterns in the data.

These patterns are certainly a target of interest for the biologist, but they are not interest-dependent in McAllister's sense, since no interest-loaded information goes into the pattern-detection algorithm. Before moving on to my case study that will help clarify this claim, I will spend some time introducing and discussing the notion of "agnostic science" as introduced in a recent paper (Napoletani et al., 2011). In this article the authors identify what they think is a novel perspective or trend in data analysis that is emerging typically -though not solely- in the kind of large molecular biology experiments described above. As pointed out, the analysis and interpretation of data from high throughput experiments rests essentially on statistical techniques and learning algorithms in order to screen for patterns in the data. Napoletani et al. contend that the patterns identified in this manner do not correspond to models of data in Suppes' sense. They do not deny that biological theories do play a part in building the experiment and providing the data; what they claim is that the statistical and computational techniques that are used to interpret the results of high-throughput experiments do not have a modelling role. The tools of statistics and computation are often applied to the data without the underlying conviction that real features of the object are being represented and modelled. Quite the contrary, the scientist is "agnostic" about the structure that is revealed by the high-throughput experiment, as she approaches it with a limited set of assumptions. From this discussion it follows that, just as much as the phenomena extracted from high-throughput experiments are not interest-dependent patterns in McAllister's sense, they seem not to be theory-dependent models in Harris' sense either.

I turn now to a typical technique in the new experimental biology, mass spectrometry for proteomics, in order to examine in more detail how "agnostic science" works, and the relation it bears to the data-phenomena distinction we started from.

4. A Case-Study: Mass Spectrometry of Proteins

Quantitative Mass Spectrometry using stable isotope labelling (SILAC-MS) [2] is a fairly novel technique that allows for the detection and quantification of pro-

[2] see Mann (2006)

teins contained in a cell. Since the quantification is a relative one, this technique provides a way to compare two expression profiles, as for instance of normal and tumour cells, in search for differences in the quantity and type of protein they synthesize. The readout from the experiment is a set of numbers that represent, for every protein, its ratio of abundance between the two types of cells. If all proteins were expressed at the same level in both kinds of cells, the ratios would be expected to be narrowly distributed around one. Differential expression of some protein between populations of cells is identified as an outlier in this distribution of ratios. The technique of statistical significance test is used to discriminate real hits from proteins that show a difference in expression just by chance. This relies mainly on a feature of the data population, the variance of the distribution: a large difference may not be significant if the variance is also large, while a small difference coupled with a very small variance could be significant.

There are a few highlights to the kind of experiment described above. As a first thing, it is evident that no biological information is used to discriminate the differential responders, and the statistical factors that play a role are unrelated to the theoretical frame of reference. This is in line with BW's account: "The scientific and methodological problems that arise in connection with the example are problems of data-analysis and statistical inference" (p. 310)

There is a second, and possibly more relevant sense, in which the example above might serve to illuminate the theory-free level of data analysis postulated by BW. This is the observation that the statistical assumptions do not have any modelling function in this example. In other words, the statistical treatment that is applied to the quantitative mass spectrum rests only on statistical assumptions about the distribution of the ratios. To adopt a particular statistical model does not entail the commitment to any hypothesis on the biological role of the proteins that are being detected; for what matters, the same procedure can be applied to totally different experiments involving completely different entities, for example in a medical trial aimed at detecting differential response to a treatment in two population of patients. As pointed out by Napoletani et al. (2011) "data are chosen and measured according to a basic description of a certain phenomenon, but no theory [...] is available for transforming them in a model in Suppes' sense" (p. 11).

The two highlighted points suggest that the distinction data and phenomena as Bogen and Woodward propose it is useful to dissect the way high-throughput experiments are designed and analysed. If we consider, along with McAllister, phenomena to be observer-dependent patterns that are influenced by the scientist's aims and interests, we are not able to understand a relevant part of how biologists successfully use algorithms and statistical techniques to identify patterns in mass spectrometry -and high-throughput- data. On the other hand, if we conceive of phenomena as data models in Suppes' sense we are also missing a part of the story: according to Harris, data models provide an organization of the data informed by theory, but this is not what happens with the statistical models that allows the biologist to identify differentially expressed proteins.

5. Conclusion

With the help of the case study proposed, it is possible to see that the level of statistical analysis represents a novel and interesting point of view to understand high-throughput experiments in modern biology. The treatment that is applied to data is not laden by reference to the theoretical background, nor it has the objective to model or isolate a theoretically-related feature of the experimental object. In conclusion we see that, on one hand, Bogen and Woodward's data-phenomena distinction provides a useful tool for understanding the goal and methods of modern experimental biology. On the other hand, the case of high-throughput experiments seems to support the idea that the distinction that Bogen and Woodward make is a tenable and a fruitful one.

REFERENCES

BOGEN, J. (2010), "Noise in the World", *Philosophy of Science*, 77(5), pp. 778– 791.

BOGEN, J. and WOODWARD, J. (1988), "Saving the Phenomena", *The Philosophical Review,* 97(3), pp. 303–352.

HARRIS, T. (2003), "Data Models and the Acquisition and Manipulation of Data", *Philosophy of Science,* 70, pp. 1508–1517.

MANN, M. (2006), "Functional and quantitative proteomics using SILAC", *Nature Reviews Molecular Cell Biology,* 7(12), pp. 952–958.

MCALLISTER, J. (1997), "Phenomena and Patterns in Data Sets", *Erkenntnis,* 47(2), pp. 217–228.

MCALLISTER, J. (2010), "The Ontology of Patterns in Empirical Data", *Phi-losophy of Science*, 77(5), pp. 804–814.

MORGAN, M. and MORRISON, M. (1999), *Models as Mediators*, Cambridge: Cambridge University Press.

NAPOLETANI, D., PANZA, M., and STRUPPA, D. (2011). "Agnostic Science. Towards a Philosophy of Data Analysis", *Foundations of Science*, 16(1), pp. 1–20.

SUPPES, P. (1962), "Models of Data" in NAGEL, E., SUPPES, P., and TARSI, A. (editors) *Logic, Methodology, and Philosophy of Science*, Stanford University Press.

WOODWARD, J. (2009), "Data and Phenomena: a Restatement and Defense", *Synthèse*, pp. 1–15.