November/December 2003 · Vol. 5 · No. 6

Genomics and inductive reasoning: Revolution, renaissance, or rhetoric?

To the Editor:

The genomics era sprang on much of molecular biology with unexpected force. This era has brought new technologies that have altered the focus of scientific inquiry and potentially the theoretical methodology for approaching scientific problem solving as well. Now that multiple genomes have been sequenced and microarray technologies are becoming commonplace, it is time for scientists to reflect on the methods of "postgenomic" science and ask if they have in fact "revolutionized" science as the rhetoric commonly goes. We propose that the goals and theoretical methods of the genomics era are in many ways similar to the descriptive scientific projects of the 19th century, a time when many explorers and scientists set out to map a new landscape by delineating its borders while characterizing and classifying its elements (Table 1). However, unlike this earlier exploration, the new tools of genomics offer the potential of truly completing the map. In this letter, we examine the philosophical and methodological consequences of the genomics era on scientific inquiry. In specific, we suggest that postgenomics science may grant access to a novel inductive problem solving logic and that this may be the most lasting effect of the genomics era on science.

It has been nearly 60 years since philosophers of science such as Carl Hempel and Karl Popper formally rejected the possibility that inductive logic (deriving general principles from specific cases) is justifiable in scientific methodology. Both Hempel and Popper proposed instead that scientists engage inquiry and problem solving with hypotheticodeductive reasoning: first inductively jumping from initial observations to a logically unjustifiable general principle or "hypothesis" and second testing outcomes predicted or deduced from this hypothesis. Thus, the hypothesis, as a testable statement, is used during scientific inquiry not because it can be verified, but because it can be falsified and, having withstood sufficient testing, comes to be believed.

We believe it is time to revisit the critiques of inductive logic as "postgenomic" science appears to be inventing a problem-solving inductive methodology. Briefly, Carl Hempel's critique identified practical problems for the inductive scientist.[1] The inductive scientist would approach a problem by first observing and recording all facts "without selection or a priori guess as to their relative importance." This inductive scientist would then analyze, compare, and classify these facts "without hypothesis or postulates other than those necessarily involved in the logic of thought." From these facts, generalizations would then be drawn that might be tested in a more deductive manner. However, this scientist is doomed from the start. As Hempel points out, no one has the time, money, or patience to gather "all" facts. Scientists must limit themselves to gathering "relevant" facts, particularly if engaged in solving a specific problem. But relevant to what? A priori the scientist, as problem solver, must have a tentative hypothesis, influenced by a matrix of paradigms and assumptions, in order to prioritize the fact-finding that will take place.

In contrast to Carl Hempel, Karl Popper addressed the legitimacy of inductive logic in solving problems or establishing truth by exposing the incongruity of a small closed set of singular observations being sufficient to derive general principles that will apply to all members of an open set.[2] For example, no matter how many swans a scientist observes, this scientist has not observed all swans and is therefore logically unable to justify the general statement, "All swans are white."

What if we are no longer interested in a fictive infinite set, such as all swans, but would rather generate general statements concerning a closed set, such as gene expression patterns under defined experimental conditions? The promise of genomics is that only a limited number of genes can be expressed, and that a limited number of splice variants exist for each gene and finally, that we should be able to identify and measure all of them. This number has been too large for anyone to seriously tangle with until recently. However, if the promise of genomics is fulfilled, a scientist ought to be able to characterize and measure the expression patterns of all genes across an experiment. Furthermore, using mathematical algorithms and the computer power now available, that same scientist is already able to view, organize, and examine the expression patterns in all such data without a preliminary hypothesis to guide the clustering priorities. From these observations, this scientist, who had been previously engaged in a simple descriptive project, should be justified in inductively generating universal statements derived from empirical observations concerning the effects of specific treatments and conditions on the expression of all genes in the system being studied.

Interestingly, this may offer a significant shift in the epistemological value of the information generated by the biological sciences. In contrast to a hypotheticodeductive scientist, who provides a hypothesis and supportive evidence but lacks conclusive proof of a thesis, an inductive scientist should be justified in the proof of the matter because he or she will have observed (to the limit of measurement) each potential result for that experimental system. Although establishing the epistemological truth-value of scientific results has rarely been a pressing issue for members of the biological sciences, biologists may now find themselves situated at a unique epistemological moment in the history of science.

There remain important limitations to a truly inductive methodology. Will it be possible to exhaustively identify all genes? It is tempting to believe that because the genome is finite (as it must be in order to exist on a finite number of chromosomes), then the number of genes that can be expressed from a single genome will also be finite, and having sequenced the entire genome, scientists will be able to identify every gene in the genome. Will it be possible to know when every atypical gene and splice variant has been properly identified?

Likewise, the heterogeneous nature of a species' genome, which is collectively engaged in an ongoing dynamic process of change, development, and evolution, cannot be represented by any of the currently available methods. At best, such heteroge-

Copyright © American College of Medical Genetics. Unauthorized reproduction of this article is prohibited.

Table 1
Comparison of descriptive scientific projects

Mapping	
	19th Century: the globe
	21st Century: genomic projects
Cataloging	
	19th Century: plants and animals into genus and species
	21st Century: G.O. consortium and Alliance for Cell Signaling
Circumscribing	
	19th Century: periodical table of the elements
	21st Century: microarray

neity can be represented by subsets of individuals to which scientists have access. Thus, the representation of a species' genome as a single, closed set is just that: a representation that serves well only when the restrictions that generated the set are acknowledged.

More important is a definitional problem: does generating a general principle for a closed set based on the observation of each member of that set constitute an inductive process or does this remain a descriptive project? Hempel's inductive scientist begins by gathering all facts. Under this definition, empirical observation of all members of a closed set would be included as an inductive methodology. In contrast, other philosophers, such as Georg van Wright, have defined induction as deriving attributes of uncharacterized members of a set from characterized members of that same set.[3]

Finally, although the bias of a preconceived hypothesis may partially be removed from an expression profiling experiment by measuring the pattern of "all" genes across the experiment, bias of expected outcome will necessarily remain in the selection of experimental model, treatments, and time-points. Thus, a hypotheticoinductive problem solving methodology will be more appropriately attributed to genomics experiments rather than a strict inductive one.

Whether or not genomics technology will eventually allow scientists to engage in truly inductive methodologies, this new field has already begun to alter scientific rhetoric. According to Ludwig Wittgenstein,[4] our language and jargon both reflect and shape our understanding of reality. We believe we are now witnessing a transformation in the scientific rhetoric used to describe the purpose and justification of the experimental process and that this has reciprocally affected both the vocabulary and practice of science. Three consistent claims are new to the scientific rhetoric.

The first common claim is that biological elements exist in limited numbers and thus identification of the members of such sets can be pushed to completion. This claim is the basis of all genome projects; once a genome project has been completed, all possible genes can be identified and complete expression analysis can be accomplished as this involves studying a closed set.[5–7]

The second common claim is that when experiments are performed in a tissue or cell type where a genome project has yet to exhaustively describe all genes, studying the expression pattern of a large but limited number of genes will represent the general pattern of all genes. Authors commonly refer to such results as representing or monitoring "global,"[8] "genome-wide,"[5] or "comprehensive"[9] expression patterns even when the array in use contains only a fraction of the total suspected genes that could be expressed in the tissue being studied. The principle used to justify such a belief is that genes will be coordinately regulated in functional clusters. Thus, the measurement of a limited number of genes within a single gene bank will be sufficient to predict the expression pattern of other members of that same functional cluster. This would represent a truly inductive logic. As this derivation is logically dubious, it will be interesting to see in which situations this principle will be ultimately scientifically sound.

The third and perhaps most important claim is that this technology has "revolutionized the power of unbiased"[5,9] generation of data and the unbiased evaluation of this data. Alfred Goodman Gilman, founder of the Alliance for Cell Signaling states, "I think now we need to get a bit away from this glorification of hypothesis-driven research. Hypothesis-driven research is quite wonderful, but it's not the only way."[10] Kari Stefansson, founder of Iceland's DeCode, has gone a significant step further to say that his data are "mined systematically and is unbiased and unblinded by hypothesis. Hypothesis is something that you have to avoid."[11] Both of these statements suggest a very fundamental shift in the acceptability of descriptive projects where the bias of hypothesis becomes an impediment to complete exploration, accurate characterization, and consistent mapping of a new topography.

All of these rhetorical claims reflect elements of inductive reasoning and are used by their authors to suggest that they have engaged inductive methodologies. If nothing else, this represents a return to the rhetoric of inductive methodology despite a deep commitment of granting agencies to direct scientific thought and activity into hypotheticodeductive channels. This shift in rhetoric marks a change not only in the rhetoric of science, but also, as Wittgenstein suggests, a change in the methods of science. It indicates a reformatting in scientific approaches to problem solving and a redirection in the types of problems scientists are engaging. The genomics era has reopened pathways of exploration and discovery, stimulating a renaissance of the 19th century scientific projects focused, like Hempel's inductive scientist, on the "unbiased" exploration of a new world with the goal of identifying, naming, and categorizing all its elements into circumscribed sets that can then be manipulated according to future scientific needs (Table 1). However, unlike the descriptive projects of the 19th century, the new genomics projects may offer the possibility of transitioning from simple description to hypotheticoinduction as these projects are pushed to completion, thus allowing a scientist to generate justified universal statements from empirical data.

Copyright © American College of Medical Genetics. Unauthorized reproduction of this article is prohibited.

In conclusion, we would like to propose a further area for thought and critique. Like the great flourishing of descriptive projects in the 19th century, our modern renaissance of descriptive projects appears to be dependent on new technologies. During the enlightenment, natural philosophers set about to explore and circumscribe the world, map and name its geography, identify and delimit its plants and animals into categories of genus and species, and finally to contain all this within the intellectual borders of an encyclopedia—a collection of all knowledge. In the 19th century, it was the technology of shipbuilding and open sea navigation in the context of nationalism that gave scientists access to a new world and facilitated the immense travel and exploration essential to these projects. Likewise, recent descriptive projects have been facilitated by technologies that grant access to a new landscape; in this case, it is the technologies of high-throughput sequencing and computational power in the context of a wealthy, aging population concerned about its health. Unlike the technologies of earlier explorers, these new technologies promise a certain completeness in map-making, a glimpse of the new world en toto, as if mapping the continents from a satellite in space. This analysis highlights the scientific opportunities created by novel technology. As technologies grant scientists access to new landscapes, descriptive projects of naming, mapping, and classifying necessarily follow. To what extent will scientific methodologies and focus then always be determined by the contemporary technologies? The difference this time is that genomics technologies tantalizingly promise completeness in mapping the new topography, thus granting scientists access to a new hypotheticoinductive science.

References

1. Hempel CG. Philosophy of Natural Science. Englewood Cliffs, NJ: Prentice-Hall, 1966.
2. Popper KR. The Logic of Scientific Discovery. San Francisco: Harper Torchbooks, 1968.
3. von Wright GH. The Logical Problem of Induction. New York: Macmillan, 1957.
4. Eden T. Lebenswelt und Sprache. eine Studie zu Husserl, Quine und Wittgenstein. In, Phänomenologische Untersuchungen München: Fink, 1999;20.
5. Tanaka TS, Jaradat SA, Lim MK, Kargul GJ, Wang X, Grahovac MJ et al. Genome-wide expression profiling of mid-gestation placenta and embryo using a 15,000 mouse developmental cDNA microarray. *Proc Natl Acad Sci U S A* 2000;97:9127–9132.
6. Ye RW, Tao W, Bedzyk L, Young T, Chen M, Li L. Global gene expression profiles of Bacillus subtilis grown under anaerobic conditions. *J Bacteriol* 2000;182:4458–4465.
7. Bienz M, Clevers H. Linking colorectal cancer to Wnt signaling. *Cell* 2000;103:311–320.
8. Khan J, Bittner ML, Saal LH, Teichmann U, Azorsa DO, Gooden GC et al. cDNA microarrays detect activation of a myogenic transcription program by the PAX3-FKHR fusion oncogene. *Proc Natl Acad Sci U S A* 1999;96:13264–13269.
9. Shaffer AL, Yu X, He Y, Boldrick J, Chan EP, Staudt LM. BCL-6 represses genes that function in lymphocyte differentiation, inflammation, and cell cycle control. *Immunity* 2000;13:199–212.
10. Gilman AG. Please check EGO at door. *Mol Intervent* 2001;1:14–21.
11. Rudra S. Scientist turns homeland into one-of-a-kind lab. *Genom Proteom* 2001;1:23.

John S. Welch, PhD
University of California at San Diego
La Jolla, California
Gerhard Rogler, MD, PhD
Klinik und Poliklinik für Innere Medizin I
Universitätsklinik Regensburg
Regensburg, Germany

Erratum

In the article "The Stickler syndrome: Genotype/phenotype correlation in 10 families with Stickler syndrome resulting from seven mutations in the type II collagen gene locus COL2A1"[1] in the January/February 2003 issue of *Genetics in Medicine*, the names of Ekaterina Tsilou, MD and Benjamin I. Rubin, MD, of the National Eye Institute, National Institutes of Health, Baltimore, Maryland, was unintentionally omitted from the list of authors. The authors regret this omission.

Reference

1. Liberfarb RM, Levy HP, Rose PS, Wilkin DJ, Davis J, Balog JZ, Griffith AJ, Szymko-Bennett YM, Johnston JJ, Tsilous E, Rubin BI, Francomano CA. The Stickler syndrome: Genotype/phenotype correlation in 10 families with Stickler syndrome resulting from seven mutations in the type II collagen gene locus COL2A1. *Genet Med* 2003;5:21–27.

Erratum

In the article by Mascarello et al. in the September/December 2003 issue of *Genetics in Medicine*, the title was incorrectly printed in the article. The correct title should be as follows: Problems with ISCN FISH Nomenclature make it not practical for use in clinical test reports or cytogenetic databases. The title appears correctly in the Table of Contents.

Reference

1. Mascarello JT, Cooley LD, Davison K, Dewald GW, Brothman AR, Herrman M, Park JP, Persons DL, Rao KW, Schneider NR, Vance GH. Problems with ISCN FISH Nomenclature make it not practical for use in clinical test reports or cytogenetic databases. *Genet Med* 2003;5:370–377.

Copyright © American College of Medical Genetics. Unauthorized reproduction of this article is prohibited.