



Newport Biotech Consultants

Volume 2 Issue 5 – May 2009



K. John Morrow, Jr. is president of Newport Biotech Consultants and publishes this monthly newsletter.

He is a biotechnology consultant and writer in the field of molecular immunology. He offers his talents in speaking, writing and consulting in molecular immunodiagnostics and related fields.

Visit the archives for previous newsletter issues www.newportbiotech.com

In this issue

Data Proliferation: How to Manage it?.....1

Data Proliferation: How to Manage it

By K. John Morrow, Jr. President
[Newport Biotech Consultants, Inc.](http://www.newportbiotech.com)

The unprecedented amount of data flowing out of computer systems affects every aspect of our society. While the problems of data overload originally referred to challenges associated with the storage of hardcopy on paper, the question now turns on the issue of electronic storage.

Chris Anderson of Wired Magazine put forth an interesting proposition that looks at the bright side of the data proliferation issue. He begins his argument with what appears to be a dubious proposition: "All models are wrong, but some are useful."

According to Anderson in an era of massively abundant data "we don't have to settle for wrong models. Indeed, we don't have to settle for models at all."

Anderson's argument is based on the example of the search engine Google, which he stated conquered the advertising world without knowing anything about the products, just collecting huge amounts of data and following them to their conclusion, which in this case was a superior search engine and advertising vehicle.

Anderson quotes Peter Norvig, Google's research director, who strongly endorses George Box's maxim: "All models are wrong, and increasingly you can succeed without them."

For instance, Anderson offers the case of Newton's Laws of Motion,

which were eventually replaced by Einstein's more accurate description of the Laws of Relativity, claiming that Newtonian Mechanics is a wrong model, which was replaced by the more appropriate (but not completely correct) Einsteinian views of the universe. However, he appears to have missed the point. Newton's model wasn't wrong; his model fit the data that he had available at the time precisely. It was an approximation of reality that served the world of science and engineering extremely well for hundreds of years, and still does, for calculations we make in our day-to-day world.

It may well be that people building business models such as the best search engines don't need any model, because they're looking for a profit, not an increased level of scientific understanding. But the big target in Bioinformatics isn't advertising, though, it's science. The scientific method is built around testable hypotheses. These models, for the most part, are systems visualized in the minds of scientists. The models are then evaluated against the data, and experiments confirm or falsify theoretical models of how the world works. This is the way science has operated since the time of the Greek philosophers.

Anderson argues in the Wired article, "There is now a better way. Petabytes allow us to say: 'Correlation is enough.' We can stop looking for models. We can analyze the data without hypotheses about what it might show. We can throw the numbers into the biggest computing clusters the world has ever seen and let statistical algorithms find patterns where science cannot."

Anderson cites J. Craig Venter's huge ongoing shotgun sequencing project, in which he is collecting sequences from organisms taken from all over the world. Anderson states that Venter

has discovered thousands of previously unknown species of bacteria and other life-forms.

This is, literally, a "fishing expedition" and for years this term has been the kiss of death for grant applicants who came forward with proposals which did not present a testable model. Now it may well be that Venter's approach may produce valuable information, but so far it has only produced mountains of gene sequence data, which may or may not find a valued application.

However the history of the last decade of genomics and proteomics analysis is not encouraging for the "data without a model" concept of doing science. For years pharma and biotechnology companies and research labs have cranked out results of genomic and proteomic screening, in which they searched for viable targets for new therapies. After spending millions and millions of dollars and countless man-years of effort they came up with essentially nothing. The work was not well controlled, it was not thoroughly researched, the experimental designs were poor and there was no hypothesis or model, just an endless search for targets.

There were many papers in the peer reviewed literature as well as countless reports at meetings and conferences in which hundreds or thousands of new potential targets were described. But these all vanished when they could not be confirmed or when drugs that attacked these targets were found to have serious side effects. Current surveys of the biomarker literature indicate that researchers are now examining more complex "models" (there's that word again!) in which serum samples from cancer patients are tested for a profile of biomarkers, with the hope of finding consistent patterns of gene expression. But such studies are quantitative, and will require extremely precise, standardized conditions in

order to detect differences in LEVELS of proteins, rather than all-or-none observations in which the presence or absence of a protein signals a cancer or other pathological condition.

The conclusion at this point is that the proposition that “all you need is massive amounts of data to do science” must be taken with many grains of salt.

We need your feedback!

- Forward it to a friend
- Email us to let us know what you think
- Have an interesting product or concept for a news story?

Email us today at kjohnmorrowjr@insightbb.com

© Copyright 2008-2009 by Newport Biotech Consultants