John Weinstein discusses information-intensive approaches to cancer drug discovery

Interviewed by Joanna Owens

John N. Weinstein, Senior Research Investigator, National Cancer Institute

What made you decide to work in cancer research?
I think most of us have friends and family who have had cancer. I originally came to the NIH to study another abstract subject – the lipophysical chemistry of membranes – but increasingly became interested in significant medical problems, including cancer.

Can you tell me about the work you are doing at the National Cancer Institute?
I am working closely with the Developmental Therapeutics Program at the National Cancer Institute (NCI) to integrate several buzzword disciplines – genomics, proteomics, bioinformatics and chemoinformatics – in the context of drug discovery for cancer. Since 1990, the Developmental Therapeutics Program has been screening drugs for anticancer activity in 60 different cancer cell lines in culture. The patterns of activity of the compounds against these 60 cell lines are rich in information about mechanisms of drug action and resistance. Informatic analysis of that system was initiated by Kenneth Paull, who tragically died three years ago, and we are trying our best to follow in his footsteps. The patterns of drug activity are useful particularly because we can map them into molecular structures of the compounds tested and also into molecular characteristics of the cells used for testing. Our principal enterprise is to characterize those cells comprehensively at the DNA, RNA and protein levels, that is, 'omically'.

Have you identified any striking correlations between drug-type and the gene-expression profiles in the cell lines studied?
Yes, we have defined quite a few. One that is particularly interesting relates to the gene for asparagine synthetase and the drug asparagine, a bacterial enzyme that has been used since the early 1970s to treat lymphocytic leukaemia. We found there is a strong negative correlation between expression of asparagine synthetase in leukaemia cells and the activity of asparagine against those cells. This observation prompted us to ask whether the same would be true for any other cancer cell type. We found that it was true for ovarian cancer cell lines. This correlation suggested the possibility (but it is only a possibility) that a subset of ovarian cancer patients whose tumors express only low levels of asparagine synthetase might be candidates for therapy with asparaginase. This possibility is being followed up in a collaboration with Daniel von Hoff, Director of the Arizona Cancer Center (Tucson, AZ, USA).

The challenge is to figure out which [relationships] are causally interesting, which are epiphenomenal and which are statistical coincidence.

Have you found any surprising correlations?
The correlation involving ovarian cancer that I just mentioned was a surprise. There have been other unexpected findings at the basic scientific level, but we are still faced with the fact that we are dealing with hundreds of thousands of individual relationships between drugs and genes. The challenge is to figure out which are causally interesting, which are epiphenomenal and which are statistical coincidence. To meet this challenge we make use of the literature – which is based largely on the efforts of those who study one gene, protein or process at a time. To facilitate searching of the literature on gene–gene and gene–drug relationships, we have developed a program called MedMiner, which is freely available on our website, (http://www.discover.nci.nih.com). Those who use it say it speeds up the processes of searching and organizing literature by 5–10-fold.

Those with major professional expertise in informatics have to be thought of as rock stars.

What are the limitations of your research approach, and what experiments will you do to overcome these?
The first obvious limitation is that these are cells in culture, and it is well known that cells in culture are not truly representative of cancer cells *in vivo*. Even primary cultured cancer cells have been extracted from their society of cytokines and other cells in the body. So, when we project towards clinical materials we do so with humility and with the understanding that we are developing clues to formulate hypotheses to carry to validation. It is worth remembering, however, that most of our knowledge of molecular biology and molecular pharmacology has come, historically, from studying cultured cells rather than clinical samples.

What experiments could you undertake to validate your findings?
There are several types of experiment and several meanings of validation. If we are trying to validate basic results from gene expression arrays, for instance, we can resequence clones used in making the arrays and also use a very different technique, such as real-time RT-PCR, to verify the expression levels obtained from the arrays. We are also very directly comparing cDNA arrays and oligonucleotide arrays for the same set of 60 cell lines. Testing of hypotheses about clinical relationships must be done with clinical materials. This we are doing, in collaboration with Mark Raffeld of the NCI Laboratory of Pathology, using tissue arrays from the Tissue Array Resource Program (TARP) at the NCI.

1359-6446/01/$ – see front matter PII: S1359-6446(01)02046-3

How will you go from identifying a relationship between drug sensitivity and expression of a gene, to using this in cancer drug discovery?
Generically, the aim of pharmacogenomics is to define subpopulations on the basis of their gene expression, or their gene or protein pattern, to lead to candidates for a particular therapy. Once an association is noted it will suggest either looking over existing clinical data to see if the correlation holds up there, or else, planning prospectively to treat patients.

So is it possible that your work could lead to the screening of patients for certain genes or markers that could provide personalized treatment options?
Yes, it is a possibility. For example, we are working with Mark Raffeld to identify diagnostic markers of metastases of unknown origin. Such markers could have implications for therapy.

Could your approach to gene-expression profiling lead to better and earlier screens for cancer?
I would say an emphatic 'I don't know'. It is the kind of thing that can arise from research into patterns of expression, but we are not explicitly looking at the differences between normal tissues and cancer cells. Many other laboratories are doing that, and we had to decide early on whether we join in with that definitely important endeavour or to focus on different, more adventurous projects.

How will the completion of the first draft of the human genome sequence impact on your work, and on cancer research in general?
It will have a major technical benefit for us because our ability to identify particular cDNA clones of genes is still incomplete. Although we rely on UniGene database clusterings, we can increasingly cross-check our findings against the developing genome-sequence information. Once we have the full sequence of the genome the scientific community will have a much easier time trying to establish functional relationships, find whole gene sequences and assign genes to particular chromosomal locations.

What do you think are the biggest hurdles for cancer drug discovery in the post-genomic era?
Let me cite just one. If you ask scientists at pharma and biotech companies it is clear that large numbers of targets for cancer therapy have been identified, but perhaps are not the right targets. Everyone is having difficulty figuring out which targets will be most useful.

What will really be the equivalent of the Human Genome Project is the complete wiring diagram of the relationship of the proteins in a mammalian cell.

Which new technologies do you think have been key to the development of cancer therapies? Which technologies still need to be developed?
Clearly gene sequencing technologies have been seminal, as have those for gene expression profiling. And there's PCR, of course.

As for technologies still to be developed, the first that I would like to see – which may or may not be feasible – is a way of amplifying proteins in the same way that we amplify nucleic acids; that is, PCR for proteins. The closest methods we have at the moment are those based on *in vitro* translation. Second, and an extension of that idea, is that we have perhaps 500,000 interesting protein states making up the 'wiring diagram' of the cell, and yet we have no technology capable of analyzing their expression levels comprehensively. This is what we need. There are continuing innovations in such technologies as mass spectrometry and microfluidics, but still they can't begin to approach a comprehensive characterization of the whole cell.

There is definitely a need for more training in informatics.

Do you think that there needs to be more integration of chemoinformatics and bioinformatics for optimal use of genomic and chemical data? Will they eventually become one science?
My laboratory has been trying for years to merge chemoinformatics and bioinformatics. Currently, we're working with Paul Blower from Leadscope (Columbus, OH, USA) on algorithms and software directed towards that goal. Although the statistical and analytical issues in the two areas are quite similar, the basic subject matters and expertise of biologists and chemists are different.

Do you think there is a need for more training of people in informatics expertise?
Yes, there is definitely a need for more training. This problem has been a major subject of discussions both at the NCI and elsewhere. In my view there are different flavours of bioinformatics. I would differentiate between applied bioinformatics, which is principally the province of those trained in biology, and developmental bioinformatics, which is principally the province of those trained in statistics, artificial intelligence and computer science. The former are using the tools (the databases and methods of analyses); the latter are developing them. Currently, large numbers of students with biological backgrounds want to get into bioinformatics; fewer computer scientists are available to carry out the developmental work and to guide the interpretation of the data. So that is a problem. One of the arguments I've made is that those with major professional expertise in developmental informatics have to be thought of as rock stars. It's not up to us to decide whether they are worth higher salaries or higher stipends – either we pay or we don't play.

Who do you think should be responsible for informatics training: academia or industry?
The majority of companies will do what they see as necessary for their own operation and, clearly, informatics is a major focus of their recruiting. At the *American Statistical Association Joint Statistical Meeting* (5–9 August 2001, Atlanta, GA, USA), many companies were interviewing hundreds of potential candidates with statistical backgrounds. Academia has, historically, been less organized and proactive about it, but the situation is improving.

Do you think that, in the long-term, research will focus more on the development of diagnostics and prevention, rather than drugs to treat cancer?
I hope that we get to a point at which treatment of cancer is unnecessary because we have sufficiently early diagnosis and the ability to prevent cancer. But, at this point,

the research has to go on at all three levels, because we really don't know at this point which is going to prevent the most suffering and death.

The completion of the Human Genome Project has been a tremendous milestone in life sciences research. What do you predict will be the next similar milestone?
That's an interesting question. An equivalent of the Human Genome Project's product – but much harder to achieve – is the complete wiring diagram of protein states in a mammalian cell. There are several-100,000 such states, it's going to take a long time. Still, 15 years ago when the Human Genome Project was faced with three-billion base pairs, sequencing the genome seemed rather a far-off goal. In that case, the goal was well-defined, and a public–private enterprise developed the accelerated technologies to make it happen. It will be interesting to see how soon the same sort of focus, public and/or private, on the wiring diagram will make what is currently a pipe-dream into reality.

I hope that we get to a point at which treatment of cancer is unnecessary.

What do you think is the greatest unanswered scientific question?
Not cancer. In the 19th century it was the origin of the species. In the 20th century it was the molecular nature of life. In the 21st century it is the working of the human mind. Never in history has it been such an exciting privilege to be a scientist.

What has been your greatest achievement?
My greatest preoccupation – perhaps achievement – has been to indicate, in prototype, ways in which information on DNA, RNA and protein levels in our cells can be coupled with information on the pharmacology of potentially therapeutic compounds in the service of drug discovery for cancer.

What is your ultimate career goal?
My ultimate career goal is to cure cancer. That is my first goal; my second is to help train the generation who will succeed after I have failed.

National Cancer Institute
National Institutes of Health
9000 Rockville Pike
Bethesda
MD 20892, USA

The *Discussion Forum* provides a medium for airing your views on any issues related to the pharmaceutical industry and obtaining feedback and discussion on these views from others in the field. You can discuss issues that get you hot under the collar, practical problems at the bench, recently published literature, or just something bizarre or humorous that you wish to share. Publication of letters in this section is subject to editorial discretion and company-promotional letters will be rejected immediately. Furthermore, the views provided are those of the authors and are not intended to represent the views of the companies they work for. Moreover, these views do not reflect those of Elsevier, *Drug Discovery Today* or its editorial team. Please submit all letters to Rebecca Lawrence, News & Features Editor, *Drug Discovery Today*, e-mail: Rebecca.Lawrence@drugdiscoverytoday.com

The commercial use of structural genomics ▼

There has been a good deal of debate recently about various structural genomics initiatives. It is becoming clear that the goals of these initiatives vary depending upon the nature of the organization(s) involved. From an academic perspective, structural genomics is seen as a way of obtaining the structure of as many novel proteins as possible to fill out further structural-fold space and enable better predictive methods. This is a major goal of the public initiatives funded by the National Institute of General Medical Sciences (Bethesda, MD, USA). For such public domain efforts, it does not really matter whether the proteins are from eukaryotic or prokaryotic sources and pragmatism is generally used to select the proteins.

From a commercial perspective the imperatives are somewhat different. There is an implicit need to concentrate on discovering the structures of proteins involved in specific pathways, or families of proteins believed to be important for drug discovery (for example, proteins involved in bacterial cell-wall biosynthesis, or for eukaryotic cells, the various protein kinases and so on). Once obtained, new crystal structures are used as templates for drug discovery and design by looking at how compounds bind to the proteins and using structural data to guide modification of the compounds for optimized potency, selectivity and, in some cases, reduced toxicity. Several biotechnology companies engaged in structural genomics are building or accessing the means to embark upon programs of structure-enhanced drug discovery. High-performance protein-modeling software can be used to augment structural genomics efforts. Importantly, computational predictions must be subjected to the rigors of experimental testing and methods. There is a great deal of power in being able to check the products of virtual screening by looking at the crystal structures of complexes between the selected compound and the protein. It is this integration of computational and experimental technologies that will be key in fully developing the value of structural genomics for drug discovery.

Access to core technologies that enable structural genomics initiatives is anything but trivial. In the first place, suitable bioinformatics capabilities have to be available to enable selection of the most appealing targets and to choose the most appropriate constructs to make. Most people are now doing structural genomics by using X-ray crystallography as the preferred method for structure determination. Today, NMR is simply not competitive from a high-throughput perspective: 85% of those structures in the Protein Data Bank[1] (PDB; http://rcsb.org/pdb) come from X-ray structure determination.

Crystallography involves setting up rapid and automated crystallization trials. Most structural genomics efforts are using Multiwavelength Anomalous Diffraction (MAD) phasing for crystallography, a technique pioneered by Wayne Hendrickson in which selenomethionine is incorporated into the proteins to enable rapid data-interpretation[2]. Although it is possible to use MAD phasing for data collected at most synchrotron radiation sources, it is becoming widely acknowledged that the best data are being collected from third-generation sources of X-rays, such as the Advanced Photo Source in Chicago (IL, USA), because of the speed and efficiency with which they are obtained. It is now almost routine to collect data at the APS and use automated interpretation methods to go from diffraction pattern to refined novel structures in less than eight hours. Molecular replacement methods are even quicker. The commercial structural-genomics efforts will enable the provision of three-dimensional data on the way compounds bind to their cognate targets to medicinal chemist in real time. This will be important for increasing the efficiency of lead discovery and optimization.

References

1 Berman, H.M. *et al.* (2000) The Protein Data Bank. *Nucleic Acids Res.* 28, 235–242

2 Hendrickson, W.A. (2000) Synchrotron chromatography. *Trends Biochem. Sci.* 25, 637–643

Tim Harris
Structural GenomiX
11099 N. Torrey Pines Rd
Suite 160
La Jolla
CA 92037
USA

PII: S1359-6446(01)02048-7