W Prinz, M Jarke, Y Rogers, K Schmidt, and V Wulf (eds), *Proceedings of the Seventh European Conference on Computer-Supported Cooperative Work, 16-20 September 2001, Bonn, Germany*, pp 399-418

When Worlds Collide: Molecular Biology as Interdisciplinary Collaboration

Vicki O'Day, Annette Adler, Allan Kuchinsky, Anna Bouch
University of California, Santa Cruz, USA; Agilent Technologies, USA; University College London, UK
oday@calterra.com; [annette_adler, allan_kuchinsky]@agilent.com; A.Bouch@cs.ucl.ac.uk

Abstract. The field of molecular biology is in a remarkably rapid period of change, as the genome sequencing projects and new experimental technologies have generated an explosion of data To analyze and draw insights from the vast amounts of information, biologists use a new generation of bioinformatics software tools, often working closely with mathematicians and computer scientists There are elements of both *collision* and *convergence* in these interdisciplinary encounters We conducted user studies with biologists engaged in investigating the molecular basis of disease. We describe several issues that arise in this collision/convergence of disciplines, drawing on the notion of *boundary objects in-the-making*. We provide recommendations on building technology for people whose work now sits at the crossroads of diverse and rapidly changing scientific fields.

Introduction

The field of molecular biology is in a remarkably rapid period of change. One notable characteristic of current genomics research is its increasing reliance on computational tools, including genomic databases (public and proprietary), online scientific literature, and data analysis software. This has led to immense interest and investment in bioinformatics information and tools. In addition to this

proliferation of information and tools in the genomics research community at large, some molecular biologists are generating huge amounts of information in their own labs using a new technology called DNA microarrays. Microarrays allow biologists to simultaneously probe the activities of thousands of genes under diverse experimental conditions, which is useful for investigating relationships within and across families of genes. Microarray experiments can produce terabytes of data, and it is simply not possible to analyze these data without significant computational support. It is likely that even larger data sets will arise as data are shared among many academic and industrial labs, just as genomic and other databases arose from distributed efforts. One of the central themes of computer-supported cooperative work (CSCW) is the study of collaborative encounters, especially when new technologies are involved. We are interested here in a collaborative encounter between *disciplines*, as biologists and computational experts work together to solve hard biological problems. Interdisciplinary collaboration raises challenging issues for practitioners and for technology designers who wish to support their work.

Several years ago, a leader in the biotechnology industry declared that "biology is now an information science" (Williams, 1995). More recently, an article in the *New York Times* announced that " all science is computer science." (Johnson, 2001) These are provocative claims, and most biologists probably would not characterize the changes in their field in quite those terms. However, these statements do point to a current tension for molecular biologists. Information models are not new to biology, but the work of biologists is changing through its contact with informatics (and vice versa). On the one hand, biologists are clearly working on problems that emerge from biology; they want to identify genes and understand how they function in living organisms. On the other hand, it is increasingly necessary to address these problems using information visualization, statistics, and other techniques for manipulating large amounts of data. But these techniques do not come ready-made for biological applications. To provide effective computational support, computational experts have to understand biological questions and work with biologists to try out new forms of analysis, using real biological data. Biological systems have a different character than computational or physical systems. They are less well-behaved; exceptions are as common as rules. It is harder to make simplifying assumptions than it is in the physical and computational sciences. Computational experts have to adapt their approaches to the way living systems work.

What makes an interdisciplinary encounter difficult (and interesting) is that it involves different worlds, or different systems of meaning. People learn through their disciplines to formulate and solve problems in particular ways. They learn what *counts*—as a valid object of attention, a good method of analysis, or a reasonable solution. In CSCW, we are accustomed to thinking about people's different (and possibly competing) activities, responsibilities, locations, and work

styles But here we must also consider people's different ideas about what information *is* and what makes it reliable and meaningful.

We emphasize here the interdisciplinary collaborations between molecular biologists and what we call "computational experts," but each of these categories is diverse in itself. Molecular biologists work in different problem areas that shape their perspectives in particular ways. Computational experts include (at least) mathematicians, statisticians, and software developers, each with different backgrounds and skills. These fields converge in bioinformatics and computational biology and in industry terms such as "biocomputing," which are emblematic of the hybrid character of this work. However, at this time most practitioners are still either biologists or computational experts, but not both. We focus here on the challenge of collaborating across this significant disciplinary boundary, while keeping in mind the many differences within each group.

In some settings, biologists and computational experts work together directly on the analysis of experimental data, especially when the biologists are just beginning to use computational packages. Later, it is more common for biologists to encounter statistical methods and other computational techniques primarily through data analysis software. That is, the tools act as intermediaries between biologists and computational experts. Eventually, analytical techniques may become a kind of black box—something a biologist knows how to apply and need not understand in depth. But at this stage, biologists and computational experts must build on one another's knowledge and intuitions to develop workable methods for finding and interpreting patterns in biological data. Although the title of our paper refers only to worlds colliding, we see an interplay between collision and convergence in the new biology.

The central question addressed in this paper is: How should technology be designed for people whose work now sits at the crossroads of different disciplines? This is a situation in which the technology mediates between disciplines with different ways of looking at the world. What issues arise for participants in an interdisciplinary collaboration? How do people collaborate across disciplines when the ground is shifting on each side? To explore these questions, we discuss interdisciplinary encounters between biologists and computational experts. This discussion draws on a series of ethnographic interviews with research biologists and microarray designers. The discussion is lopsided in favor of biologists, since the issue of interdisciplinary collaboration emerged through a set of interviews with biologists which were originally meant to look at their particular technology uses and needs.

We find the notion of *boundary object* (Star & Griesemer, 1989, Bowker & Star, 2000) to be particularly helpful in thinking through the issues raised in this example. In our view, it is not a perfect fit for the objects we see migrating across community boundaries, but that makes it even more interesting—it adds to our understanding of the idea of boundary objects, as well as of the objects held in

common by molecular biologists and computational experts. We use boundary objects in discussing three issues that arise in this interdisciplinary collaboration: contrasting biological and computational stories, contrasting notions of biological and statistical significance, and changing work practices. We end with a discussion of software design implications.

Method

Our group at Agilent is engaged in software research to support molecular biologists, with a particular focus on the problems involved in disease and drug discovery. To ground our group's software design projects in an understanding of users' needs, we conducted a series of ethnographic interviews with molecular biologists. Most of our interviews (about twenty) took place in a single laboratory at the (U.S.) National Institutes of Health (NIH), where biologists are investigating the molecular basis for different types of cancer. We also interviewed several biologists and computational experts in university labs and local biotechnology companies. In addition, we interviewed colleagues in our own company who were trained as molecular biologists and who now work as developers of our company's microarray technology.

We conducted our interviews in the biologists' labs. The interviews were open-ended and informal; we asked the scientists to explain to us how they formulated and carried out their research projects and what they were learning from them. In most cases, people turned to their computers and walked us through their data, showing us which software tools they used, how each tool fit into their analytic strategy, and how they interpreted the information presented in each tool. For the most part, we chose to talk with people who used microarrays in their experiments. Agilent makes microarray products (including arrays, scanners, and analysis software) and we were interested in how people handle the large volumes of data generated by microarrays.

Microarrays in Molecular Biology

Molecular biologists seek to identify and understand the relationships of genes, proteins, and pathways in living organisms. An increasingly important tool for research in molecular biology is the DNA microarray, or gene chip. Using microarrays, biologists shift from examining the way a single or a few genes change as cells move from state to state to simultaneously monitoring thousands of genes across different conditions. Microarrays are a new technology, and only a few labs have experience in using them (DeRisi et al, 1996) The NIH lab we visited (which is among the earliest microarray users) has been using microarrays for about four years. During this time, members of the lab have developed basic protocols for microarray use and established lab-wide software standards.

A microarray gene expression experiment starts with an array and a sample. The array is a glass slide on which thousands of *probes* have been deposited in a grid-like arrangement. Each probe consists of a small sequence of DNA that complements a particular gene from a particular organism. Many researchers at the NIH lab use an array with about 6000 probes that represent a cross-section of the 30,000 genes in the human genome.

To perform an array analysis, researchers collect samples of the biological materials whose genetic activity they want to study. For example, a sample might consist of tumor cells from cancer patients or cells from a particular kind of tissue. Then RNA is extracted from the sample, the RNA molecules are produced by active genes and are specific to those genes, so they indicate which genes are expressed or "turned on" in a cell. The RNA is used as a template for synthesizing a form of DNA called complementary DNA, or cDNA, which is used because it is more stable than RNA. This cDNA is in turn labeled with a fluorescent dye. Then a solution containing these labeled molecules is distributed over the slide containing the probes. If the sample contains cDNA that matches any of the probes, the cDNA will bind (or hybridize) to those spots on the array.

The slide is scanned and the amount of fluorescence is measured at each spot (a measurement which requires considerable data processing in itself). The different levels of fluorescence at different array locations give information about which genes are being expressed in the sample and at what levels. The brighter the fluorescence, the more cDNA has attached to the probe, so the more active the gene corresponding to that probe must be in that sample. Each time a sample is hybridized to an array, thousands of data points are generated. It is a common practice to include labeled reference material along with the sample, so each of the thousands of probes actually gives two data points—one for the experimental sample and one for the reference. Relative gene expression patterns are determined by comparing the expression levels in different experimental samples to the same reference sample.

Once the data have been generated, researchers use a variety of tools to look at them, including spreadsheets (for basic number-crunching), off-the-shelf database programs (to query the data along different dimensions), and special-purpose bioinformatics tools (for more complex algorithms to find patterns in the data). Each of these tools is a way of filtering the thousands of data points, to identify a small number of genes (usually fewer than a hundred) that seem to be interesting and worth looking at further. Then the biologists usually consult genomic databases (such as GenBank) and scientific literature databases (such as Medline) to see what is already known about these genes.

There are many different kinds of gene expression experiments that can be done using microarrays. Though we cannot characterize a typical experiment, we can give examples of the projects being done by researchers in the NIH lab Several researchers are trying to find genetic markers for particular kinds of

cancer For example, certain cancers are quite rare and difficult to diagnose. By analyzing the gene expression activity associated with similar cancers, researchers hope to find a set of genes whose pattern of expression is unique to each cancer. They hope these unique genetic profiles could be used to develop diagnostic tests. Other researchers are trying to learn the functions of particular genes. They can modify a gene of interest in a cell line and then take snapshots of more general gene expression activity at several time intervals after the modification. They hope to figure out what the "downstream" effects of that gene are, to begin to piece together the biological pathways in which it participates. An important aspect of using microarrays is figuring out how to translate research questions into choices for arrays and samples in such a way that the data analysis will yield useful answers, in light of current data analysis approaches and tools.

There is a good deal of formal and informal collaboration among people working on related projects. Visitors and post-docs share lab space in very tight conditions—each person has only a few feet of lab bench and nearby desk space (with computer) People are aware of each other's projects, especially where there are overlapping interests in particular diseases, and they exchange information and advice about lab protocols and data interpretion. Most papers are co-authored by a long list of researchers.

There are two people in the NIH lab we visited who play a special role in supporting the molecular biologists in their microarray experiments. These people (one biochemist and one image processing expert) choose which software tools should be used in the lab, teach people how to use them, consult with people on data analysis for particular experiments, and write custom software when needed. This kind of contribution has been well-documented in the human-computer interaction literature (Mackay,1990, Gantt & Nardi, 1992, Williams, 1993). We draw attention to it here because it is a vitally important part of the interdisciplinary collaboration between biologists and computational experts—most of the biologists we interviewed emphasized that they could not have done their experiments without the help of their on-call consultants.

Boundary Objects In-the-Making

The biologists we interviewed had adopted a large number of computational resources, especially since they had begun using microarrays. Each person had assembled a working set of information tools and services to help them throughout the course of their research projects. Despite the biologists' adeptness in using their tools, it was clear from our conversations that there were aspects of these technologies that puzzled or bothered them at times, as we will describe in more detail in the following sections. There is something about the way these information tools work that seems a little askew—they don't quite fit the way biologists think about their world. We want to look more closely at this problem

of fit, and we find the concept of *boundary object* helpful in thinking about the gaps left open by bioinformatics technologies.

Star uses boundary objects as a way to talk about objects that circulate among different communities of practice, taking on distinct local meanings and uses in each one (Star & Griesemer, 1989, Bowker & Star, 2000) (A *community of practice* is one in which people develop a sense of shared activities and membership through sustained participation (Lave & Wenger, 1991).) A boundary object retains some common structure and is recognizably the same object across communities, but it has "different meanings in different social worlds" (Bowker & Star, 2000, p. 297). In the examples of Bowker and Star, boundary objects travel well and facilitate collaboration across communities in part because their local differences don't have to be confronted or reconciled.

The mobility of objects such as software tools, algorithms, and data sets allows biologists and computational experts to cooperate on solving analytical problems in molecular biology. However, several questions arise: Do computational experts and biologists agree on what counts as important in data? If you translate biological data into computational data and perform mathematical operations on them to find patterns, are the results biologically meaningful? Are anomalies in one domain also anomalies in the other domain? How much common structure is retained, and to what extent do local differences of interpretation matter? The interdisciplinary collaboration would be more comfortable if analytical tools and data worked as boundary objects whose local interpretations were not called into question, but this is not quite the case.

Some features of this collaboration depart from the boundary object scenarios described by Bowker and Star. First, this is not a situation in which people in *different communities of practice are focused on their own activities and* problems. On the contrary, it is a more intimate collaboration, in which people from different disciplinary communities are trying to work together on a common problem. They must enter into each other's worlds, shift their own practices, and accommodate unfamiliar points of view. They have to achieve a kind of double vision, to see common objects both from their own disciplinary perspective and from the perspectives of their colleagues. Second, Bowker and Star point out that boundary objects arise over time in durable cooperative arrangements. However, the interdisciplinary encounter between biologists and computational experts is new and not yet stable. It is far from a durable cooperation, although it may grow into such a relationship over time.

It may be more useful to look at the objects in this interdisciplinary collaboration as boundary objects *in-the-making*—they are circulating across communities, but sometimes it is necessary to confront and reconcile their different local meanings. What we want to draw attention to here is that these unstable objects can still work to facilitate collaboration across communities—they give people common ground for discussion and negotiation.

The interdisciplinary collaboration we describe has features in common with several earlier projects in the CSCW literature Bannon and Bødker discuss the issues of designing and using "common information spaces" (Bannon & Bødker, 1997). They include situations in which cooperative work is mediated by a database, since the person who records information in a database and the person who accesses that information try to understand each other's context. Bannon and Bødker point to the tensions and tradeoffs in creating common information spaces, and they draw attention to the importance of human mediators in helping people from different communities make use of common information. As we have mentioned, we also found that human mediators play key facilitator roles.

Van House, Butler, and Schiff describe a digital library project that involves sharing environmental planning data sets (mostly measurement data) on the web (Van House et al, 1998). They describe users' concerns about the ways data might be misused or misunderstood when it is dissociated from its original contexts and communities of expert practitioners. Harper reports on an ethnographic study of "missions" sent by the International Monetary Fund to member countries, during which economists gather and analyze data and prepare reports on national economies (Harper, 1997). His fascinating study shows how meetings between the visiting economists and their hosts help to make the numbers "count." These meetings are a "social process that converts *speechless* numbers into ones that have a *voice*" (Harper, 1997, p. 363). Through conversation and negotiation, "raw numbers" are converted into meaningful and useful information. As in Harper's case, we must pay attention to whether and how numbers are adopted, not just how they are generated.

Each of these projects raises the important issue of trust and reliability when information objects travel across communities of practice. We emphasize the complexity of trust in the sections below, since it emerges as a central theme of the interdisciplinary collaboration between biologists and computational experts. We do not mean to imply that people are suspicious of each other, but rather that both biologists and computational experts are still trying to bring their problems and methods into alignment, so they can both feel confident of their results.

Comparing Stories

As we listened to molecular biologists talking about their research, we were struck by how often they described their activities in terms of *telling a story* Each story is an interpretive framework—a way of making sense of experimental data and situating it in a context of earlier work By looking at how these stories are put together, we can see some of the strangeness bioinformatics tools have for biologists. In this section, we revisit our account of microarray experiments to look at collisions and convergences in biological and computational stories.

Biological Stories

The stories told by molecular biologists usually focus on how genes and proteins interact with other genes and proteins in biological pathways (such as energy metabolism or cell growth). When genes are expressed, they encode proteins In turn, proteins can catalyze biochemical reactions in the body, provide cell structure, transport nutrients, or regulate further gene expression. Molecular biologists explain biological phenomena by narrating the interrelationships of genes and proteins.

As we have discussed, microarray experiments yield data about the relative expression levels of genes under specific conditions, such as which genes are expressed in a set of breast tumor samples. Usually, the researchers in the NIH lab examine a series of samples using their arrays and cross-compare the results. The data live in huge spreadsheets, where each of the 6000 rows corresponds to one gene and each column corresponds to an experimental condition; it is not uncommon to have 20 to 40 experimental conditions. It is very challenging to see the traces of biological phenomena in this sea of numbers. People are looking for patterns, usually quite subtle, that imply interrelationships among the genes. They hope to infer a network of cause and effect relationships among genes—including the coordinated effects of multiple genes acting together—since such networks are the basis of biological pathways. Deciphering pathways is what most molecular biologists refer to when they speak of "putting together the story."

Biological stories are told from diverse points of view. When biologists search the scientific literature for references to a gene they are investigating, they are most interested in what is known about the gene in a context similar to their own. But they are likely to come across references to quite different contexts, which tell a different story about the gene—in fact, the gene might have quite different functions across these contexts. For example, a prostate cancer researcher might find information about a gene of interest, but it is told from the perspective of a biologist studying liver function. Or two or more researchers may independently discover the same gene, but refer to it by different names. Since there is a pride of ownership involved in naming a gene, it is not easy to standardize gene names. Biologists have to be aware of the different aliases for a gene under study.

In general, biological stories need to accommodate multiple hypotheses and alternative explanations, which change as new data are obtained. Biologists use bioinformatics tools or literature searches to generate and check out their ideas, and they find it challenging to organize and manage the links between computational results and the biological stories they support.

Computational Stories

There is no way to analyse spreadsheet data in thousands of rows and a series of columns without some computational help. Biologists use bioinformatics tools to filter, sort, and find patterns. That is, they use tools that put together a *computational* story for those data.

One important character in a computational story is a *cluster*. The clustering algorithms used in bioinformatics tools work like document clustering—they reorganise a large set of elements into groups whose elements are somehow similar to one another. One row of a spreadsheet (the expression levels of a particular gene over a number of experimental conditions) can be thought of as an expression "profile" for one gene, and different gene profiles can be compared to see how similar they are, based on mathematical notions of similarity. Sets of genes with similar expression profiles are grouped together in clusters. In other words, the clustering algorithm surveys the numerical values of expression levels and looks for non-random correlations. If clusters have biological as well as mathematical meaning, then each cluster describes a group of genes that may be co-regulated, implying their involvement in the same biological pathway.

People often combine clustering with visual inspection Gene expression levels are encoded into shades of red and green, and biologists look at array data that is mapped in this way to see if they can see patterns in the arrangement of colors. They can use clustering algorithms to reorder the data, to allow for new patterns to emerge through visual inspection.

Double Vision: Making Stories Converge

Although clustering tools are common for molecular biologists who use microarrays, we found that biologists are reluctant to trust the clusters identified by the tools. Biologists draw a sharp distinction between computation and biology and do not take for granted that meanings can be translated in a straightforward way from one domain to the other. Nor do they trust the findings they read in the scientific literature when these results are based on mathematical analyses that the paper's authors don't appear to understand.

Many of the biologists we interviewed try to adopt a mathematician's perspective, to look into the inner workings of the algorithms and convince themselves that each step makes sense from a biological perspective. This brings the biologists into contact with someone else's unfamiliar way of looking at things. Consider this biologist's description:

> I think we spend two thirds of the time thinking about the biology and a third thinking about this kind of logic which feeds into it eventually Because it makes you—it determines whether or not you can believe what all these computer manipulations are telling you

We are struck here by the alien quality of "this kind of logic." The computational logic "feeds into" the biology eventually, and there is no way to get to the biology except through this logic. Despite its strangeness, biologists have to work with it to make the different stories converge.

There is an ironic note to the labor-saving possibilities of bioinformatics analyses. Although in principle these tools reduce analytical effort by suggesting mathematical relationships that may be useful, biologists tend to run through extra analyses so the analyses will check up on each other. For example, they may run a clustering algorithm repeatedly, using different similarity measures to see if the sets of clusters that come up are consistent. If they get the same results from several different methods, they have some confidence that the results may be biologically significant. Of course, checking results and calibrating tools are always part of good scientific practice. But in this situation, biologists emphasized to us their uncertainty about how much cross-checking was enough—how much it would take to make the computational patterns biologically convincing.

We emphasize that microarrays are new, and both the biologists and the computational experts are unsure of how to bring their disciplinary strengths together. Both are learning, and both are nervous about the prospect of producing unreliable findings. A mathematically-derived set of clusters in gene expression data is a new kind of common object that emerges through interdisciplinary collaboration, belonging to both disciplines and not fully to one or the other. The analytical algorithms are themselves a topic of research, and they continue to be tested and revised with new sets of biological data. A computational expert at the lab, whose skills are very much appreciated by his biologist colleagues, talked about some of his doubts:

> When you run the program and see all those relationships, it tends to mean something but basically means nothing So you don't want to make a big story out of nothing You really want to make sure that everything goes smooth, as if there is a real story there So that part of the abstraction is usually—we are all learning and trying to figure out

For both computational experts and biologists, it is not enough to have a story that makes sense only in one perspective or the other. It is important for the stories to converge, since biologists rely on computational tools in sorting through their data. One of the biologists talked in similar terms when he described a conversation with some mathematicians about a new algorithm:

> One caveat they pointed out, this analysis is susceptible to meaninglessness. They showed us some data that just didn't correlate with anything—no clinical characteristics, no laboratory measurement, no demographic information, nothing Instead their strongest association was with a particular date in the year, of all things It turns out that Cluster A was done prior, statistically significant So what happened? I'm willing to go that route, because they may have had some technical problems, we don't know What we can do—we can then look at the genes and we want the genes to tell us . "

He wants to be able to trust the convergence between biological and computational features of the data (so the genes will be able to speak), but he is also realistic about the possibility of misalignment.

Stories as Boundary Objects

Stories are a kind of boundary object. As they travel across communities of practice, they are more or less successful in helping communities to collaborate with one another. They work to focus people's collective attention on something of common interest: an account of how genes seem to be related to one another. In the interdisciplinary encounters between biologists and computational experts, these stories are rather unstable—it is hard to be sure whether a particular story is a good one or not. As boundary objects in-the-making, stories about genes based on computational analysis are subject to scrutiny, and their different interpretations call for explanation. But even in these circumstances, the stories work to facilitate the collaborative analysis of experimental data across disciplines.

Stories are boundary objects between biologists, as well as across disciplines. They circulate through different biological contexts, and as people collaborate on data analysis, alternative stories come into play. As one post-doc said, "Most people in the lab analyze each other's data and recognize things. Because no two people had the same training." When biologists look at data, they look for familiar characters—genes they know well from other projects and other contexts. A gene that is quite unknown to one biologist may be an old friend to another. In general, biologists have to look through and understand other people's points of view (both biological and computational) to compose a good biological story. As one scientist at NIH put it, they try to develop a "consensus hallucination as to what the data is trying to indicate to us," pulling together the knowledge and ideas of people with diverse experience and expertise.

Comparing Biological and Statistical Significance

One of the central tensions that emerges in the interdisciplinary work of biologists and computational experts is how to decide which data are biologically significant. The discipline of statistics has developed mathematical methods for establishing significance in data analysis. Biologists use statistics, but they also rely on experimental know-how and a sense of what is biologically interesting and believable. Sometimes their approaches and criteria lead in different directions than those of computational experts.

Deciding what is significant is a way of organizing what can be seen. When data are labeled significant, they become more visible—they *count*. When data are labeled not significant, they become invisible and no longer receive attention

Notions of significance are deeply embedded in many data analysis tools; this is not a concept that can be readily customized for different users. In this section, we discuss how molecular biologists grapple with differences between statistical and biological significance, as they try to tease apart the signal and noise generated in biological experiments.

Outliers and Anomalies

To make sure that only good data are presented to users, software tools often set high statistical cut-offs on measurement data (such as the level of signal intensity required to indicate the influence of a gene). However, biological phenomena are often subtle and may be lost when statistical cut-offs are too stringent. Statistical assessments of significance rely on looking at relatively few variables over large samples. But there are many dimensions of interest in biological array data, which may not all have the large number of measurements needed by traditional statistical methods. These data call for new kinds of analytical approaches from computational experts. Also, it is important to consider what constitutes a new finding: an effect that is slight but unexpected can be much more informative than an effect that is pronounced but already understood by biologists.

Many biologists talked about wanting to look around the edges of their tools' statistical cut-offs, to see what is "under the shadow." They want to see shades of gray, rather than just black and white. When they can't see past a statistical wall, it raises concerns about the integrity of the data analysis. It is especially difficult to figure out how to handle outliers—those bits of data that don't seem to line up with the rest. The problem is that there are different ways to account for the categorization of data as outliers. They might be effects of the mathematical algorithms—perhaps the algorithms rendered marginal some data that are worth paying attention to biologically. (An odd variation in the pattern might be just the thing to look at.) Or an anomalous piece of data might be the outcome of unintentional variations in the experimental set-up, such as a smudged slide. Or it might be just a sideline to the important biological story whose patterns are beginning to take shape in the data In this case, biologists want to set it aside. Depending on the explanation, outliers are either *extra* data that can be safely discarded or *missing* data that should not be discarded. Of course, biologists don't know how to explain outliers unless they take a close look at them, which may be difficult with tools that make them hard to see. This is not a problem of biologists not understanding statistics; they do understand them. It is rather a difference in how to interpret and work with categories such as "outlier."

Here is how one biologist discussed outliers in his data set:

> So what defines that data set, and what our gut feeling about it was that I think it was nice that there were these overlaps of genes, and I think if you shrink the data set down too much where you have only one copy of the gene, if it shows up and it's out there in the middle of nowhere and there's nothing questionable around it, you don't know if it's a red herring or if it's just the

most important gene That's why it should be looked at . How do you look at it? These are distributed here because they're meaningful and that's the reason they're here? Or were they distributed here as some type of statistical variation? I think that if a chip has multiple homologs [variants] for a certain gene, you start becoming more comfortable with these outliers

The biologist is pointing here to the useful strategy of taking statistical significance into account in his experimental design. In particular, he finds that it works best if there are variants of the same gene on the array. When the sample is hybridized to the array, the expression levels for that gene should be about the same for each of the variants. If that turns out to be true, then the biologist can at least feel assured that the result is correct, even if it can't yet be explained biologically. Here is a description from another biologist along the same lines:

So as your sample size increases, the need to view the actual image [fluorescence data] decreases But these experiments are only one of the two ways people are doing things The other way people are doing things is they are taking one sample and doing it in triplicate and looking at the genes that change over that series of experiments, so in that case your N is somewhat smaller. Small is relative, whether it's small or not What's too small?

In this case, the biologist describes a strategy of replicating sample runs rather than replicating genes on the array, but it's the same idea of building in redundancy. Yet even with redundant designs, it is still hard to know how much is enough.

Flexibility and Rigidity

Biologists in the NIH lab are also attuned to the material features of the microarray technology and how these might affect their statistical analyses. The microarray technology includes printers and scanners, and clogs or other problems can lead to a "dirty" hybridization, where one portion of the array is no good Here, a biologist points out that analysis tools can't always cope with situations like this:

You can have non-specific hybridization going on across the whole slide, which was missed by the analytical measurement What you can do is, you can look at the ratio outliers to see if they're uniformly distributed or non-uniformly distributed The histogram and scatterplots will give you some idea also, but it's the whole picture, it's all the parts And likewise you can also look for some dirty hyb. It could be confined to one quadrant or it could be the whole thing If it's the whole thing, the statistics usually tell you, but if it's a portion of the array, the statistics sometimes don't pick it up

In circumstances like this, it may be the biologist who can judge what is going on and the statistical tools that have limited information. In this case, the biologist wants to be able to tell the tool, rather than the other way around—for example, to mark off one quadrant as unsalvageable, while keeping the remaining three quadrants in the analysis. This is a complex scenario; in some

circumstances, several different tools used in combination can elucidate the problem. However, sometimes the whole hybridization has to be thrown away.

Biologists in the NIH lab draw a contrast between biological and computational styles of thinking in terms of flexibility In their view, statistics are unnecessarily rigid. As we watched people using their information tools, people offered a number of suggestions for how the tools could be made more usable. Most of the time, the suggestions had to do with loosening up restrictions and making hard-wired operations tailorable. What about letting the user adjust the angle that separates clusters, or allowing more cross-experiment comparisons?

Besides offering usability feedback, the biologists' comments reflected more generally on their sense that they wanted to steer the tools differently, as biologists. One person explained as he showed us an operation he wanted to do:

> So if I'm the biologist and I want to say: 'I think for this experiment, I think these spots are important and also these, only these things are not important so I want to draw a line like this '

We notice in these comments how the biologist claims his authority to direct the tool. He says, "if I'm the biologist," because he wants to remind us that the biologist should be in charge. This same person named the central issue as one of interdisciplinary differences:

> Sometimes I find that the statistics thing, the hard statistics world, is not well applied to biology because we're not that rigid somehow Because life is always very flexible So it's really hard to say, "Cut-off is this " What about the red over there? So that's the thing we deal with constantly

It is important to keep in mind that there are strengths in the statistical analysis tools, as well as uncomfortable mismatches. Biologists do want to let the statistics inform the biology, as well as the other way around—the analytical tools have capabilities they want to take advantage of. The design problem is to find a balance that works to suggest new directions and at the same time support biologists' ideas about what makes sense biologically.

Changing Work Practice

We have discussed the different stories and types of significance produced by biologists and computational experts as they work with data. In this section, we turn to the question of how microarray technology rearranges lab practice. Microarrays are a disruptive technology—they perturb the customary rhythms of the research projects in which they are used.

Molecular biologists in the NIH lab have experienced a change in the pace in their experimental work when they use microarrays. There is a difference in what people do and for how long. In particular, data analysis has assumed a much larger part of the picture, to a degree that was unexpected by the biologists.

Consider this account of a visiting biologist in the NIH lab, in answer to a question about how long she had been working on her project:

> Biologist Oh yeah! Almost two years since I started doing the first hybs I think maybe this case is a bit unusual because I had a very limited number of samples, and they're so precious I had to be sure that it was going to work, if I was going to do a hyb I was actually working in [my home lab], and I just came here to do these hybs. We stupidly thought that I was going to do it really quickly and then just go back home. It didn't work that way. So I decided to come back—so there was actually a break, when I went home and stuff. And then we did this with fewer samples and started doing MDS [multi-dimensional scaling] plots and everything We knew that we had to do more tumors, so I did them last—oh, around Christmas, I guess But since then, and even meanwhile, when we didn't have that many samples, everybody was working on the data frames, looking at them in different ways It's definitely the largest portion of the work, analyzing the data.

The project is not neatly divided into temporal phases. The biologist and her colleagues continue to explore the data produced in early hybridizations while she looks for more samples to use in later hybridizations. The insights they get from these plots and clusters suggest new things to try with the next sets of data. For this biologist, the work of "analyzing the data" seems to encompass more and more of the project.

We are interested in how people experience these changes. One of the biologists in the NIH lab described data analysis as a kind of "downtime":

> I guess one consequence of working out there on the edge—there's a lot of downtime in the sense that—what do I mean by downtime You know, when you're used to working at the bench 8, 10 hours a day, there's no waiting at the bench Things happen, there's a protocol, there's a script you have to follow In some cases with this data analysis, we're waiting for it You know, it's a process, and we're all willing participants I'm a biologist, a human geneticist by degree, and actively involved in the bioinformatics and the data analysis, and that's where we want to go

There is considerable ambivalence in this biologist's story of what it feels like to be working out on the edge. The time away from the bench is unstructured (there's no protocol or script to follow), and he feels it as *downtime*, a time when things are *not* happening, rather than a busy and productive time. This particular biologist is notable for the breadth of his tool set But more options lead to less certainty about which paths or protocols he should follow, and some of the analysis options take a very long time to run. The biologist is convinced of the utility of microarrays and plans to continue using them in future projects. Microarrays, with their huge data sets and attendant data analysis tools, are "where we want to go." But microarrays also interfere with the usual protocols for lab work and experimentation, and biologists do not yet have stable new protocols that capture the particular rhythms of microarray work

It is not just the introduction of microarray technology that produces uncertainty in lab practice. There are continued changes in technologies and experimental methods (such as protein and tissue arrays), so biologists must

repeatedly adjust their practices. Also, biologists sense that their data may have a short shelf life, depending on changes in tools and instrumentation. When tools change, old data may no longer be analyzable. Comparison is at the heart of data analysis: As we have seen, biologists compare one set of tumors to another, or normal cells to tumor cells, or cells at one moment to cells at another moment. Cross-experiment comparisons are made possible when each experiment is first compared to a common reference of some kind. But when the instrumentation changes, this throws up a barrier between past and future experiments—the reference changes too, and data can no longer be compared. New software can lead to the same problem, if data formats change.

Computational approaches change the kinds of research questions people ask in molecular biology. The rapid updates to computational technologies increase the sense of urgency people feel in designing experiments to address their new questions. We have offered only a brief discussion of changing work practices here, but our interviews suggest that this would be a fruitful area to explore through extended observations of lab practice.

Implications for Software Technology

In this paper we have described the work of molecular biologists and some of the current changes in their science and practice. These are turbulent times for molecular biologists: As new understandings and new experimental approaches emerge, new questions and new ways of considering data emerge too. Microarrays in particular offer potential for great insight, but they do so by generating phenomenal amounts of data. These changes have led biologists to a collaboration with computational experts (statisticians, mathematicians, and computer scientists) to help in interpreting the resulting mountains of data. We have characterized this collaboration in two ways. It is a *collision*, in light of the felt impact of new methods of analysis and new colleagues who have quite different ways of making sense of data. It is also a *convergence*, as biologists and their analytical allies strive for a shared understanding that shapes the new biology.

We have described how people from each discipline bring their own interpretative frame to the data and negotiate to find a mutually understandable and workable perspective—a "consensus hallucination." At the end of the day, people in biological and computational disciplines try to produce biological understanding by bringing their distinctive interpretive frames together. But as we have discussed, it is likely that there will be an ongoing need for negotiation between disciplines. It is not the case that biologists can simply learn how to run the numbers, the numbers and ways to run them continue to be problematic as biologists ask new questions and encounter new forms of data. Similarly, mathematicians and computer scientists are challenged to develop new analytical

methods to deal with the flexibility and multi-dimensionality of living systems. Biologists and computational experts need to continue their collaboration.

This collaboration is largely mediated through software. Software plays a central role in representations of data, analytical tools, and databases of genomic information and research publications. There is currently a proliferation of new tools. People are trying out a variety of approaches, from natural language analysis of scientific literature to machine learning and probabilistic reasoning over data sets. In this environment, we have to assume that collaborative capabilities must coexist comfortably with many different analytical tools and databases. Support for rapid updates, user customization, and mix-and-match tool strategies is needed. Drawing on our understanding of the difficulties experienced by biologists working with computational experts, we offer several recommendations for software to support their collaboration.

Support exploratory thinking for groups. When biologists work with computational tools, they try out different ideas—changing parameters to see how clusters shift, going out to the gene databases to see what information comes up about some genes that appear in an interesting pattern, using a different visualization technique to see what new relationships it highlights. This is an exploratory process, and it is easy to lose track of the trail one has followed and the links that have emerged along the way to developing a biological story. It would be useful to have explicit support for the earliest phases of observing relationships and developing hypotheses from array data. This support might look like design rationale tools (e.g., Conklin & Begeman, 1988, Moran & Carroll, 1996), but the kind of software we envision would emphasize the short-term capture of emergent lines of thinking, rather than long-term archives. The idea is to capture a developing understanding in a lightweight way—to help people remember (for now) that this gene should be explored further or that cluster looks familiar from another context. Annotations of both data and actions (across diverse tools) would be useful for this purpose.

Take advantage of local experience. In addition to helping a small group to track its own developing ideas, it would be useful if people could piggyback on the ongoing activities of larger groups, such as those sharing a lab or a particular research focus. As people find that certain articles or gene database entries are more relevant for their purposes than others, they need ways to make those more salient to those around them and to their own computational environments. Previous work on adaptive indexing and collaborative filtering of large databases in light of their patterns of use among groups (e.g., Furnas, 1985, Maes, 1995) might be fruitfully reapplied in this domain.

Support drill-downability. Array data span many layers of representation. As the data are visualized and manipulated using data anlysis tools, it is often the case that one layer of representation obscures others. For example, after image processing software has resolved the locations and magnitudes of spots on a slide,

spot data are usually replaced by numeric expression levels. There are sideways layers too—a multi-dimensional scaling program shows one view, and a dendogram shows alternative relationships among the same data points. Different perspectives are built into these views. When people in an interdisciplinary collaboration are trying to achieve the kind of double vision we have described, they need to juxtapose multiple views and see how they relate to one another—flexibly drilling down and across layers of representation to build a bigger picture and see how layers (and perspectives) link up. There are analogies here to coordinated multiple view systems, such as Spotfire (Schneiderman, 1999). Supporting navigation through multiple views is one way of handing more control over to the biologists who use computational tools, to enhance their ability to apply their own intuitions while working with computational algorithms.

Acknowledgements

We thank all of the scientists who generously allowed us to visit their labs, peer at their data, and learn about their work. We also thank our management at Agilent for their support, and we thank Robert Ach, Mike Bittner, Patricia Collins, Bonnie Nardi, and several anonymous reviewers for their helpful feedback on this paper

References

Bannon, L & Bødker, S (1997) 'Constructing common information spaces', *J Hughes et al (eds), Proceedings of the Fifth European Conference on Computer Supported Cooperative Work*, Kluwer Academic Publishers, pp 81-96

Bowker, G C & Star, S L (2000) *Sorting Things Out Classification and its Consequences* MIT Press, Cambridge, Massachusetts

Conklin, J & Begeman, M L (1988) 'gIBIS A Hypertext Tool for Exploratory Policy Discussion', *ACM Transactions on Office Information Systems*, vol. 6, no 4, pp. 303-331

DeRisi J , Penland L , Brown P O , Bittner M.L , Meltzer, P S , Ray M , Chen Y., Su Y A , Trent J M (1996) 'Use of a cDNA microarray to analyse gene expression patterns in human cancer', *Nature Genetics*, 14, pp 457-460

Furnas, G. (1985) 'Experience with an adaptive indexing scheme', *Proceedings of ACM CHI '85 Conference on Human Factors in Computing Systems*, pp 131-135.

Gantt, M & Nardi, B A. (1992) 'Gardeners and gurus: patterns of cooperation among CAD users', *Proceedings of the Conference on Human Factors in Computing Systems (CHI '92)*, pp 107-117

Harper, R H.R (1997) 'Gatherers of information. the mission process at the International Monetary Fund', *J Hughes et al (eds), Proceedings of the Fifth European Conference on Computer Supported Cooperative Work*, Kluwer Academic Publishers, pp 361-376.

Lave, J & Wenger, E (1991) *Situated Learning*, Cambridge University Press, Cambridge, England

Mackay, W E (1990) 'Patterns of sharing customizable software', *Proceedings of the Conference on Computer-Supported Cooperative Work*, Los Angeles, California

Maes, P (1995) 'Social information filtering: algorithms for automating "word of mouth"', *Proceedings of ACM CHI '95 Conference on Human Factors in Computing Systems*, pp 210-217

Moran, T P & Carroll, J M (Eds) (1996). *Design Rationale Concepts, Techniques, and Use.* Hillsdale, NJ Lawrence Erlbaum Associates

Schneiderman, B (1999) 'Dynamic queries, starfield displays, and the path to spotfire', <http //www cs umd edu/hcil/spotfire>

Star, S L & Griesemer, J (1989) 'Institutional ecology, translations, and boundary objects amateurs and professionals in Berkeley's Museum of Vertebrate Zoology, 1907-39', *Social Studies of Science,* 1, pp 387-420

Van House, N A , Butler, M H , & Schiff, L R (1998) 'Cooperative knowledge work and practices of trust sharing environmental planning data sets', *Proceedings of the Conference on Computer-Supported Cooperative Work,* Seattle, Washington, pp. 335-343.

Williams, M G & Begg, V (1993) 'Translation in participatory design lessons from a workshop', *Proceedings of the Conference on Human Factors in Computing Systems—Adjunct Proceedings (ACM INTERCHI '93)*, pp 55-56.

Williams, N (1995) 'Europe opens institute to deal with gene data deluge', *Science*, v 269 (Aug 4 '95), p 630.

Johnson, George (2001). 'In Silica Fertilization; All Science is Computer Science', *New York Times*, March 25, 2001.