Available online at www.sciencedirect.com
ScienceDirect
Stud. Hist. Phil. Sci. 38 (2007) 642–656
Studies in History and Philosophy of Science
www.elsevier.com/locate/shpsa

The evolution of BioBike: Community adaptation of a biocomputing platform

Jeff Shrager [1]

Stanford University, Symbolic Systems Program (consulting), Stanford, CA, 94305, USA

Abstract

Programming languages are, at the same time, instruments and communicative artifacts that evolve rapidly through use. In this paper I describe an online computing platform called BioBike. BioBike is a trading zone where biologists and programmers collaborate in the development of an extended vocabulary and functionality for computational genomics. In the course of this work they develop interactional expertise with one another's domains. The extended BioBike vocabulary operates on two planes: as a working programming language, and as a pidgin in the conversation between the biologists and engineers. The flexibility that permits this community to dynamically extend BioBike's working vocabulary—to form new pidgins—makes BioBike unique among computational tools, which usually are not themselves adapted through the collaborations that they facilitate. Thus BioBike is itself a crucial feature—which it is tempting to refer to as a participant—in the developing interaction.

Keywords: Trading zones; Collaboration; Interactional expertise; Genomics; Computer programming languages; Lisp

1. Having been becoming a molecular biologist

Computers as technical tools pervade modern science. Uniquely flexible devices, they can simulate any imaginable instrument, and programming has become a standard skill in many sciences such as physics and chemistry.[2] Biologists realized the importance of computers only after the Web had popularized the 'vending machine' model of computation, where everything is expected to be available at the push of a button. As a result, the current generation of working biologists is not trained in programming and must instead rely upon programmers who are usually not trained in biology. Indeed, whole new fields, including 'bioinformatics' and 'computational biology', have grown up to fill this niche.[3]

About 1996, armed only with training in computer science and cognitive science, I decided that I wanted to be a molecular biologist and work on cyanobacteria, a class

E-mail address: jshrager@stanford.edu
[1] Current address: CommerceNet, 169 University Ave., Palo Alto, CA, 94301, USA.
[2] Galison (1997), Ch. 8, provides a detailed analysis of some of the many complexities of computers as instruments and as simulations.
[3] Bioinformatics represents more of the database side of biological programming, whereas computational biology represents more of the scientific and mathematical side. One can even get vocational degrees in bioinformatics. Here I will use the term 'biocomputing', to avoid the subtle differences between these.

0039-3681/$ - see front matter © 2007 Elsevier Ltd. All rights reserved.
doi:10.1016/j.shpsa.2007.09.009

of environmentally important marine photosynthetic bacteria.[4] Fortuitously, at about that time a friend asked me to join his computational drug discovery start-up, which offered me both an excuse to learn organic chemistry and biochemistry, and a possible path to the funding I would need to be able to volunteer in a lab. While commuting to that job by CalTrain, I studied chemistry and biology from books, and in 1999 I took the Cell, Molecular Biology, and Biochemistry Graduate Record Examination. Soon after, in 2000, I volunteered in the laboratory of Dr. Arthur Grossman at the Carnegie Institution of Washington, Department of Plant Biology.

When I joined the lab I began to keep a 'cognitive diary' of insights about my learning about 'real' laboratory biology.[5] Although it runs only a few months, many kinds of learning are described, categorizable roughly as: (a) concrete procedures including both physical skills and ways of attending and perceiving, (b) explicit knowledge about concrete procedures, such as the reasons things work or not in various situations, (c) abstract or concrete facts and relations about organisms or biological systems, (d) extra-domain knowledge and skills, such as what procedures are preferred in which labs, which journals are best to read or submit to, how to speak deferentially to faculty v. post docs v. grad students, how not to invade people's bench and freezer space, how to keep notebooks, and (e) ways of speaking, reading, writing, and interacting. Here is an example entry that crosses a number of these types of expertise (all emphasis is as the source text):

> *20000709*: [. . .] although I can run gels now easily, I am not yet sure what in all, or even in most cases I'm supposed to get as a result. So I asked -LZ-, and she showed me what to expect. HOWEVER, in showing me how to do this, she asked me which vector I was using, and I told her that I was using the T-Gem kit. ''But which vector?'' she persisted. Is there more than one?! I had only the vaguest idea of what she was talking about. So she showed me in the manual that there are two vectors, one of which is cut by EcoRI and one of which isn't. Which one had I used? Now, I had studied the manual for this kit VERY VERY CAREFULLY before beginning into this process. But it was suddenly clear to me that I hadn't understood a word that it was saying! There are two vectors?! Um, well, whatever one you handed me. As it turns out, I was lucky and had used the vector that is, in fact, cut by EcoRI, and the gel actually worked—or, more precisely, the restriction worked—or, even more precisely, one of the four copies of the restriction worked (I've learned to do everything in four copies because three of them won't work. I'm hoping against hope that after a few years I'll be good enough at this that I won't have to do this all the time.) So this time it happened to work out, and I learned a small but important fact about the use of this kit, which is that you have to keep track of which vector you're using.
>
> But [. . .] nearly at the moment at which -LZ- explained to me about the two vectors, something much larger clicked into place for me. I don't know quite how this happened, but somehow I had all the pieces of the puzzle (well, this local puzzle anyhow) in hand and identified, but hadn't put them into the frame. When -LZ- showed me the picture in the manual of the two vectors, with their various restriction sites, that was the frame for the whole procedure, and all the pieces fell right into it, and I very suddenly—literally in a matter of a few seconds—''saw'' what I had been doing for the past day: I could see why we were cutting the vector and amplifying the gene, and ligating them together and why I had to use EcoRI. And then I understood, all in that same perceptual unit, how to figure out what to expect from the gel. Maybe this was just the first time I had actually had time to think, as opposed to feverishly cooking and being lost, but it doesn't feel like that. I think that I've been trying to think all the way along, but there just wasn't enough material to think with, or there were crucial pieces missing, or the frame was missing, or something.

Much of this could be described as learning how to 'talk biologist'—what Collins and Evans term 'interactional expertise'.[6] A more literal example of language learning is found in the following extract:

> *20000724*: [Note added 20000830: Everyone says EcoRI: ''Eco-R-one,'' (not ''Eco-R-Eye'') but there seems to be disagreement about some others. For example, -DB- calls BglI ''Bagel-one'', which sounds dumb to me. And -LZ- calls SmaI ''Smaaa-one'', which also sounds dumb. Generally, unless the thing is very common and very pronounceable (like EcoRI), I prefer to spell out the letters, as ''B-G-L-one'' or ''S-M-A-one'', and I've heard people do this. Generally, saying things wrong is among the most embarrassing beginner mistakes one can make. It marks you as not knowing what you are talking about, unless you are a foreigner, even though these are all made-up words, so who cares! . . .]

The point to take away from these examples is that after nearly three years of, on average, a few hours/day of study of biochemistry and biology, I got to the lab armed with a great deal of book knowledge yet still had little idea how to do or talk molecular biology. This I learned 'on the job',

[4] B.S.E./M.S.E 1980/1981 in Computer Science, M.S./Ph.D. 1982/1985 in cognitive psychology. In my Ph.D. work I studied human learning and simulated it with computer models. I subsequently spent nine years conducting research along those lines, and then did a post doc in cognitive neuroscience.

[5] Shrager (2000). This formed the basis of Shrager (2004), and Sahdra & Thagard (2003).

[6] Collins & Evans (2002).