



The Library of Maynard-Smith: My Search for Meaning in the Protein Universe

Borges's infinite, and unsearchable, library is reflected in the Maynard-Smith collection of all possible protein sequences

Frances H. Arnold

I'll never forget reading Jorge Luis Borges' short story "The Library of Babel" when I was working in Madrid in the summer of 1976. I was thrilled by Borges' description of this collection of all possible books, and deeply struck by the despair of the librarians, who searched for meaning in the essentially infinite stacks of "senseless cacophony, verbal nonsense, and incoherency." I found it delicious that the librarian's certainty that one of those books contained the answers to all the fundamental mysteries of mankind was accompanied by the unbearable realization that he could never locate such a treasure. He had to accept finding a few phrases embedded in typographical gibberish as the sum of his life's work. I could not imagine spending life wandering that desolate maze.

Two decades later, Dan Dennett's wonderful book, *Darwin's Dangerous Idea*, brought back these youthful memories and put them into context for me in a powerful way. Dennett's variation on Borges' library, the Library of Mendel, is the collection of all possible genomes (whose written description, by the way, is a subset of the Library of Babel). And what glorious creatures Mendel's library contains! But the vast majority of possible genomes does not encode life—as has been noted, there are far many more ways to be dead than alive.

I have spent my years searching for meaning in one of the virtual rooms of the Library of Mendel, the Maynard-Smith Collection, which houses the set of all possible protein sequences. Like the Libraries of Babel and Mendel, the

Maynard-Smith Collection is incomprehensibly large, and the vast majority of its sequences encode rubbish. Instead of 500 pages, a typical protein is only 500 letters long, and there are only 20 letters (amino acids) in the protein alphabet (for now, at least). But that gives 20^{500} different combinations, already well beyond the number of particles in the universe. Meaningful proteins—proteins that DO something—are extremely rare, because there are far more ways to be meaningless (e.g., unfolded and unable to function). I can revel in all the gorgeous proteins that must be contained within this collection—cures for cancer, the answer to the energy crisis. But have I joined the Babel librarians in a hopeless search for these magnificent, but extremely rare books?

Maynard-Smith's library is different from Borges' in two very important ways. These differences have made it a far more satisfying universe to explore. The Babel librarians' deep depression came from their inability to find even a comprehensible sentence, much less a book, even a bad one, written in any recognizable language. They were also frustrated by the apparent random order of the books. Reading one book would give no clue as to the contents of the next, and therefore no directions as to where one should go in order to find a better book.

In the libraries of Mendel and Maynard Smith, meaningful books jump right out at us. They are alive! You can literally scrape some of these rare genomes, ones that actually encode life, and all their glorious proteins, from the bottom of your shoe! (Of course there are many

Frances H. Arnold is the Dick and Barbara Dickinson Professor of Chemical Engineering, Bioengineering and Biochemistry at the California Institute of Technology, Pasadena. This article is one of a series that are adapted from an upcoming ASM Press book on *Darwin, evolution, and microbiology*.

other interesting and meaningful genomes, and proteins, that you cannot find on your shoe, or anywhere on earth for that matter, but at least we are not throwing ourselves off the balconies of the Babel Library for never having had the pleasure of reading such a book.) This is the great gift of natural selection, which has done the search for meaning for us. Natural selection relentlessly washes out the weak stories and supports the strong. And woe to those who do not heed natural selection's editorial dicta! They might become recipients of the infamous Darwin Awards, conferred—posthumously in most cases—on those who “do a service to humanity by removing themselves from the gene pool.”

The other key difference between the Maynard-Smith (virtual) Library and the hopelessly unorganized Babel warehouse is that there is order in the former. Because he was not limited to three paltry physical dimensions, Maynard-Smith could organize his protein books in a very special way. Each sequence is surrounded by its one-mutant neighbors, that is, by all the proteins that differ from it by a change in single amino acid letter. Never mind that there may be thousands of ways to change a single letter in a protein book—remember, this is a virtual library, not a physical one. Maynard-Smith set up his library with as many dimensions as there are ways to make a mutation.

Why is this ordering so important? These neighbors are the ones that evolution explores. When a random mutation—a step to a neighboring book—is made, natural selection “reads” that book. And, with natural selection inexorably sweeping away the debris, only the meaningful ones are left behind. A step to a bad book just sets you back to where you were before. Over millions of years, this is just what evolution has done, taken meandering paths through this vast space of possible proteins. Look at all the interesting molecules that have emerged! And we only see the ones that still exist, on Earth, in 2011. So many beautiful proteins have been discovered and incorporated into the living world. So many more have been tested and forgotten, or found wanting and thrown away. And many, many more have never been tested at all, for there has not been enough time, since the beginning of life on Earth, for evolution to have tested even a small fraction of this vast space of possibilities.

Now, imagine that *I* get to decide which

books in this library are meaningful. I'm on a search for new proteins, proteins unlikely to be found in the natural world. Proteins that might be useful to someone or that will tell me something unconventional that natural proteins, the few that survived editing by natural selection, cannot. I'm the pitiless editor now, and I get to judge each book by my own criteria. I read each one, and if I don't like it, I throw it away and go back to my previous book. But if I do like it, I move on, taking another step in a random direction. With a single change, a mutation, the book may look a lot like the one before, but then again it may not if that word is critical. (Dennett's example from *Moby Dick*: consider the difference between “Call me Ishmael.” and “Call me, Ishmael.”) If I have a particular story in mind, I can even choose to move to books that begin to build that story. I can accumulate my choice of one-letter changes, and if I am patient I might even be able to follow this path to a pretty good new story. It may not be the “best” book in the collection, but remember that this is a very big library. I can't hope to find the best. But what I find may be good enough to keep me happy. Especially if the alternative is the book I started with, or no book at all.

My students and I have spent the better part of 20 years taking sometimes straight and sometimes slightly crooked paths in Maynard-Smith's library of proteins. Maybe “forced marches” is a better way of describing our walks: these experiments can be tiring. It's a lot of work to read all these sequences—make all these proteins—and decide which are interesting and which are not. So my forays have been short. And few.

While I have explored but the tiniest fraction of this remarkable place, I can tell you some interesting things I have learned during my travels. First, even more than books do, protein properties can change quite significantly with even a single mutation. Proteins can become more stable, they can change their biological functions, they can change color. Most of the possible mutations don't have a big effect, as far as I can tell with my limited eyes. And of course some of the mutations can greatly damage or even destroy a protein (and remove it from the gene pool). But we can often find mutations that allow the protein to do new things, such as catalyze a reaction on a new substrate, or adapt to environments that it had not previously tol-



erated. I was gratified, perhaps even surprised, to see how “evolvable” these lovely molecules are, how quickly they could “learn” new tricks! Of course, now I understand that this property of adaptability is precisely what allowed protein-based life to take over and diversify into the many forms we enjoy today.

I also learned that proteins can do lots of things that nature probably never asked them to do, including things that humans might benefit from. A reviewer of one of my earliest proposals (back in the previous millennium) criticized my idea to direct the evolution of proteins to do nonnatural things, like function in an organic solvent. He said that proteins simply couldn’t do that, because nature had never gone there before. Of course, this is exactly the wrong reasoning. It’s precisely *because* nature did not ask for these behaviors that we were so successful in finding mutations that allowed them. In fact, it’s much harder to find mutations that will improve a protein doing its natural job, because nature has been looking for those for a while. But improvements in something that the protein does not naturally do well are usually far easier to find. There’s much more room for improvement, and there are many more paths that lead to improved sequences, provided you don’t ask for too much too fast.

I also learned that there are plenty of places that a hike in Maynard-Smith’s library isn’t likely to take you. Since I don’t have the lifetimes to devote to my searches that Borges’ librarians did, I have to be careful about which steps I take. There are lots of steps that take me away from the protein I am looking for, and only a very few that appear to take me closer. I say “appear” because I cannot see beyond the protein in front of me; it’s as if I am climbing a mountain in the clouds and can only feel or see one step ahead. I choose, therefore, to take the step that moves me closer to the top, but I can never know if this particular route will ultimately lead me there. So it takes some faith as well as patience.

This blind walk has nonetheless been highly productive. Since we understand so very little about how the sequence of a protein encodes meaning, that is, how its sequence determines what a protein does, we cannot write our own books. That, of course, is the dream of many a protein engineer, but I fear that it is a distant if not impossible one. We have to discover interesting new proteins by making them and seeing what they do. I have found that a walk in Maynard-Smith’s collection is a wonderful way to discover these new proteins. There are wonderful stories just waiting to be discovered.

SUGGESTED READING

Borges, J. L. 1941 The Library of Babel. Editorial Sur.

Dennett, D. C. 1996. Darwin’s dangerous idea: evolution and the meanings of life. Simon and Schuster, New York.

Smith, J. M. 1970. Natural selection and the concept of a protein space. *Nature* 225: 563–564.