How do you cross an armadillo with a porcupine and other problems that arise from naming proteins

OK. Now it’s getting really weird. Not so long ago Alpine Cave Dweller sent me some rock-mail about the role of editors in rejecting papers just as I was writing a Sticky Wicket on that topic, and now Caveman Tony does it again by sending one on naming proteins as I was starting a piece on that too! Here is my piece in the form of a response to his mail.

Dear Caveman,

Upon emerging from your winter cave, you surely noticed that spring is here and the 2001 baseball season has begun. Last season was particularly exciting from the standpoint of individual achievement. Mike Piazza of the New York Mets was in pursuit of the record for most consecutive games (18) with an RBI (runs batted in). Most interesting was the fact that few fans and media were aware of the record holder, a little-known player from the 1930s. Conversely, most fans knew that Joe Dimaggio holds the record for most consecutive games (56) with a hit.

At this juncture I know that you are probably asking: “Where is this troglodyte going with this?” Well it has to do with the naming of proteins and the significance that is frequently attached to these names. It is what I call the name is the name of the game. In my scientific youth, proteins with a known function were named appropriately (e.g. sodium channel, potassium channel). Otherwise, they were given a more descriptive name that reflected their molecular weight or cell of origin. Now, with numerous proteins being identified by molecular cloning, there has been a proliferation of protein names that appear to be designed to catch the reader’s attention in an ever more competitive market. There are protein families, novel proteins, signaling proteins, cytoskeletal proteins, proteins with functional domains, and so on. Many of these have clever names (my favorites are the Drosophila proteins) that clearly catch one’s attention.

Caveman, do we have to give our protein(s) interesting names (even if they are presently uninteresting since their function is unknown) so that our data do not get lost in an ever expanding protein universe? Are we, in essence, headed towards a situation like the baseball analogy presented above? If the protein (or baseball record) is highly recognizable, then its importance may take on a significance greater than that of a less recognizable one? Perhaps the solution to all this is to form a ‘universal

STICKY WICKET

An occasional column, in which Caveman and other troglodytes involved in cell science emerge to share their views on various aspects of life-science research. Messages for Caveman and other contributors can be left at caveman@biologists.com. Any correspondence may be published in forthcoming issues.

protein naming group’. My own personal favorite would be Your Own Darn Antigen (YODA), although adoption would probably require permission from George Lucas.

Sincerely yours,

Tony the Trilobite (a.k.a. George Ojakian)

Dear Caveman Tony,

Welcome back into the sunlight! For some of us, winter has been over for some time! But, I perceive that you equate the end of winter with the start of the baseball (a.k.a. rounders) season. I am afraid that your discussion of the prowess of one Mike Piazza is wasted on me. If this were what I was waiting for at the end of winter, I would turn over and go back to sleep!

I am always impressed with the significance of statistics in baseball: most consecutive games (period), most consecutive games with an RBI, most consecutive games chewing tobacco, etc. Anyone would think that this was a form of slave labor: poor Mike Piazza taken from his parents when he was three, forced to learn how to hit a small ball with a bat and then run, and doing all of this while holding down two janitorial jobs and being paid like a postdoc. I’m sorry; he gets paid millions and has nothing else to do all day except hit a ball and run – in my opinion, he (and the rest of them) should be doing everything on a consecutive basis. Mind you, at least a baseball game only lasts a few hours, unlike another bat and ball game that can go on for five days!

Despite this digression, you make an excellent point in your rock-mail about naming proteins. It is quite interesting to consider what is done in different organisms. In yeast, it is the genes that are ‘named’ rather than the proteins, which have a diminutive ‘p’ placed after the name. But the gene name is sort of matter of fact, rather than imaginative (see below). Genes involved in protein secretion are termed ‘SEC’ followed by a number - SEC18 – for example, and the corresponding protein is Sec18p. With the completion of the yeast genome sequence several years ago, there is a movement to annotate the genes in a more systematic manner. Similarly, genes identified in C. elegans are also named on the basis of the type of screen that was used to identify mutants - hence UNC or LIN - and, as in the case of yeast, genes in these different categories are given numbers (e.g. UNC5). Very logical and functional, albeit a little boring, but that is perhaps what you’d expect from geneticists.

In other organisms and cells, a considerable amount of imagination has gone into the naming of genes and proteins. In Drosophila, genes have been named on the basis of a sort of description (with a lot of latitude) of the phenotype of the mutant. In the initial screens, larval mutants were classified according to the pattern of surface denticles, which gave rise to names such as hedgehog, armadillo, porcupine, etc. This set the trend, and every gene has a name that sometimes has relevance to function - breathless (a gene required for formation of trachea), shaker (K+ channel; the flies actually shake so much they vibrate), and my personal favorites, grim and reaper (genes that control apoptosis, of course) - I am also waiting for the ‘not the salmon mousse’ mutant (for those of you who remember your Monty Python!). I agree with you Caveman Tony that these names are kinda fun, although I know how much contempt yeast and worm geneticists have for them.

In mammalian cells, many different approaches have been taken. Sometimes a good classical education is helpful (some of the intermediate filament proteins come to mind, like vimentin and desmin). Other times, a simple knowledge of the function of the protein is sufficient for one to coin a name - channel, transporter, pump or antiport. And then again, a receptor is a receptor, unless it happens to be an orphan. Then, there is the acronym - Arf, SNARE, or perhaps the most banal, JAK, which stands for ‘just another kinase!’ Of course, the simplest solution to all of this is to call every protein ‘CD’ followed by a number, the inspirational approach taken by immunologists! However, this is one better than those who have no idea of the function of a protein that they have identified and are relegated to using ‘p’ followed by a number to designate the molecular weight (e.g. p175). Of course the fact that different organisms express homologs (or orthologs) of proteins furthers the confusion. Is it Sec18p or NSF, or armadillo or β-catenin (and UNC??), and which CD?? is the same as α3β1 integrin?

It is all very confusing, Tony. But, I am not sure what the solution is. YODA is a possibility, but that smacks too much of a scientific ‘Thought Police’, but then again perhaps we should let the geneticists take charge.

Caveman

P.S. I am still wondering what baseball, Mike Piazza and Joe Dimaggio have to do with the naming of a protein. I can only surmise that the winter was particularly hard in New York and you have a prehistoric form of cabin (cave?) fever.