Back to Basics

The *Ka/K*s ratio: diagnosing the form of sequence evolution

What is *Ka/K*s?

The ratio of the number of nonsynonymous substitutions per nonsynonymous site (Ka) to the number of synonymous substitutions per synonymous site (Ks).

Sorry, what was that?

Let us start from the beginning. Imagine you align the sequences of the same gene from two species. There will usually be differences between the sequences (evolution!). Some of these will lead to differences in the amino acids of the encoded protein (nonsynonymous changes) and some, because of the degeneracy of the genetic code, leave the protein unchanged (synonymous, or silent changes). Counting up the number of each gives us a measure of the amount of change of the sequence. Then we have to adjust these figures.

I see, we now have measures of the rates of evolution. So why adjust anything?

Due to the degenerate nature of the code, only about 25% of the possible changes in our sequence are synonymous. Imagine that our gene is not under selection in either species; that is, it is evolving neutrally with chance alone determining whether a new mutation goes from rare to common. Most such mutations are lost by chance, but the chance that a given new neutral mutation will go to fixation in a diploid population is $1/2N$, where N is the population size. In this case, the likelihood that a nonsynonymous mutation would go to fixation is the same as that for a synonymous mutation. So, if I correct for the degeneracy of the code, I should have a method that reports that the number of nonsynonymous changes at each possible nonsynonymous site is the same as the number of synonymous changes per synonymous site; that is, $Ka/Ks = 1$. Deviations from a ratio of one will then tell me something about the selective forces acting on the protein, given that Ks is telling me the background rate of evolution.

Sounds simple enough…

Unfortunately not. For example, consider the codons specifying aspartic acid and lysine: both start AA, lysine ends A or G, and aspartic acid ends T or C. So, if the rate at which C changes to T is higher

from the rate that C changes to G or A (as is often the case), then more of the changes at the third position will be synonymous than might be expected. Many of the methods to calculate Ka and Ks differ in the way they make the correction needed to take account of this bias.

Doesn't the time that has lapsed since the two species separated matter?

Good point. As sequences diverge over time, the observed number of changes underestimates the real number of changes. Imagine a nucleotide that begins as A. In one lineage it is replaced by a C and then again by a T, the two changes would still only cause one difference in the alignment. Also, sites that match in the alignment could have changed independently, but ended up the same. Fortunately, the extent of real divergence can be estimated from the total observed amount of divergence. These 'multi-hit correction' methods also differ in their sophistication. However, none can work miracles: as the number of changes increases, the amount of information from the alignment decreases and we approach saturation, in which case the data are useless.

I can see this is getting a bit complicated. Does all this really matter?

Unfortunately it often does. For example, the relationship between GC content and Ks is important in the debate about the evolution of GC content in humans, but the pattern is very sensitive to the method of estimation.
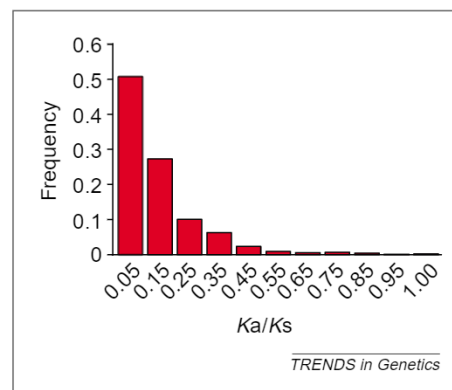
Fig. 1. The frequency of different values of *Ka/K*s for 835 mouse–rat orthologous genes. Figures on the *x* axis represent the middle figure of each bin; that is, the 0.05 bin collects data from 0 to 0.1

OK, so I have *Ka* and *K*s for my gene in a given comparison between species. So what?

Well, you now have figures for the amount of protein evolution (Ka) that is comparable between genes from the same species comparison. If selection does not act on silent sites (a big if!), from the neutral theory of evolution, Ks should also be proportional to the mutation rate of the gene. This is because, while the probability that a new neutral mutation goes to fixation is $1/2N$, they are created each generation at a rate $2N\propto$, where $\propto$ is the neutral mutation rate per generation. Hence, the rate of neutral evolution is $2N\propto/2N = \propto$ and it is this rate that Ks is measuring. If our method has gone to plan, the ratio of the Ka and Ks also now tells us about the way the gene has evolved.

What we usually find is that Ka is much less than Ks (i.e. $Ka/Ks \ll 1$; Fig. 1) because a mutation that changes a protein is much less likely to be different between two species than one which is silent; that is, most of the time selection eliminates deleterious mutations, keeping the protein as it is (purifying selection).

Sort of what you expect, given that proteins evolved to do what they do and do it rather well.

Yes, but this is not always the case. In a few instances (often when immune system genes co-evolve with parasites), we find that Ka is much greater than Ks (i.e. $Ka/Ks \gg 1$). This is strong evidence that selection has acted to change the protein (positive selection).

I get it. So if *Ka* equals *K*s then evolution of the sequence must be neutral?

Not so fast. Neutral evolution is a possibility that cannot be excluded. But, what if one part of the gene (one protein domain, say) was under positive selection, but other parts under purifying selection? Then you might get an average $Ka/Ks = 1$. There are recent methods that allow you to take a multiple sequence alignment, the phylogeny of the species involved and work out a Ka/Ks ratio for each codon. If in all lineages a codon is undergoing positive selection, this is a powerful method to

detect it. Alternatively, you can ask whether there is a different ratio in one lineage, suggesting that something happened peculiar to that species. These new sorts of analyses are revealing much more positive selection than we suspected.

Forget the details, how can I convert my alignment into *Ka* and *Ks* estimates?

There are several different ways you could do this. There is an excellent free package for PCs called MEGA2 that implements loads of different methods:

http://www.megasoftware.net/. If you have GCG on a UNIX machine, the 'diverge' package calculates the Li93 protocol (not the best, but OK). The codon-based or lineage-based methods are implemented in PAML: http://abacus.gene.ucl.ac.uk/software/paml.html. This will also calculate an old measure (Nei and Gojobori) and a new estimate to the maximum-likelihood method (Yang and Nielsen). PAML can be implemented on Unix, linux, PCs and Macs (OS X and before).

Where can I read more?

• Nei, M. and Kumar, S. (2000) *Molecular Evolution and Phylogenetics*, Oxford University Press. The book that goes with MEGA2.
• Yang, Z.H. and Bielawski, J.P. (2000) Statistical methods for detecting molecular adaptation. *Trends Ecol. Evol.* 15, 496–503
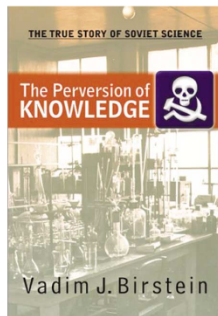
Laurence D. Hurst

Dept of Biology and Biochemistry, University of Bath, Bath, UK BA2 7AY.
e-mail: bssldh@bath.ac.uk

Book Review

Science, a social product?

The Perversion of Knowledge: The True Story of Soviet Science

by Vadim J. Birstein
Westview, 2001. US$ 32.50 (512 pages)
ISBN 0 8133 3907 3

This is a disquieting book. Vadim J. Birstein recounts the story of soviet science (~1920–1960) by listing the fates of soviet scientists and their persecutors. The story begins in the twenties, when soviet politicians tried to create a new type of science, a science that was no longer done by the sons of the bourgeoisie, but by the sons and daughters of the working class. For them, the class to which a scientist belonged was more important than their work. The politicians did not care whether science was right or wrong, merely that it followed or claimed to follow party policy. The party and the secret police were active in enforcing this policy, and many scientists who spoke out against it were shot. One of the defenders of the 'old-fashioned' science was Nobel prize winner Pawlow (1849–1936), famous for studying reflexes in dogs. The old man was fearless; in 1934 he wrote to Bucharin, one of the organizers of revolution who by then had lost all power, 'My Lord, it has become so hard for any decent person to live in your socialist paradise.' However, Pawlow was too famous for the secret service to act against him.

Next, Birstein concentrates on the attack of Lysenko and his supporters on biological science. The general story is well known. Lysenko, a peasant, maintained that he could 'educate' plants, heritably altering their characteristics. He posed as a saviour, claiming to increase agrarian production gigantically. Yet it was all fraud. It is deeply moving to read of those who did not accept Lysenko's lies; most had to leave science and many were murdered. The fates of the plant breeder Vavilov and others are described in detail. Some objectors survived; for example, Iosif Rapoport, an excellent geneticist who had shown that certain chemicals are mutagenic in *Drosophila*. In the World War II, he lost an eye and received the highest medal a Jewish officer could receive in the Red Army. In August 1948, he publicly called one of the defenders of Lysenko an 'obscurantist' at the notorious meeting of the Academy of Agricultural Sciences. He lost his job, but he survived. It is unsettling to read that Alexander Oparin, well known for his ideas on the origin of life, praised Lysenko.

Throughout the book one aspect is missing – the actual research done by the scientists. The reader would like to know their relevant discoveries. Instead, we learn a mass of interesting new details about the persecutions in the Lysenko affair. The fate of scientists is not the entire history of science. The scientific thoughts, experiments and breakthroughs should also be documented, although this is certainly more laborious.

But Birstein has also discovered something essentially new: a biochemist, Grigory Mairanowsky, used political prisoners under sentence of death to test various poisons. Mairanowsky began his human experiments in 1935, although the lab books of his experiments have disappeared, so one can only guess which poisons he and his collaborators tested. In any case, the experiments seem rather similar to the experiments done in German concentration camps; for example, those by Professor Hirt in the concentration camp of Natzweiler. Mairanovsky was jailed from 1951 to 1961, but later his career recovered, and he became director of a provincial institute. What a story. It destroys the illusion that that murderous experiments with humans were the speciality of the Germans or Japanese; the Soviets did them as well.

Genetics was in worse shape in the Soviet Union than in Nazi Germany, where genetics was only partially demolished. In the Soviet Union, genetics was almost completely destroyed. Truth had completely evaporated. In this context the author makes an interesting comparison. He compares the work of the GESTAPO (the German secret service) with the work of the GPU/NKVD (the Soviet secret service). The former tried to get the truth about the enemies of Nazism to annihilate all of them. The latter did not care about truth, just wanting signatures on prefabricated confessions. Both institutions worked differently, but were equally despicable.

In retrospect, it is astonishing that biological science in the USSR survived to begin again in the sixties. Apparently, it is difficult to destroy science: it might almost be dead, but it will recover. What do we learn from the book? Truth is the essence of science. Secrecy and lies signal its ill