GENETICS

Genomic Clues to DNA Treasure Sometimes Lead Nowhere

Geneticists used to think that important DNA sequences always reject mutations. Now they are not so sure what sequence conservation really means

For many biologists, evolution does more than organize the history of life. It also guides them to hidden treasures in our DNA. When a gene works, evolution holds on to it, keeping its sequence intact even as bases around it change over time. Genome researchers had come to depend on this conservation to steer them to critical regions in the genome: If a stretch of DNA remains unchanged across different species, that DNA is probably performing a vital function. But as Eddy Rubin found out, that's not always the case.

For several years, Rubin, Len Pennacchio, and their colleagues at the Lawrence Berkeley National Laboratory (LBNL) in California have combed the genome for regions that regulate genetic activity. Because the so-called enhancers that they study can influence genes thousands of bases away, there are few obvious landmarks to help locate them. So the researchers looked at "ultraconserved" 200-base-long sequences previously found to be identical in rats, mice, and humans and at others that were similar even in fish. The strategy worked—or so they thought. When they inserted those sequences into mice, more than half turned on an accompanying reporter gene in particular tissues at a specific developmental stage.

But when the LBNL team looked deeper at four promising candidates, they were surprised that none of them caused any obvious problem when deleted from the mouse genome. "There are a lot of [sequences] that we thought if we knocked [one] out, it would kill the animal," Rubin recalls—but that didn't happen.

Results like these are causing Rubin and others to take a closer look at just how tightly conservation and function are linked. A growing number of examples show that not all conserved sequences are important and, worse, that not all important sequences are conserved. That second observation—which would have been considered heresy until about a decade ago—means that researchers who had typically relied on conservation to guide them could have missed critical genes or unknown regulatory regions. "It does question an awful lot about what's going on," says Laurence Hurst, an evolutionary geneticist at the University of Bath in the United Kingdom. But even as he and others scramble to understand how the "conservation equals function" rule has failed them, they are uncovering profound new subtleties in how genes are controlled and how they adapt during evolution.

Missing function

The most extensive data relating function and conservation come from the 2007 results of the pilot phase of the ENCyclopedia Of DNA Elements (ENCODE) consortium, which examined a selected 1% of the human genome. Along with many other tests, the researchers evaluated conservation of these human DNA sequences by comparing them with related regions in other vertebrates or between people.

For most regions, mutations that have accumulated over time have resulted in many differences between the bases. The longer it's been since two species parted ways, the more differences there are. But some sequences, particularly in genes, differ less than others. If a sequence is more conserved than expected, researchers ascribe the difference to "constraint," inferring that mutations were rejected during evolution because they reduced the organism's fitness. In genes, those mutations could be particularly deleterious and often are quickly weeded out.

The ENCODE team estimated that about 5% of the human genome is constrained to some degree, as hinted by previous studies. Of this, only about 25% to 30% matched with protein-coding regions. (Overall, protein-coding genes represented only about 1.5% of the DNA in the ENCODE regions.) Most of the remaining constrained sequence was transcribed into RNA—despite being "noncoding" DNA. The constraint suggested that this RNA might help regulate genetic activity.

To test this idea, the ENCODE team assessed biochemical activity throughout the chosen regions, including the constrained noncoding sequences. The researchers looked at whether the DNA binds transcription factors and whether either the DNA or the proteins that package it are chemically altered to silence or stimulate its activity.

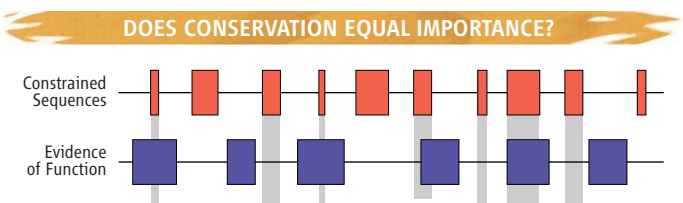
CREDIT: MATTHEW TWOMBLY

Published by AAAS

"We expected all the other [biochemically] functional sequences that we identified to start overlapping the remaining 75% [of the constrained regions]," says Elliott Margulies of the National Human Genome Research Institute in Rockville, Maryland, but only about 60% showed any clear signal in their assays. That leaves 15% of the sequences showing some constraint for no apparent reason.

Some researchers have suggested that the missing functionality is a laboratory artifact: The sequences' true role would be apparent only in a more challenging real-life environment. But in work published in the January issue of *PLoS Genetics*, Jianzhi Zhang and his colleagues at the University of Michigan, Ann Arbor, found no correlation between the degree of conservation in a sequence and its function, even for yeast genes that proved essential in 400 highly varied conditions. "It was not due to the mismatch between lab and environment," concludes Zhang.

Indeed, some conserved regions may truly have no function. "Simply because a sequence is conserved, one should not jump to conclusions," cautions Eugene Koonin of the National Center for Biotechnology Information in Bethesda, Maryland, especially if the conservation is weak. Conserved noncoding introns within eukaryotic genes, for example, may have survived not because they do anything but because "selective pressure might not have been sufficient over all this span of evolution to get rid of them," he says.



Imperfect correlation. Not all DNA conserved between species (red) coincides with functional sequence (blue) and vice versa.

Lack of constraint

Researchers can rationalize the existence of constrained sequences that have no detected function, but they are truly baffled when clearly important sequences seem hardly more conserved than the rest of the genome. In one early example, Hurst and his Bath colleague Nick Smith showed 10 years ago that dozens of essential genes, without which mice die, have accumulated as many mutations since mice diverged from rats as have nonessential genes, whose absence is tolerated.

Hurst recalls that when he began, "molecular biologists said, 'Why would you want to do that?' " The reason, he says, is that he had begun to realize that the widespread confidence in the connection between conservation and function had virtually no experimental backing. The study is now regarded as seminal, but for many years, the result "just seemed to be completely overlooked," Hurst says.

Since then, other worrisome data have appeared. In a 2003 paper, for example, Koonin and his colleagues found only a modest correlation between sequence conservation and function, a connection detectable only after they sampled enough genes to make it statistically significant. And the ENCODE pilot results have confirmed Hurst's suspicions. "Of all of these functional sequences, a large portion showed no evidence of evolutionary constraint," says Margulies. "That was the big surprise." Ongoing ENCODE analyses have not shed any new light on this mystery.

Researchers are keen to understand how important sequences evade evolutionary pressure, and they have proposed many explanations. One is that a sequence may play a vital role in only one of the compared species but not in others. Sequences that fish use for fins, for example, may mutate without penalty in other vertebrates. Comparing many different species, some close and some distant, reveals such lineage-specific genes, says Ross Hardison of Pennsylvania State University, University Park. Comparing many species should also reveal when important genes are free to mutate because another gene picks up the slack. Rubin, for one, thinks this is a common occurrence. "I think there's a lot of redundancy," he says.

Sometimes, biochemical assays may detect activity that has no cellular impact, making nonconserved sequences—such as those in the ENCODE data set—seem important when they really are not. "You might get reproducible transcription, or reproducible protein binding to DNA at specific locations, but they have no biological consequence to the organism," says Margulies.

Beyond sequence

Other, more subtle types of constraints exist that researchers are only now coming to appreciate. "Our view of functional sequences and evolutionary constraint in some ways has been tainted by the first functional sequence that we've known about—namely, protein-coding genes," says Margulies. For example, mutations that create a new three-base codon for the same amino acid and thus leave the protein intact were long thought to be unconstrained because they supposedly have no consequence. (TTT and TTC both code for phenylalanine, for example.) But researchers have found that even these "synonymous" mutations make an evolutionary difference. In one recent example, Joshua Plotkin of the University of Pennsylvania and his colleagues made more than 150 versions of a gene for green fluorescent protein, varying the sequence at synonymous sites. In *Escherichia coli*, the amount of protein varied 250-fold, in large part because codons differentially affected the stability of the messenger RNA produced, they reported in the 10 April issue of *Science* (p. 255).

For noncoding regions, constraint may depend on other properties that are still only partially understood. For example, if the DNA (and thus the RNA transcribed from it) includes nearly complementary segments oriented in opposite directions, the two resulting RNA segments can fold together to form a "stem-loop" structure. Such structures form a key piece of many regulatory RNA molecules and thus tend to be conserved, but their sequence signature is completely different from that of proteins.

The DNA sequence also modifies the interactions with regulatory, DNA-binding proteins. In the 17 April issue of *Science* (p. 389), Margulies, Boston University chemist Thomas Tullius, and their colleagues explored how the local sequence of DNA alters its shape and thus its accessibility to solvent molecules. This approach, Margulies says, "can identify roughly twice as much sequence that's under evolutionary constraint as some of these other methods that look at primary sequence alone."

Clearly, assessing the importance of a DNA sequence is harder than just comparing its bases between species. What researchers need is more data, both genetic and functional, in a variety of species, individuals, and tissues, says Ewan Birney of the European Bioinformatics Institute in Hinxton, U.K., to understand the ways that the conservation-function link breaks and, from there, to discern both the mechanisms of genetic regulation and the complex ways that evolution creates and preserves functions. "Constraint is still an enormously useful tool to identify important sequences in the human genome," notes Greg Cooper of the University of Washington, Seattle. But it doesn't find everything. "Truth be told," he adds, "we really don't know what we're missing." **–DON MONROE**

Don Monroe is a freelance writer based in New Jersey.

CREDIT: ADAPTED FROM MARGULIES *ET AL.*, *GENOME RESEARCH* 2007

Published by AAAS