Calculating the Original Family—of Exons

Walter Gilbert's estimate that there were fewer than 7000 original exons draws high praise—and claims of "naivete"

THE TENS OF THOUSANDS OF PROTEINS found in humans and other animals are a diverse lot—making up everything from toenails to hormones. Despite such diversity, today's proteins are constructed from a surprisingly small number of genetic building blocks that have been around for 2 billion years. That, at least, is the conclusion of a Harvard University team led by Nobel prize–winning molecular biologist Walter Gilbert, who offer their conclusion on page 1377 of this issue of *Science*.

Using a bold approach, the Harvard team calculates that a few pieces of genetic material (Gilbert and his colleagues leave open the question of whether it was DNA or RNA) won out in early molecular competition. They became the modules used to build proteins in eukaryotes (organisms whose cells have a nucleus). Only 1000 to 7000 of those initial exons, or coding regions, were then shuffled and linked over millennia to form the array of proteins needed for contemporary life forms. "This is a scandalously small number," says Gilbert. "Before doing this calculation, even in my wildest dreams I would have thought that on the order of a million to tens of millions of sequences would be involved."

Scandalous is a word that might appeal to many of Gilbert's peers, since even before publication the paper has, begun to cause a stir. Some scientists say privately that its mathematical model is "naive" and that Gilbert is out on a limb. They predict that his calculated range of exons won't hold up as more is learned about protein sequences. But those who like the paper say the exact number doesn't matter so much. "The specific number has to be taken with at least grams of salt," says Eric Lander, a geneticist and mathematician at the Massachusetts Institute of Technology's Whitehead Institute. "Nonetheless, that doesn't undermine the tremendous value of the paper: to make us begin to think seriously about the finiteness of the universe of exons."

> **"** *The specific number has to be taken with at least grams of salt. Nonetheless, that doesn't undermine the tremendous value of the paper.*
> —Eric Lander **"**

Gilbert's work is the outgrowth of a long search—specialists in molecular evolution have been trying for years to trace the ancestry of proteins. Most workers believe proteins evolved from a common set of early modules, or "motifs" (although not all agree those modules were today's exons). Much attention has focused on the early modules, with researchers estimating their number and form—usually coming up with families containing 1000 or fewer members. Those early estimates, however, were largely "intuitive," says the author of one, Emile Zuckerkandl, a molecular biologist at the Linus Pauling Institute of Science and Medicine in Palo Alto. "I think this is one of the first quantitative attempts at establishing that number."

Gilbert, with molecular evolutionist Robert L. Dorit and computer scientist Lloyd Schoenbach, approached the question in an unusual way. They collected more than 2500 amino acid sequences of known exons stored in the GenBank and European Molecular Biology Laboratory computer databases. Then they wrote computer programs that recognized duplicate or highly similar sequences, so they could purge repetition. Once they had distilled the database, they identified exon matches, and showed them to be statistically significant by using a computational technique known as the Monte Carlo method. They made two estimates, which were combined to come up with the universe of 1000 to 7000 original, nonrepeating exons.

The implications of the estimate are far-reaching, says Gilbert. If proteins were built from prefabricated modules, they probably evolved more quickly than if they had been constructed from scratch. But, efficiency carried a price tag: If Nature preferred a few winning motifs from the start, it gave up diversity of protein shape and structure for speed of assembly. As a result, the great variety of contemporary proteins is just a glimpse of all the possible shapes.

"The thrust of the paper makes evolution a lot easier," says Ford Doolittle, a biochemist who is a fellow of the Canadian Institute for Advanced Research. "It saves you an immense amount of time searching through amino acid sequences for protein shapes. What it doesn't show is what fraction of the possibilities there were for protein diversity."

Several specialists in protein evolution contacted by *Science* had serious reservations about Gilbert's paper, though only one wanted to express them on the record. The gist of their criticisms is that the way Gilbert and his colleagues attempt to detect common ancestry among exons and eliminate duplication is flawed. Russell Doolittle, a well-known protein chemist at the University of California at San Diego, was willing to comment for the record. Although Gilbert's team uses a standard mathematical method, it is "misapplied," he says, because it fails to identify the original exons correctly. It misses sequences that are known duplicates, and identifies repeats that are not related. Doolittle adds that he was disturbed to recognize several protein sequences in the "distilled" set of ancient exons that were purported to be dissimilar but that today are known to be derived from a common ancestral molecule. That undermined the credibility of the Harvard group's model, says Doolittle, who adds that the value of the work has been "exaggerated."

Other critics add that the paper relies on a random model that gives all of the original 1000 to 7000 exons an equal chance at combining to form proteins—a postulate that is at odds with known mechanisms. It is more likely that a few exons formed families of proteins that, in turn, became predominant motifs. And the critics note that only about half the eukaryotic proteins have so far been examined. As others are sequenced, new exons will emerge and the number of ancestral exons will have to be adjusted. But Gilbert notes that the stgrength of the model is that it preducts how many exons will be found in the future—and that will be testable.

As a result of such uncertainties, even one of the scientists who likes the Gilbert paper, Ford Doolittle, admits: "It could be wrong." He seems to speak for many when he concludes: "There are probably a lot of places where the analysis could fall apart. But there's value in this approach. The conclusion is interesting and possibly quite true. I sort of believe it." ■ ANN GIBBONS