Copyright © 2001 by the Genetics Society of America

Perspectives

Anecdotal, Historical and Critical Commentaries on Genetics

Edited by James F. Crow and William F. Dove

Shannon's Brief Foray into Genetics

James F. Crow

Laboratory of Genetics, University of Wisconsin, Madison, Wisconsin 53706

CLAUDE Shannon, 1916–2001, was the father of the digital communication age. He laid the mathematical foundations for communication theory and devised a precise definition for the vague concept of information. Although the word "bit" was invented by John Tukey, Shannon made it a household word among scientists, including geneticists. Yet, what is not generally known is that Shannon's Ph.D. thesis dealt with population genetics. Immediately after receiving the degree, he went to work for the Bell Telephone Laboratories and began his path-breaking studies of communication. He never returned to genetics and the thesis was never published. After half a century it was finally reprinted along with most of Shannon's major papers (Sloane and Wyner 1993). The thesis is now readily available for any who are interested in population genetics and its history.

Not many people have a master's thesis that is more famous than their Ph.D., but Shannon was one. His master's thesis was entitled "A symbolic analysis of relay and switching circuits" (Shannon 1938). Regarding this, Goldstein (1972) says, "Claude E. Shannon, the founder of what is often called Information Theory, in his master's thesis showed in a masterful way how the analysis of complicated circuits for switching could be effected by the use of Boolean algebra. This surely must be one of the most important master's theses ever written . . . . The paper was a landmark in that it helped to change digital circuit design from an art to a science." A few years later Shannon wrote a second, even more famous paper, "A mathematical theory of communication" (Shannon 1948), which gave birth to the science of information theory. This was republished in book form and included a nontechnical introduction and exposition by Warren Weaver (Shannon and Weaver 1963). It immediately became a best seller.

In this paper Shannon showed that, with the proper definition of information, all information sources have a source rate, measured in bits per second. The measure of information was $\Sigma P \log P$, in which P is the probability of choosing a particular message from among the alternatives, which is of the same form as entropy, long used as a measure of disorder in physical systems. For information theory it is natural to measure information in logs to the base 2. Thus, a simple system with two equally likely alternatives has $\log_2 2 = 1$ bit of information. The information in 1 bp, if all four pairs were equally frequent, is 2 bits. It was very much in vogue in the 1950s to speak of DNA as a molecule with 2 bits of information per nucleotide pair. I might note that "information" is used in a way that differs from ordinary English. It is a measure of the number of alternatives from which a message may be chosen. Entropy has since been adopted by many fields, for example ecology, where it has been fashionable as a measure of diversity. Whether it is the best measure of diversity has been questioned (see May 1981, p. 218). In the desire to be *au courant*, other fields have used the word with various degrees of imprecision of meaning.

One of Shannon's most surprising results was that a noisy system can send an undistorted signal provided that the appropriate error corrections or redundancy are built in. An interesting example that Shannon explored is the English language, which is about 50 percent redundant. It is this redundancy that permits us to understand from imperfect hearing what is being said in a noisy party. It also makes crossword puzzles feasible. Here is a Shannon example of a sentence with incomplete information—the vowels are omitted—but which is perfectly clear:

MST PPL HV LTTL DFFCLTY N RDNG THS SNTNC

Shannon extended this work in several directions. He made major contributions to cryptography and developed a general theory. He developed the theory of two-way communication channels. These papers are all regarded as original, substantial, and thorough. They and others are included in the nearly complete collection by Sloane and Wyner (1993).

Address for correspondence: Laboratory of Genetics, University of Wisconsin, 445 Henry Mall, Madison, WI 53706-1574.

SHANNON'S PH.D. THESIS

Between Shannon's two landmarks came an unpublished thesis in genetics. Shannon had been associated with Vannevar Bush at MIT in developing the differential analyzer, an analog computer for solving differential equations. His master's thesis grew out of the need to understand the complicated system of switches and relays involved in the analyzer; there were more than 100 relays. Bush was impressed by Shannon and his master's thesis and suggested he change to a mathematics major. Bush was also president of the Carnegie Institution of Washington, which included the Cold Spring Harbor Laboratory. He thought that Shannon's algebra might be useful in genetics. On this advice, Shannon spent the summer of 1939 at Cold Spring Harbor, working with Barbara Burks. Out of this grew his 1940 Ph.D. thesis in Mathematics at MIT.

The main purpose of the thesis was to develop a genetic algebra. Shannon's formalism was original and quite different from any previous work. The idea was to predict the genetic makeup in future generations of a population starting with arbitrary frequencies. He introduced a set of symbols for populations of multilocus genotypes and a set of rules for manipulating them. The result for three loci was new at the time. Most of the thesis, however, was not new. But it is clear that his main object was not to find new results but to introduce a new methodology. In his words, "In this paper an attempt will be made to develop an algebra especially suited to problems in the dynamics of Mendelian populations. Many of the results presented here are old in the theory of genetics, but are included because the method of proof is novel, and usually simpler and more general than those used previously" (Sloan and Wyner 1993, p. 892). He erred in the criteria for stability of a multi-allelic locus under selection, wrongly asserting that the necessary condition is heterozygote superiority in fitness for every pair of alleles, a conclusion that is not necessary. (I should like to use this opportunity to confess an earlier error of mine in interpreting this theorem. See Sloane and Wyner 1993, p. 921.)

Apparently, Shannon spent only a few months on the thesis. Perhaps if the work had been extended, either by him or by others, it might have led to significant discoveries. One gets the impression that he regarded this not as an end but as a beginning of a new methodology. Whether this is correct or not, Shannon went to work at the Bell Labs immediately after receiving his degree. There he found a stimulating environment with outstanding engineers, physicists, and mathematicians interested in communication. This got him started on a new career, and genetics was dropped. The thesis lay buried and unnoticed. In an interview in 1987, he said, "I set up an algebra which described this complicated process [of genetic changes in an evolving population]. One could calculate, if one wanted to (although not many people have wanted to in spite of my work), the kind of population you would have after a number of generations" (Sloane and Wyner 1993, p. xxvii).

Because the thesis was unpublished, it had no impact on the genetics community. In its obscurity, Shannon's thesis joins the work of two other researchers. One was Charles Cotterman, whose unpublished Ph.D. thesis was also submitted in 1940 (Cotterman 1940). The two theses were similar in devising a genetic algebra and placing great emphasis on a suitably suggestive notational system. Cotterman was more restrictive in one way by confining his studies to a single locus, but he included genetic relationships and inbreeding. His development of the concept of derivative genes, now called "identical by descent" (Crow 1954) or IBD, has had a great, though belated influence. His *k*-coefficients for specifying genotypic relationships have become standard in human genetics, as has his useful distinction between unilineal and bilineal relatives. Throughout his life, Cotterman had a curious reluctance to publish (Crow and Denniston 1989). His perfectionism meant that most of his best work was left unfinished because he was never fully satisfied. If his thesis had been published and recognized, the field of population genetics "would have advanced the progress of genetics by a decade or more on several fronts" (Ballonoff 1974, p. 156). Cotterman also provided a classification of inbreeding systems. Characteristically, this was not published during his lifetime and was instigated by his coauthor (Boucher and Cotterman 1990).

At about the same time, even a bit earlier, Gustave Malécot was publishing path-breaking papers of a more mathematical sort. In particular, he considered stochastic processes and, more than anyone else, ushered in modern population genetics. For a thorough and thoughtful review of his work, see Nagylaki (1989). Malécot's papers had three strikes against them. First, they were published in French, which greatly reduced their accessibility to unilingual English-speaking geneticists. Second, the papers were highly mathematical, which again limited the readership. Third, they were published mainly in journals little read outside France. In comparing Shannon's thesis with the others, Nagylaki (Sloane and Wyner 1993, p. 921) says, "This dissertation is in the spirit of Cotterman's, but the latter's is far more penetrating and important, and presents many more applications. Malécot's has all the qualities of Cotterman's and is also mathematically powerful."

If all three of these authors had been published in widely circulated journals, what would have been the consequence? Clearly, Cotterman's treatment of inbreeding and relationship would have caught on immediately. Equally clearly, Malécot would have brought stochastic processes and diffusion theory into the theory of population genetics. What consequence would the Shannon thesis have had? The answer, I believe, is rather little in comparison with the other two, although it might have had more influence if it had been carried further. With his creativity, if Shannon had stayed in population genetics, he would surely have made some important contributions. Nevertheless, I think it is fair to say that the world is far better off for his having concentrated on communication theory, where his work was revolutionary.

SHANNON'S LATER WORK

Shannon's interests were unusually diverse. Seemingly, he was motivated entirely by curiosity. He was adept, not only in mathematics, but in gadgeteering and invented many kinds of models and toys. After 15 years he retired from the Bell Telephone Laboratories, much to the regret of his colleagues who had come to count on his quick apprehension of problems and original approach to solutions. He moved to MIT in 1958 where he became Donner Professor. He had a few students and continued to refine his ideas on information theory. Over the years, he more often worked at home. His wife, Elizabeth, was also a mathematician and shared many of his interests.

Increasingly, before and after his retirement in 1978, he devoted time to an astonishing variety of games, toys, and hobbies. The game of Nim can be analyzed in terms of binary numbers. Therefore, it was an easy step to translate the mathematical strategy into a relay circuit. Shannon used judiciously applied voltage differences to construct a device for playing Hex. Of more lasting influence, he was one of the first to develop a chess-playing program. Although limited by the computer power of the time—this was in the day of vacuum tubes—it played a strong game. Shannon himself was an excellent chess player. On a visit to Russia he enjoyed a game with world champion Botvinnik; he was ahead for a while, but finally succumbed to the champion's superior prowess.

Shannon must have been physically adept, for he enjoyed juggling and riding a unicycle. He designed a unicycle with an eccentric wheel. He also rode a unicycle and juggled at the same time, causing astonishment and amusement in the halls of the Bell labs. He wrote a theoretical article on juggling and contrived a diorama in which three miniature clowns juggled record numbers of balls, clubs, and rings. The backstage mechanism was concealed by judicious use of ultraviolet light and fluorescent foreground objects.

He devised a machine, THROBAC, which did calculations in Roman numerals. He designed a machine to solve the Rubik Cube. He developed a maze-learning device and constructed a mouse that would discover the way through a maze by trial and error, but once it succeeded would never fail again and could be started at any intermediate point. This was one of the first devices capable of learning. Another clever idea was a "mind reading machine" that played a game of matching pennies. It worked by discerning patterns in the opponent. Since any human being eventually displays some sort of pattern, a sufficiently alert machine with sufficient time can detect this and produce a winning strategy.

These, along with other ideas, both playful and deep, are included in the collection of Shannon's major papers (Sloane and Wyner 1993), which also includes a biography and interview. Much of the material in this essay comes from that source. There is a wealth of happy reading—not all of it easy—for anyone interested in this remarkable man. In Shannon's words, "I've spent lots of time on totally useless things." But he also did profoundly deep and important things.

Finally, let me mention a personal favorite, Shannon's "ultimate machine," based on an idea from Marvin Minsky. I was fortunate in the 1950s to see Shannon demonstrate this on a television program. The memory is still vivid. The machine was a small closed box with a toggle switch on the front. Shannon flipped the switch. Then the lid opened, with whirring noises in the box, and a small hand emerged and shut off the switch, whereupon the noises stopped and the lid snapped shut. To quote Arthur Clarke (Sloane and Wyner 1993, p. xiv), "There is something unspeakably sinister about a machine that does nothing—absolutely nothing—except switch itself off."

I am indebted to N. J. A. Sloane for bringing this thesis to my attention some years ago. Much of the material in this essay is from the collection put together by him and Wyner. I thank Tom Nagylaki for several useful suggestions.

LITERATURE CITED

Ballonoff, P. (Editor), 1974 *Genetics and Social Structure.* Dowden, Hutchinson & Ross, Stroudsburg, PA.

Boucher, W., and C. W. Cotterman, 1990 On the classification of regular systems of inbreeding. J. Math. Biol. **28:** 293–305.

Cotterman, C. W., 1940 A calculus for statistico-genetics. Dissertation, Ohio State University, Columbus, OH (reprinted pp. 157–272 in *Genetics and Social Structure,* edited by P. Ballonoff. Dowden, Hutchinson & Ross, Stroudsburg, PA, 1974).

Crow, J. F., 1954 Breeding structure of populations. II. Effective population number, pp 543–556 in *Statistics and Mathematics in Biology,* edited by O. Kempthorne, T. A. Bancroft, J. W. Gowen, and J. L. Lush. Iowa State College Press, Ames, IA.

Crow, J. F., and C. Denniston, 1989 In memory of Charles W. Cotterman, 1914–89. Am. J. Hum. Genet. **44:** 903–904.

Goldstein, H. H., 1972 *The Computer From Pascal to von Neumann.* Princeton University Press, Princeton, NJ.

May, R. M., 1981 Patterns in multi-species communities, pp 197–227 in *Theoretical Ecology,* edited by R. M. May. Sinauer Associates, Sunderland, MA.

Nagylaki, T., 1989 Gustave Malécot and the transition from classical to modern population genetics. Genetics **122:** 253–268.

Shannon, C. E., 1938 A symbolic analysis of relay and switching circuits. Trans. Am. Inst. Elect. Eng. **57:** 713–723.

Shannon, C. E., 1948 A mathematical theory of communication. Bell Syst. Tech. J. **27:** 379–423, 623–656.

Shannon, C. E., and W. Weaver, 1963 *The Mathematical Theory of Communication.* University of Illinois Press, Urbana/Chicago, IL.

Sloane, N. J. A., and A. D. Wyner, 1993 *Claude Elwood Shannon Collected Papers.* IEEE Press, Piscataway, NJ.