Russell F. Doolittle, born 1931 in Connecticut, is currently a research professor at the Center for Molecular Genetics, University of California, San Diego. His principal research interests center around the evolution of protein structure and function. He has a PhD in biochemistry from Harvard (1962) and did postdoctoral work in Sweden. He was an early advocate of using computers as an aid to characterizing proteins.

For some it may be difficult to envision a time when the World Wide Web did not exist and every academician did not have a computer terminal on his or her desk. It may be even harder to imagine the primitive state of computer hardware and software at the time of the recombinant DNA revolution, which dates back to about 1978. It was in this period that Russell Doolittle, using a DEC PDP11 computer and a suite of home-grown programs, began systematically searching sequences in an effort to find evolutionary and other biological relationships. In 1983 he stunned cancer biologists when he reported that a newly reported sequence for platelet derived growth factor (PDGF) was virtually identical to a previously reported sequence for the oncogene known as ν-sis.[13] This was big news, and the finding served as a wake-up call to molecular biologists: searching all new sequences against up-to-date databases is your first order of business.

Doolittle had actually begun his computer studies on protein sequences much earlier. Fascinated by the idea that the history of all life might be traceable by sequence analysis, he had begun determining and aligning sequences in the early 1960s. When he landed a job at UCSD in 1964, he tried to interest consultants at the university computer center in the problem, but it was clear that the language and cultural divide between them was too great. Because computer people were not interested in learning molecular biology, he would have to learn about computing. He took an elementary course in FORTRAN

13. *Oncogenes* are genes in viruses that cause a cancer-like transformation of infected cells. Oncogene ν-sis in the *simian sarcoma virus* causes uncontrolled cell growth and leads to cancer in monkeys. The seemingly unrelated *growth factor* PDGF is a protein that stimulates cell growth.

programming, and, with the help of his older son, developed some simple programs for comparing sequences. These were the days when one used a keypunch machine to enter data on eighty-column cards, packs of which were dropped off at the computer center with the hope that the output could be collected the next day.

In the mid-1960s, Richard Eck and Margaret Dayhoff had begun the Atlas of Protein Sequence and Structure, the forerunner of the Protein Identification Resource (PIR) database. Their original intention was to publish an annual volume of "all the sequences that could fit between two covers." Clearly, no one foresaw the deluge of sequences that was to come once methods had been developed for directly sequencing DNA. In 1978, for example, the entire holding of the atlas, which could be purchased on magnetic tape, amounted to 1081 entries. Realizing that this was a very biased collection of protein sequences, Doolittle began his own database, which, because it followed the format of the atlas, he called NEWAT ("new atlas"). At about the same time he acquired a PDP11 computer, the maximum capacity of which was only 100 kilobytes, much of that occupied by a mini-UNIX operating system. With the help of his secretary and his younger son (eleven years old at the time), Doolittle began typing in every new sequence he could get his hands on, searching each against every other sequence in the collection as they went. This was in keeping with his view that all new proteins come from old proteins, mostly by way of gene duplications. In the first few years of their small enterprise, Doolittle & Son established a number of unexpected connections.

Doolittle admits that in 1978 he knew hardly anything about cancer viruses, but a number of chance happenings put him in touch with the field. For one, Ted Friedmann and Gernot Walter (who was then at the Salk Institute), had sought Doolittle's aid in comparing the sequences of two DNA tumor viruses, simian virus 40 (SV40) and the polyoma virus. This led indirectly to contacts with Inder Verma's group at Salk, which was studying retroviruses and had sequenced an "oncogene" called ν-mos in a retrovirus that caused sarcomas in mice. They asked Doolittle to search it for them, but no significant matches were found. Not long afterward (in 1980), Doolittle read an article reporting the nucleotide sequence of an oncogene from an avian sarcoma virus—the famous *Rous sarcoma virus*. It was noted in that article that the Salk team had provided the authors with a copy of their still unpublished mouse sarcoma gene sequence, but no resemblances had been detected. In line with his own project, Doolittle promptly typed the new avian sequence into his computer to see if it might match anything else. He was astonished to find that in fact a match quickly appeared with the still unpublished Salk

sequence for the mouse retrovirus oncogene. He immediately telephoned Inder Verma; "Hey, these two sequences are in fact homologous. These proteins must be doing the same thing." Verma, who had just packaged up a manuscript describing the new sequence, promptly unwrapped it and added the new feature. He was so pleased with the outcome that he added Doolittle's name as one of the coauthors.

How was it that the group studying the Rous sarcoma virus had missed this match? It's a reflection on how people were thinking at the time. They had compared the DNA sequences of the two genes without translating them into the corresponding amino acid sequences, losing most of the information as a result. It was another simple but urgent message to the community about how to think about sequence comparisons.

In May of 1983, an article appeared in *Science* describing the characterization of a growth factor isolated from human blood platelets. Harry Antoniades and Michael Hunkapiller had determined 28 amino acid residues from the N-terminal end of PDGF. (It had taken almost 100,000 units of human blood to obtain enough of the growth factor material to get this much sequence.) The article noted that the authors had conducted a limited search of known sequences and hadn't found any similar proteins.

By this time, Doolittle had modem access to a department VAX computer where he now stored his database. He typed in the PDGF partial sequence and set it searching. Twenty minutes later he had the results of the search; human PDGF had a sequence that was virtually identical to that of an oncogene isolated from a woolly monkey. Doolittle describes it as an electrifying moment, enriched greatly by his prior experiences with the other oncogenes. He remembers remarking to his then fifteen-year old son, "Will, this experiment took us five years and twenty minutes." As it happened, he was not alone in enjoying the thrill of this discovery. Workers at the Imperial Cancer Laboratory in London were also sequencing PDGF, and in the spring of 1983 had written to Doolittle asking for a tape of his sequence collection. He had sent them his newest version, fortuitously containing the ν-sis sequence from the woolly monkey. Just a few weeks before the *Science* article appeared, Antoniades and Hunkapiller replied with an effusive letter of thanks, not mentioning just why the tape had been so valuable to them. Meanwhile, Doolittle had written to both the PDGF workers and the ν-sis team, suggesting that they compare notes. As a result, the news of the match was quickly made known, and a spirited race to publication occurred, the report from the Americans appearing in *Science* only a week ahead of the British effort in *Nature*. Doolittle went on to make many other matches during the mid-

1980s, including several more involving oncogenes. For example, he found a relationship between the oncogene ν-jun and the gene regulator GCN4. He describes those days as unusual in that an amateur could still occasionally compete with the professionals. Although he continued with his interests in protein evolution, he increasingly retreated to the laboratory and left bioinformatics to those more formally trained in the field.