A reprint from
American Scientist
the magazine of Sigma Xi, The Scientific Research Society

This reprint is provided for personal and noncommercial use. For any other use, please send a request to Permissions, American Scientist, P.O. Box 13975, Research Triangle Park, NC, 27709, U.S.A., or by electronic mail to perms@amsci.org. ©Sigma Xi, The Scientific Research Society and other rightsholders

permission only. Contact perms@amsci.org.

A Cipher to Thomas Jefferson

A collection of decryption techniques and the analysis of various texts combine in the breaking of a 200-year-old code

Lawren M. Smithline

A year or so ago, I started talking to my neighbor, Amy Speckart, about Thomas Jefferson. She had taken a leave of absence from William & Mary to write her dissertation on early American history. During that time, Speckart worked at The Papers of Thomas Jefferson. This decades-long project at Princeton University—and its twin at Monticello, Jefferson's home—collects and publishes all of the correspondence and papers of Jefferson. Late in the winter of 2007, Speckart told me that they'd found several letters using ciphers, or secret codes. That intrigued me, because I am a mathematician at the Center for Communications Research in Princeton, New Jersey, and this center deals with modern communications, including cryptology. Despite my interest, I didn't pursue the ciphers at that time. Then, in June 2007, Speckart told me, "We have a letter in cipher, and we can't read it." Immediately, I asked for a copy.

Speckart provided a link to the archives at the Library of Congress, and I soon obtained a copy of the letter. It was dated December 19, 1801, and sent from Robert Patterson to Jefferson. At that time, Jefferson served as the president of the American Philosophical Society, and Patterson was the vice president. The two men corresponded often and on a range of topics, including cryptography.

Patterson started this particular letter by defining four features of what he called a "perfect cypher." It should be adaptable to all languages, easy to memorize and simple to perform. Last—but "most essential" in Patterson's view—he wrote that a perfect cipher should be "absolutely inscrutable to all unacquainted with the particular key or secret for decyphering."

In this letter to Jefferson, Patterson described a technique that he believed met those four criteria. In addition, Patterson included an enciphered message in the letter, which no one—to my knowledge—had deciphered. As Patterson wrote: "I shall conclude this paper with a specimen of such writing, which I may safely defy the united ingenuity of the whole human race to decypher to the end of time.…" Nonetheless, I took on Patterson's cryptogram with a collection of tools, among them one common in other fields, including computational biology.

Enhancing the Secrecy of Ciphers

For centuries, people encrypted messages through substitution ciphers, which substitute one letter of the alphabet for another. Solving such a cipher, though, does not prove absolutely inscrutable—Patterson's cardinal parameter—because frequency analysis exposes the hidden text. Frequency analysis, or counting the number of occurrences of each letter of the alphabet in a message, can be used to reconstruct the key. In English, for example, the most-common letter is "e." Thus, the most-common letter in an English-language text enciphered by substitution probably substitutes for "e." The observed letter counts might not conform exactly to a frequency table, yet they indicate a small set of good choices to try for the most-common letters. In *The Codebreakers*, David Kahn suggests that European culture knew about frequency analysis no later than the 15th century.

The diffusion of the frequency-analysis technique likely precipitated an industry of developing new ciphers, such as the nomenclator. A nomenclator is a catalog of numbers, each standing for a word, phrase, name, syllable or even a letter. The operation of the nomenclator is simple and intuitive. Although this method is susceptible to frequency analysis, an extensive codebook vocabulary makes such an attack difficult. The earliest examples of nomenclators are from the 1400s, and Jefferson's correspondence shows that he used several codebooks.

Patterson would have known about nomenclators and objected to them because they cannot be memorized. Consequently, a nomenclator's security relied on carefully controlled possession of a single thing, the codebook. Instead of any sort of substitution, Patterson's letter described

Lawren M. Smithline is a mathematician at the Center for Communications Research. In 2000, he earned a Ph.D. from the University of California, Berkeley, for his thesis on p-adic modular forms. He then worked at Cornell University for several years, where he shifted his focus to computational biology. At the Center for Communications Research, he acquired an interest in signal processing. He continues to work on a spectrum of applied- and theoretical-math problems. Address: Center for Communications Research, 805 Bunn Drive, Princeton, NJ 08540. Internet: lawren.smithline@idaccr.org

© 2009 Sigma Xi, The Scientific Research Society. Reproduction with permission only. Contact perms@amsci.org.

Philadelphia Dec 19th 1801

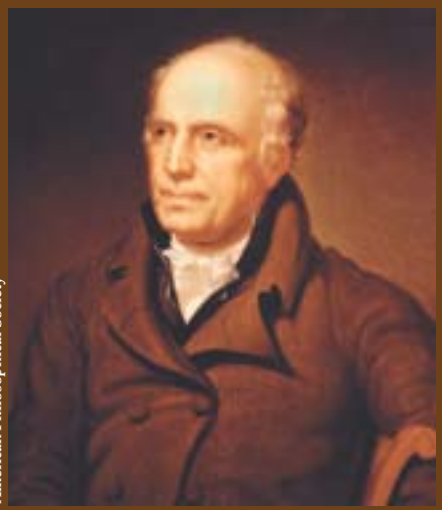
Sir)

The art of secret writing, or, as it is usually termed, writing in cypher, has occasionally engaged the attention both of the Statesman & philosopher for many ages; and yet I believe it will be acknowledged, by all who are acquainted with the present state of this art, that it is still far short of perfection. A perfect cypher, as it appears to me, should possess the following properties.—

1. It should be equally adapted to all languages
2. It should be easily learned & retained in memory.
3. It should be written and read with facility & dispatch
4. (which is the most essential property) it should be absolutely inscrutable to all unacquainted with the particular key or secret for decyphering.

I shall not enter into a tedious details of the various systems of secret writing, or point out their several defects, but shall immediately proceed to lay before you a system which, I flatter myself, will be

Mathematics was my first muse,



American Philosophical Society

The Art Archive/Laurie Platt Winfrey

Figure 1. On December 19, 1801, Robert Patterson *(far left)*—a professor of mathematics at the University of Pennsylvania—wrote a letter to Thomas Jefferson *(immediate left)* about cryptography. In this letter *(above)*, Patterson described his vision of a "perfect cipher," which required four elements: adaptable to all languages, easy to memorize, simple to perform and inscrutable without the key. Patterson also described an encryption technique that he believed met these criteria. In addition, he included encrypted text, which he said could never be decrypted. There is no evidence that Jefferson was able to decode the text. The author took on Patterson's challenge using techniques that could have been applied—if laboriously—in the early 19th century. (All letter reproductions courtesy of the Library of Congress.)

© 2009 Sigma Xi, The Scientific Research Society. Reproduction with permission only. Contact perms@amsci.org.

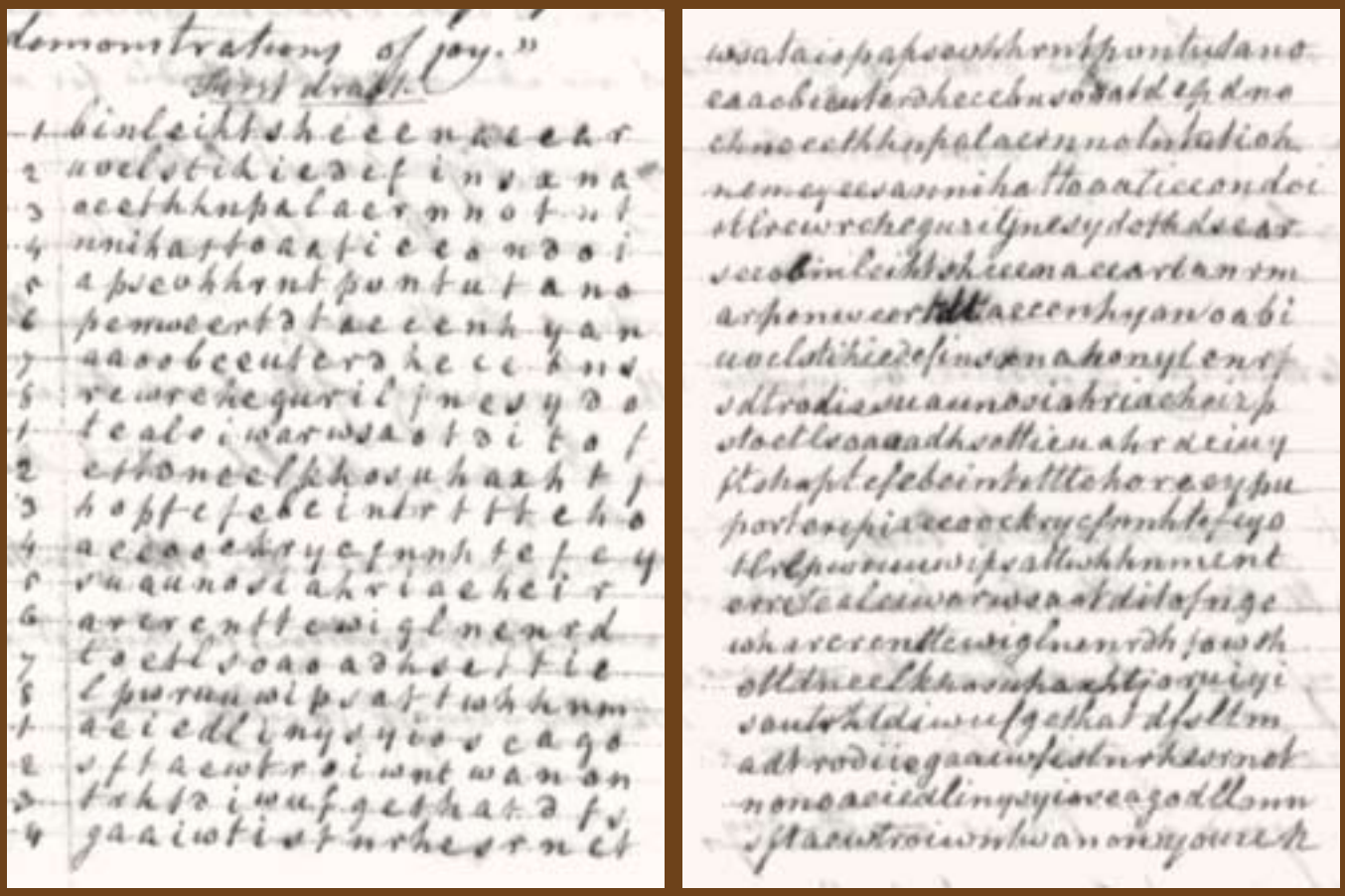


Figure 2. A worked example in Patterson's letter demonstrates his transposition technique. He started by writing the message in columns, following letters placed beneath the preceding letters, like Chinese writing, and starting new rows as needed *(left)*. His worked example began: "Buonaparte has at last given peace to Europe." Patterson also included an encrypted version of this text *(right)*. He broke the rows into sections of nines lines or less, scrambled the lines within the sections—done the same in each section—and added an arbitrary number of letters to the beginning of each line. The number of added letters remained the same for each line throughout the encryption, such as, say, adding 3 letters to line 8 in every section of the encrypted text.

a transposition cipher, which changes the order of characters from the original text to conceal a message. As Patterson wrote:

> In this system, there is no substitution of one letter or character for another; but every word is to be written at large, in its proper alphabetical characters, as in common writing: only that there need be no use of capitals, pointing, nor spaces between words; since any piece of writing may be easily read without these distinctions.

He continued:

> Let the writer rule on his paper as many pencil lines as will be sufficient to contain the whole writing.… Then, instead of placing the letters one after the other, as in common writing, let them be placed one under the other, in the Chinese manner, namely, the first letter at the beginning of the first line, the second letter at the beginning of the second line, and so on, writing column after column, from left to right, till the whole is written.

To demonstrate the approach, Patterson included an example that began: "Buonaparte has at last given peace to Europe," and he explained how to encipher it:

> This writing is then to be distributed into sections of not more than nine lines in each section, and these are to be numbered 1. 2. 3 &c 1. 2. 3 &c (from top to bottom). The whole is then to be transcribed, section after section, taking the lines of each section in any order at pleasure, inserting at the beginning of each line respectively any number of arbitrary or insignificant letters, not exceeding nine; & also filling up the vacant spaces at the

© 2009 Sigma Xi, The Scientific Research Society. Reproduction with permission only. Contact perms@amsci.org.

> end of the lines with like letters. Now the key or secret for decyphering will consist in knowing—the number of lines in each section, the order in which these are transcribed, and the number of insignificant letters at the beginning of each line.…

A column of two-digit numbers provides the key to Patterson's cipher. For each pair of digits, the first represents a line number within a section, and the order of the first digits indicates how to rearrange the lines. The second digit in each pair indicates how many extra letters to add to the beginning of that line.

Crunching Patterson's Challenge

In describing this cipher to Jefferson, Patterson wrote, "It will be absolutely impossible, even for one perfectly acquainted with the general system, ever to desypher the writing of another without his key." Moreover, Patterson estimated the number of keys available for his cipher at more than "ninety millions of millions." Jefferson might have simply accepted Patterson's warning—"the utter impossibility of decyphering will be readily acknowledged"—and Jefferson probably never cracked the enciphered portion of the letter. Still, Jefferson was so taken by the cipher's apparent efficacy that he forwarded the method to Robert Livingston, ambassador to France. Nonetheless, Livingston continued to use a nomenclator.

Others also bypassed Patterson's cipher. For example, when Ralph E. Weber—a scholar in residence at the U.S. Central Intelligence Agency and National Security Agency—described Patterson's cipher method in 1979 in *United States Diplomatic Codes and Ciphers 1775–1938*, Weber dealt only with the worked example, completely skipping the challenge cipher.

Is Patterson's cipher truly unsolvable? Although the analysis of the frequencies of single letters cannot break Patterson's code, I suspected that analyzing groups of letters might. Like the frequencies of single letters in text, digraph frequencies—the likelihood of specific pairs of letters appearing together—are not uniform and therefore might help to break Patterson's cipher.

To test this idea, I needed a table of digraph frequencies of English made from text that was contemporary with Patterson's cipher. To build such a table, I used the 80,000 letters that make up Jefferson's State of the Union addresses—with spaces and punctuation removed, capitalization ignored—and counted the occurrences of "aa," "ab,""ac" and so on through "zz." This created a table with 26 columns and 26 rows of digraph counts. Then, dividing each digraph count by the total number of letters used in the text gave the frequencies. I also built a digraph-frequency table from a much larger collection of writing from Patterson's era. In both cases, the digraph frequencies came out virtually the same.

1	binlei	58	wsataispapsevh …
2	uvclst	71	eaaoobc …
3	oeethh	33	chnoeeth …
4	nnihat	49	nemeyeesannihat …
5	apsevh	83	stlrcwreh …
6	penwee	14	seesbinlei …
7	aaoobc	62	arpenwee …
8	rcwreh	20	uvclst …
1	tealei	58	sdtrodiesuauno …
2	ettdne	71	stoetls …
3	hopfcf	33	ptohopfcf …
4	aeeooc	49	porterepiaeeooc …
5	suauno	83	tlrlpwruu …
6	arcrcn	14	etretealei
7	toetls	62	wharcrcn …
8	lpwruu	20	ettdne …
1	aeiedl	33	sautrhtdi …
2	sftaew	49	adtradiiegaaiwt …
3	tvhtdi	14	nonsaeiedl …
4	gaaiwt	20	sftaewtvoiw …

Figure 3. A column of two-digit numbers provided the method for encrypting and the key. The first digit indicated the line number within a section and the second was the number of letters added to the beginning of that row. In Patterson's worked example, the key was 58, 71, 33, 49, 83, 14, 62, 20. To encrypt the first section of the example text, which is shown in part *(left)*, Patterson moved row 5 to the first line *(right)* and added 8 letters, moved row 7 to row 2 and added 1 letter, and so on. Then, he made the same transpositions for the following sections. This example shows the encryption for "Buonaparte *(red)* has *(green)* at *(purple)* last *(gold)* given *(blue)*…." In the second line of the cipher, the ɵ indicates an "o" that Patterson left out when transcribing row 7 *(left)* to row 2 *(right)*.

Next, I guessed at five things: the number of rows in a section size, two rows that belong next to each other and the number of extra letters inserted at the beginning of those two rows. So instead of trying to figure out Patterson's entire key, I just guessed at part of it. For example, I could guess that each section consists of 8 rows, and that rows 7 and 3 belong next to each other. That would mean that the pattern would repeat every 8 rows—making row 15 (8 rows after 7) and 11 (8 rows after 3) lie next to each other, and the same for rows 23 and 19, and so on. Given

© 2009 Sigma Xi, The Scientific Research Society. Reproduction with permission only. Contact perms@amsci.org.

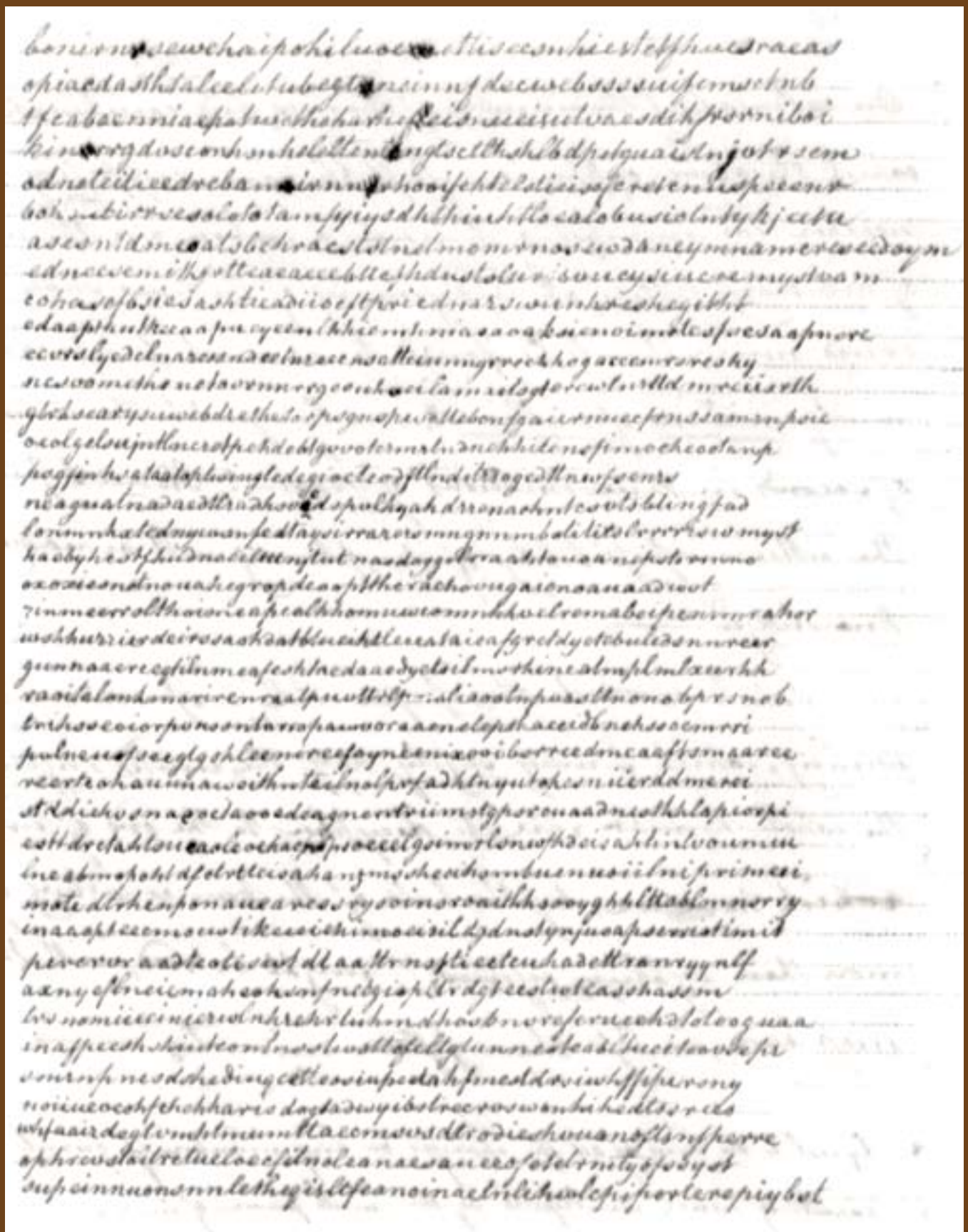


Figure 4. Patterson wrote that his challenge cipher, shown here, was "absolutely impossible, even for one perfectly acquainted with the general system, ever to desypher.…" He added that the number of possible keys was more than "ninety millions of millions. " In fact, no record indicates that anyone had decrypted Patterson's challenge cipher.

these guesses, I matched the pairs of rows and aligned them by columns based on the guesses at the number of random letters added to the start of each.

If the combination of section size, row pair and extra letters is right, that leads to better digraphs than if the combination is wrong. For instance, the letter pair "vj" is impossible in English, so that excludes any alignment that creates that digraph. Alternatively, the letter pair "qu" is rare, but when there is a "q," it must line up with a "u." When "q" and "u" do line up, that is strong evidence in favor of that alignment. Once this approach reveals how one pair of rows lines up, I guess about how another row might line up with one of the two that I already have. Once I get that, I add more rows, until I solve the entire key. (As a quick aside, this can also be done with trigraph frequencies—the likelihood of specific triplets of letters—but that isn't necessary for this problem.)

Distinguishing Digraphs

Above, I mention looking for "better" digraphs, but what makes one better than another? Think of this as the search for the most-likely digraphs, which would increase the likelihood that the selection of section size, adjacent rows and added letters is correct. Distinguishing one digraph as better than another can be done in more than one way, and I wanted one that would show me whether the computations were feasible by turn-of-the-19th-century technology.

In addition to a table of digraph frequencies, I also needed the frequencies of single letters. Then for any particular digraph, I asked: Did I ever see it in the text used to build the frequency tables? If yes, I asked: Is the frequency of the digraph greater than the product of the frequencies of the individual letters. For example, if the digraph is "wi," is the frequency of "wi" great than the frequency of "w" times the frequency of "i"? That is, does seeing "w" predict that the next letter is more likely to be "i" than it would be at random? If yes again, I called the digraph "favorable." Otherwise, the digraph was classified as "unfavorable" or "nonexistent." For the text in Jefferson's State of the Union Addresses, some favorable digraphs were "nt," "qu" and "se," while "et," "ls" and "od" were unfavorable, and "dx," "gq" and "wd" were nonexistent.

By the way, it might appear counterintuitive that the digraph "et" rates as unfavorable. Although this digraph is very common, upon seeing the letter "e," it is less likely that the next letter is "t" than it would be if we just looked at a single letter at random with no knowledge of the letter before. Also, "wd" is not impossible in English; it just doesn't show up in any of Jefferson's State of the Union addresses.

rating	score	examples
favorable	+1	wi ve nt in se qu
unfavorable	−1	od ls tq sk ei et
nonexistent	−5	wd lj pd dx gq vz

Figure 5. Likelihoods of specific pairs of letters appearing together—derived from so-called digraph frequencies—can break Patterson's cipher. The author used a table of digraph frequencies made from Jefferson's State of the Union addresses to assess the promise of guesses at the key. If a guess at the organization of rows in a section and the number of letters added to each line produced digraphs that were more likely than the two letters just happening to appear side by side—such as "wi" and "qu"—they were marked as favorable and given a +1 rating. Digraphs that were less likely than the random pairing of the letters—such as "od" and "et"—were classified as unfavorable and given a –1 rating. Digraphs that didn't appear in Jefferson's State of the Union addresses at all—such as "wd" and "vz"—were called nonexistent and rated as –5.

Then, given the digraphs created by a particular guess of section size, adjacent rows and added letters, I calculated a score built from: +1 for each favorable digraph; –1 for each unfavorable digraph; and –5 for each nonexistent digraph. Since the number of random letters added to rows varies, some rows extend beyond others when aligned by column, and any letters that stick out with no mating letter get scored as 0.

At that point, I still faced two challenges: mistranscribing some letters and organizing

K	R	C	S	D	score
5	2	6	5	2	26
6	4	5	3	0	26
7	1	3	5	2	60
8	6	1	2	1	22
9	4	6	8	6	28

Figure 6. Dynamic programming used the digraph frequencies to generate top-scoring guesses for a key to Patterson's encrypted message. Specifically, the author guessed at section size *(K)* and row pair *(R* and *S)*—initially limited to guesses that matched the "q" in cipher row 22 with the letter "u"—and the program calculated the best number of extra letters: *C* and *D*, for rows *R* and *S*, respectively. The combination of best guesses produced the highest scores. The author recorded the best combinations for each value of *K*. Here, for example, the combination for *K*= 7, which scored 60, was the best of the best. After deciding on the section size of 7 rows, the table indicated that cipher row 1 belongs above cipher row 5, row 1 gets 3 extra letters at the start, and row 5 gets 2 extra letters. From that point, the author guessed at another row, and another, until he determined the entire key.

© 2009 Sigma Xi, The Scientific Research Society. Reproduction with permission only. Contact perms@amsci.org.

row 1	13	bonivnsewe	1 ivnsewe
row 2	34	opiacdasth	2 neteidie
row 3	57	tfcabaenni	3 cdasth
row 4	65	kinrrgdosc	4 o
row 5	22	adneteidie	5 nni
row 6	78	boksutirrs	6 gdosc
row 7	49	asesntdmeo	7 rs
row 8	13	edneesemit	1 eesemit
row 9	34	cohasefbsi	2 svomethe
row 10	57	edaaprhutk	3 sefbsi
row 11	65	eevrslyege	4 j
row 12	22	resvomethe	5 utk
row 13	78	gbrksearys	6 lyege
row 14	49	oeolgelsuj	7 ys

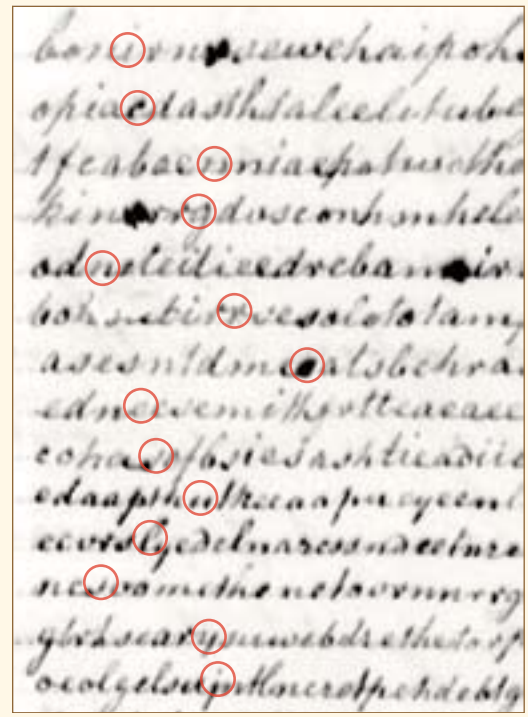


Figure 7. The key to Patterson's cipher was 13, 34, 57, 65, 22, 78, 49. As shown here *(left)*, the first row of the encrypted text—also shown in Patterson's letter *(right)*—stayed in row 1 but 3 extra letters were added, so the first letter of the decrypted text *(middle)* is "i" *(red)*. Row 5 provides row 2 of the decrypted text and it has 2 letters added at the start, making the decrypted letter "n" *(red)*. Stringing the letters one on top of the other begins to expose the message.

this apparently massive computation. For the first problem, as soon as I saw Patterson's letter, I realized that it would be difficult to make a perfect transcription. Amy Speckart assured me that one gets used to the antique script, which is true, but plain language is easier to read than a cipher, because the letters make words. I knew this was a problem for Patterson, too, because he made a mistake in his worked example and—as I would learn—in his challenge cipher, too. Nonetheless, my scoring technique is forgiving enough, as long as the transcription is largely correct. Rather than immediately discarding an alignment that produces "wd," for example, it gets rated very poorly. In addition, I designed my technique to allow the occasional insertion of a blank space, accounting for things like copying the letter "w" as "ui."

Adding Programming Power

For the computation, I turned to dynamic programming—the engine that solves the scoring of all the possibilities and efficiently determines the best guesses. Dynamic programming solves a large problem by systematically solving constituent small problems and then knitting together the solutions.

A classic dynamic-program example is Dutch computer scientist Edsger W. Dijkstra's route-finding algorithm. Suppose I want to travel from New York City to San Francisco by car on roads mapped by my favorite atlas, and I want to make the journey in the shortest distance. I do not have to compute the distance for every possible route between New York City and San Francisco. Instead, I can calculate the shortest path from New York City to every

i n c o n g r e s s j u l y f o u r t h o n e t h s u s a n d s e v e n h u n d
v e d a n d s e v e n t y s i x s d e e l a r a t i s n b y t h e r e p r s e
n t a t i o e s o f t k e u n i t e d s t a t l s o f a m e r i c a i n c o n g
s e s s a s s e m b l e d w k e n i n t h c e o u r s p o f h u m a n e v e n t
e i t b e c o m e s n e e e s s a r y f o r a n e p e a p l e t o d i s s s l v
w d h e p o l i t i c a l b a n d s i h i e h d a v e c o n n c u t e d t h e m

Figure 8. Patterson's decrypted message starts with: "In Congress July Fourth." It goes on to provide the preamble to the Declaration of Independence, which was written by Thomas Jefferson. Even with mistakes in interpreting Patterson's handwriting, the author's technique finds the correct key. The message can be read and the errors corrected along the way.

© 2009 Sigma Xi, The Scientific Research Society. Reproduction with permission only. Contact perms@amsci.org.

crossing of the New York State line, and likewise from San Francisco to the California border. For each state, I can calculate the shortest routes between road entry points. The shortest route across the country and its total distance can be assembled from these data. Within a state, I can solve the same problem by dividing up routes on the county level, and so on, down to the scale of turn-by-turn directions at every intersection.

Like route finding, I compose my dynamic program to help me make top-scoring guesses about the key to Patterson's cipher. As mentioned above, I guess at section size, row pair and extra letters, but this is a slight fib. I guess section size and row pair, and the dynamic program tells me the best number of extra letters, as well as whether and where I should insert a blank space. Formally, I represent the variables as: *K* for section size; *R* and *S* for rows tested for lying one over the other in a section; and *C* and *D* for the extra letters at the beginning of rows *R* and *S*, respectively. Based on the digraph frequencies, the dynamic program computes the best *C* and *D* to go with *K*, *R* and *S*. Here, "best" means the *C* and *D* that generate the best score in the dynamic program. The program also tells me what that score is, so I pick the best scoring *K*, *R* and *S*, and unravel the cipher key row by row from there.

Patterson's cipher offered one opportunity to simplify the decoding. Row 22 of Patterson's cipher includes a "q" at position 11, and this "q" has the fewest nearby possibilities for a following "u." So in guessing at section size and rows that go one above the other, I used the combinations that put this "q" next to a "u." Moreover, rather than transcribing the entire length of every line in Patterson's cipher, I started with the first 30 columns of each line.

These constraints reduced the overall computational load to fewer than 100,000 simple sums—tedious in the 19th century, but doable. As a result, one guess at the partial key stands out, and it is: $K = 7$ rows; cipher row 1 belongs above cipher row 5, and those rows include 3 and 2 extra letters at the start, respectively. Those rows turn out to be rows 1 and 2 of the deciphered message. Adding one row at a time, the key appears: 13, 34, 57, 65, 22, 78, 49.

Revealing Insights

That key quickly unveils Patterson's hidden message, beginning with: "In Congress July Fourth." In fact, the complete decryption recites the preamble to the Declaration of Independence, authored by Thomas Jefferson.

Beyond deciphering Patterson's message, this work offers other lessons. For instance, assessing the similarity of two biological sequences resembles the challenge in aligning cipher text. For example, the Smith-Waterman algorithm—developed in 1981 by Temple Smith of Boston University and Michael Waterman of the University of Southern California—looks for similar regions in two sequences, instead of looking at each sequence as a whole, much like looking for pieces to the cipher solution. In fact, I constructed my dynamic program as a mimic to biological-sequence comparison. The logical structure designed for one field—biology—applies to another field, cryptanalysis. The mathematical justification for digraph analysis as a means of solving a cipher comes for free with the translation.

Patterson's letter also teaches us about cryptology ahead of its time. Although Patterson overlooked digraph properties when constructing his cipher, he did point out a crucial property of cryptology: Decryption of a cipher is difficult "even for one acquainted with the general system." This presages a principle published in 1883 by the Dutch cryptographer Auguste Kerckhoffs. Although no one argues Kerckhoffs's priority in publishing, the modesty that he expressed in his writing might indicate that, by 1883, the concept, still called Kerckhoffs' Principle, was not novel. Furthermore, this concept—the antithesis of security through obscurity—continues as a maxim to the present day. As stated so simply by Claude Shannon, known as the father of information theory: "The enemy knows the system."

As this journey to decrypt the cipher sent to Jefferson shows, Patterson adopted Shannon's maxim. Even knowing the system, however, the solution is not simple. Nonetheless, insight from the past two centuries of scientific development opens the path to this decryption and continued exploration across many fields.

Bibliography

Kahn, David. 1996. *The Codebreakers*. New York: Scribner.

Kerckhoffs, A. 1883. La cryptographie militaire. *Journal des Sciences Militaires* 9:5–38.

Smith, T. F., and M. S. Waterman. 1981. Identification of common molecular subsequences. *Journal of Molecular Biology* 147:195–197.

The Thomas Jefferson Paper, 1606–1827. The Library of Congress (http://memory.loc.gov/ammem/collections/jefferson_papers/).

Weber, Ralph E. 1979. *United States Diplomatic Codes and Ciphers 1775–1938*. Chicago: Precedent Publishing.

For relevant Web links, consult this issue of *American Scientist Online:*

http://www.americanscientist.org/issues/id.77/past.aspx

© 2009 Sigma Xi, The Scientific Research Society. Reproduction with permission only. Contact perms@amsci.org.