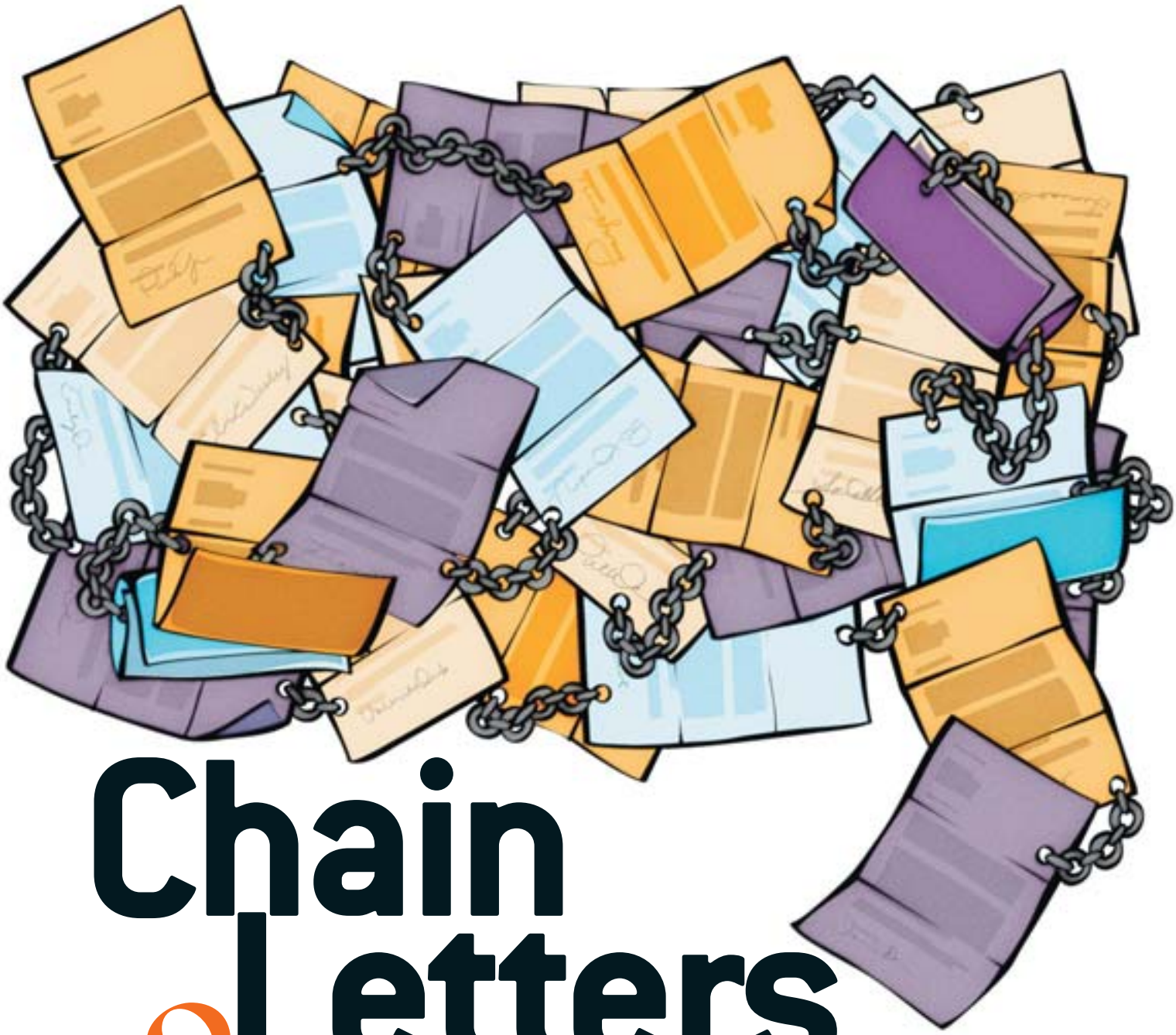


Chain Letters & Evolutionary Histories

A study of chain letters shows how to infer the family tree of anything that evolves over time, from biological genomes to languages to plagiarized schoolwork

BY CHARLES H. BENNETT, MING LI AND BIN MA

IN OUR HANDS ARE 33 VERSIONS OF A CHAIN LETTER,

collected between 1980 and 1995, when photocopiers, but not e-mail, were in widespread use by the general public. These letters have passed from host to host, mutating and evolving. Like a gene, their average length is about 2,000 characters. Like a potent virus, the letter threatens to kill you and induces you to pass it on to your "friends and associates"—some variation of this letter has probably reached millions of people. Like an inheritable trait, it promises benefits for you and the people you pass it on to. Like genomes, chain letters undergo natural selection and sometimes parts even get transferred between coexisting "species." Unlike DNA, however, these letters are easy to read. Indeed, their readability makes them especially suitable for classroom teaching of phylogeny (evolutionary history) free from the arcana of molecular biology.

The letters are an intriguing social phenomenon, but we are also interested in them because they provide a test bed for the algorithms used in molecular biology to infer phylogenetic trees from the genomes of existing organisms. We believe that if these algorithms are to be trusted, they should produce good results when applied to chain letters. Using a new algorithm that is general enough to have wide applicability to such problems, we have reconstructed the evolutionary history of our 33 letters [*see illustration on page 79*]. The standard methods do not work as well on these letters. Originally developed for genomes, our algorithm has also been applied to languages and used to detect plagiarism in student assignments: anything involving a sequence of symbols is grist for its mill.

A Mind Virus

THE 33 CHAIN LETTERS are a fascinating collection. We labeled them arbitrarily from L1 to L33. The letters differ significantly. There are 15 titles, 23 names for "an office employee," and 25 names for the original author of the letter. Misspellings, swapped sentences, and missing or added phrases, sentences and paragraphs are common. (A typical letter is shown on the next page, along with some of the many variations.) Nearly all are more or less faded photocopies of typescripts, leading us to surmise that the mutations arose by an intermittent process, whereby a letter would be photocopied for several generations until its legibility was so reduced that the next recipient decided to retype it, introducing new errors and variations.

All but three of the letters we received were unique; for L4, L6 and L22, a second copy arrived within a few months of the first. Besides the 33 English-language letters, we received (but did not include in our study) four in French and one each in Dutch and German, all clearly sharing a common ancestry with the ones in English.

To analyze the letters, we retyped them into computer files entirely in lowercase, ignoring extra information such as dates and marginal notes as well as the division of the text into lines and paragraphs. Each letter became one continuous string of characters.

Before we applied our new algorithm, we tried analyzing the letters with a procedure called multiple alignment, which is widely used for examining genes to infer phylogeny. This method attempts to line up as many matching sections of all the letters as possible. The amount of matching between any pair of letters defines their similarity, and from that data another algorithm constructs an evolutionary tree. Unfortunately, multiple alignment only finds matches with everything remaining in the same order, so it gets confused by L12 and L26, in which the order of sentences has been rearranged. For the same reason, the technique is known to work better within individual genes than for whole genomes, in which such translocations are more common.

We tried omitting L12 and L26 and then performing multiple alignment on the remaining 31 letters. Even with this truncated set, the resulting tree seemed wrong, classifying L6, L7 and L13 as closely related. This error occurred because those three letters are all relatively short, giving them a correspondingly small number of differences. This problem can arise in genetics as well: merely counting differences can overestimate the similarity of short genomes while underestimating that of long ones. A proper measure should give more weight to a small difference in a small genome than to a small difference in a big genome.

We turned to devising our own similarity measure, one that could be applied to genomes, chain letters or any other type of data that might be stored as a computer file. We wanted to make our new similarity measure insensitive to minor mix-ups such as translocations, which represent only a small loss of informational similarity. To cope with differences in length, we wanted our measure to assign two completely dissimilar data files a score of 0 and two identical ones a score of 1, regardless of their sizes.

The natural measure of the information content in a data file is not its raw size in bits but rather the smallest size it can be compressed to by a file compression program such as zip or StuffIt. These programs are designed to save space on hard disks by finding and squeezing out the most common kinds of redundancy (for instance, repeated phrases), resulting in a smaller file from which the original can be perfectly reconstructed when needed.

Something interesting happens if we compress two files together so that both can be regenerated from the compressed file. If the two files share no information at all, the joint compressed file will be as big as the two individual compressed files combined. But if the two files contain some of the same information,

SARA CHEN

THE CHAIN LETTERS ARE ESPECIALLY SUITABLE FOR CLASSROOM TEACHING OF PHYLOGENY.

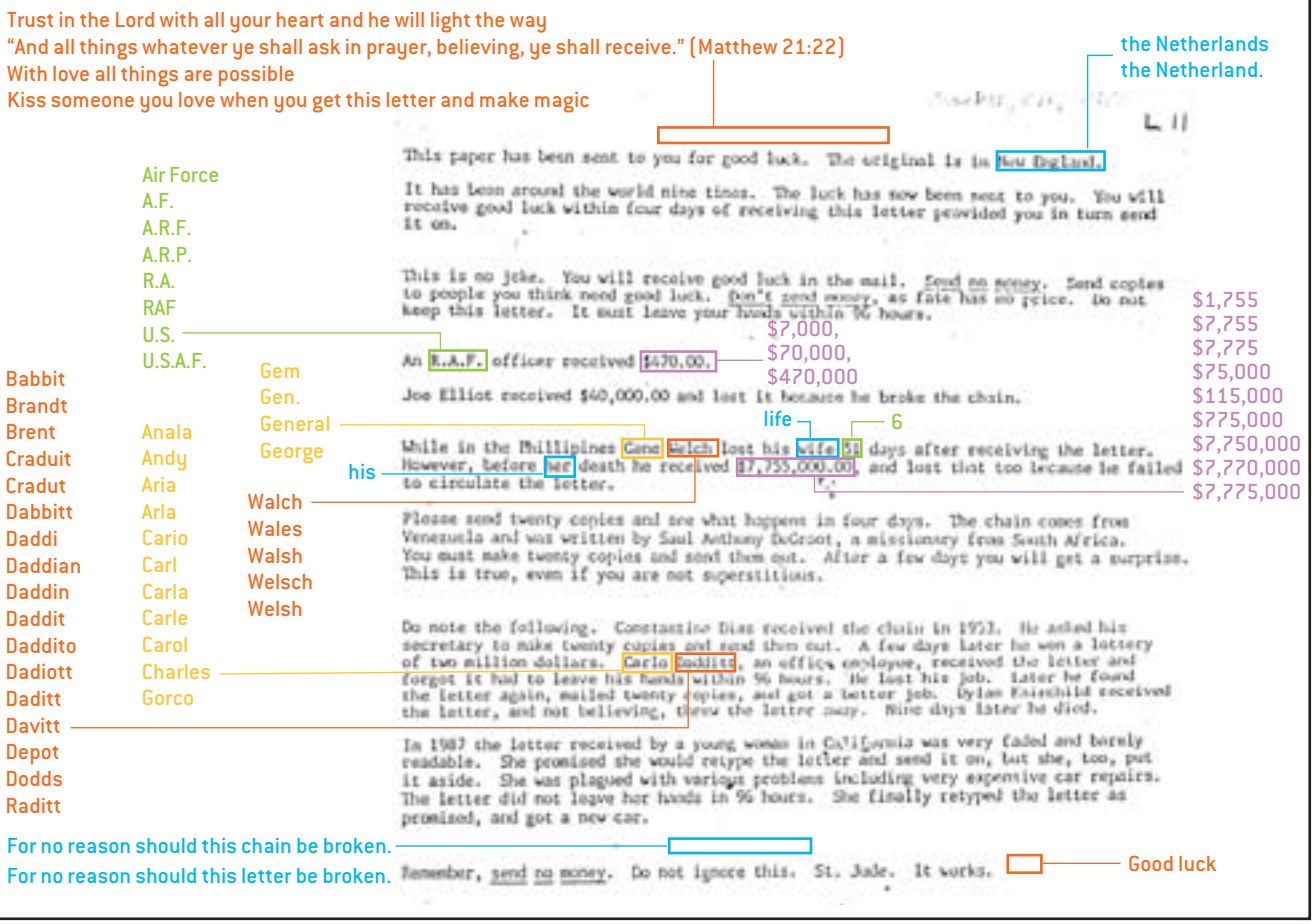
that repetition will be detected by a good compression program, and the joint compressed file will be smaller. In this way, the size of the joint compressed file compared with the sum of the individual compressed files provides a measure of the files' similarity.

That measure is not yet a good one for our purposes, because two large files will tend to have greater similarity than two small files. To correct this problem, we define our "relatedness" measure to be the *proportion* of shared information—that is, the percentage by which the sum of the separately compressed files exceeds the size of the jointly compressed file. This makes the relatedness range from 0 for unrelated files to 1 (or 100 percent) for identical files, regardless of length.

Which compression program should we use? Obviously, our relatedness measure will depend on that. Ideally, we would want to use a program that compresses every file to the smallest possible size. The study of information measures defined in terms of such ultimate compressibility forms an elegant branch of information theory known as algorithmic information theory or Kolmogorov complexity (after the late Russian mathematician Andrei N. Kolmogorov, one of its founders). Unfortunately, information theorists have proved that such an ideal zip program would take essentially an infinitely long time to perform its task. For our purposes, then, we decided to use a particular compression algorithm called GenCompress, created by Xin Chen of the University of California at Santa Barbara. GenCompress was designed for genomes and works well on them. As we shall see, it also works well on chain letters.

THEME AND VARIATIONS

A SAMPLE CHAIN LETTER, labeled L11, illustrates some of the ways that related letters changed when they were retyped (presumably after photocopies became illegible). The greatest variations occurred with unfamiliar names and quantities of money—errors in those elements are easily overlooked because they do not change the meaning of the letter.



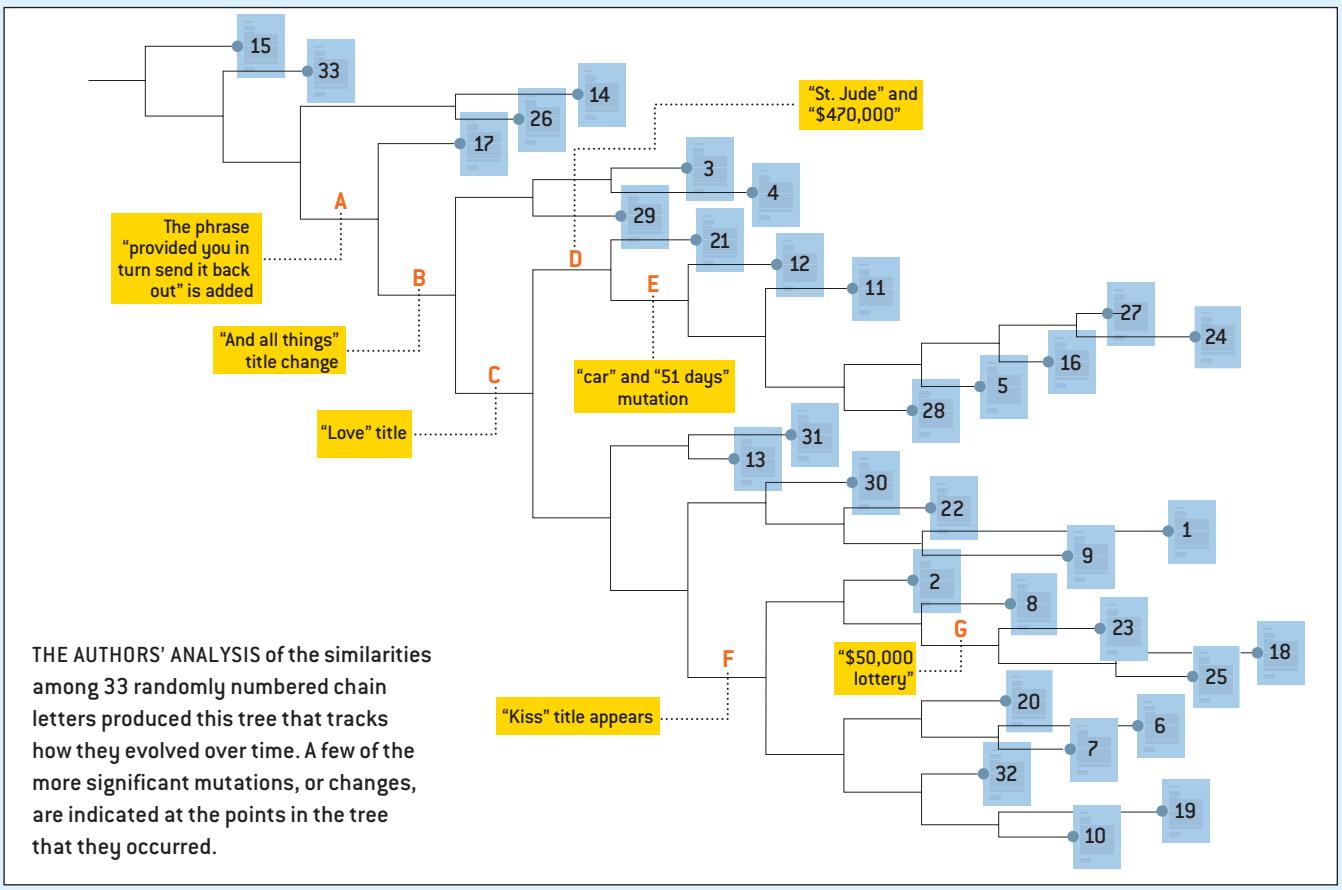
How the Letters Evolved

THE EVOLUTIONARY TREE of the chain letters deduced by the authors' automated "relatedness" measure shows a number of interesting features. (Some other features are discussed in the main text.) At point A, a phrase, "provided you in turn send it back out," was adopted. At point B, a new title evolved: "And all things whatever ye shall ask in prayer, believing, ye shall receive." At point C, the title further mutated to "With love all things are possible." Also at point C, "The Netherlands" changed to "New England," and "General Welch" became "Gene Welch." The sentence "For no reason should it be broken" got lost.

At point F, the title mutated to "Kiss someone you love when you get this letter and make magic." After we had finished our analysis, we found a very interesting and comprehensive study of more than 460 chain letters of many varieties by mathematician Daniel W. VanArsdale. His study raised the question, Which title came first: the "Kiss" or the "Love" one? We have not yet applied our algorithm to the many chain letters in his collection, but judging from our phylogeny, the "Love" title came first. This conclusion is further supported by the amount of money Gene Welch received: in all the "Kiss" letters (except for the mutation at G), he received $7,755, whereas in the "Love" group he got $7,755,000. In the group before C, the figure was $775,000. The mutation sequence $775,000 → $7,755,000 → $7,755 is clearly more parsimonious than $775,000 → $7,755 → $7,755,000.

At point D, $70,000 grew to $470,000, and "St. Jude" was adopted. Neither $470,000 nor St. Jude appears outside of this group. At point E, two interesting alterations happened concurrently: the California woman's car story was added and the time between Gene Welch's receiving the letter and losing his wife changes from six days to 51 days. This is consistent across all letters in the group rooted at E, except for L28, which requires a special explanation. The "car and 51" mutation doesn't appear anywhere outside of the E group.

L28 provides evidence of horizontal transfer—that is, the transfer of information from one "organism" to another in addition to simple inheritance. In the group rooted at D, every letter has each R.A.F. officer receive $470,000 or $470, except for L28, in which the amount is $70,000. L28 also has Gen. Welch, who otherwise features only in letters before point C. All the letters in the D group except L21 have the car story, and all but L21 and L28 have the "51 days" mutation. It seems unbelievable to assume that L28 gained "$70,000" and "Gen. Welch" by mutation independently of the other instances of these mutations elsewhere in the tree. One might try placing the "car and 51" mutation before the "$470,000 and St. Jude" mutation, but then L21 must undergo a very implausible genesis: either it must lose the whole car story and mutate "51 days" back to "six days," or it must gain "$470,000" and "St. Jude" independently. Apparently somebody had two letters in his or her possession while composing L28 (or L21) and introduced a foreign gene from a letter before C. *—C.H.B., M.L. and B.M.*



THE AUTHORS' ANALYSIS of the similarities among 33 randomly numbered chain letters produced this tree that tracks how they evolved over time. A few of the more significant mutations, or changes, are indicated at the points in the tree that they occurred.

SARA CHEN

THE ACCUMULATION OF MUTATIONS IN A MITOCHONDRIAL GENOME (OR A CHAIN LETTER) ACTS AS A CLOCK.

Given a set of chain letters, it is a straightforward and entirely automatic process to calculate the relatedness of each pair using the GenCompress program. The next step, converting the relatedness data into an evolutionary tree, is also largely automatic (many software packages exist for this purpose). The result can either be a simple tree diagram with arbitrary branch lengths of the kind shown on the preceding page, indicating simply the qualitative pattern of descent, or a more detailed diagram, with branch lengths that represent relatedness quantitatively.

In either case, the main human input is deciding where to put the root of the tree, which represents the hypothesized common ancestor of all the letters (or species). In biological phylogenies, the root stands for an extinct species from millions of years ago, so it should not be too closely related to any of the branches denoting organisms alive today. In our study, the chain letters were collected over a 15-year period, and some of them were dated near the beginning of that time, so we chose to put the root near one of these (L15). Unfortunately, most of the letters were collected without recording their postmark or date of receipt, reflecting the fact that this project began as a hobby and only belatedly evolved into scientific research.

St. Jude's Phylogeny

THE EVOLUTIONARY TREE that was inferred for the chain letters appears to be almost a "perfect" phylogeny, in the sense that documents that share the same characteristic are always grouped together. After the tree was built, we were able to use it to help make numerous hypotheses about how the letters evolved.

First we judge that the letters before point C in our phylogeny are the oldest [*see illustration on preceding page*]. The chief evidence is that the name "Carlo Dadditt" and the title of the letter had the most mutations in this group of letters. We expect such errors to be more common in the oldest letters because photocopiers were less available at that time and retyping more frequent. In addition, among the 14 dated letters, the two that occur in the pre-C group (L4 and L15) are the oldest. These older letters are all titled with religious prayers, come from "The Netherlands" and contain the sentence "For no reason should it be broken."

Next we see an effect familiar from molecular biology, in which different parts of the genome have quite different mutation rates. Active sites of enzymes mutate scarcely at all, whereas parts far from the active site continually undergo random drift. Similarly, with chain letters the parts required for "viability" do not mutate at all, but more arbitrary parts, such as the types of mishaps that would befall those who did not propagate the chain, mutate more. Parts with little intrinsic meaning to help catch errors—for instance, unfamiliar names like "Gem Walsh" and "Carlo Craduit"—mutate the most.

Another biological phenomenon appearing in the chain letters is the occurrence of parallel, compensating mutations: two mutations that would be detrimental individually and that must occur together to be neutral or beneficial. Excluding L12 and L26, in which no one dies, all letters before point C (except nearby L29) read:

> General Welsh (or a variation) lost his life … however before his death …

On the other hand, letters after point C read:

> Gene Welch (or a variation) lost his wife … however before her death …

To preserve the sense, "his" mutates to "her" when "life" mutates to "wife." [*See box on preceding page for more of these observations.*]

Mammals and Plagiarism

BESIDES ANALYZING chain letters, our relatedness measure has been used in a wide range of settings. In bioinformatics itself, we used it to analyze the mitochondrial genomes of 18 mammals. Mitochondria are energy-producing organelles within cells whose genes are inherited solely from the mother (similar to how a chain letter inherits from a single "parent"). Because no reshuffling of maternal and paternal genes occurs, the accumulation of mutations in the mitochondrial genome acts as a clock measuring when an organism's ancestors diverged from related species.

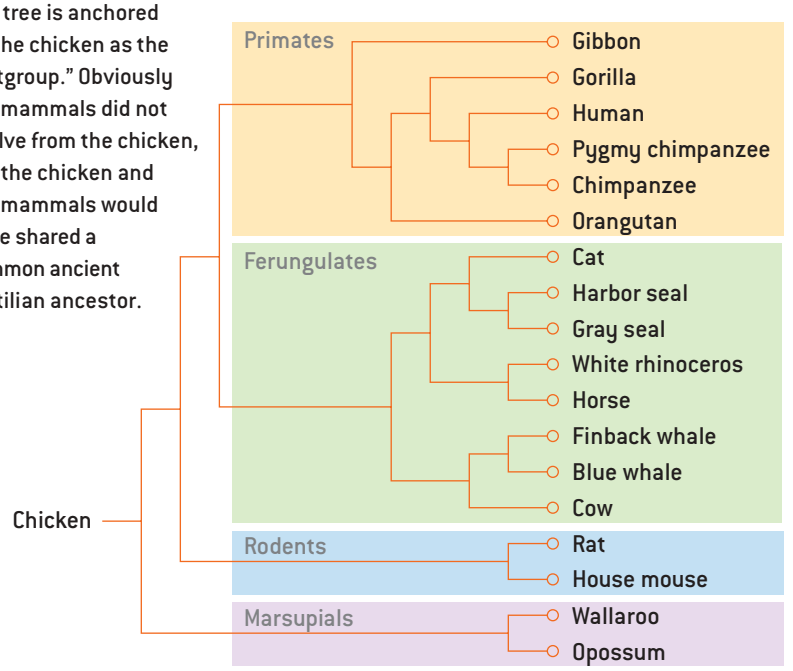
Traditional methods applied to different mitochondrial genes often give conflicting evolutionary trees, and many, unlike our new measure, cannot be applied successfully to an entire genome because of problems such as translocations. For example, using the traditional methods, about half a dozen mitochondrial genes imply that primates, such as ourselves, are more closely related to rodents than to ferungulates, a diverse group that includes cows, horses, whales, cats and dogs. Another half a dozen genes imply that primates and ferungulates are the more closely related pair, which is generally believed to be correct based on other multiple lines of evidence, such as nonmitochondrial genes and the fossil record.

THE AUTHORS

CHARLES H. BENNETT, MING LI and *BIN MA*'s joint work on chain letters began as a result of a hike in the Hong Kong mountains during which Bennett mentioned his collection to Li. Bennett is an IBM Fellow at the company's research division in Yorktown Heights, N.Y. He focuses on the physics of information processing and is a co-discoverer of quantum teleportation. Li studies bioinformatics and Kolmogorov complexity as professor of computer science at the University of Waterloo in Ontario. Ma works on bioinformatics and algorithm design as assistant professor of computer science at the University of Western Ontario.

RELATEDNESS OF MAMMALS

DIVERSE PROBLEMS can be analyzed with the authors' relatedness measure. Applied to whole mitochondrial genomes, it produced this phylogeny of mammals. Note how primates are more closely related to ferungulates than to rodents—which is believed to be true. This degree of kinship is ambiguous when using traditional techniques. The tree is anchored by the chicken as the "outgroup." Obviously the mammals did not evolve from the chicken, but the chicken and the mammals would have shared a common ancient reptilian ancestor.



When our method is applied to whole mitochondrial genomes, it produces this latter evolutionary tree without needing any ad hoc tinkering to resolve ambiguities or contradictions [*see illustration above*].

Taking the art of phylogenetic inference to an audacious extreme, Dario Benedetto, Emanuele Caglioti and Vittorio Loreto of La Sapienza University in Rome tried inferring a phylogeny of human languages not by analyzing the languages' known literatures or history but merely by applying a method similar to ours to 52 translations of the Universal Declaration of Human Rights. The result was surprisingly good, considering the tiny body of evidence on which it was based. One notable mistake was the classification of English as a romance language, closely related to French, whereas historically English evolved within the Germanic group. This error arises because of the great many French words that English acquired after the Norman Conquest (an example of parallel transfer).

Another application of our measure has been the detection of plagiarism in students' homework assignments. In one instance, two assignments in a computer programming class were flagged as being unusually alike, but the instructor could not see any obvious evidence of copying when he examined them himself. The two students were approached and, in the interest of research, given immunity to plagiarism charges in return for an honest account of whether they had collaborated. Apparently the two students had discussed the problem and how they planned to tackle it but had not worked together beyond that level. If that is what happened, our distance algorithm detected the subtle similarities engendered by their discussion!

The automatic nature of our procedure is both an advantage and a disadvantage. On the one hand, it yields objective answers, free from the need to weigh various lines of evidence (such as DNA versus the fossil record) or to take account of which parts of the genome mutate fastest. On the other hand, it does not benefit from the insights that might come from such additional data. All methods of phylogenetic inference are imperfect and will sometimes mistakenly infer a phylogeny that differs from what actually took place historically. Like historians and paleontologists, evolutionary molecular biologists have come to accept that no matter how many lines of evidence they consider, the full truth about the past can never be reconstructed. This is especially true with regard to extinct species. Many species that once existed will never be known, because they left neither fossils nor descendants. Likewise many languages have become extinct, vanishing without a trace even in the past century.

In the realm of chain letters, it is certain that many letters became extinct when too many recipients broke the chain. Like the lost plays of Sophocles, these letters' texts may never be recovered, and even their existence can only be surmised from circumstantial evidence, such as a rash of unemployment and expensive car repairs occurring in California for no apparent reason. SA

MORE TO EXPLORE

An Introduction to Kolmogorov Complexity and Its Applications. Second edition. Ming Li and Paul M. Vitányi. Springer-Verlag, 1997.

An Information-Based Sequence Distance and Its Application to Whole Mitochondrial Genome Phylogeny. Ming Li, Xin Chen, Jonathan H. Badger, Sam Kwong, Paul Kearney and Haoyong Zhang in *Bioinformatics,* Vol. 17, No. 2, pages 149–154; February 2001.

Language Trees and Zipping. Dario Benedetto, Emanuele Caglioti and Vittorio Loreto in *Physical Review Letters,* Vol. 88, No. 4, pages 048702-1–048702-4; January 28, 2002.

Chain Letter Evolution. Daniel W. VanArsdale: **www.silcom.com/~barnowl/chain-letter/evolution.html**

The chain letters used in this article and other data are at **www.math.uwaterloo.ca/~mli/chain.html**

A discussion of phylogenetic inference methods is at **helix.biology.mcmaster.ca/721/outline2/node47.html**

Kolmogorov complexity is discussed at **www.wikipedia.org/wiki/kolmogorov_complexity**

The program for plagiarism detection is at **http://dna.cs.ucsb.edu/SID/**

SARA CHEN