Linking Chain Letters

Charles Bennett[*] IBM Research Division
Ming Li[†] University of Waterloo
Bin Ma[‡] University of Waterloo

In our hands are 33 versions of a chain letter, collected during 1980 – 1995, when photocopiers, but not email, were in widespread use by the general population. These ghostly chain letters have passed from host to host, mutating and evolving. We have reconstructed their evolutionary history by our computer programs, Figure 1. Why are we interested in them? Because they provide a test bed for phylogenetic inference algorithms in molecular biology and potential similar applications (eg phylogeny of historic documents), in addition to their being an interesting social phenomenon [11]. Like a gene, the average length of these letters is about 2000 characters; like a virus [11], 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 hundreds of millions of people; like an inheritable trait, it promises you and people you pass it on to prosperity; like a genome, it mutates (fast) and sometimes even undergoes a little horizontal transfer [12, 8]. On the other hand, unlike the genes, these letters are humanly readable. Indeed, the readability of these letters makes them especially suitable for classroom teaching of phylogeny and phylogenetic inference, free from the arcana of molecular biology. We believe that if algorithms used to infer phylogentic trees from genetic sequences are to be trusted, they should also produce reasonable results when applied to chain letters.

The letters are arbitrarily labelled from L1 to L33; a typical letter (L24) is given in the appendix. All 33 letters can be found at our website [17]. No two letters are identical. In fact, most of them differ significantly. There are 15 different titles, 23 different names for "an office employee", 25 different names for the author of the article. Misspells, missing or added phrases, sentences and paragraphs, and swapped sentences are commonplace. The following table lists some distinguishing features of these letters. In the table, T_1 is "With love all things are possible", T_2 is "Kiss someone you love when you get this letter and make magic", T_3 is "And all things whatever ye shall ask in prayer, believing, ye shall receive (Matt. 21:22)", T_4 is "Trust in the Lord with all your heart and he will light the way". And T_i' denotes variants of T_i. Nearly all the specimens 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.

[*]T.J. Watson IBM Research Center, Yorktown Heights, NY 10598, USA. Email: bennetc@watson.ibm.com.
[†]Address: Department of Computer Science, University of Waterloo, Waterloo, ON, N2L 3G1, Canada. E-mail: mli@wh.math.uwaterloo.ca
[‡]Address: Same as ML. E-mail: b3ma@wh.math.uwaterloo.ca

We have derived an optimal multiple alignment of all 33 letters. ¿From this alignment we constructed trees using various programs including Neighbor Joining (NJ) and Fitch-Margoliash method in PHYLIP and PAUP [9, 15, 10] as well as the hypercleaning program [4]. However, these trees are obviously wrong for one simple reason: in L12 and L26 several sentences translocated to different locations. Such translocations preclude the multiple alignment approach – including maximum likelihood methods even if we knew how to define mutation probabilities for chain letters.

	Title	The original copy is in	What officer received $	How much did he receive	Who lost life or wife	Author of the letter	Who lost his job	California woman	St. Jude	Date received
L1	T_1	New England	RAF	$70k	Gene Walsh, wife	Decroup	Carle Dadditt	No	No	
L2	T_2'	New England	officer	$70k	Gene Welch, wife	de Croole	Carlo Dodds	No	No	9/23/85
L3	T_3	Netherland	ex-officer	$70k	General Welch, life	De Vold	Carlo Daditt	No	No	
L4	T_3'	Netherlands	U.S.A.F.	$70k	General Wales, life	DeCaiod	Carlo Depot	No	No	1981
L5	T_1	New England	R.A.F.	$470k	Gene Walch, wife	de Group	Carlo Dadditt	Yes	Yes	Summer 1989
L6	T_2	New England	A.R.P. office	$70k	Gene Welch, wife	DECROOF	Carla Daddin	No	No	1986
L7	T_2	New England	A.R.F. office	$70k	Gene Welch, wife	deCroof	Arla Daddit	No	No	
L8	T_2	New England	R.A.F.	$70k	Gene Welch, wife	deCroof	Arla Daddit	No	No	
L9	T_1	New England	RAF	$70k	Gene Welch, wife	De Croup	Carlo Daddit	No	No	1988
L10	T_2'	New England	R.A.F.	$70k	Gene Welch, wife	DeCroof	Anala Daddit	No	No	
L11		New England	R.A.F.	$470	Gene Welch, wife	DeGroot	Carlo Dadditt	Yes	Yes	1988
L12	T_1'				Gene Welch		Carlo Dadiott	Yes	Yes	Fall 1990
L13	T_1	New England	RAF	$70k	he, wife	de Group	Carlo Dabbitt	No	No	
L14	Thank Prayer	Netherlands	U.S.	$7k	Gem Walsh, life	de Cader	Charles Brent	No	No	
L15	T_4'	Netherlands	RAF	$70k	General Welch, life	de Cadif	Carlo Craduit	No	No	early 1980's
L16	T_1'	New England	RAF	$470k	George Welsch, wife	DeGrou	Carl Daddit	Yes	Yes	7/92
L17	T_4	Netherlands	RAF(?)	$70k	General Welch, life	Da Dadif	Carlo Raditt	No	No	
L18	T_2	New England	R.A.F.	$70k	Gene Welch, wife	de Croff	Aria Daddit	No	No	Jan. 1986
L19	T_2	New England	R.A.F.	$70k	Gene Welsh, wife	DeCroof	Aria Daddito	No	No	1986
L20	T_2	New England	R.A.F.	$70k	Gene Welch, wife	deCroof	Arla Daddit	No	No	
L21	T_1	New England	R.A.F.	$470k	Gene Welch, wife	DE GROUP	Carlo Daddit	No	Yes	
L22	T_1'	New England	RAF	$70k	Gene Walsh, wife	De Croup	Carlo Daddian	No	No	
L23	T_2'	New England	Air Force	$70k	Gene Welch, wife	Decroe	Andy Babbit	No	No	2/92
L24	T_1	New England	R.A.F.	$470k	George Welch, wife	De Grou	Carlo Daddit	Yes	Yes	10/91
L25	T_2	New England	R.A.F.	$70k	Gene Welch, wife	de Creff	Aria Daddit	No	No	4/1986
L26	T_4'	Netherlands	R.A.	$70k		De Codi	Carol Brandit	No	No	
L27	T_1	New England	R.A.F.	$470k	George Welch, wife	DeGrou	Cario Daddit	Yes	Yes	
L28	T_1	New England	R.A.P.	$70k	Gen. Welch, wife	DECROUP	Carlo Daddit	Yes	Yes	
L29	Matthew 21:22	Netherlands	R.A.F.	$70k	General Welch, wife	Dowell	Carlo Davitt	No	No	
L30	T_1	New England	R.A.F.	$70k	Gene Welch, wife	DE GROUP	Carlo Dadditt	No	No	
L31	T_1	New England	an officer	$40k	Gene Welch, wife	De Croup	Carlo Daddit	No	No	
L32	T_2	New England	R.A.F.	$70k	Gene Welch, wife	DECROOF	Arla Daddit	No	No	
L33	T_4'	Netherlands	R.A.F.	$70k	General Welsh, life	de Cadif	Corco Cradut	No	No	

The phylogeny of these letters in Figure 1 was constructed by first computing the amount of information $d(x, y)$ shared by each pair of letters x and y, then constructing the tree from these distances using the quartet method and the hypercleaning algorithm [4], then rooting at letter L15. Finally the edge lengths are optimized according to the distance matrix minimizing sum of squared differences. Using the same distance matrix, Fitch-Margoliash method in PHYLIP gives a similar tree, so does the NJ program, modulo some minor inconsistencies. *After* the tree was built, we have amusingly observed:

- The group before point A is the oldest. This is supported by the clear evidence that the name Carlo Dadditt and the title had the most mutations in that group. Two arguments present themselves: (a) there were fewer photocopiers and retyping resulted mutations; (b) these letters had longer time to mutate. This is of course decisively supported by the fact that L4 and L15 are the oldest-dated letters (among the 14 dated ones). These older letters all have religious prayers as the titles, and they were all from "The Netherlands" and all have the sentence "For no reason should it be broken."

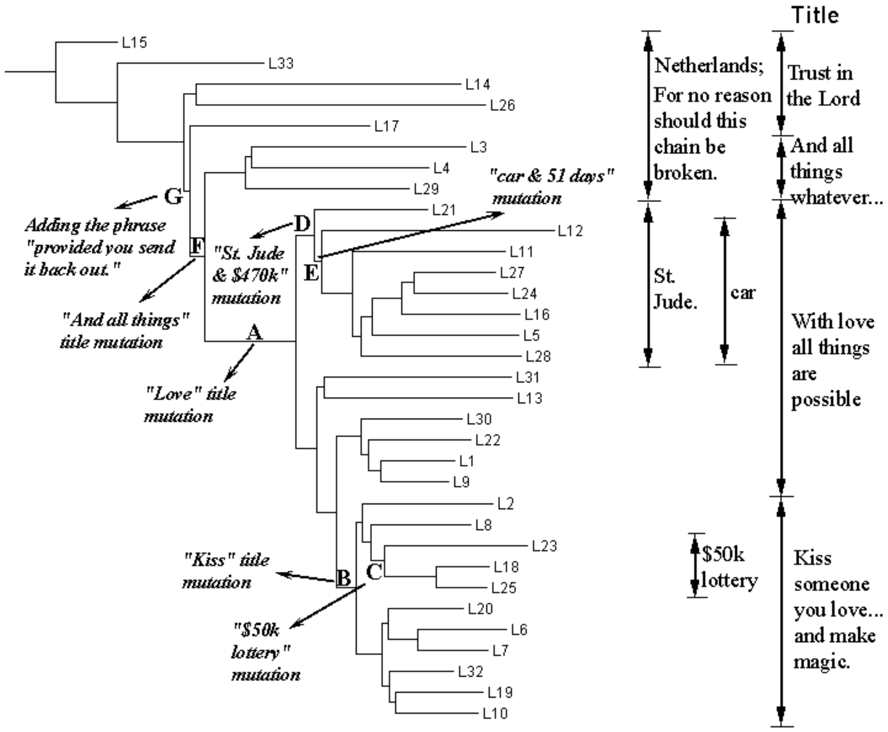

Figure 1: An Evolutionary Tree For 33 Chain Letters

- At point G, a phrase "provided you in turn send it back out" was adopted. At point F, a new title evolved: 'And all things whatever ye shall ask in prayer, believing, ye shall receive". At point A, the title further mutated to "With love all things are possible". The tree does not tell from which title this title has evolved from. At this point "The Netherlands" mutated to "New England". "General Welch" became "Gene Welch". The sentence "For no reason it should be broken" got lost.

- At point B, the title mutated to "Kiss someone you love when you get this letter and make magic". After we had finished this study, we found [7], a very interesting and comprehensive study of over 300 hundred chain letters. [7] raised an open question about which title came first: "Kiss" title or "Love" title? Our phylogeny shows "Love" title came first. This is further supported by the fact that in all the "Kiss" letters Gene Welch recieved $7755 (except the mutation at C), while in the Love group this figure is $7,755,000. In the group before A, this figure was $775,000. The mutation sequence $775,000 $\rightarrow$ $7755,000 $\rightarrow$ $7755 is clearly more parsimonious than $775,000 $\rightarrow$ $7755 $\rightarrow$ $7755,000.

- At point D, $70k mutated to $470k and "St. Jude" was adopted. Outside this group, there is neither $470k nor St. Jude. At point E, two more interesting mutations clearly happened concurrently: adding the California woman's "Car" story, and Gene Welch lost his wife after "51 days" instead of "6 days". This is consistent with all letters in the group rooted at E except for L28 which will be explained, outside this group there is no "Car & 51" mutation anywhere.

- Evidence of *horizontal transfer*: in the group rooted at D, each R.A.F. officer received \$470k (or \$470), except for L28 having \$70k and "Gen. Welch" which is a feature for only and most letters before point A. They all have the "Car" story except for L21, and "51 days" mutation except for L28 and L21 having 6 days. It seems that one cannot escape from the horizontal transfer theory here. It would be unbelievable to assume that L28 has independently gained \$70k and "Gen. Welch". If one argues that the "Car & 51" mutation had happened before the "\$470k & St. Jude" mutation, then one must argue that L21 has somehow lost the whole "California woman" story and more remarkably independently mutated 51 back to 6, or it has gained \$470k and St. Jude independently, both very unlikely. Apparently, somebody had two letters in her hands when she copied, or composed, L28 (or L21), with a foreign gene from a letter before A.

- In molecular biology, it is well-known that different parts of the genome have quite different mutation rates, with active sites of enzymes mutating scarcely at all, while parts far from the active site undergo random drift. This also naturally happens with chain letters. The parts required for "viability" did not mutate at all, more arbitrary parts, like the mishaps that befell those who did not propagate the chain, mutated more, and the parts with little intrinsic error-correction ability, e.g. unfamiliar names like Gem Walsh and Carlo Craduit, mutated the most.

  Another biological phenomenon also appearing in the chain letters is the occurence of parallel compensating mutations: two mutations either of which would be detrimental by itself, but when both occur the effect is neutral or beneficial. Excluding L12 and L26, in which no one dies, all letters before point A (except nearby L29) read:

  ```
  General Welsh (or its variation) lost his life ... however before his death ...
  ```

  On the other hand, letters after point A read:

  ```
  Gene Welch (or its variation) lost his wife ... however before her death ...
  ```

  To preserve the sense, "his" mutates to "her" when "life" mutates to "wife".

It is interesting to note that this tree is almost a *perfect* phylogeny in the sense that documents that share the same characteristic, many of them listed above, are always grouped together.

We have at our disposal a remarkable set of test data allowing us to see, in broad daylight, what mistakes a phylogeny program is making and how, hopefully, to repair them. We have pointed out clearly why multiple alignment approach does not work here. We in fact have delayed writing this paper for two years until we found a suitable phylogenetic method.

These chain letters not only provide a wonderful test bed for our phylogenetic analysis tools, but also open the door for applying these tools in bioinformatics to other areas from analyzing internet homepages to classifying historical records [1], from phylogeny of computer viruses or search engines to comparing student programming assignments.

Methods

All letters were exactly typed in as is, all in lower case letters. The title was separated by a % sign which does not appear elsewhere. The division of the original letters into lines, paragraphs, and (in a few cases) pages was not preserved. Different ways of writing Arabic numbers were unified. Extra information like dates and marginal notations were ignored. Besides the 33 English-language letters included in our study we received four in French, and one each in Dutch and German, all clearly having a common ancestry with the English-language letters. These were excluded from the study because we doubted their degree of similarity could be adequately assessed by our algorithms.

To compute the distance between chain letters x and y, we used a new distance measure defined in [6, 13]:

$$d(x,y) = 1 - \frac{K(x) + K(y) - K(xy)}{K(xy)}$$

where $K(x)$ is the algorithmic entropy (Kolmogorov complexity) of the string x, ie the the length in bits of the shortest program causing a standard universal computer to compute x as its unique output. The numerator of the fraction is the mutual algorithmic information between x and y, equal (to within a logarithmic error term [14]) to the number of bits x tells about how to compute y, or vice versa. Mutual algorithmic information is not itself a distance and does not satisfy the triangle inequality, but $d(x,y)$ defined above does [13], ranging from a minimum of 0 when $x = y$ to a maximum of 1 when x and y are independent strings of equal or differing length. By contrast with our present d, the Information Distance defined in [2] can become arbitrarily large, and thus is not suitable in the present aplication because it would over-punish long deletions. Of course it is well known that K is an uncomputable function [3, 5, 14]; accordingly we used our program *GenCompress* developed in [6] (which gives the best compression ratios for DNA sequences) to heuristically approximate it. The distance matrix computed from the 33 letters can also be found at [17]. Then the hypercleaning algorithm in [4] as well as NJ and Fitch-Margoliash programs are used to construct the final phylogeny in Figure 1. The distance $d(x,y)$ has also been successfully used to construct whole genome phylogenies from complete mitochondrial genomes of eukaryotes [13], one such tree, constructed by J.H. Badger using this method, is given in Figure 2. Figure 3 gives an evolutionary tree of Indo-European languages, constructed by Warnow ?? using a different method.

Acknowledgement. We gratefully thank J.H. Badger, B. Brejova, X. Chen, P. Kearney, S. Kwong, T. Vinar, H. Zhang, L.X. Zhang, W.H. Zou for their collaboration and help on related projects.

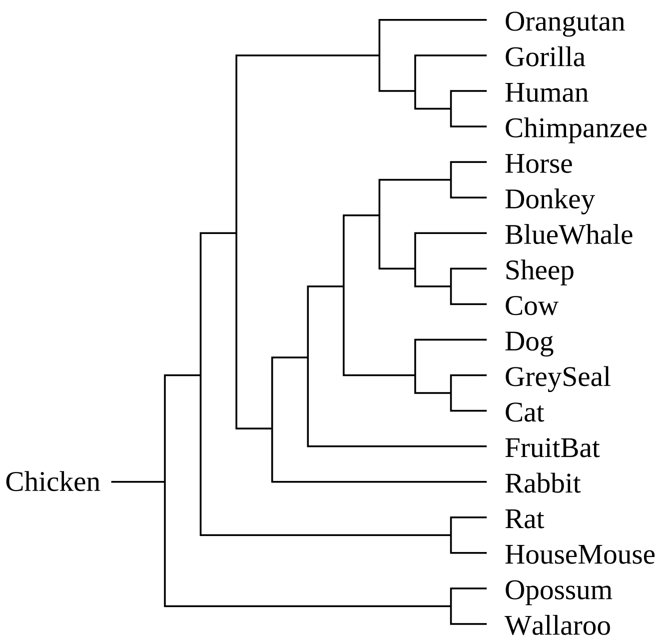

Figure 2: A phylogeny of some mammals using chicken as outgroup

References

[1] Barbrook, A., Howe, C., Blake, N., & Robinson, P., The phylogeny of the Canterbury tales. *Nature* 394, 830 (1998).

[2] Bennett, C.H., Gács, P., Li, M., Vitányi, P., & Zurek, W., Information distance. *IEEE Trans. Inform. Theory*, 44:4, 1407-1423 (1998).

[3] Bennett, C.H. & Gardner, M., The random number omega bids fair to hold the mysteries of the universe. *Scientific American*, 241, 20-34, (May 1979).

[4] Berry, V., Bryant, D., Jiang, T., Kearney, P., Li, M., Wareham, T., & Zhang, H. A practical algorithm for recovering the best supported edges in an evolutionary tree, *Proc. 11th ACM-SIAM Sypm. Disc. Alg.*, 2000.

[5] Chaitin, G.J., Randomness and mathematical proof. *Scientific American*, 232, 47-52, (May 1975).

[6] Chen, X., Kwong, S. & Li, M. A Compression Algorithm for DNA Sequences and Its Applications in Genome Comparison. *Genome Information Workshop, 1999*, Tokyo, Dec. 1999 (Best Paper Award). Also in *5th Int'l Conf. Computational Molecular Biology*, 2000.

[7] VanArsdale, D.W., Chain letter evolution. At `http://www.silcom.com/~barnowl/clevo/start.htm`, 1998.

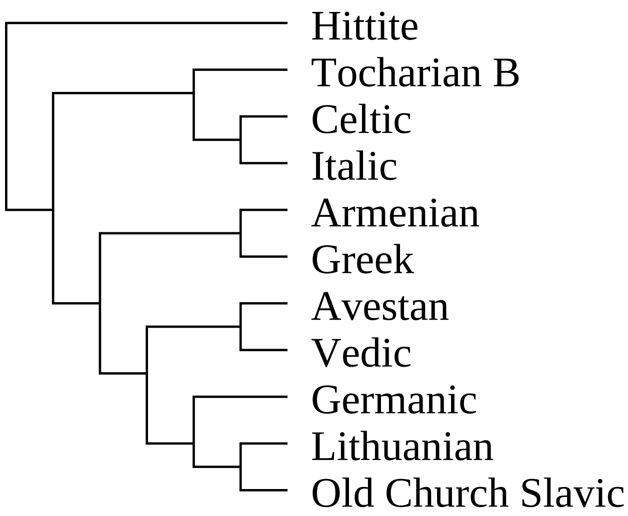

Figure 3: A phylogeny of languages

[8] Doolittle, W.F., Phylogenetic classification and the universal tree, *Science* **284**, 2124-2129 (1999).

[9] Felsenstein, J., PHYLIP (Phylogeny Inference Package) and manual. version 3.5c, 1993.

[10] Fitch, W.M. & Margoliash, E., Construction of phylogenetic trees. *Science*, 155, 279-284(1967).

[11] Goodenough, O.R. & Dawkins, R., The 'St. Jude' mind virus. *Nature* 371, 23-24(1994).

[12] Lawrence, L.G. & Ochman, H., Molecular archaeology of Escherichia coli genome, *Proc. Natl. Acad. Sci. U.S.A.* 95, 9413(1998)

[13] Li, M., Chen, X., Badger, J., Kwong, S., Kearney, P., & Zhang, H., An information based sequence distance and its application to whole genome phylogeny. *Manuscript*, 1999.

[14] Li, M. & Vitányi, P. *An introduction to Kolmogorov complexity and its applications.* Springer-Verlag, 1st Edition 1993, 2nd Edition 1997.

[15] Swofford, D.L., PAUP Version 3.1.1. (Smithsonian Institute, Wash. D.C., 1993).

[16] Warnow, T., Mathematical approaches to comparative linguistics. *Proc. Natl. Acad. Sci. U.S.A.* 94, 6586–6590(1998)

[17] Website for chain letters and the distance matrix used in this paper: `http://www.math.uwaterloo.ca/~mli/chain.html`, 1999.