OPEN ACCESS Freely available online

PLoS BIOLOGY

Essay

Phylomemetics—Evolutionary Analysis beyond the Gene

Christopher J. Howe*, Heather F. Windram

Department of Biochemistry, University of Cambridge, Cambridge, United Kingdom

Abstract: Genes are propagated by error-prone copying, and the resulting variation provides the basis for phylogenetic reconstruction of evolutionary relationships. Horizontal gene transfer may be superimposed on a tree-like evolutionary pattern, with some relationships better depicted as networks. The copying of manuscripts by scribes is very similar to the replication of genes, and phylogenetic inference programs can be used directly for reconstructing the copying history of different versions of a manuscript text. Phylogenetic methods have also been used for some time to analyse the evolution of languages and the development of physical cultural artefacts. These studies can help to answer a range of anthropological questions. We propose the adoption of the term "phylomemetics" for phylogenetic analysis of reproducing non-genetic elements.

Darwin (1809–1882) saw evolution resulting in species being related in a way that could be depicted as a tree. He famously included such a tree as the only figure in *On the Origin of Species by Means of Natural Selection*. However, he was not the first to suggest that species were not immutable, or to depict their relationships in one of a number of possible tree-like ways [1]. Lamarck (1744–1829), for example, had done both of those, and scholars in several other disciplines used trees to represent the relationships among the objects of their study [2]. People studying manuscript texts used the changes incorporated (accidentally or deliberately) when the texts were copied to determine the copying history of extant versions. Those copied from the same earlier version would share variants present in that earlier version, and the copying history was often depicted as a tree. The first recorded example of such a tree (termed a "stemma"—plural stemmata—by manuscript scholars) was probably the one published by Collins and Schlyter in 1827 showing the relationships between a group of medieval Swedish legal texts (reviewed in [2]), and Karl Lachmann (1793–1851) developed principles for the categorisation of errors for this kind of analysis. August Schleicher (1821–1868) published trees of languages from the 1850s onwards. Although there is no evidence that he communicated directly with Darwin, his *Die Darwin'sche Theorie und die Sprachwissenschaft*, published in 1863, referred to *The Origin* as an inspiration, and was addressed to Ernst Haeckel (1834–1919) who worked at Jena, like Schleicher, and was one of the leading proponents of Darwinism in Germany. Schleicher argued that historical linguistic information, such as written texts in Latin, provided a direct demonstration of how languages had developed—something that was not available to the biologist studying the evolution of species. Indeed, the English translation by Bikkers, published in 1869, of his *Darwin'sche Theorie* was called *Darwinism Tested by the Science of Language* [3].

Just over a hundred years after the publication of *The Origin*, in the early 1960s, computer-based methods for reconstructing phylogenetic trees from biological data became available (reviewed in [4]). Numerical taxonomy developed around the same time, and also drew on the increasing availability of computers. Although numerical taxonomy as originally described by Sneath and Sokal did not attempt to draw evolutionary conclusions [5], this followed shortly after [4,6]. The last few years have seen a major expansion in the application of computer-based phylogenetic methods to the study of texts, languages, and other non-genetic datasets. We will give examples of how the methods are applied to such datasets. We argue that the process of replication with the incorporation of changes is a fundamental one in human cultural activity and beyond. Given the use of the word "meme" to refer to a non-genetic principle that behaves in a genetic way [7], we argue for the adoption of the term "phylomemetics" to refer to the phylogenetic analysis of non-genetic data.

Phylogenetic Analysis of Manuscripts

The copying of a manuscript by a scribe with the incorporation of changes that were then propagated when that copy was in turn copied shows clear parallels to the error-prone replication of DNA. Inspired by the development of numerical taxonomy, many scholars started to attempt to apply its methods to questions of classification in the humanities [8]. So, for example, Griffith applied the principles to, among others, the works of Juvenal and Gospel manuscripts [9,10]. Platnick and Cameron [11] discussed the similarities between cladistics (the basis of parsimony analysis), and the evolution of texts and languages. In the 1980s, Lee applied cladistic software (MacClade and PHYLIP) to St Augustine's *Quaestiones in Heptateuchum* [12]. Robinson and O'Hara used PAUP in the early 1990s for an analysis of the Old Norse narrative, Svipdagsmal [13]. This demonstrated a very good agreement between a stemma produced by parsimony and one produced by traditional means including, unusually, scribal documentation. The parsimony approach was then applied to parts of Chaucer's *Canterbury Tales* [14] and in 1998, Barbrook et al. used a phylogenetic

Essays articulate a specific perspective on a topic of broad interest to scientists.

Citation: Howe CJ, Windram HF (2011) Phylomemetics—Evolutionary Analysis beyond the Gene. PLoS Biol 9(5): e1001069. doi:10.1371/journal.pbio.1001069

Published May 31, 2011

Funding: The work has been supported by the Leverhulme Trust (http://www.leverhulme.ac.uk/ Grant number F/09 641/G) and is currently supported by the Isaac Newton Trust, University of Cambridge (http://www.newtontrust.cam.ac.uk/). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: ch26@mole.bio.cam.ac.uk

network method, Split Decomposition, in an analysis of the Prologue to *The Wife of Bath's Tale* [15]. This also showed good agreement between a stemma produced by phylogenetic analysis and one derived by conventional means. The approach for applying phylogenetic methods to texts is simple in principle (Figures 1 and 2). The texts are aligned and then encoded as a string of characters, usually with each character corresponding to a word. The character strings are then used to build a file in exactly the same format as used by phylogenetic tree-building programs, and the file is submitted to the same programs, unaltered. The method has been used to build stemmata for a large number of sets of manuscripts including, in addition to those already mentioned, the Lanseloet van Denemerken story [16], the medieval German legend Parzival [17], parts of the New Testament [18], treatises on the use of the astrolabe [19], writings of St Gregory of Nazianzus [20], historical poems on the Kings of England [21], Dante's *Monarchia* [22], the Mahabharata [23], and the Finnish legend of St. Henry [24]. In general, the conclusions drawn using phylogenetic programs are in agreement with those from conventional scholarship. The method has also been tested using "artificial" traditions, in which volunteers copy a section of text in a predetermined copying history that is then analysed "blind." Again, the results are generally in agreement with the known copying history [24–26].

The use of phylogenetic computer programs in textual analysis has not been without its critics (e.g., [27]). One of the objections often made to the approach is that it does not deal adequately with what scholars call "contamination." This is where a scribe used more than one copy of a text when making his or her own. Broadly, contamination falls into two varieties. In one, a scribe switched from one copy to another at a particular point. In the other, the scribe used multiple copies simultaneously, to make a patchwork. Contamination has clear parallels in biology, where horizontal gene transfer can result in the incorporation into one organism's genome of a gene from distantly related organisms, or where recombination leads to a sequence that is a hybrid between two parental forms. It is still possible to use phylogenetic analyses with these sets of manuscripts. One approach is to infer trees using subsections of the text and look for individual manuscripts whose position in the tree changes according to the subsection studied [28]. In cases where a scribe switched at a reasonably well-defined point, a method developed by Maynard Smith for mapping recombination sites at the se-

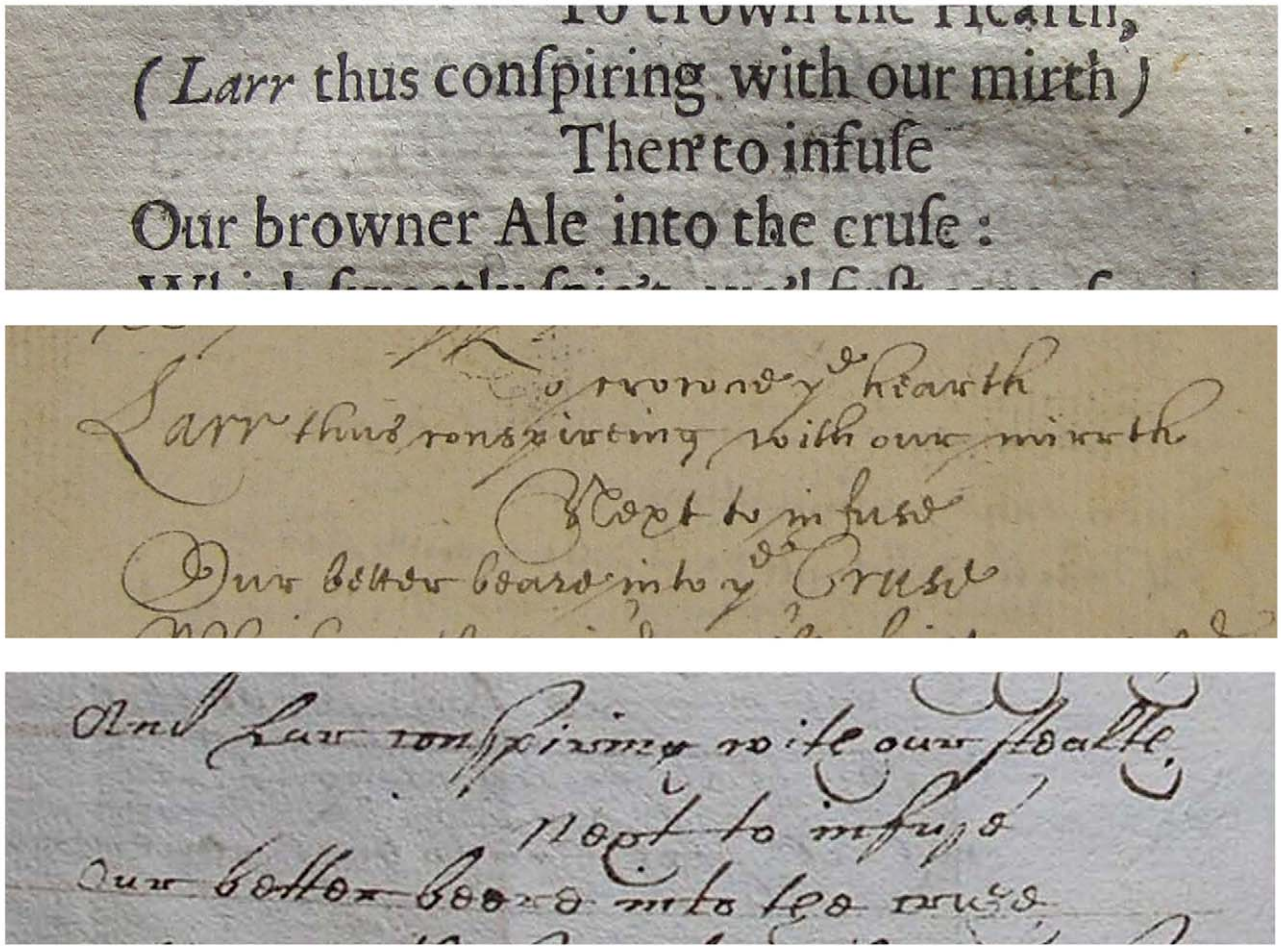

Figure 1. Extracts from the poem "His Age" by Robert Herrick. Figure 2 uses this piece of text as an example of the alignment process. Top panel (Hes in Figure 2) is a printed version from Hesperides, published in 1648 (copy owned by Professor Tom Cain). Middle panel (Ros in Figure 2) is from the Poetical Manuscript Commonplace Book MS 239/23, Rosenbach Museum & Library, Philadelphia. The bottom panel (SJC in Figure 2) is from a verse miscellany, MS S.23, by permission of the Master and Fellows of St John's College, Cambridge.
doi:10.1371/journal.pbio.1001069.g001

A set of texts is aligned so that that any given column contains the equivalent word from all the texts. An example is given below for a short sequence of words from the poem "His Age"by Robert Herrick (1591-1674) in eleven different texts (denoted by three-letter abbreviations). Examples of three of the texts are given in Figure 1. Aligning the texts is equivalent to aligning biological sequence data so that each column in an alignment contains the homologous amino acid or nucleotide residue, and a method for textual alignment has been developed based explicitly on alignment of biological data [44].

```
BLE thus conspiringe with our mirth   Next to infuse Our better  beere
BLH thus conspiring  with our mirth   next to infuse our better  beere
BoE thus conspiringe wth  our mirth   next to infuse Our better  beare
BoF thus conspiringe with our mirth   Next to infuse Our better  beere
Gre ~    ~           ~    ~   ~       ~    ~  ~      ~   ~       ~
Har thus conspiring  with our mirth   Next to infuse Our better  beere
Hes thus conspiring  with our mirth)  Then to infuse Our browner Ale
Hun this conspiring  with our mirth   next to infuse our better  beare
Osb ~    conspiring  wth  our stealth Next to enfuse Our better  beare
Ros thus conspiring  with our mirth   Next to infuse Our better  beare
SJC ~    conspiringe with our stealth Next to infuse Our better  beere
```

For many texts, spelling variants are not considered to be informative. In which case, the text may be "regularized" by correcting such variants to read the same. Punctuation may similarly be removed.

```
BLE thus conspiring with our mirth   next to  infuse our better  beere
BLH thus conspiring with our mirth   next to  infuse our better  beere
BoE thus conspiring with our mirth   next to  infuse our better  beere
BoF thus conspiring with our mirth   next to  infuse our better  beere
Gre ~    ~          ~    ~   ~       ~    ~   ~      ~   ~       ~
Har thus conspiring with our mirth   next to  infuse our better  beere
Hes thus conspiring with our mirth   Then to  infuse our browner Ale
Hun this conspiring with our mirth   next to  infuse our better  beere
Osb ~    conspiring with our stealth next to  infuse our better  beere
Ros thus conspiring with our mirth   next to  infuse our better  beere
SJC ~    conspiring with our stealth next to  infuse our better  beere
```

Finally, each aligned and regularized text is converted to a string of characters, with each text represented as a row of characters, with each character usually corresponding to a word in the text. Sometimes a character corresponds to a more complex feature, such as a rearrangement of word order in a sentence. Strings of characters are then converted to a NEXUS file, the format widely used by phylogenetic programs. In this file ? represents missing data, - is a deletion and 0 or 1 are different character states. This file can then be used directly for phylogenetic analysis.

```
BLE  00000000000
BLH  00000000000
BoE  00000000000
BoF  00000000000
Gre  ???????????
Har  00000000000
Hes  00000100011
Hun  10000000000
Osb  -0001000000
Ros  00000000000
SJC  -0001000000
```

Figure 2. Phylogenetic analysis of texts.
doi:10.1371/journal.pbio.1001069.g002

quence level has been used very successfully for mapping the position in a text where a scribe changed copying source [28]. An alternative approach is to use phylogenetic methods such as Neighbornet or Splitstree, which allow reconstruction of phylogenetic networks. This approach may also be helpful when a scribe used multiple versions simultaneously to make his or her own copy.

Phylogenetic analysis of texts offers scholars a tool for rapid and flexible analysis of texts. Once the primary textual data have been encoded and aligned, it allows scholars to answer in seconds questions such as how the copying history of one chapter compares with another. Its success lies in the fact that copying with incorporation of heritable changes, together with a degree of horizontal transfer, is a reasonable model for the development of manuscripts. But other things evolve in a similar way.

Phylogenetic Analysis of Languages

Just as 18th century scholars depicted the relationships among languages (as well as the relationships among texts or species) as trees, phylogenetic tree-building programs have also been applied to languages [29,30]. A widely used approach uses "Swadesh" lists, named after the 20th century linguistic scholar Morris Swadesh, that comprise words with a counterpart in essentially all languages. A set of words is picked from the list and examined in the languages under study. A word that is essentially the same in two languages is counted as conserved. Other words are counted as a substitution. So, for example, "water" in English and "Wasser" in German would be counted as conserved; "eau" in French would be counted as a difference. Datasets built up in this way can then be analysed with the usual phylogenetic inference programs. As well as providing information on tree topology, i.e. which languages form groups to the exclusion of others, these studies often lead to more quantitative conclusions. Just as biological data are sometimes assumed to be evolving in a clock-like fashion, allowing evolutionary divergence times to be estimated, time-calibration of linguistic trees using known divergence times of different languages also allows inferences to be made about, for example, rates of substitution of words [31]. Time calibration of selected points on a tree can also be used to infer dates of important linguistic and anthropological developments, such as the origins of particular languages and timings of population movements [32,33]. Although some of these inferences with regard to dates are controversial, the same is often true with sequence data [34]. And just as biological data show horizontal gene transfer and texts show contamination, the same is true for linguistic data, which can show "borrowing" or transfer of words between different languages.

Phylogenetic Analysis of Cultural Artefacts

A number of studies have applied phylogenetic analysis to physical cultural artefacts as well as to languages ([35] and references therein). A challenge here has been to find appropriate ways of coding the features of the artefacts in a way that is appropriate for phylogenetic analysis. Often, an important question has been to determine how well characters can be described by a tree-like evolutionary pattern, or whether other patterns are more appropriate, indicating transfer among different cultural groups. Tëmkin and Eldredge analysed the evolution of two musical instruments, the Baltic psaltery and the cornet [36]. For the Baltic

psaltery they included characters such as the presence or absence of a hand-hole, the nature of the ornamentation and the shape of the sound-hole. They recovered a topology that had Slavic and Finnic psalteries as sister groups, with Baltic ones (Latvian and Lithuanian) as a basal group. Given that Slavic and Baltic languages had previously been shown to be sister groups, Tëmkin and Eldredge interpreted this as indicating that the practices underlying instrument building followed geographical rather than linguistic proximity, although the fact that a number of characters showed a distribution that was not congruent with the overall tree indicated examples of convergent evolution or cultural exchange. Analysis of the cornet, by contrast, was much more complex. There was a high degree of reticulation, with fusion of some branches of the tree to form a network, and reconstruction of an unambiguous topology was possible only with the incorporation of historical information. This indicated a large amount of interaction among different instrument builders.

Tehrani and Collard used the degree of reticulation as a measure of cultural contact in elegant analyses of the design and construction of textiles produced in Iran and neighbouring regions [37,38]. They aimed to test whether these features were passed in a linear way from one generation to the next, or whether there was significant influence, commercial or military, from other sources. They encoded a large number of features, including aspects of the methods used for weaving, and elements of the design such as the use of particular geometric borders, birds, stars, and trees, and assessed the overall quality of fit of the data to a maximum parsimony tree by calculating the retention index (which gives an indication of the number of homoplastic or convergent changes across the tree). They also tested if particular character types (such as technical features of production) gave stronger support for groupings within the tree than other character types (such as motifs in the design). The overall fit to a tree was found to be good, and different character types gave similarly strong support, consistent with the proposal that there was little exchange of these cultural characteristics among tribes.

General Conclusions

In addition to those described here, there are many other examples of application of phylogenetic analysis to non-genetic data with the aim of recovering evolutionary history. They include studies of written scripts [39] and physical artefacts, such as arrowheads and pottery designs [40,41], animal behaviour [42], and human organizations and manufacturing structures [43]. In principle, phylogenetic methods can be applied to model the history of any system in which (i) elements can be replicated with the incorporation of changes and (ii) any change between a progeny element and its parent is stably transmitted in subsequent generations. A degree of "horizontal" transfer among elements and/or convergent changes in different lineages may also take place. Horizontal transfer and convergent changes may be recognized by a poor fit between the data and the preferred recovered tree, and can in principle be modelled using network methods of phylogenetic reconstruction. Given the use of the term "meme" to describe reproducing non-genetic elements [7], and units of cultural transmission in particular, we believe the term "phylomemetics" is an appropriate one to refer to phylogenetic analysis of objects other than genes (and their direct products). A search of the web showed occasional uses of this term (e.g., http://papers.ssrn.com/sol3/papers.cfm?abstract_id=1481394), although it did not appear in a search of ISI Web of Knowledge. We believe that it should be formally recognized to refer to this rapidly expanding field.

Acknowledgments

We thank Adrian Barbrook, Barbara Bordalejo, Matthew Spencer, and Daniel Apollon for helpful discussions, and Ruth Connolly, Tom Cain, Karen Schoenewaldt, and Kathryn McKee for help with obtaining the images.

References

1. Penny D, Hendy MD, Poole AM (2003) Testing fundamental evolutionary hypotheses. J Theoret Biol 223: 377–385.
2. van Wyhe J (2005) The descent of words: Evolutionary thinking 1780–1880. Endeavour 29: 94–100.
3. Schleicher A, Bikkers AVW transl (1869) Darwinism tested by the science of language. London: J C Hotten.
4. Edwards AWF (2009) Statistical methods for evolutionary trees. Genetics 183: 5–12.
5. Sneath PHA, Sokal RR (1962) Numerical taxonomy. Nature 193: 855–860.
6. Camin JH, Sokal RR (1965) A method for deducing branching sequences in phylogeny. Evolution 1: 311–326.
7. Dawkins R (1976) The selfish gene. Oxford: Oxford University Press.
8. Sneath PHA, Sokal RR (1973) Numerical taxonomy. San Francisco: W H Freeman.
9. Griffith JG (1968) A taxonomic study of the manuscript tradition of Juvenal. Museum Helveticum 25: 101–138.
10. Griffith JG (1969) Numerical taxonomy and some primary manuscripts of the Gospels. J Theol Stud 20: 389–406.
11. Platnick NI, Cameron HD (1977) Cladistic methods in textual, linguistic and phylogenetic analysis. Syst Zool 26: 380–385.
12. Lee AR (1989) Numerical taxonomy revisited: John Griffith, cladistic analysis and St. Augustine's Quaestiones in Heptateuchum. Studia Patristica 20: 24–32.
13. Robinson PMW, O'Hara RJ (1996) Cladistic analysis of an old Norse manuscript tradition. Res Hum Comput 4: 115–137.
14. O'Hara R, Robinson P (1993) Computer-assisted methods of stemmatic analysis. Canterbury Tales Project Occasional Papers I: 53–74.
15. Barbrook AC, Howe CJ, Blake N, Robinson P (1998) The phylogeny of the Canterbury Tales. Nature 394: 839.
16. Salemans BJP (2000) Building stemmas with the computer in a cladistic, neo-Lachmannian way [PhD thesis]. University of Nijmegen.
17. Stolz M (2003) New Philology and New Phylogeny: Aspects of a Critical Electronic Edition of Wolfram's Parzival. Literary and Linguistic Computing 18: 139–150.
18. Spencer M, Wachtel K, Howe CJ (2004) Representing multiple pathways of textual flow in the Greek manuscripts of the Letter of James using reduced median networks. Comput Hum 38: 1–14.
19. Eagleton C, Spencer M (2006) Copying and conflation in Geoffrey Chaucer's treatise on the astrolabe: a stemmatic analysis using phylogenetic software. Stud Hist Philos Sci 37: 237–268.
20. Macé C, Baret P, Lantin A (2004) Philologie et phylogénétique: regards croisés en vue d'une édition critique d'une homélie de Grégoire de Nazianze. In: Digital technology, philological disciplines. Bozzi A, Cignoni L, Lebrave J-L, eds. Pisa: Istituti editoriali e poligrafici internazionali. pp 305–341.
21. Mooney LR, Barbrook AC, Howe CJ, Spencer M (2001) Stemmatic analysis of Lydgate's *Kings of England*: a test case for the application of software developed for evolutionary biology to manuscript stemmatics. Revue d'Histoire des Textes 31: 275–297.
22. Windram H, Shaw P, Robinson P, Howe CJ (2008) Dante's *Monarchia* as a test case for the use of phylogenetic methods in stemmatic analysis. Literary and Linguistic Computing 23: 443–463.
23. Phillips-Rodriguez WJ, Howe CJ, Windram HF (2010) Some considerations about bifurcation in diagrams representing the written tradition of the Mahabharata. Vienna Journal of South Asian Studies 52–53: 29–43.
24. Roos T, Heikkilä T (2009) Evaluating methods for computer-assisted stemmatology using artificial benchmark datasets. Literary and Linguistic Computing 24: 417–433.
25. Spencer M, Davidson EA, Barbrook AC, Howe CJ (2004) Phylogenetics of artificial manuscripts. J Theoret Biol 227: 503–511.
26. Macé C, Baret P, Robinson P (2006) Testing methods on an artificially created textual tradition. Linguistica Computazionale 24-25: 255–283.
27. Hanna R (2000) The application of thought to textual criticism in all modes - with apologies to A. E. Housman. Studies in Bibliography 53: 163–172.
28. Windram HF, Howe CJ, Spencer M (2005) The identification of exemplar change in the Wife of Bath's Prologue using the maximum chi-squared method. Literary and Linguistic Computing 20: 189–204.

29. Forster P, Renfrew C, eds (2006) Phylogenetic methods and the prehistory of language. Cambridge: McDonald Institute.
30. Atkinson QD, Gray RD (2005) Curious parallels and curious connections - phylogenetic thinking in biology and historical linguistics. Syst Biol 54: 513–526.
31. Pagel M (2009) Human language as a culturally transmitted replicator. Nat Rev Genet 10: 405–415.
32. Gray RD, Atkinson QD (2003) Language-tree divergence times support the Anatolian theory of Indo-European origin. Nature 426: 435–439.
33. Gray RD, Drummond AJ, Greenhill SJ (2009) Language phylogenies reveal expansion pulses and pauses in Pacific settlement. Science 323: 479–483.
34. Graur D, Martin W (2004) Reading the entrails of chickens: molecular timescales of evolution and the illusion of precision. Trends Genet 20: 80–86.
35. Steele J, Jordan P, Cochrane E (2010) Evolutionary approaches to cultural and linguistic diversity. Philos Trans R Soc Lond B Biol Sci 365: 3781–3785.
36. Tëmkin I, Eldredge N (2007) Phylogenetics and material cultural evolution. Curr Anthropol 48: 146–153.
37. Tehrani J, Collard M (2002) Investigating cultural evolution through biological phylogenetic analysis of Turkmen textiles. J Anthropol Archaeol 21: 443–463.
38. Tehrani JJ, Collard M (2009) On the relationship between interindividual cultural transmission and population-level cultural diversity: a case study of weaving in Iranian tribal populations. Evol Hum Behav 30: 286–300.
39. Skelton C (2008) Methods of using phylogenetic systematics to reconstruct the history of the Linear B script. Archaeometry 50: 158–176.
40. Mesoudi A, Whiten A, Laland KN (2006) Towards a unified science of cultural evolution. Behav Brain Sci 29: 329–383.
41. Shennan S (2008) Evolution in archeology. Annu Rev Anthropol 37: 75–91.
42. Lycett SJ, Collard M, McGrew WC (2007) Phylogenetic analyses of behavior support existence of culture among wild chimpanzees. Proc Natl Acad Sci U S A 104: 17588–17592.
43. Baldwin JS Allen PM, Winder B, Ridgway K (2005) Modelling manufacturing evolution: thoughts on sustainable industrial development. Journal of Cleaner Production 13: 887–902.
44. Spencer M, Howe CJ (2004) Collating texts using progressive multiple alignment. Comput Hum 38: 253–270.