CAES Vol. 1, № 4 (December 2015)

A bioinformatic perspective on linguistic relatedness

Simon Brown

PhD in biochemistry, Deviot Institute; Deviot, Tasmania, Australia;
e-mail: Simon.Brown@deviotinstitute.org

Abstract

It is usually assumed that all languages are ultimately derived from the same proto-language. If this is the case then all languages are related, however distantly. However, relatedness is only defined with reference to unrelatedness, so it must be possible for languages to be unrelated. This is reminiscent of the search for genes 'missing' from genomes using sequence analysis, which is based on measurements that are related to lexical distance. The 'relatedness' of gene sequences can only be established probabilistically because relatedness lies on a continuum that ranges from 'identical' to 'completely different'. This is also true of languages irrespective of the basis or bases of the distance measurement.

Key words: gene; genome; language; relatedness; uncertainty

Introduction

Recently, Akulov (2015a) considered the 'unrelatedness' of languages. He argued that there is an inconsistency inherent in the usual view: languages from a common stock are related, but those not in a common stock are still related because all languages are considered to be derived from the same proto-language, although this has yet to be demonstrated. If this is the case, then all languages are necessarily related. He argued from a set theoretic perspective that unrelatedness is not just abstract, it is intrinsic to relatedness.

Such issues are not unique to linguistics: analogous questions arise in the analysis of genes and genomes (the collection of genes in an organism). The connection between language distance and genetic distance has previously been considered by Harding and Sokal (1988) who related genetic distance to geographical distance between "... subjectively chosen centers of language-family regions" as a proxy for linguistic distance, despite the problems of such an approach. However, much has been learnt about both genomics and linguistic distance since then.

Linguistic and genetic distance

While no single measure is entirely satisfactory, language relatedness has been assessed on the basis of phonology (Sanders & Chin 2009), morphology (Schepens *et al.* 2013), semantics (Cooper 2008), mutual intelligibility (Voegelin & Harris 1951), the time needed to learn a language (Chiswick & Miller 2005) and much else, but most often lists of cognate words have been used to estimate the lexical distance between languages. For example, one commonly used measure of the distance (D) between words is the number (N) of single character substitutions, deletions and insertions required to convert one word (w_1) into another (w_2) (Levenshtein 1966) which is then expressed relative to the length (L) of the longer word

$$D(w_1, w_2) = \frac{N(w_1, w_2)}{\max(L(w_1, w_2))}, \quad (1)$$

so that D varies between 0 and 1. The distance is usually measured using a list of n words (Petroni & Serva 2010, Pompei *et al.* 2011, Serva & Petroni 2008), so

$$\langle D \rangle = \frac{1}{n} \sum_{k=1}^{n} \frac{N_k(w_1, w_2)}{\max(L_k(w_1, w_2))}, \quad (2)$$

is the average distance and the subscript k simply indicates that function is applied to the kth pair of words. While measures of this sort are very commonly used, it has been argued that lexical distance is at best unreliable (Akulov 2015b, Hoijer 1956, Matisoff 1990) and certainly it can never be better than the data on which it relies. Despite these reservations, I concentrate on lexical distance here because it bears strong similarity to the analysis of gene and protein sequences (Gray & Atkinson 2003, Gray & Jordan 2000).

Genes can also be written as a sequence of characters[1], each of which represents a particular chemical compound (a nucleotide or a base), and the distance between genes is estimated from the differences between the sequences. In this respect a gene (or the protein it might encode) is analogous to a word and the 'spelling' differences between genes (or proteins) are measured in a manner that is comparable[2] to the measurement of distances between languages using (1) (Lipman *et al.* 1989). The analogy can be extended beyond such 'spelling' differences because each gene encodes information (such as, but not limited to, a protein) and genes encoding the equivalent information in different species or individuals need not be the same length. Futhermore, the effect (expression) of a gene is modulated by regulatory elements, located upstream of (before), within and downstream of (after) a gene and elsewhere in the genome, which have natural linguistic analogues: prefixes, infixes, suffixes and other elements. Finally, each gene is located within the context of other genes in the genome, some regions of which are more accessible than others, which also influences the frequency and rate of expression of the gene. The importance of the combination of sequence and regulatory information and the inherent biological function has lead to the use of more than just sequences to analyse the relatedness of systems and species (Guzzi et al. 2011, Pesquita et al. 2009).

An example of a simple analysis of a short protein sequence is shown in Figure 1 (the gene sequences would yield similar results). The sequence of insulin chain B is used because it is short and was the first ever determined (Sanger & Tuppy 1951). In the alignment shown in Figure 1A the human sequence is identical to that in the other three species in 17 of 30 positions (characters on a black background), but it differs to varying extents from the other sequences (measured using N_k) in the other positions as indicated in the lower triangle of Figure 1B. An alternative representation of the relatedness between the sequences in Figure

[1] The alphabet used to write a gene sequence comprises only four letters (A, C, G, T), but genes usually encode proteins that are also represented as a sequence of letters (corresponding to amino acids) and the alphabet in this case is larger (A, C, D, E, F, G, H, I, K, L, M, N, P, Q, R, S, T, V, W, Y). Note that the fact that the former alphabet is a subset of the latter is not significant, the characters represent completely different chemical compounds.

[2] The availability of very large numbers of gene, protein and genome sequences leads naturally to the comparison of more than two sequences at a time and the use of a probabilistic approach that necessitate somewhat different calculations (Carrillo & Lipman 1988, Dayhoff *et al.* 1983) that have been applied to the development of language (Gray & Atkinson 2003, Gray & Jordan 2000). The calculation is based on the idea of a 'cost' that is not increased when two sequences agree at a particular position in the sequence and is increased when an insertion or substitution is required. However, not all substitutions have equal effect: in some cases the chemical change is small, but in others it is much larger and the effect on the cost varies accordingly.

1A is shown in Figure 1C. From this it is apparent that the most distantly related sequences are the ratsnake and the cavy (= 0.161 + 0.039 + 0.176 = 0.376 substitutions per site) and the most closely related sequences are the alligator and the human (= 0.024 + 0.039 + 0.031 = 0.094 substitutions per site). The phylogram obtained from the lexical distances shown in the lower triangle of Figure 1B is similar (Figure 1D).

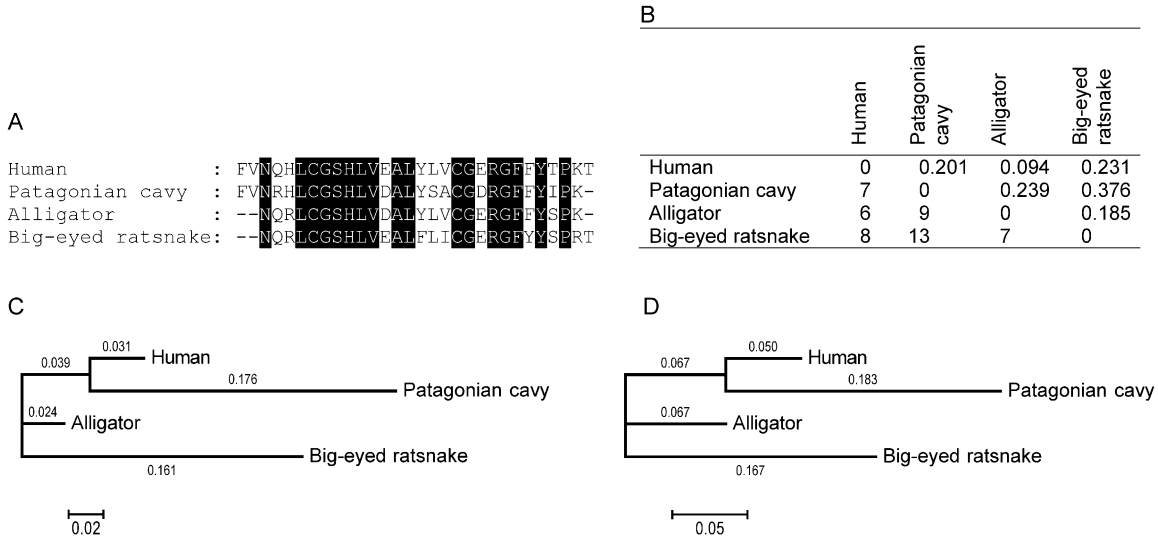

Figure 1. Example of an analysis of protein sequences showing (A) an alignment of four insulin chain B sequences, (B) the pairwise differences between the sequences (lower triangle is N_k and the upper triangle is the total number of substitutions per site derived from the phylogram shown in (C)), (C) the phylogram estimated from the alignment using Mega 5 (Tamura *et al.* 2011) and (D) the approximate phylogram estimated using <*D*> (2). The numbers shown in (C) represent the number of substitutions per site and those in (D) are the average number of insertions, deletions and substitutions per site. The arbitrarily selected sequences of insulin chain B are: human - P01308; Patagonian cavy (*Dolichotis patagonum*) - Q5BVE9; alligator (*Alligator mississippiensis*) - P12703; and big-eyed ratsnake (*Ptyas dhumnades*) - P12708.

Genomics

If a gene is analogous to a word, then a genome is perhaps analogous to an exhaustive and highly structured word list. However, there are three distinct types of genome, distinguished by their location in the nucleus, chloroplasts or mitochondria within a cell. Given this, plants have three genomes, humans have two and bacteria have just one. Each genome has a particular complement of genes, but the distribution of genes between them varies between species. Extending the analogy, this would mean than some species might have a word in one word list but not another while it might be in a completely different list in other species.

The mitochondrial genome is small and relatively simple. For example the human mitochondrial genome is circular and has 16569 nucleotides (Figure 2). It has 13 protein-coding genes and 2 ribosomal RNA genes, as well as several small transfer RNA genes that are not shown in Figure 2.

It is common for the sequence of each of the genes in the mitochondrial genome to differ between individuals as is indicated by the number of changes identified in Figure 2. Fortunately, such sequence changes are often benign, but in some cases the effect is devastating. It follows that those who have large-scale deletions (about 30% of the genome) have a poor prognosis. In comparison, the entire mitochondrial genome is often lost by yeast, which are then smaller and grow more slowly, but still survive (Contamine & Picard 2000). Similarly, bacteria function 'normally' without genes that are essential to other species (Cordwell 1999).

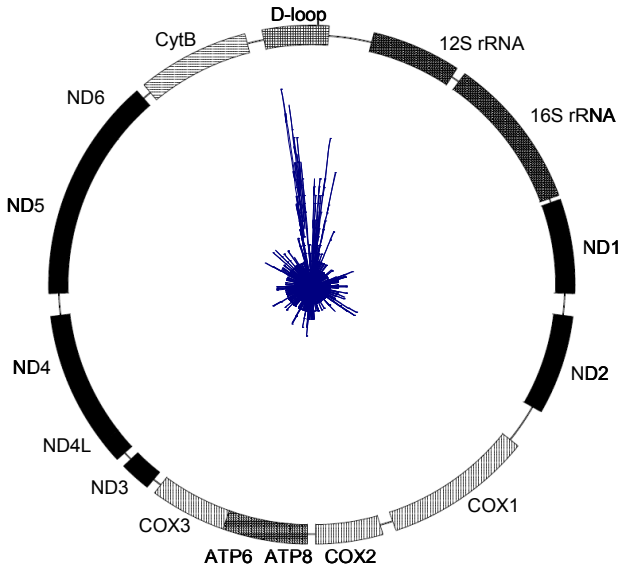

Figure 2. A map of the human mitochondrial genome and the distribution of identified single base changes. The length of each of the radiating lines represents the number of changes reported at the indicated position in the genome. The map of the genome (J01415) indicates the positions of the the D-loop (a control region) and most of the genes (the small transfer RNA genes have been omitted for clarity).

Not only are there 'missing' genes, there are also many genes that are not "understood". While a gene has specific features that facilitate its identification, the initial assignment of a likely function is based on the similarity of its sequence to those of other genes. If there are no similar sequences, no suggestion can be made about the probable function, and the same result is reached if no function has been assigned to any of those sequences to which a gene is similar. This is not unusual. For example, *Methanobacterium thermoautotrophicum* is a bacterium that has 1858 genes, although fewer were identified initially. Of these, 844 were assigned functions based on sequence analysis, 514 could be related to sequences with no known function and 496 had little or no similarity to any other sequence (Smith *et al.* 1997). In effect, less than half of the genes could be assigned a function based on sequence analysis. In this case, the word list may have been developed, but there is no translation in a known language for many of the words and others are merely lexically similar to incomprehensible words in other languages.

One strategy adopted to overcome this has been to determine the three-dimensional structure of proteins that are 'not understood' in the hope that this might give some idea about function. This is predicated on two observations: (i) proteins of similar function tend to have similar

structures and (ii) protein structure is not entirely apparent from the sequence. Christendat *et al.* (2002) applied this strategy to a protein known as MTH1491 from *M. thermoautotrophicum*. The structure of the protein that best matched it is shown in Figure 3A and that of MTH1491 is shown in Figure 3B. As MTH1491 is only about one third the size of the matching protein, the extraneous part of the former was removed and the two structures were superimposed (Figure 3C). There may be some structural similarity between them, but it is limited (compare Figures 3B and 3C), which is consistent with the significant lexical distance between the matching sequences ($D = 0.93$). Even if one ignores this, it does prompt contemplation of the significance of the two thirds of the matched protein that was removed. When the same method was applied to another protein from *M. thermoautotrophicum* (MTH1020) a clear structural match was identified, but MTH1020 corresponded to only $1/15^{th}$ of the matching protein and the two proteins did not have the same function (Saradakis *et al.* 2002).

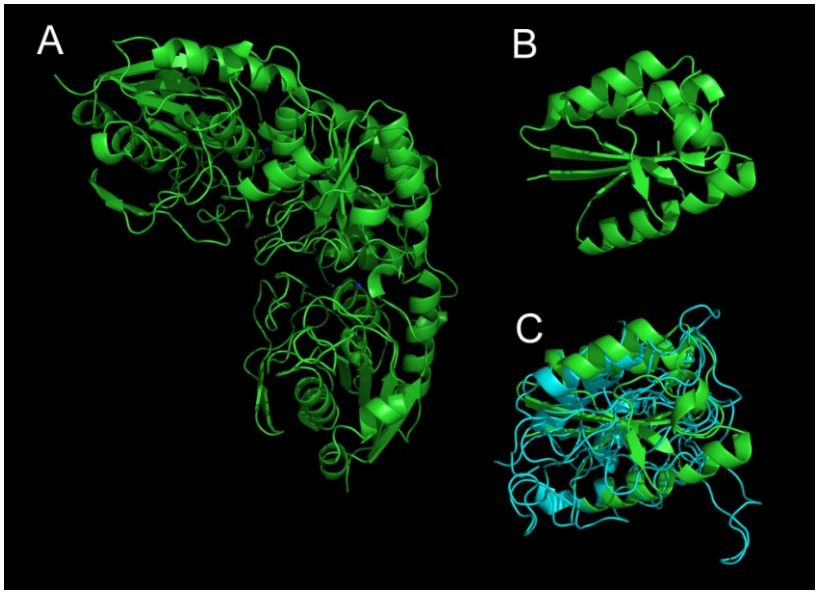

Figure 3. A representation of the structures of (A) the best structural match identified by Christendat *et al.* (2002) and (B) MTH1491. In (C) the two structures are aligned and superimposed after removing the extraneous part of the match. Notice that the background is clearly visible through the central region of MTH1491 (B), but this is obscured by the superimposed match in (C). If the structures matched perfectly there would be very little difference between (B) and (C) in the extent to which the background is visible in this region. Neglecting the non-matching region, $D = 0.93$, the sequences were identical in 7% of sites and similar in 18% of sites, and the average physical distance (rmsd) between the structures in (C) is 4.9 Å over 111 amino acids (although Christendat *et al.* (2002) report a slightly better rmsd of 3.7 Å for the match over only 89 amino acids). The image was created using PyMOL [http://sourceforge.net/projects/pymol/].

The number of genomes, genes and protein structures available for study increases very rapidly, so there is good reason to expect that the function of MTH1491 will eventually be determined. It is unsurprising, therefore, that the analysis carried out by Christendat *et al.* (2002) using the structure of MTH1491 no longer identifies the protein shown in Figure 3A as the best structural match available. That protein now ranks at number 51 in the list, but the

top 50 possibilities are proteins that have no known function. The dangers associated with the use of lexical similarity are well known, but new issues inevitably arise as analyses are developed and applied.

Implications for the detection of unrelatedness

The foregoing raises several points that are relevant to the relatedness of both languages and species. First, there is a continuum of relatedness between species, as there is between languages: at one end sequences may be identical ($D = 0$) and at the other a gene may be absent from one species ($D = 1$) or they might be entirely different ($D = 1$), despite having identical biological effects when expressed. The inevitable inference is that there must be uncertainty in the strength of 'relatedness'. But if many genes are considered, as is the case for words in measuring lexical distance, it is unlikely that $D = 1$ for every gene and so $\langle D \rangle < 1$. From a statistical perspective, there is some critical value of $\langle D \rangle$ (D_{crit}) above which the strength of the relatedness is so low, that species may be regarded as being unrelated rather than distantly related. The size of D_{crit} is yet to be determined.

Second, the mere presence of a recognisable genome is not necessarily sufficient to infer that the genome is related to those of all organisms. Viruses, for example, have genes in a genome, but are not usually considered to be 'alive', although this is under discussion (Forterre 2010), nor do viruses have a place on Woese's (1990) 'tree of life', although they can be related to one another. Recently, even this long-standing model has been questioned based on sequence analysis (Nasir & Caetano-Anollés 2015, Nasir *et al.* 2012). Whether this is just an example of Matisoff's (1990) *columbicubiculomania*[3] is to be seen. Similarly, there are language isolates, some of which may once have had relatives, others that might yet be shown to have relatives and relatives may never be identified for a third group (Campbell 2010). Of course, the third category of isolates may simply reflect a lack of sufficient data, but it is also conceivable that they represent 'unrelated' languages.

Third, how it is possible to be certain that a gene is really 'missing' (Cordwell 1999). The identification of a (possibly lengthy) sequence of characters in a genome that can be very large is now a relatively simple task. However, the problem is greatly complicated by the facts that (i) the match need not be perfect, (ii) the length may not be the same as the model provided because there may be insertions in one sequence relative to another and (iii) there may be quite different sequences that encode functionally equivalent proteins. The software currently used to carry out such searches is focussed on locating matches (and providing statistical estimates of their significance) rather than identifying the absent. The identification of a 'missing' sequence of any length is rather like confirming that a specific grain is not present in a handful of sand when you are not entirely certain what the grain in question looks like.

A single sequence change (a 'spelling mistake'), the loss of a single gene (loss of a word from the word list) and the loss of the entire mitochondrial genome (the loss of a [small] volume from the word list) do not lead to the conclusion that a new species has been generated. Correspondingly, differences in spelling (such as 'color/colour' or 'tire/tyre' in American and British English) do not prompt the conclusion that different languages are involved. Nor does the absence of particular words (the Faroese *starilsi* 'the act of standing and staring out into the blue' has no equivalent in English) or even several (the many versions of the definite

[3] His definition is 'a compulsion to stick things into pigeonholes, to leave nothing unclassified'.

article in Icelandic have not even one equivalent in Russian) cast significant doubt that these languages are related (Gray & Atkinson 2003, Petroni & Serva 2010, Serva & Petroni 2008).

On the other hand, it is likely that natural languages might be completely unrelated to constructs such as computer programming 'languages' and mathematics. The latter is often said to be the 'language of science', but, while it may have some of the properties, it is not a language (Ford & Peat 1988, Peat 1990). Bertrand Russell, for one, saw a difference:

> Ordinary language is totally unsuited for expressing what physics really asserts, since the words of everyday life are not sufficiently abstract. Only mathematics and mathematical logic can say as little as the physicist means to say. As soon as he translates his symbols into words, he inevitably says something much too concrete, and gives his readers a cheerful impression of something imaginable and intelligible, which is much more pleasant and everyday than what he is trying to convey. (Russell 1954: 85)

Similarly, no programming language is really a language, no matter how much poetry might be written in Perl. In each case, there is no obvious basis on which to calculate a lexical distance, for example, they do not possess enough of the characteristics of the natural languages with which we are familiar. A similar problem might arise when extraterrestrial life is encountered.

Conclusions

Relatedness forms a continuum from completely different to identical. At one end two items (words or genes) have nothing in common ($D = 1$) and at the other there is no difference ($D = 0$). Because of the inevitable variability between individuals or locations, it is unlikely that many items will all be identical and it takes only one pair to have a single difference to yield $\langle D \rangle > 0$. The same argument indicates that it is rare to observe $\langle D \rangle = 1$, but this does not preclude the possibility that languages or species are unrelated. Somewhere on the continuum the probability of 'unrelatedness' becomes overwhelming, but the measures used to estimate D must be reliable.

References

Akulov A. 2015a. Whether it is possible to prove genetic unrelatedness of certain languages? *Cultural Anthropology and Ethnosemiotics*, vol. 1 no. 3; pp.: 2 – 4

Akulov A. 2015b. Why conclusions about genetic affiliation of certain language should be based on comparison of grammar but not on comparison of lexis? *Cultural Anthropology and Ethnosemiotics*, vol. 1 no. 3; pp.: 5 – 9

Campbell L. (2010). Language isolates and their history, or, what's weird, anyway? In *36th Annual Meeting of the Berkeley Linguistics Society*, in press, Berkeley

Carrillo H., Lipman D. 1988. The multiple sequence alignment problem in biology. *SIAM Review*, vol. 48 no. 5; pp.: 1073 – 1082

Chiswick B. R., Miller P. W. 2005. Linguistic distance: a quantitative measure of distance between English and other languages. *Journal of Multilinguial and Multicultural Development*, vol. 26 no. 1; pp.: 1 – 11

Christendat D., Saridakis V., Kim Y., Kumar P. A., Xu X., Semesi A., Joachimiak A., Arrowsmith C. H., Edwards A. M. 2002. The crystal structure of hypothetical protein MTH1491 from *Methanobacterium thermoautotrophicum*. *Protein Science*, vol. 11; pp.: 1409 – 1414

Contamine V., Picard M. 2000. Maintenance and integrity of the mitochondrial genome: a plethora of nuclear genes in the budding yeast. *Microbiology and Molecular Biology Reviews*, vol. 64 no. 2; pp.: 281 – 315

Cooper M. C. 2008. Measuring semantic distance between languages from a statistical analysis of bilingual dictionaries. *Journal of Quantitative Linguistics*, vol. 15 no. 1; pp.: 1 – 33

Cordwell S. J. 1999. Microbial genomes and "missing" enzymes: redefining biochemical pathways. *Archives of Microbiology*, vol. 172; pp.: 269 – 279

Dayhoff M. O., Barker W. C., Hunt L. T. 1983. Establishing homologies in protein sequences. *Methods in Enzymology*, vol. 91; pp.: 524 – 545

Ford A., Peat F. D. 1988. The role of language in science. *Foundations of Physics*, vol. 18 no. 12; pp.: 1233 – 1242

Forterre P. 2010. Defining life: the virus viewpoint. *Origins of Life and Evolution of Biospheres*, vol. 40 no. 2; pp.: 151 – 160

Gray R. D., Atkinson Q. D. 2003. Language tree divergence times support the Anatolian theory of Indo-European origin. *Nature*, vol. 426; pp.: 435 – 439

Gray R. D., Jordan F. M. 2000. Language trees support the express-train sequence of Austronesian expansion. *Nature*, vol. 405; pp.: 1052 – 1055

Guzzi P. H., Mina M., Guerra C., Cannataro M. 2011. Semantic similarity analysis of protein data: assessment with biological features and issues. *Briefings in Bioinformatics*, vol. 13 no. 5; pp.: 569 – 585

Harding R. M., Sokal R. R. 1988. Classification of European language families by genetic distance. *Proceedings of the National Academy of Sciences of the USA*, vol. 85; pp.: 9370 – 9372

Hoijer H. 1956. Lexicostatistics: a critique. *Language*, vol. 32 no. 1; pp.: 49 – 60

Levenshtein V. I. 1966. Binary codes capable of correcting deletions, insertions, and reversals. *Soviet Physics - Doklady*, vol. 10 no. 8; pp.: 707 – 710

Lipman D. J., Altschul S. F., Kececioglu J. D. 1989. A tool for multiple sequence alignment. *Proceedings of the National Academy of Sciences of the USA*, vol. 86; pp.: 4412 – 4415

Matisoff J. A. 1990. On megalocomparison. *Language*, vol. 66 no. 1; pp.: 106 – 120

Nasir A., Caetano-Anollés G. 2015. A phylogenomic data-driven exploration of viral origins and evolution. *Science Advances*, vol. 2015 no. 1; pp.: e1500527

Nasir A., Kim K. M., Caetano-Anolles G. 2012. Giant viruses coexisted with the cellular ancestors and represent a distinct supergroup along with superkingdoms Archaea, Bacteria and Eukarya. *BMC Evolutionary Biology*, vol. 12; pp.: 156

Peat F. D. 1990. Mathematics and the language of nature, in Mickens R. E. (ed.) *Mathematics and Science*. World Scientific, Singapore; pp.: 154 – 172

Pesquita C., Faria D., Falcão A. O., Lord P., Couto F. M. 2009. Semantic similarity in biomedical ontologies. *PLoS Computational Biology*, vol. 5 no. 7; pp.: e1000443

Petroni F., Serva M. 2010. Measures of lexical distance between languages. *Physica A*, vol. 389; pp.: 2280 – 2283

Pompei S., Loreto V., Tria F. 2011. On the accuracy of language trees. *PLoS ONE*, vol. 6 no. 6; pp.: e20109

Russell B. 1954. *The scientific outlook*. George Allen and Unwin Ltd, London

Sanders N. C., Chin S. B. 2009. Phonological distance measures. *Journal of Quantitative Linguistics*, vol. 16 no. 1; pp.: 96 – 114

Sanger F., Tuppy H. 1951. The amino-acid sequence in the phenylalanyl chain of insulin. 2. The investigation of peptides from enzymic hydrolysates. *Biochemical Journal*, vol. 49; pp.: 481 – 490

Saridakis V., Christendat D., Thygesen A., Arrowsmith C. H., Edwards A. M., Pai E. F. 2002. Crystal structure of *Methanobacterium thermoautotrophicum* conserved protein MTH1020 reveals an NTN-hydrolase fold. *Proteins: Structure, Function, and Genetics*, vol. 48; pp.: 141 – 143

Schepens J., van der Slik F., van Hout R. 2013. Learning complex features: a morphological account of L2 learnability. *Language Dynamics and Change*, vol. 3; pp.: 218 – 244

Serva M., Petroni F. 2008. Indo-European languages tree by Levenshtein distance. *Europhysics Letters*, vol. 81; pp.: 68005

Smith D. R., Doucette-Stamm L. A., Deloughery C., Lee H., Dubois J., Aldredge T., Bashirzadeh R., Blakely D., Cook R., Gilbert K., Harrison D., Hoang L., Keagle P., Lumm W., Pothier B., Qiu D., Spadafora R., Vicaire R., Wang Y., Wierzbowski J., Gibson R., Jiwani N., Caruso A., Bush D., Safer H., Patwell D., Prabhakar S., McDougall S., Shimer G., Goyal A., Pietrokovski S., Church G. M., Daniels C. J., Mao J., Rice P., Nölling J., Reeve J. N. 1997. Complete genome sequence of *Methanobacterium thermoautotrophicum* ΔH: functional analysis and comparative genomics. *Journal of Bacteriology*, vol. 179 no. 22; pp.: 7135 – 7155

Tamura K., Peterson D., Peterson N., Stecher G., Nei M., Kumar S. 2011. MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. *Molecular Biology and Evolution*, vol. 28 no. 10; pp.: 2731 – 2739

Voegelin C. F., Harris Z. S. 1951. Methods for determining intelligibility among dialects of natural languages. *Proceedings of the American Philosophical Society*, vol. 95 no. 3**;** pp.: 322 – 329

Woese C. R., Kandler O., Wheelis M. L. 1990. Towards a natural system of organisms: proposal for the domains archaea, bacteria, and eucarya. *Proceedings of the National Academy of Sciences of the USA*, vol. 87; pp.: 4576 – 4579