

Analyzing Genetic Connections between Languages by Matching Consonant Classes

Peter Turchin,^a Ilia Peiros,^b Murray Gell-Mann^{b,1}

^aDepartment of Ecology and Evolutionary Biology, University of Connecticut, Storrs, Connecticut 06269, USA

^bSanta Fe Institute, 1399 Hyde Park Road, Santa Fe, New Mexico 87501, USA

¹To whom correspondence should be addressed. E-mail: mgm@santafe.edu

Corresponding author: Murray Gell-Mann, Santa Fe Institute, 1399 Hyde Park Road, Santa Fe, New Mexico 87501, USA. Tel. 505-946-2745; fax: 505-982-0565. E-mail: mgm@santafe.edu.

Author contributions: PT, IP, and MG-M designed research; IP provided linguistic data; PT performed statistical analyses; and PT, IP, and MG-M wrote the paper.

On-line Supporting Information:

http://cliodynamics.info/data/ConsClass_SI.doc

<http://cliodynamics.info/data/SuppInfo.xls>

Abstract

The idea that the Turkic, Mongolian, Tungusic, Korean, and Japanese languages are genetically related (the “Altaic hypothesis”) remains controversial within the linguistic community. In an effort to resolve such controversies, we propose a simple approach to analyzing genetic connections between languages. The Consonant Class Matching (CCM) method uses strict phonological identification and permits no changes in meanings. This allows us to estimate the probability that the observed similarities between a pair (or more) of languages occurred by chance alone. The CCM procedure yields reliable statistical inferences about historical connections between languages: it classifies languages correctly for well-known families (Indo-European and Semitic) and does not appear to yield false positives. The quantitative patterns of similarity that we document for languages within the Altaic family are similar to those in the non-controversial Indo-European family. Thus, if the Indo-European family is accepted as real, the same conclusion should also apply to the Altaic family.

Introduction

Tracing “genetic” relationships between languages is sometimes a source of controversy in comparative linguistics. For example, within the linguistic community there is not universal acceptance of the Altaic family, i.e., the idea that the Turkic, Mongolian, Tungusic, Korean, and Japanese languages are genetically related (share a common ancestor) (1). Even the recent publication of the *Etymological Dictionary of the Altaic Languages* (2) did not put an end to this controversy (3, 4). The critics claim that the observed similarities can be due either to chance resemblances or to “areal convergence”—borrowing resulting from cultural contacts (discussion in Ref. (5)).

To demonstrate that languages belong to the same linguistic family it is best to trace them back to their common ancestor (= proto-language of this family), with known sound system, grammar, and partial lexicon. In most cases such proto-languages have to be reconstructed. According to the standard methods of comparative linguistics this can be done only if potentially related languages preserve a sufficient number of proto-language morphemes. Through analysis of such morphemes linguists establish a system of correspondences between the sound systems of daughter languages. For example, many German words beginning in /c/ (z in orthography) have the same meaning as English words beginning in /t/—*Zunge:tongue*, *Zahn:tooth*, etc., while initial German /t/ corresponds to English /d/, as in *trinken:drink* or *trocken:dry*. A set of such observations is used to reconstruct the phonological system of the proto-language and the forms of its individual morphemes (phonological reconstruction). The meanings of the morphemes are reconstructed using much less rigorous methods. One problem here is that there can be a substantial semantic shift between two related words (cognates). An example is English *clean* and German *klein* ‘small’—although these words are known to be cognates (the original meaning was ‘neat, clean’) they now have rather different meanings.

So far, proto-languages of only a limited number of language families have been properly reconstructed, thus demonstrating that the languages forming these families are related. As proto-languages of most proposed families are yet to be reconstructed, linguists still lack convincing evidence on possible relationships between languages. To compensate for the lack of information linguists use a variety of provisional methods ranging from inspection-based judgments to more formalized *lexicostatistics*. The assumption here is that if languages are related they should have lexical morphemes of common origin having identical meanings from the Swadesh 100-item list (6). Since no changes in meanings are accepted, semantic connections between the morphemes are straightforward. Still, phonological identification of relatedness is not based in this case on a system of correspondences* and therefore is not strict enough, with some similarities being possibly due to chance.

Here we propose a procedure based on lexicostatistics that does use strict phonological identification and permits no exceptions. This approach allows us to estimate the probability that the observed similarities between a pair (or more) of languages occurred by chance alone (7, 8). By design the proposed method is “conservative”: we go to great lengths to minimize the possibility of false positives

* Another application of lexicostatistics requires good knowledge of comparative phonology and etymologies and is used to generate linguistic families classifications, based on the amounts of etymologically identical words revealed by each pair of languages studied.

(concluding that languages are related when in fact they are not). Such an approach, which places a heavy burden of proof on anyone favoring a genetic relationship, is far from optimal, but we adopt it to avoid polemical controversies while applying our method to cases such as that of the Altaic family. The method is not a substitute for the more sophisticated approaches of comparative linguistics. Rather, it provides a procedure for testing hypotheses of genetic relationships without relying on matters of choice or judgment.

Methods

Linguistic data

The linguistic data (lexicostatistical lists of individual languages) are taken from a collection of databases prepared by participants of the Evolution of Human Languages Project (Santa Fe Institute, USA) and the Tower of Babel Project (Moscow, Russia). We code each root (= main lexical morpheme) in the 100-word list for each language by replacing its first two consonants with generic consonant classes, following a suggestion put forward by Dolgopolsky (14). Table 1 gives the mapping of consonants to the nine classes. We have performed this procedure for 53 Eurasian and North African languages (see *Supporting Information* for the list of languages).

The measure of similarity between two languages is the proportion of roots of the same meaning whose first two consonant classes match. For example, English *nose* and German *Nase* (both coded NS) are classified as similar while *dog* and *Hund* (TK versus #N) are classified as dissimilar. The German *Zunge*, coded CN, and English *tongue*, coded TN, are also classified as dissimilar, even though they are cognates. Our measure of similarity misses systematic sound correspondences that cut across our consonant classes. (In addition, it omits information contained in vowels and in any consonants other than the first two.)

Statistical Analyses

The next step after determining the proportion of matches between two 100-word lists is to estimate the statistical significance of this result. A naïve approach assumes that the probability of a match between the first consonants or the second ones is one in nine (the number of consonant classes) and the probability of both consonants matching is 9^{-2} or one in eighty-one. With this method we would expect, on average, a bit more than one match ($100/81=1.2$) in a list of 100 words. This approach is, however, flawed in several ways. First, some consonant classes are more common than others, and therefore the random chance of both consonants matching is, on average, greater than 1:81. Second, presence of a certain consonant in one position may affect the probability of finding another consonant in the other position. In other words, the assumption of independence may not be warranted. Finally, we must deal with such irregularities as missing or multiple words in some positions.

We use the bootstrap method (15) to estimate the statistical significance of the observed proportion of matches between word lists of two languages (the Similarity Index, SI). The procedure works as follows. We randomly select a root from List 1 and match it with a random root from List 2 (there are two alternative methods of random selection, see the next paragraph for the explanation). Repeating this step 100 times, we calculate the “bootstrap SI” (the proportion of matches between two random 100-word

lists). Next, we replicate this procedure many times (e.g., 10,000 iterations) and use the 10,000 bootstrap SIs to approximate the probability distribution of the SI under the null hypothesis (that any matches are due to chance). Finally, we determine the proportion of bootstrapped SIs that is equal to or greater than the index calculated for the original lists. This gives us an estimate of the probability of observing this value (or a larger one) under the null hypothesis. The smaller this estimated probability, the greater our degree of belief that the proportion of observed matches could not arise by mere chance.

There are two ways to perform random selection: with or without replacement. In the first case (the classic bootstrap) after a word is chosen from the list, and matched with a word from the other language's list, the word is put back. In other words, the same word can be chosen several times (and, therefore, some other words are never chosen). The alternative procedure (known as the permutation test) is to sample without replacement, so that each word is selected once. We repeated our analyses using both the bootstrap and the permutation test and obtained similar results. However, the permutation test was slightly more permissive (it gave a greater proportion of false positives), and therefore we report only the bootstrap results. We routinely used 10,000 bootstrap iterations to construct the probability distribution of the SI, but in cases where all bootstrapped SI were smaller than the observed one, we reran analysis with 1 million iterations. Thus, $P < 10^{-6}$ means that the observed SI was greater than *all* of 1 million bootstrapped SIs.

Our approach allows for missing words. Thus, the SI is the number of matches divided by the number of possible matches (subtracting observations with missing values). Missing values are handled during the bootstrap in exactly the same manner. That is, a bootstrapped SI may also have a number less than 100 in the denominator, if missing values happened to be chosen during the sampling process.

Results

Testing the Method on the Indo-European and Semitic Families

Before tackling the Altaic family, we test how well this Consonant Class Matching (CCM) method works on the well-studied Indo-European family. We distinguish between using modern languages for this purpose and using attested or reconstructed ancient languages. Applying the procedure to 21 modern Indo-European (IE) languages (additional tables are in *Supporting Information*) we find that it reliably identifies such branches as Indic, Slavic, Germanic, and Romance (SIs varying between 45 and 77%, all statistically significant at $P < 10^{-6}$). By contrast, similarity between languages belonging to different branches is much lower (between 1 and 21%). A particularly interesting comparison is between Germanic and Indic languages (Table 2). The SIs are very low, between 1 and 7%. Half of the comparisons are not significant at the 0.05 level, while all but one of the rest are weakly significant at $0.05 < P < 0.01$.

Both the Indic and the Germanic groups reveal themselves beyond any doubt, while the genetic relation between these two groups is not convincingly demonstrated by Table 2. We recall that the validity of the IE family was originally established not on the basis of modern languages but rather by comparing ancient ones, which are much closer to each other. The results of the CCM method (Table 3a) reflect the greater degree of similarity (all comparisons are significant at least at $P < 0.02$ level, and most at much

higher significance levels). The SI between Old High German and Old Indian, in particular, is 14%. The probability of this overlap happening by chance is vanishingly small ($<10^{-6}$). When we apply the CCM approach to several ancient Semitic languages (Table 3b) we find that SIs for all comparisons are highly significant ($P \ll 10^{-6}$).

The improved resolution obtained with ancient languages is not surprising. The longer the period since the two languages diverged, the more opportunity there has been for roots in the 100-item list to “mutate” and become dissimilar (that is, cross into a different phonetic class) or to be replaced (as a result of a semantic shift). As time passes, the degree of similarity between any two genetically related languages should eventually decline to the point where in direct comparison it is indistinguishable from random noise. However, if we keep applying the procedure of reconstructing proto-languages we may be able to defeat that phenomenon.

The Indo-European and Semitic families are unusual in that they enjoy such a rich abundance of attested ancient languages. Does that mean that we cannot investigate genetic relationships when ancient written sources are lacking? As suggested just above, one possible approach to this problem is to use reconstructed proto-languages. When we apply the CCM method to the proto-languages of four IE branches, we obtain the same pattern as for attested ancient languages (Table 4a). For example, the SI between the Proto-Iranian and the Proto-Germanic languages is 13%. By contrast, in pairwise comparisons between five modern Germanic languages (German, English, Dutch, Icelandic, and Swedish) and two modern Iranian languages (Kurdish and Ossetian) it ranges between 5 and 10% (average = 7%).

Using reconstructed proto-languages can sometimes yield even better results than using attested old languages, as is shown in the Iranian–Germanic comparison. The SIs between Old High German and Avestan or Classical Persian are only 9–10%, whereas the overlap between Proto-Germanic and Proto-Iranian is 13% (and the statistical significance of the result increases by several orders of magnitude). This improvement is at least partially due to the greater age of Proto-Germanic and Proto-Iranian compared with Old High German and Classical Persian respectively.

It should be mentioned, however, that the main issue is not the age of the languages, but the degree to which they resemble their proto-languages. Ancient languages are usually more archaic in this sense, as they retain many features of their proto-languages, both in phonology and lexicon. At the same time some modern languages are also quite archaic, for example, Lithuanian. Therefore the role played by this language in Indo-European studies is similar to that of Ancient Greek, Latin and other ancient languages. In some cases a proto-language can be only a thousand years old, but because of its archaic character its relations with other (proto-)languages can be identified even by the CCM method.

Applying the methodology to the Altaic family

Next, we use the CCM approach to test the reality of the Altaic family. We have four independent reconstructions (2, 9): Proto-Turkic, Proto-Mongolian, Proto-Tungus, and Proto-Japanese (Korean dialects are too similar to one another to justify a reconstruction of Proto-Korean). We also calculated the degree of similarity between these four languages and Proto-Eskimo, because Mudrak (9, 10) proposed that Eskimo languages are closely related to the Altaic family. The SIs for the four Altaic proto-languages (Table

4b) range between 6 and 11% (average = 8.7%). This range of values is lower than that for the IE family. Nevertheless, the significance levels range between 0.01 and $<10^{-5}$, and this is strong evidence for historical connections among the four linguistic groups. Note that when we run the test on modern languages, the degree of similarity between them is greatly attenuated. For example, comparing five modern Turkic languages (Turkish, Tatar, Chuvash, Yakut, and Tuvinian) with two modern Japanese ones (Tokyo and Nasa) we detect a statistically significant relationship only in two out of ten cases (P -values are 0.03 and 0.01). The SI between the proto-languages, however, is significant at $P < 0.001$ level. This is the same pattern that we have already noted in the context of the IE family. Interestingly, we find support for the hypothesis of Mudrak that there is a relationship between Altaic and Eskimo (Table 4b; significant SIs in 3 out of 4 cases).

We can now reject the explanation that the observed similarities between Altaic languages are due merely to chance. What remains, however, is the second objection: that the proto-languages of these families could have acquired similar lexicons “due to a prolonged history of areal convergence” (3). One possible response to this alternative explanation is that borrowings into the basic lexicon (100-word lists) are rare (11). Thus, we expect that languages belonging to different linguistic families will have low SIs, even when they have coexisted in the same region for a long period of time. We test this proposition empirically.

First, we looked at comparisons of languages belonging to different families that were located in spatial proximity: (a) Old Chinese vs. the proto-languages within Altaic; (b) Turkish vs. modern languages of people that inhabited the Ottoman Empire (1378–1914); and (c) Turkish vs. Classical Persian and Arabic (Table 5). The last comparison is particularly interesting because these three languages have coexisted in close cultural interaction at least since the Seljuk Sultanate (eleventh century), and many educated persons in the Middle East were trilingual.

The SIs in Table 5 are somewhat higher than expected under the null hypothesis: three out of eleven comparisons are significant at 0.05 level, and the maximum SI is 6%. What is important for our purposes, however, is that prolonged contact yields much lower SIs than those observed between proto-languages within Altaic (such as the SIs of 11% observed in comparisons of Proto-Mongolic with Proto-Turkic or Proto-Tungus). This observation is contrary to the hypothesis that the observed similarities between Altaic languages are entirely due to borrowings.

More generally, in the 66 comparisons between Altaic and Semitic languages the SIs ranged between 0 and 5% and there were only two significant P -values (whereas we expect 3.3; more generally, the distribution of P -values is not significantly different from the uniform, $\chi^2_9 = 9.55$, $P = 0.39$). This pattern is precisely what should happen when languages are so distantly related that most “signal” has been lost and there were no cross-borrowings into the basic lexicon. In the 363 comparisons between Altaic and IE languages, however, there were 45 significant values (versus the expected 18). There is, thus, evidence for either some limited degree of cross-family borrowing or else deeper genetic connections between the Altaic and Indo-European families, as was proposed by Illich-Svitych in the context of his Nostratic superfamily (12), or both. The main point, however, is that the evidence for internal connections between the Altaic languages is orders of magnitude stronger. (To test the superfamily idea properly using CCM it will be necessary to compare the reconstructed proto-languages of Indo-European, Altaic, and so

forth.) The maximum observed SI in comparisons of modern languages or proto-languages within Altaic to those within IE was 8% (between Albanian and Nasa, no doubt caused by chance: the bootstrap-estimated probability of getting at least one SI=8% or better in the 363 comparisons is $P > 0.7$). By contrast, in the comparisons between the proto-languages within Altaic we observe SIs up to 11%. The bootstrap-estimated probability of getting *two* SIs of 11% in six comparisons (Table 4b) by chance is much less than 10^{-6} .

Discussion

In summary, the Consonant Class Matching approach classifies languages correctly for well-known families and does not appear to yield false positives. It gives reliable statistical inferences about historical connections between languages recorded a relatively short time (say 3,000–4,000 years) after their divergence. Greater time depth (say 6,000–7,000 years) can degrade the signal to the point where it is not detected by the method. We can circumvent this problem, however, using proto-languages, which act like attested ancient languages.

The quantitative patterns of similarity that we documented for languages within the Altaic family are somewhat similar to those in the noncontroversial Indo-European family. The evidence for the common origin of the Altaic languages, at least with respect to word-list comparisons, is thus nearly as strong as that for the Indo-European languages. If the Indo-European family is accepted as real, the same conclusion should also apply to the Altaic family.

However, we do not make a stronger claim that the Altaic languages we analyzed form a monophyletic group (13), because in order to do so, we would need to use the method to construct a phylogenetic tree for these languages. More generally, it should be strongly emphasized that the CCM method is not seen by the authors as a substitute for the standard procedures of comparative linguistics. Properly reconstructed proto-languages remain the principal tool for demonstrating that the daughter languages are genetically related. In the absence of such reconstructions the CCM method can be used as a “short-cut” approach for finding non-random relationships among languages.

ACKNOWLEDGMENTS

The authors thank Tanmoy Bhattacharya, Roy Andrew Miller, Mark Pagel, and Eric Smith for their comments that greatly improved the manuscript.

References

1. Campbell L & Poser WJ (2008) *Language Classification: History and Method* (Cambridge University Press, Cambridge).
2. Starostin S, Dybo A, & Mudrak O (2003) *Etymological Dictionary of the Altaic Languages* (Brill, Leiden).
3. Georg S (2004) Review of Etymological Dictionary of the Altaic Languages. *Diachronica* 21:445-450.
4. Vovin A (2005) The end of the Altaic controversy. *Central Asiatic Journal* 49:71-132.
5. Dybo AV & Starostin GS (2008) In Defense of the Comparative Method, or the End of the Vovin Controversy. *Aspects of Comparative Linguistics* 3:109-258.
6. Swadesh M (1955) Towards greater accuracy in lexicostatistic dating. *International Journal of American Linguistics* 21:121-137.
7. Ringe DA (1992) On calculating the factor of chance in language comparison. *Transactions of the American Philosophical Society* 82:1-110.
8. Kessler B (2001) *The Significance of Word Lists* (CSLI Publications, Stanford, CA).
9. Mudrak OA (1984) K voprosu o vneshnikh svyazyakh eskimosskikh yazykov [On the external relations of Eskimo languages]. *Linguistic reconstruction and the prehistory of the East*, (Institute of Oriental Studies, Moscow).
10. Mudrak OA (2009) Some Altaic Etymologies Found in the 100-word Etymostatistics List. Manuscript.
11. Starostin SA (2000) Comparative-historical Linguistics and Lexicostatistics. *Time Depth in Historical Linguistics*, eds Renfrew C, McMahon A, & Trask L (McDonald Institute for Archaeological Research, Cambridge), pp 223-259.
12. Illich-Svitych VM (1971-84) *Opyt sravneniya nostraticheskikh yazykov (semito-hamitskij, kartvel'skij, indoevropeyskij, ural'skij, dravidijskij, altajskij). Sravnitel'nyj slovar'. Vols 1-3.* (Nauka, Moscow).
13. Pagel M (2009) Human language as a culturally transmitted replicator. *Nature Reviews Genetics* Published online 7 May 2009:1-11.
14. Dolgopolsky A (1986) A probabilistic hypothesis concerning the oldest relationships among the language families. *Typology relationship and time: a collection of papers on language change and relationship by Soviet linguist*, eds Shevoroshkin V & Markey T (Karoma, Ann Arbor), pp 27-50.
15. Efron B & Tibshirani RJ (1993) *An introduction to the bootstrap* (Chapman and Hall, New York).

Table 1. Consonant classes.

	Front: labial, labiodental	Central: dental, alveoral, postalveolar, palatal, retroflex	Back: velar, uvular
Nasal	M: m, m̥, m̃	N: n, n̥, ɲ, ɳ	ŋ, N
Plosive, implosive, ejective	P: p, ph, b, p', ɓ	T: t, th, d, d̥, t'	K: k, g, k', q, G
Fricatives and approximants	ɸ, β, f, v	S: s, z, ʃ, ʒ, ʒ̥, ʃ, θ	x, ʁ, R
Affricates		C: c, ʒ, tʃ	
Laterals		L: l, l̥, ɭ	

Bold capital letters refer to the class code. r-type consonants are grouped together with retroflexes in **T**. The ninth class (#) is formed by consonants which historically can be identical with ø (lack of a consonant): ʕ, h, y, etc. For this paper we treat all vowels as one single class. As a result, each root is represented as a sequence of consonantal classes: *gataw* > KT, *xuthip* > KT, *mars* > MT, and *arbil* > #T (# represents the missing initial consonant). Altogether there are 81 (9x9) possible forms of roots.

Table 2. Similarity Indices (consonant class matches) within and between modern Indic and Germanic language groups. Below the diagonal: Similarity Indices (percentage of matches). Above the diagonal: the bootstrap-estimated probability of the observed SI or a larger one under the null hypothesis.

	Hindi	Beng.	Nep.	German	Engl.	Dutch	Ice.	Swed.
Hindi	–	<10 ⁻⁶	<10 ⁻⁶	0.02	0.6	0.2	0.3	0.04
Bengali	52	–	<10 ⁻⁶	0.02	0.01	0.01	0.14	0.01
Nepali	56	49	–	0.005	0.3	0.08	0.7	0.01
German	6	6	7	–	<10 ⁻⁶	<10 ⁻⁶	<10 ⁻⁶	<10 ⁻⁶
English	2	7	3	46	–	<10 ⁻⁶	<10 ⁻⁶	<10 ⁻⁶
Dutch	4	7	5	76	59	–	<10 ⁻⁶	<10 ⁻⁶
Icelandic	3	4	1	46	45	54	–	<10 ⁻⁶
Swedish	6	7	7	64	57	76	72	–

Table 3. CCM results for ancient languages (SIs below the diagonal, *P*-values above the diagonal).

(a) Indo-European languages.

	Old Ind.	Avest.	Class. Pers.	OH Germ.	Latin	Old Irish	Anc. Greek	Hitt.
Old Indian 1000 BCE	–	<10 ⁻⁶	<10 ⁻⁶	<10 ⁻⁶	<10 ⁻⁶	0.0003	0.02	<10 ⁻⁴
Avestan 600 BCE	42	–	<10 ⁻⁶	0.001	<10 ⁻⁴	0.01	0.01	0.01
Classical Persian, 1000 CE	23	40	–	0.0002	<10 ⁻⁴	0.0002	0.004	0.001
O.H.German 900 CE	14	10	9	–	<10 ⁻⁶	<10 ⁻⁶	0.0001	0.006
Latin 300 BCE	19	15	13	17	–	<10 ⁻⁴	<10 ⁻⁴	<10 ⁻⁴
Old Irish 900 CE	9	8	9	14	13	–	0.001	0.009
Ancient Greek 600 BCE	7	8	7	11	22	9	–	<10 ⁻⁶
Hittite 1500 BCE	16	8	9	8	16	8	20	–

(b) Semitic languages.

	Akkad.	Hebrew	Aramaic	Arabic	Ge'ez
Akkadian 1600 BCE	–	<10 ⁻⁶	<10 ⁻⁶	<10 ⁻⁶	<10 ⁻⁶
Hebrew 700 BCE	42	–	<10 ⁻⁶	<10 ⁻⁶	<10 ⁻⁶
Aramaic 300 CE	37	49	–	<10 ⁻⁶	<10 ⁻⁶
Arabic 600 CE	24	33	34	–	<10 ⁻⁶
Ge'ez 400 CE	32	37	29	33	–

Table 4. CCM results for reconstructed protolanguages (SIs below the diagonal, *P*-values above the diagonal).

(a) Indo-European

	P-Iranian	P-Slavic	P-Baltic	P-Germanic
Proto-Iranian	–	$<10^{-6}$	$<10^{-6}$	$<10^{-6}$
Proto-Slavic	20	–	$<10^{-6}$	$<10^{-6}$
Proto-Baltic	13	35	–	$<10^{-6}$
Proto-Germanic	13	19	21	–

(b) Altaic and Eskimo

	P-Turkic	P-Mong.	P-Tungus	P-Japanese	P-Eskimo
Proto-Turkic	–	$<10^{-4}$	0.002	$<10^{-4}$	0.04
Proto-Mongolian	11	–	$<10^{-5}$	0.0008	0.004
Proto-Tungus	8	11	–	0.02	0.00003
Proto-Japanese	9	7	6	–	0.41
Proto-Eskimo	6	8	10	2	–

Table 5. CCM results for inter-family comparisons

(a) Altaic proto-languages vs. Old Chinese

	Proto-Turkic	Proto-Mong.	Proto-Tungus	Proto-Japanese
Similarity Index	2	2	5	6
<i>P</i> -value	0.28	0.21	0.02	0.03

(b) Languages of the Ottoman Empire vs. Turkish

	Kurdish	Serbian	Albanian	Greek	Armenian
Similarity Index	4	4	5	1	1
<i>P</i> -value	0.10	0.08	0.01	0.72	0.76

(c) Turkish vs. Persian and Arabic

	Persian	Arabic
Similarity Index	4	1
<i>P</i> -value	0.06	0.79