Bioinformatics Evolutionary Tree Algorithms Reveal the History of the Cretan Script Family

Peter Z. Revesz

***Abstract*—** This paper shows that Crete is the likely origin of a family of related scripts that includes the Cretan Hieroglyph, Linear A, Linear B and Cypriot syllabaries and the Greek, Phoenician, Old Hungarian, South Arabic and Tifinagh alphabets. The paper develops a novel similarity measure between pairs of script symbols. The similarity measure is used as an aid to develop a comparison table of the nine scripts. The paper presents a method to translate comparison tables into DNA encodings, thereby enabling the use of bioinformatics algorithms that construct hypothetical evolutionary trees. Applying the method to the nine scripts yields a script evolutionary tree with two main branches. The first branch is composed of Cretan Hieroglyph, Cypriot, Linear A, Linear B, Old Hungarian and Tifinagh, while the second branch is composed of Greek, Phoenician and South Arabic. It is also considered how Proto-Sinaitic and Ugaritic may belong to this script family.

***Keywords*—**Cretan Hieroglyph, Linear A, Linear B, Evolution, Neighbor Joining, Old Hungarian, Phylogenetics, UPGMA, Tifinagh.

I. Introduction

CRETE was the birthplace of several ancient writings that were first categorized by Arthur Evans, the explorer of Knossos Palace, as the Cretan Hieroglyph, the Linear A and the Linear B scripts [5]. Linear A, which dates back to about 2500 BC, was the main script used in the Minoan palaces of ancient Crete. The Cretan Hieroglyph script, which may predate Linear A, was used for centuries simultaneously with Linear A. Linear A was replaced around 1450 BC by Linear B, which was used in Mycenaean Greece and is the oldest known Greek writing [10]. As described in Chadwick [2], despite Evan's decades long attempt to decipher Linear B, it remained a mystery until 1952 when Michael Ventris gave a decipherment of Linear B showing that it is an archaic version of Greek. However, the Cretan Hieroglyph and the Linear A scripts are still not deciphered.

In order to understand better these three ancient Cretan scripts, in this paper we study their relationship with six other scripts. The other scripts are the Cypriot syllabary, and the Phoenician, the South Arabic, the Greek, the Old Hungarian and the Tifinagh alphabets.

The Cypriot syllabary [26], which was used between the 11th and the 4th centuries BC, was deciphered by George Smith, who was aided by a bilingual Phoenician-Cypriot inscription. The Cypriot syllabary derives from the earlier Cypro-Minoan syllabary, whose similarity with Linear A was noted by Evans.

The Phoenician alphabet [28] was a major influence on the development of many other alphabets due to the Phoenicians' widespread commercial influence in the Mediterranean area. The Phoenician and the South Arabic [30] alphabets are assumed to derive from the Proto-Sinaitic alphabet, which originated in the Sinai Peninsula sometime between the mid-19th and mid-16th century BC [29]. Phoenician represents the northern branch, while South Arabic represents the southern branch of Proto-Sinaitic.

The classical Greek alphabet from about 800 BC had a major influence for many other European alphabets. The classical Greek alphabet derives from the Phoenician alphabet except for the letters Φ, X, Ψ and Ω [27].

The Old Hungarian alphabet is the alphabet used by Hungarians before the adoption of the Latin alphabet. Parallel with the Latin, it was used sporadically until the 20th century. The origin of Old Hungarian is still debated. Hosszú [11] presents a detailed view of the development from Phoenician via Aramaic and Turkish and Proto-Rovas scripts. In contrast, Forrai [8] and Varga [25] claim that the Old Hungarian script already existed in the Bronze Age and cite putative translations of engraved artifacts going back to 1000 BC.

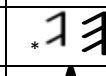
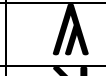
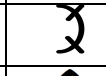
Tifinagh is another ancient script that was used by Berber language speakers in North Africa and on the Canary Islands. Tifinagh is attested in writing from at least the 3rd century BC. The origin of the Tifinagh alphabet is also unknown, although it is often assumed to derive from Phoenician [31].

Using bioinformatics methods, we show in this paper that the above nine scripts are members in the same script family that spans across language families. In computational biology, the study of evolutionary relationships is greatly facilitated by the use of phylogenetic tree construction algorithms, such as Saitou and Nei's *neighbor-joining* method [21] and Sokal and Michener's UPGMA method [23]. The books by Baum and Smith [1], Hall [9] and Lerney et al. [12] review the *maximum likelihood* and several other methods. Recently, Revesz [15] also proposed the *Common Mutations Similarity Matrix* or CMSM method for phylogenetic tree construction. The CMSM method derives from a series of previous evolutionary biology studies, including [14], [18]-[20], [22] and [24].

Some of the efficient phylogenetic tree algorithms are able to reconstruct hypothetical evolutionary trees in a few minutes of computational time. Moreover, they are based on statistical techniques that are free of human bias, which sometimes prevents the objective evaluation of linguistic artifacts.

Peter Z. Revesz is with the Department of Computer Science and Engineering, University of Nebraska-Lincoln, Lincoln, NE 68588, USA (revesz@cse.unl.edu). A preliminary version of this paper was presented at the AMCSE conference in Agios Nikolaos, Crete, Greece in October 2015 [16].

Table I A script comparison with each symbol's known sound value to the right (blue) except for Tifinagh, which is similar to South Arabic's.

Hieroglyph and Phaistos	Linear A	Linear B		Cypriot		Phoeni -cian		South Arabic		Greek	Old Hungarian		Tifin -agh
			A		I		ʔ		ʔ	**A**		A	
			SE		SE		B		B	**B**		P	
					KO		G		G	**Γ**		G	
			DA		TA, TE		D		D	**Δ**		D, T	
			KU		XE		H		H	**E**		E	
					MO		W		W	**Υ**		US	
					MU		Z		Z	**Z**		U	
			PA		PA		Ħ		ḍ ʕ	**H**		ɟ	
			KA		E		Tˢ		Tˢ	**Θ**		C	
			MA		WO		Y		Y	**I**		J	
			WE		NU		K		K	**K**		K	
			PU		LI		L		L	**Λ**		L	
			TWE		MI		M		M	**Μ**		M	
			NE		U, NA		N		N	**Ν**		NT	
			TE		TO		S		S	**Ξ**		H	
			QE		JA		ʕ		ʕ	**Ο**		J, L	
			PO		PO		P			**Π**			
			ZO		TI		Ṣ		Ṣ			t͡s	
			QA		KI		Q		Q			K	
					RA		R		R	**Ρ**		*R	
			TI		SA		š		š	**Σ**		š	
			RO		LO		T		T	**Τ**		D	
					A				F			F	
										Χ		H	
			RE		RI				Ħ	**Ψ**		ʒ	
			TA		LU					**Ω**		O	

Human translation attempts are inherently prone to error. For example, the Phaistos Disk, which contains some form of Cretan Hieroglyph writing, was translated in numerous contradictory ways by a large number of professional and amateur linguists. Faucounau [6] and Fisher [7] are examples of decipherment attempts, and Duhoux [4] is a critique of previous decipherment attempts. In spite of these problems, human evaluation of the similarity of script symbols is still common in linguistics and is a source of bias. In contrast, bioinformatics developed many sophisticated mathematical measures of the similarity of DNA and proteins [33]. In this paper we develop a mathematical measure for the similarity of pairs of script symbols. Our similarity measure is particularly applicable to the linear scripts studied in this paper. The other source of bias is the human construction of linguistic evolutionary trees. Instead of that approach we use the UPGMA algorithm to construct script evolutionary trees. The idea of using bioinformatics tools is attractive, but the issue is to figure out how to translate the linguistic problem of scripts into a bioinformatics problem. In this paper we present a method of translating script syllabaries and alphabets into a DNA-like encoding. This DNA-like encoding of related scripts can be passed into the evolutionary tree algorithms used to reconstruct hypothetical evolutionary scripts. In this translation, each script becomes like the DNA of a species studied by bioinformatics.

This paper is organized as follows. Section II presents a similarity measure for script symbols. Section III uses the similarity measure to make a comparative table of the script symbols. Section IV describes the DNA encoding of the scripts. Section V presents a computational reconstruction of the evolutionary tree of the scripts. Section VI discusses the results and presents a hypothesis of the spread of the scripts from a common putative source in Crete. Finally Section VII gives some conclusions and directions for future work.

II. A Similarity Measure for Script Symbols

Many researchers have given subjective opinions about the similarities of symbols. It seems better, however, to use an objective similarity measure that is applied uniformly and avoids some possibilities for bias in deciding which symbols are similar to each other. We present below a similarity measure that formalizes the process of making these decisions.

The similarity of script symbols can be measured in many ways. One general technique would be to take the Hausdorff distance between two scripts letters. However, the Hausdorff distance varies according to how the two script letters are written. In particular, the Hausdorff distance tends to increase if one letter is fixed while the other letter is written in increasingly bigger font. The measure that we propose below avoids this problem and is particularly suitable to linear scripts, whose script symbols are mostly composed of straight lines and a few curves. In Table I, all the scripts except the Cretan Hieroglyph and Phaistos script in the first column are linear scripts.

The rarity of curved lines makes the scripts symbols that contain them stand out from the ones that contain only straight lines. Hence we divide script symbols into two groups:

Group 1. Symbols that contain some curved lines
Group 2. Symbols that contain only straight lines

For example, for the Cypriot script, Group 1 contains:

All the symbols in the first group contain a significant part of an arc of a circle or oval as a deliberate feature. Only one symbol seems somewhat difficult to classify, namely:

Although the above symbol contains a little bit of curve near the bottom, the curve here seems to connect two straight lines, namely the vertical line, which is the stem of the arrow, and a small horizontal line at the bottom. Hence the curve connecting these two straight lines seems only a secondary feature that was probably introduced as a writing convenience in a later stage of writing rather than an original deliberate feature of the character. In contrast, all the six Group 1 symbols listed above contain curves lines, which seem deliberate and original features of those symbols. Therefore, we classified the above script symbols and all the remaining symbols into the second group.

Our next major division of script symbols is according to whether they enclose some region.

Group I. Symbols that enclose some region
Group II. Symbols that do not enclose any region

For example, the following Cypriot symbols enclose some region and can be classified as belonging to group I.

The next to the last symbol in the above list is somewhat debatable, but it seems that the small U letter within that script symbol is enclosed under a big gate symbol. The rest of symbols can be classified as belonging to Group II.

Our third division of symbols is according to whether they contain slanted lines or only contain vertical and horizontal lines.

Group α. Symbols that contain slanted straight lines
Group β. Symbols that do not contain slanted straight lines

The classifications are not independent of each other. For example, symbols with curved lines tend to also enclose regions. Out of the 25 Cypriot script symbols, we classified six as belonging to Group 1 and also six as belonging to Group I, but four symbols belong to both Group 1 and Group I. If the two classifications were independent of each other,

then we would expect the probability of a symbol belonging to both Group 1 and Group I to be 0.0576 rather the actual probability of 0.16, which is nearly three times larger. That is:

$$\mathrm{P(1\ and\ I)} = 0.16 \neq \mathrm{P(1)P(I)} = \frac{6}{25} \times \frac{6}{25} = 0.0576$$

The same pattern seems to hold for the other scripts too. That implies that some of the Group 2 symbols that enclose a region may have been originally also Group 1 symbols whose shape was linearized over time. For example, the Phoenician

may have been originally either an figure 8 like symbol or a circle with a line division in the middle. Similarly, the following Old Hungarian symbol may have been originally an oval or a circle:

Our similarity measure is a scoring function S based on reward scores assigned for various similarities. For any pair of symbols *a* and *b*, the similarity *S(a,b)* measure is the sum of the rewards assigned based on certain rules. The reward rules are the following:

Rule 1. *If both symbols belong to Group 1, or both belong to Group 2, then the reward is two points.*

Rule 2. *If both symbols belong to Group I, or both belong to Group II, then the reward is two points.*

Rule 3. *If both symbols belong to Group* α, *or both belong to Group β, then the reward is two points.*

Rule 4. *If both symbols have two or more parallel lines, then they get one point reward for each of the shared parallel lines.*

Rule 5. *If both symbols contain a cross* X, *then the reward is two points.*

Rule 6. *If both symbols contain wedge* ∧, *then the reward is one point. If they both contain a wavy line, then the reward is two points.*

Rule 7. *If both symbols have similar sound values, then the reward is two points.*

Rule 8. *If both symbols have the same meaning, then the reward is two points.*

Rule 3 is only applicable if the symbols are not rotated. Otherwise, even a vertical or horizontal line may become slanted. The other rules allow rotation of the symbols. Rule 8 needs to be applied judiciously. Whenever it is obvious that two symbols depict the same object, Rule 8 is a legitimate rule to apply to increase their similarity measure. However, when the meaning of objects is not obvious, then it would be misleading to apply Rule 8 by imagining the symbols to depict certain objects. Hence we applied Rule 8 cautiously.

The above eight rules can be applied to determine the similarity of any pair of linear symbols. For example, consider the following pair of symbols from Cypriot on the left and Old Hungarian on the right:

The similarity of these two symbols is more striking if we rotate the first symbol:

Clearly, the two symbols both belong to Group 1 (two points by Rule 1), to Group II (two points by Rule 2) and to Group α (two points by Rule 3). The first has two and the second has three parallel lines. Hence they share two parallel lines (two points by Rule 4). In addition, both symbols contain a wedge (one point by Rule 6), and both symbols have similar sound values, LI and L, respectively, (two points by Rule 7). Hence the similarity of the two symbols is:

$$s(\text{  }, \text{  }) = 2 + 2 + 2 + 2 + 1 + 2 = 11$$

The visual rules (Rules 1-6), the sound rule (Rule 7), and the semantic rule (Rule 8) almost always support each other, although sometimes the support is not obvious. For example, consider the following pair of Cypriot and Old Hungarian symbols:

Both of these symbols belong to Group 1, Group I, and Group *β*. Hence the similarity score of the two symbols would be six based on only Rules 1, 2 and 3. With a value of six, we would expect the sound values to support each other, but the Cypriot sound value "MO" and the Old Hungarian sound value "US" are different. However, it seems apparent these symbols are related the Phoenician and South Arabic semivowel "W" sound values (see Table I) and the Tifinagh "B" sound. In languages where the "W" was not used, it was commonly translated as the vowel "U," including in ancient Greek, where the symbol was named "UPSILON." The Old Hungarian "US" may be a similar adaptation of "W" to "U." The sound "W" also changes sometimes to sounds "B" and "M." Hence while the "MO" and the "US" look different, some similarity can be found between these two sounds too. Hence Rule 7 is applicable, and the similarity score of the two symbols can be updated to eight.

As another example, the Linear B and the Phoenician symbols:

have a score of six because they both belong to Group 1 (two points by Rule 1), both belong to Group I (two points by Rule 2) and both contain a cross (two points by Rule 5). In this case the Linear B sound value is "RA" cannot be reconciled with the Phoenician sound value that corresponds to Greek Θ

or "THETA." The major difficulty here is not simply that Linear B is a syllabary while Phoenician is an alphabet. A syllabary with consonant-vowel syllable combinations can have a natural evolution into an alphabet when either the consonant or the vowel is dropped. The problem here is that the "R" sound cannot be reconciled with the "TH" sound. Hence the above pair is a rare example where a relatively high visual similarity is not accompanied with a sound similarity.

III. A Comparative Table of Script Symbols

We built a comparative table of script symbols shown in Table I. In Table I, the Phoenician alphabet and the South Arabic alphabet columns are taken from [28] with minor modifications. The Greek alphabet column is taken from [27]. The Old Hungarian and the Tifinagh [31] columns are our arrangement. The sound values of the Old Hungarian alphabet are from [8], [11] and [25]. The symbols marked with a star * are Proto-Rovas symbols that were used in the early phases of Old Hungarian according to Hosszú [11]. Our reconstruction assumed that the * symbols represent the more archaic form of Old Hungarian. It is possible that these archaic forms were changed to the latter forms due to Turkish or other influences. Our reconstruction of Old Hungarian was guided by the rules and the similarity measure described in Section II.

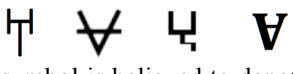
Linear B and its sound values are from Chadwick [2] and Hooker [10]. The Cretan Hieroglyph and Linear A correspondences to Linear B are our reconstructions based in part on previous observations by Evans [5], Fisher [7] and Young [32]. Since the sound values of the Cretan Hieroglyph and Linear A script symbols are unknown, their arrangement in Table I was guided by the similarity measure in Section II without using Rule 7. Not being able to use Rule 7 to rewards points for sound similarities probably slightly underestimated the similarities in many cases.

IV. The DNA Encoding of Script Symbols

A. From Script Symbols to DNA

After the alignment of the script symbols as shown in Table I, we took a careful look at each row. In each row, we divided the set of symbols into groups such that in each group the symbols were closer together than they were to members of other groups. We call the first group the A group, the second group the C group, the third group the G group, and the fourth group the T group. These groups are named after the four DNA nucleotides. If a script does not have a symbol, then we write a dash to indicate that it is not in any group. We explain the process in a few examples.

Row 1: The Linear B, Phoenician, the South Arabic and the Greek symbols can be rotated as:



The Phoenician symbol is believed to denote the head of an ox with two horns [28]. The other symbols clearly imitate this idea with two horns. Hence the semantic rule (Rule 8) establishes a similarity among all three symbols, and they are grouped into Group A. The Hieroglyph, the Linear A and the Cypriot symbols clearly denote persons:

The second form of Linear A and the Old Hungarian seem to denote the head of a person rather than the head of an animal with horns:

Hence they also are grouped together into Group C. Tifinagh, which did not have a corresponding symbol, was marked by a dash.

Row 2: For this row we consider only the first element of the Hieroglyph and the second element of Old Hungarian. None of the symbols except Greek contains curved lines. None of the symbols except Greek and Phoenician contains enclosures. None of the symbols except Hieroglyph, the Linear A, the Phoenician and the Hungarian contains slanted lines. All symbols except Phoenician and Greek have some parallel lines. The sound values of Linear B and Cypriot are similar, and the sound values of Phoenician, South Arabic, Greek and Hungarian are also similar. These generate the similarity matrix shown in Table II, where each entry for row i and column j records the similarity measure of the symbols associated with row i and column j. The similarity matrix shows that all script symbols can be placed into Group A except Phoenician and Greek, which are placed into Group C.

Table II The similarity matrix based on the second row of symbols. H = Hieroglyph, A = Linear A, B = Linear B, C = Cypriot, P = Phoenician, S = South Arabic, G = Greek and O = Old Hungarian.

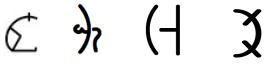
	A	B	C	P	S	G	O
H	9	7	7	4	6	0	9
A		7	7	4	6	0	9
B			9	2	6	2	7
C				6	4	2	2
P					4	4	4
S						2	4
G							2

Row 3: The Hieroglyph, the Linear A, the Cypriot, the Old Hungarian and the Tifinagh symbols denote tripods with either the three legs or the flat top clearly visible. Hence they share a semantic similarity. These plus the Phoenician symbol contain a wedge, and all but South Arabic and Greek contain slanted lines. The sound values are all similar because "K" and "G" are both velar plosives. These put the Hieroglyph, the Linear A, the Cypriot, the Old Hungarian and the Tifinagh symbols into Group A. The other symbols are put into Group C.

Row 4: The symbols denote a bow and an arrow or a slingshot. In South Arabic the arrowhead is illustrated by a triangle pointing right and the bow is simply a vertical line. The Phoenician and the Greek symbols preserve the

arrowhead, although it is pointing left and up instead of right. These symbols belong to Group I because the triangles are closed areas. The sound values are all similar because "D" and "T" are both alveolar plosives. These place Phoenician, South Arabic and Greek into Group A and the rest into Group C.

Row 5: The symbols in this row may denote the verb "hit" by depicting some hitting object, either arrow or a fist. The Linear A, Linear B, Cypriot, and Old Hungarian symbols have curved lines and seem to denote arrows and bows:



although the Linear B sign also could be a bird according to some researchers. The Cypriot sign seems to be the oldest form. It seems likely that the two separate lines for the bow and the arrow with the string touched each other in the past, forming a spatial enclosure. The Linear A symbol only suggests the presence of an arrow by having the bowstring contain an angle. The Hungarian symbol likely was similar to the Linear A sign with the two cross lines on the semicircle touching each other. Hence all four signs originally formed an enclosure and belonged to Group 1.

The Hieroglyph, the Phoenician, the South Arabic and the Greek symbols may denote a fist and are similar to each other, especially when displayed with some rotation.

The above symbols have parallel lines varying in number from five in the Hieroglyph to two in the South Arabic symbol. Both "K" and "X" are velar sounds, and "X" and "H" are both fricatives, and word initial "H" is often dropped. Hence "XE" may change to "HE" and then to "E." Hence there is some similarity in the sound values. We put the first four symbols into Group A and the second four symbols into Group G.

Row 6: The Hieroglyph, the Linear A, the Cypriot, the South Arabic, the Old Hungarian and the Tifinagh symbols contain curved lines and enclosed regions and form Group A. The Phoenician and the Greek symbols form Group C. All of the symbols have a vertical line in the middle and the sound values as discussed in Section II are also similar.

Row 7: The Hieroglyph, the Cypriot, the South Arabic and the Old Hungarian are similar with enclosed regions and crosses, resulting in hourglass shapes. The Phoenician and the Greek do not have enclosed regions. Here the Cypriot "MU" and Old Hungarian "U" are similar sound values, but they are different from Phoenician, Greek, South Arabic and Tifinagh "Z." All of these shapes share in common two parallel lines except Tifinagh. We put the Hieroglyph, Cypriot, South Arabic and Old Hungarian symbols into Group A, and Phoenician, Greek and Tifinagh into Group G. In Tifinagh the Phoenician symbol would mean /ʒ/, which is a similar sound.

Row 8: The Hieroglyph, Phoenician and South Arabic, the Tifinagh symbols have three parallel lines and all but the last contains enclosed regions:

whereas the others have only two parallel lines and no enclosed regions. Hence we put the above displayed symbols into Group A and the other five into Group C.

Row 9: The Hieroglyph and South Arabic have no curved lines, enclosed regions and share three parallel lines, forming Group A. Linear A, Linear B, Phoenician, and Greek have curved lines, and the first three have a cross in them while the Greek has only a horizontal line. We placed these into Group C. Neither the Cypriot symbol nor the Hungarian symbol encloses a region, but they share a cross. It seems that these symbols depict a spinning wheel in motion. Both symbols illustrate the counterclockwise rotation of the spinning wheel by small lines emanating from one of the spokes of the wheel. Hence the Cypriot and the Hungarian symbols form Group G.

Row 10: The symbols in this row depict the head of a cat (Rule 8) with varying degree of abstraction. The Hieroglyph, Linear A, Linear B, Cypriot and the Old Hungarian symbols have slanted lines (Rule 3) and reflect the prominent ear tip of a cat by the use of a wedge (Rule 6) and the more abstract ones of these avoid the use of curved lines. Hence they form Group A. The Phoenician, Greek and South Arabic symbols lack any wedge. They can be further divided into the Phoenician and Greek, which do not contain curves and enclosures, hence form Group C, and the South Arabic, which contains those features, hence forms Group G. In this row "M" is close to the semivowels "W" and "J" are semivowels.

Row 11: These symbols seem to denote ropes. The original form may have been the ones in Linear A and Linear B, which have curved lines and form Group A. The Hieroglyph and Old Hungarian are linearized forms of the original rope form and have slanted and parallel lines. Hence they form Group C. The Cypriot, Phoenician and Greek seem to depict the rope when it is tying together something. They have both slanted lines and wedges. Hence they form Group G. South Arabic lacks the curved lines of Group A, the slanted lines of Group C, and the wedges of Group G. Hence it is put into Group T. The sound values "K," "NU" and "WE" are also different.

Row 12: These symbols may depict a bird as in the first Hieroglyph symbol (Rule 8). In Linear A and Linear B the flexible neck of the bird turning backward is illustrated by curved lines (Rule 1), and the two legs and the tip of the tail feathers are illustrated by three parallel lines (Rule 4). Hence Hieroglyph, Linear A and Linear B form Group A. The first Hieroglyph symbol simplifies the bird into a triangular shape still with three legs. The triangular shape with some legs is continued in the Cypriot and the Old Hungarian, which have slanted lines and a wedge shape, forming Group C. The Phoenician, South Arabic and the Greek symbols are further reduced to a single wedge, forming Group G. In the Tifinagh symbol only two parallel lines remain. This symbol forms group T. The sound values "L/LI" in Groups C, G and T agree but are different from the Linear B sound value "PU."

Row 13: All of these symbols except the Tifinagh symbol have a wavy line in them (Rule 6). Linear A, Phoenician and Greek contain only a single wavy line, placing them into

Group A. The Hieroglyph, the Linear B, the South Arabic, and the Old Hungarian contain enclosures (Rule 1). The Cypriot is clearly more than a single wavy line and shows a partial enclosure of space. If we rotate the symbols and align them, then it is clear that with a simple extension of the lines in the Cypriot symbol, the symbol becomes like the other symbols.

Hence we placed these symbols into Group C. The Tifinagh symbol was an outlier and was classified as G.

Row 14: Hieroglyph, Linear A, Linear B, Cypriot, and Old Hungarian depict a bird; hence they have a semantic similarity (Rule 8). The second Cypriot symbol seems a simplification of the first Cypriot symbol. These symbols are characterized by a wavy line, which tends to be symmetric and composed of four line segments (Rule 6). Hence these were placed into Group A. The Phoenician, South Arabic and Greek symbols also have wavy lines, but those wavy lines are composed of only three line segments. These slight differences in the wavy patterns make sense in the light of Colless [3], which connects the Phoenician symbol with the Egyptian Hieroglyph for snake. Hence these symbols have their own semantic similarity (Rule 8) besides the wavy pattern (Rule 6) and are placed into Group C. The Tifinagh symbol, an outlier, was placed into Group G.

Row 15: The Linear B, Cypriot, Phoenician, and Greek may depict either a tree with branches or a fishbone. They have no spatial enclosure, no slanted lines, but they have parallel horizontal lines. Hence they belong to Group A. The Hieroglyph, South Arabic and Old Hungarian symbols probably denote a fish and have a special enclosure and slanted lines. They form Group C. The Tifinagh symbol may be a simplification of the fish to the point that only its eye remains. It is placed in Group G.

Row 16: These are essentially all circles with curved lines and enclosures except the Tifinagh symbol, which had three dots. The Cypriot and the Hungarian have a little bit of a wedge in them, and some of them contain various numbers of dots. Still they all seem to belong to Group A.

Row 17: The four symbols are similar, and in the history of the Greek alphabet some early forms of "P," for example at Euboa, were written using a curved line [27]. Hence we placed all four into one group, Group G.

Row 18: Hieroglyph, Linear A, Linear B, Cypriot and Old Hungarian have a semantic connection because they each denote an arrow (Rule 8). They have neither an enclosure (Rule 2) nor slanted lines (Rule 3) apart from a wedge on the top (Rule 6). Hence they are placed into Group A. Phoenician and Greek contain waves (Rule 6) and slanted lines (Rule 3) and are placed into Group C. South Arabic is the only symbol with a curved line and an enclosure and forms Group G. The Tifinagh we put into Group T.

Row 19: These symbols all seem to depict the head of a person (Rule 8). All the symbols have an enclosure (Rule 2) and a similar sound value (Rule 7). The Cypriot and the Old Hungarian have no curved lines (Rule 1) and form Group A, while the others form Group C.

Row 20: The Hieroglyph, Linear A, Cypriot, South Arabic, Greek, Old Hungarian and Tifinagh symbols contain curved lines (Rule 1). The Phoenician stands out form the rest by having slanted lines (Rule 3) but no curved lines (Rule 1). Hence we placed Phoenician into Group A and the rest into Group G.

Row 21: The Hieroglyph, Linear A, Linear B, Cypriot and Old Hungarian symbols consist of a wedge with the addition of a vertical line in Linear B and Old Hungarian, where it is optional. Hence these form Group A. The Phoenician, South Arabic and Greek symbols have slanted lines (Rule 3) and waves (Rule 6) and are placed into Group C. The Hieroglyph and the Tifinagh symbols contain enclosed space, in fact, little circular endings. These symbols may denote an arm with the circles forming the hand. The Tifinagh symbol is symmetric representing both hands. Since this forms a semantic similarity (Rule 8), and arguably the bends of the arms form wedges, we also placed the Tifinagh symbol into Group A.

Row 22: These symbols are very similar to each other and consist of a simple cross (Rule 5) except in the case of Greek. We placed Greek into Group A. We placed Phoenician and South Arabic into Group C because in them the cross is rotated which introduces slanted lines (Rule 3). We placed the rest of the symbols into Group G. The Linear B sound value "RO" and the Cypriot sound value "LO" are similar, but the Old Hungarian sound value "D" is closer to "T" because both of them are alveolar plosives.

Row 23: The Hieroglyph symbol has a region enclosure (Rule 2), and the Phaistos rosette form has a cross in it too. The region enclosure is shared with South Arabic, Greek and Old Hungarian, while the cross is shared with Cypriot and Old Hungarian. South Arabic, Greek and Old Hungarian share the sound value "F" while the Cypriot sound value "A" may have been "FA" originally. Hence it is a compatible sound value. These symbols are hard to divide into groups. Hence we assigned them all into Group A.

Row 24: Both the Greek and the Old Hungarian symbols have a cross in them. In fact, the Old Hungarian symbol has two crosses in it. The crosses have the same orientation. Therefore they have slanted lines. The sound values are similar. Hence both symbols were assigned to Group A.

Row 25: The apparent semantic connection (Rule 8) in these symbols is that they depict tridents. Hence in spite of some minor variations, we grouped them all into Group C.

```
Hieroglyph     CAACGAAAAACACACA-ACGAGA-C-
Linear_A       CAACAA-CCAAAAA-A-ACGAG--CC
Linear_B       AA-CA--CCAAACAAAGAC-AG--CC
Cypriot        CAACAAACGAGCCAAAGAAGAGA-C-
Phoenician     ACCAGCGACCGGACAAGCCACC----
S_Arabic       AACAGAAAAGTGCCCA-GCGCCA-C-
Greek          ACCAGCGCCCGGACAAGCCGCAAACA
O_Hungarian    CAACAAACGACCCACA-AAGAGAACA
Tifinagh       --AC-AGA---TGGGA-T-GAG----
```

Fig. 1 The DNA encoding of the seven alphabets

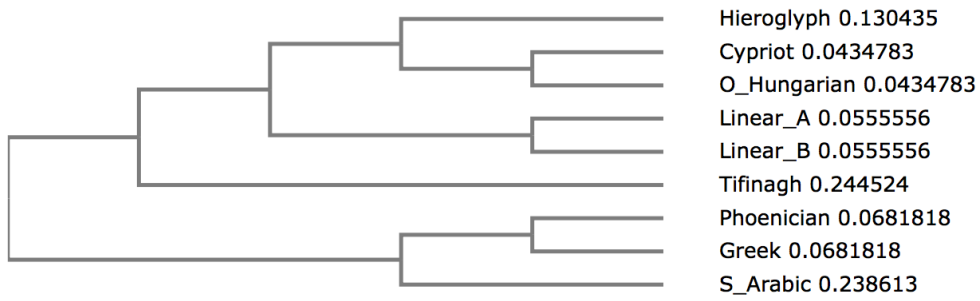


Fig. 2 The evolutionary tree generated by the UPGMA phylogeny algorithm in ClustalW2 and displayed as a cladogram.

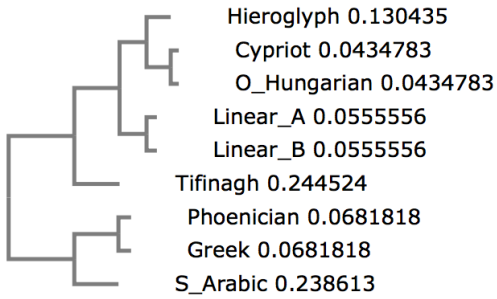


Fig. 3 The evolutionary tree generated by UPGMA phylogeny algorithm in ClustalW2 and displayed as a tree that suggests a time of origin for each script.

Fig. 4 The figure illustrates the hypothesis that Crete is the origin of a family of scripts. The first branch is colored red and the second branch is colored blue. The first branch diversified in Crete and included Hieroglyph, Linear A and Linear B and spread to North Africa where it became Tifinagh and also spread to Anatolia where it split into a northern group that reached the Black Sea area and an eastern group that reached Cyprus. It possibly spread further east to Ugarit. The second branch spread to Phoenicia and from there to Yemen and later spread back to Greece. Possibly, the second branch formed Proto-Sinatic before Phoenician, or Proto-Sinatic may have derived from Egyptian, as proposed by Colless [3], and also influenced the development of Phoenician and other languages in the second branch.

Row 26: Here the Hieroglyph, the Greek and the Old Hungarian symbols contain curved lines and the same sound value "O" for at least the latter two. Hence these were placed into Group A, and Linear A and Linear B into Group C.

After the grouping of the symbols in each row of Table 1, we wrote down the group labels in a column where the rows corresponded to the nine scripts. Fig. 1 shows the result.

B. From Script Symbols to Proteins

In case we need to use more than four groups, the groups could be named after the twenty amino acids. In that case, we would get a protein encoding of the alphabets instead of a DNA encoding. Each pair of amino acids can have a different similarity value. This allows a more precise description of similarities as needed. Although the four amino acid groups were enough for our encoding presented in Section A, we present this suggestion for future work.

V. A Computational Reconstruction of an Evolutionary Tree Using Phylogenetics

We used ClustalW2's phylogenetic algorithms because they are among the most frequently used in bioinformatics and are available free to all users from the website:
http://www.ebi.ac.uk/Tools/phylogeny/clustalw2_phylogeny/

For the DNA encoding in Fig. 1, ClustalW2 computed a hypothetical phylogenetic tree as shown in Fig. 2 and Fig. 3. Fig. 2 shows a cladogram, which is only concerned about the evolutionary relationships of the items, whereas Fig. 3 is a tree that also suggests a relative time of origin of the various items. The script evolutionary trees in Fig. 2 and Fig. 3 were generated using the UPGMA method [23].

VI. Discussion of the Results

The results shown in Fig. 2 and Fig. 3 suggest that the nine scripts had a common ancestor from which two branches descend. These branches are as follows:

1. Cretan Hieroglyph, Linear A, Linear B, Cypriot, Old Hungarian, Tifinagh
   - 1.1 Tifinagh
   - 1.2 Cretan Hieroglyph, Cypriot, Linear A, Linear B, Old Hungarian
     - 1.2.1 Linear A, Linear B
     - 1.2.2 Cretan Hieroglyph, Cypriot, Old Hungarian
2. Greek, Phoenician, South Arabic

Fig. 4 presents a hypothetical spread of the scripts. It appears that Crete was the birthplace of a family of ancient scripts. The first branch of this script family is indicated by red and the second branch is indicated by blue in Fig. 4.

According to our script evolution hypothesis, an original native Cretan script separated into two branches. Within Branch 1, sub-branch 1.1 spread to Northern Africa and was the originator of the Tifinagh alphabet [31]. Sub-branch 1.2 can be further divided into two groups. The first group (1.2.1) stayed in Crete and diversified into Linear A and Linear B. The second group (1.2.2) included Cretan Hieroglyph which likely spread from Crete to the costal regions of western Anatolia and later split into a northern group that reached the Black Sea area and included Old Hungarian and into an eastern group that reached Cyprus and developed into the Cypro-Minoan and later the Cypriot syllabary scripts.

Branch 2 shows a close similarity between Greek and Phoenician, which is due to the widely recognized ancient Greek adoption of the Phoenician script [27] centuries after Linear A and Linear B both went out of use. Phoenician and South Arabic are also supposed to have a common ancestor called Proto-Sinaitic [29]. Proto-Sinaitic could be the common root of the Greek, Phoenician and South Arabic scripts.

The script evolutionary trees shown in Fig. 2 and Fig. 3 help to settle a debate regarding the origin of Old Hungarian. The figures show that Old Hungarian and Phoenician belong to different branches of the script evolutionary tree. That result contradicts the view presented in Hosszú [11] that Old Hungarian is a late, c. 7th century, derivative of Phoenician. Instead, Fig. 2 and Fig. 3 support the view of Forrai [8] and Varga [25] that Old Hungarian is a Bronze Age script that does not derive from Phoenician. The script evolutionary tree shows that Tifinagh is also not a derivative of Phoenician.

The evolutionary trees in Fig. 2 and Fig. 3 also help illuminate the debate on the origin of the ancient Ugaritic abjad (consonants only alphabet). Colless [3] claims the Ugaritic script is derived from Proto-Sinaitic, which would place Ugaritic in Branch 2 of the script evolutionary tree. On the other hand, Naddeo [13] suggested a relationship between the Ugaritic abjad and the Old Hungarian script. Based on Naddeo's observations, we indicated by a dashed line in Fig. 4 the possibility of the spread of the Cypriot syllabary, or more likely the earlier closely related Cypro-Minoan syllabary to the Ugarit area. Cyprus and Ugarit are very close together in the Eastern Mediterranean. Hence a strong contact between the two locations can be expected. A spread from Cyprus to Ugarit would still imply a close relationship between Old Hungarian and Ugaritic because both Cypriot and Old Hungarian belong to Branch 1.2.2 of the script evolutionary tree. However, more research needs to be done to decide where exactly the Ugaritic script can be placed into the script evolutionary tree.

The Cretan script family outlined in this paper includes as a subfamily all derivatives of the Phoenician alphabet. The Phoenician script family includes the Latin alphabet, which was widely adopted by the speakers of many different languages from many language families. Script families and language families do not necessarily overlap. Hence the languages of particular Cretan Hieroglyph and Linear A writings remain undecided.

VII. Conclusions and Future Work

The application of bioinformatics evolutionary tree algorithms to the study of script evolution is a novel idea of this paper as well as the exact similarity measure for pairs of script symbols and the DNA encoding of scripts. It is an interesting future work to apply our methods to the study of other script families.

Our study has also implications for the decipherment of the unknown Cretan Hieroglyph and Linear A scripts. In particular, as a first step, the sound values of these unknown scripts need to be found. In Table I, the sound values correspond well for Phoenician, South Arabic, Greek, Old Hungarian and Tifinagh. However, the Linear B sound values are often markedly different. It has been attempted to read Cretan Hieroglyph and Linear A scripts using Linear B sound values without any fruitful result. Our script evolutionary trees suggest that the sound values of the Cretan Hieroglyph script symbols may be closer to the sound values of the corresponding Cypriot and Old Hungarian script symbols. In addition, the sound values of the Linear A symbols may be reconstructed by finding the possible common ancestor sound values of the corresponding Phoenician, Greek, South Arabic, Old Hungarian and Tifinagh alphabet symbols. In both cases, the newly predicted sound values may correspond better than the Linear B sound values to the actual sounds of the Cretan Hieroglyph and the Linear A script symbols. We hope that this realization will open a new phase in the understanding of the ancient Cretan scripts. In fact, using the new sound values, we already gave a tentative translation of the Phaistos Disk [17].

REFERENCES

[1] D. Baum and S. Smith, *Tree Thinking: An Introduction to Phylogenetic Biology*, Roberts and Company Publishers. 2012.

[2] J. Chadwick, *The Decipherment of Linear B*, Cambridge University Press, 1958.

[3] B. Colless, "Cuneiform alphabet and picto-proto-alphabet," https://sites.google.com/site/collesseum/cuneiformalphabet, downloaded July 5, 2015.

[4] Y. Duhoux, "How not to decipher the Phaistos Disc," *American Journal of Archaeology*, Vol. 104, No. 3 (2000), pp. 597–600.

[5] A. J. Evans, *Scripta Minoa: The Written Documents of Minoa Crete with Special Reference to the Archives of Knossos*, Volume II, Classic Books, 1909.

[6] J. Faucounau, *Le Déchiffrement du Disque de Phaistos: Preuves et conséquences*. L'Harmattan, Paris/Montreal 1999.

[7] S. R. Fisher, *Glyph-Breaker*, Springer, 1997.

[8] S. Forrai, The Old Hungarian Writing from Ancient Times to the Present, (in Hungarian), Antológia Kiadó, 1994.

[9] B. G. Hall, *Phylogenetic Trees Made Easy: A How to Manual*, 4th edition, Sinauer Associates, 2011.

[10] J. T. Hooker, *Linear B: An Introduction*, Bristol Classical Press, 1980.

[11] G. Hosszú, *Heritage of Scribes: The Relation of Rovas Scripts to Eurasian Writing Systems*, Rovas Foundation Hungary, 2013.

[12] P. Lerney, M. Salemi, and A.-M Vandamme, editors. *The Phylogenetic Handbook: A Practical Approach to Phylogenetic Analysis and Hypothesis Testing*, 2nd edition, Cambridge University Press, 2009.

[13] M. Naddeo, *The Ugarit Abjad ... A Rovás Alphabet*, self-published book, 2007.

[14] P. Z. Revesz, *Introduction to Databases: From Biological to Spatio-Temporal*, Springer, New York, 2010.

[15] P. Z. Revesz, "An algorithm for constructing hypothetical evolutionary trees using common mutations similarity matrices," *Proc. 4th ACM International Conference on Bioinformatics and Computational Biology*, ACM Press, Bethesda, MD, USA, September 2013, pp. 731-734.

[16] P. Z. Revesz, "A computational study of the evolution of Cretan and related scripts," *Proc. 3rd International Conference on Applied Mathematics, Computational Science and Engineering*, Agios Nikolaos, Crete, Greece, October 2015, pp. 21-25.

[17] P. Z. Revesz, "A Computational translation of the Phaistos Disk," *Proc. 3rd International Conference on Applied Mathematics, Computational Science and Engineering*, Agios Nikolaos, Crete, Greece, October 2015, pp. 53-57.

[18] P. Z. Revesz and C. J.-L. Assi, "Data mining the functional characterizations of proteins to predict their cancer relatedness," *International Journal of Biology and Biomedical Engineering*, 7 (1), 2013, pp. 7-14.

[19] P. Z. Revesz and T. Triplet, "Classification integration and reclassification using constraint databases," *Artificial Intelligence in Medicine,* 49 (2), 2010, pp. 79-91.

[20] P. Z. Revesz and T. Triplet, "Temporal data classification using linear classifiers," *Information Systems*, 36 (1), 2011, pp. 30-41.

[21] N. Saitou and M. Nei, "The neighbor-joining method: A new method for reconstructing phylogenetic trees," *Molecular Biological Evolution*, 4, 1987, pp. 406-425.

[22] M. Shortridge, T. Triplet, P. Z. Revesz, M. Griep, and R. Powers, "Bacterial protein structures reveal phylum dependent divergence," *Computational Biology and Chemistry*, 35 (1), 2011, pp. 24-33.

[23] R. R. Sokal, and C. D. Michener, "A statistical method for evaluating systematic relationships," *University of Kansas Science Bulletin*, 38, 1958, pp. 1409-1438.

[24] T. Triplet, M. Shortridge, M. Griep, J. Stark, R. Powers, and P. Z. Revesz, "PROFESS: A protein function, evolution, structure and sequence database," *Database -- The Journal of Biological Databases and Curation*, 2010, Available: http://database.oxfordjournals.org/content/2010/baq011.full.pdf+html

[25] G. Varga, *Bronzkori Magyar Irásbeliség*, Irástörténeti Kutató Intézet publication, 1993.

[26] Wikipedia, "Cypriot syllabary," downloaded October 20, 2015. Available: https://en.wikipedia.org/wiki/Cypriot_syllabary

[27] Wikipedia, "History of the Greek alphabet," downloaded July 6, 2015. Available: https://en.wikipedia.org/wiki/History_of_the_Greek_alphabet

[28] Wikipedia, "Phoenician alphabet," downloaded July 6, 2015. Available: https://en.wikipedia.org/wiki/Phoenician_alphabet

[29] Wikipedia, "Proto-Sinaitic script," downloaded July 6, 2015. Available: https://en.wikipedia.org/wiki/Proto-Sinaitic_script

[30] Wikipedia, "South Arabian alphabet", downloaded July 5 2015. Available: https://en.wikipedia.org/wiki/South_Arabian_alphabet

[31] Wikipedia, "Tifinagh", downloaded October 20, 2015. Available: https://en.wikipedia.org/wiki/Tifinagh

[32] J. G. Young, "The Cretan Hieroglyphic script: A review article," *Minos* 31-32 (1996-1997[1999]) 379-400.

[33] S. Zhang and T. Wang, "A new distance-based approach for phylogenetic analysis of protein sequences," *International Journal of Biology and Biomedical Engineering*, 3(3), 2009, pp. 35-42.

Peter Z. Revesz holds a Ph.D. degree in Computer Science from Brown University. He was a postdoctoral fellow at the University of Toronto before joining the University of Nebraska-Lincoln, where he is a professor in the Department of Computer Science and Engineering. Dr. Revesz is an expert in databases, data mining, big data analytics and bioinformatics. He is the author of *Introduction to Databases: From Biological to Spatio-Temporal* (Springer, 2010) and *Introduction to Constraint Databases* (Springer, 2002). Dr. Revesz held visiting appointments at the IBM T. J. Watson Research Center, INRIA, the Max Planck Institute for Computer Science, the University of Athens, the University of Hasselt, the U.S. Air Force Office of Scientific Research and the U.S. Department of State. He is a recipient of an AAAS Science & Technology Policy Fellowship, a J. William Fulbright Scholarship, an Alexander von Humboldt Research Fellowship, a Jefferson Science Fellowship, a National Science Foundation CAREER award, and a "Faculty International Scholar of the Year" award by *Phi Beta Delta*, the Honor Society for International Scholars.