Patterns of Equidistant Letter Sequence Pairs in Genesis

Harold J. Gans
Senior Cryptologic Mathematician(retired)
United States Department of Defense
Fort George G. Meade, Maryland

Zvi Inbal
Department of Research and Development
Dynamic Laboratories Ltd.
Ashdod, Israel

Nachum Bomboch
Department of Management
Bombach Schiff and Company
Ramat-Gan, Israel

Abstract

Witztum, Rips, and Rosenberg [12] provide strong statistical evidence for the non random coincidence (patterns) of equidistant letter sequence (ELS) pairs in the standard Hebrew text of the book of Genesis. Specifically, they show that among all the chosen appellations of a famous Jewish personality, and among the common forms of expressing a Hebrew date of birth or death of that personality, it is likely that the collection of low skip rank appellation-date ELS pairs will form a more compact geometric pattern on the surface of a cylinder than is expected by chance.

This result has been challenged on the basis that it is difficult to verify that all components of the experiment were completely a priori. We present a new experiment that is algorithmically structured so as to be demonstrably a priori, and which extends the original results. A significance level of 4×10^{-6} *is obtained for this experiment.*

1. Introduction

In [12], Witztum et al. (henceforth referred to as "WRR") provide strong statistical evidence for the non random coincidence of equidistant letter sequence (ELS) pairs in the standard Hebrew text of the book of Genesis (G). An ELS in G, (n, d, k), is defined as a sequence of letters in G found at positions $n, n+d, n+2d, \ldots, n+(k-1)d$. d is called the "skip distance". WRR show that among all the chosen appellations of a famous Jewish personality, and among the common forms of expressing a Hebrew date of birth or death of that personality, it is likely that the collection of low skip rank appellation-date ELS pairs will form a more compact geometric pattern on the surface of a cylinder than is expected by chance. The list of personalities, appellations, and matching dates selected for this study is referred to as "list 2" and is extracted from an encyclopedia of famous Jewish personalities [8], which we shall refer to as "ME". This sample is called "list 2" to distinguish it from an earlier disjoint sample, "list 1", which also appears in WRR and is extracted from ME. The p-level obtained by WRR for the compactness phenomenon on list 2 was 1.6×10^{-5}.

This result has been challenged [9] on the basis that it is difficult to verify that all components of the experiment were a priori (see [10] for a response to this challenge). The experiment consists of 4 components, viz:

1. The standard ("Textus Receptus") Hebrew text of Genesis.

2. The list of personalities derived from ME and appellations and spellings for these personalities. The appellations were obtained from the ME and reviewed/corrected by an expert, S. Z. Havlin. The spelling rules were provided by an expert, Y. Orbach. Appellations are limited to be between five and eight letters long.

3. A list of Hebrew dates of birth and death for the personalities taken from the ME (with some corrections). They are also limited to be between five and eight letters long.

4. A feature measuring the compactness of the geometric patterns formed by pairs of ELSs taken from 2 and 3 above respectively and a technique for computing the p-level of the measures obtained.

We design a new experiment in which the personalities, along with their appellations and spellings from list 1 and list 2 are combined. Components 1 and 4 of the experiment are identical to WRR. This automatically guarantees that components 1,2 and 4 are completely a priori. Thus, all questions concerning the validity, appropriateness or a priori nature of these components in WRR becomes irrelevant to the current study except that if these components are not valid, the current experiment is expected to fail. Note that it is not the objective of this study to explain or justify the various components of the WRR experiment, but rather to use them "as is" so that they are undeniably a priori in the current study. The objective is to validate the Torah codes phenomenon, not the particular choices made by WRR. For component 3, we use a list of the Jewish communities of birth and death of the personalities, as opposed to dates. So as to guarantee that the list of communities and their Hebrew spellings are correct as well as a priori, we use an algorithm to derive the information from the ME, and use the Encyclopedia Hebraica (EH) [1] instead of expert consultants to determine names and spellings. Every entry for the list is obtained by this linguistic protocol (LP) without exception, and can be checked for accuracy. The algorithm, too, can be checked for linguistic accuracy and completeness. Thus, the individual community names are not subject to errors in judgment or the non uniform application of vague rules. Furthermore, the data is completely reproducible. It is important to note that although it is possible to manipulate the list of personality appellations and spellings in WRR by not following strict a priori guidelines so as to artificially obtain a small p-level when the experiment is done on a text other than G (see [9] and [10]), it has never been demonstrated that it is possible to construct a linguistically correct LP that will do the same.

2. The Compactness Measure

We summarize the definition of the compactness measure from WRR. For further details and motivation, see [12]. Given two ELSs, $e = (n, d, k), e' = (n', d', k')$ in G, $\delta_h(e, e')$, is defined by writing G as a single helix of letters spiraling down a cylinder with h vertical columns of letters and setting $\delta_h(e, e') = f^2 + f'^2 + g^2$, where f is the usual Euclidean distance (in columns and rows of letters) between two consecutive letters of e on the surface of the cylinder, f' is the same for e', and g is the minimal Euclidean distance between a letter of e and a letter of e' on the cylinder. Then $\mu_h(e, e') = 1/\delta_h(e, e')$ is directly related to the compactness of the configuration of e and e' on the cylinder for given h. In general, setting $h = h(i) =$ the nearest integer to $|d|/i$ tends to make f small for small i, so we let $h(i) =$ nearest integer to $|d|/i$, and $h'(i) =$ nearest integer to $|d'|/i$ and define

$$\sigma(e, e') = \sum_{i=1}^{10} \mu_{h(i)}(e, e') + \sum_{i=1}^{10} \mu_{h'(i)}(e, e') \qquad (1)$$

Note that $\sigma(e, e')$ tends to be large provided that there is a relatively compact configuration of e and e' on a cylinder with $h(i)$ or $h'(i)$ columns for at least one $h(i)$ or $h'(i), i = 1, \ldots, 10$.

Suppose the letters of a word W are found as an ELS $e = (n, d, k)$ in G with $|d| \geq 2$. Let us define $e \subset T$ to mean that the letters of e are contained in the segment T of G. We define T_e as the maximal segment of G such that if $e' = (n', d', k)$ has the same letters as e then

$$e = (n, d, k) \subset T_e \text{ and } e' \subset T_e \text{ implies } |d| \leq |d'|.$$

We say that e is minimal in T_e. Let $\lambda(T)$ denote the length of a segment T of G. We then define $\omega(e, e') = \lambda(T_e \cap T_{e'})/\lambda(G)$. $\omega(e, e')$ is the fraction of G in which both e and e' are minimal.

Let $D(W)$ be the largest d for an ELS, (n, d, k), spelling the word W such that the expected cardinality of $\{(n, d, k)|2 \leq |d| \leq D(W)\}$ is less than or equal to 10. (See [12] for the explicit computation of $D(W)$). We define $\Omega(W, W') = \sum \omega(e, e')\sigma(e, e')$ where the summation is taken over all ELSs $e = (n, d, k)$ and $e' = (n', d', k')$ spelling W and W' respectively, such that $2 \leq |d| \leq D(W)$ and $2 \leq |d'| \leq D(W')$. $\Omega(W, W')$ is the unnormalized compactness measure of the patterns of pairs of ELSs for W and W' respectively. $\Omega(W, W')$ incorporates both an aggregate measure of compactness over the set of patterns formed by the ELS pairs, as well as a minimality constraint on the skip distances of the ELSs. We now normalize $\Omega(W, W')$ by defining an (x, y, z)-perturbed ELS, $(n, d, k)^{(x,y,z)}$, where x, y, and $z \in \{-2, -1, 0, 1, 2\}$, as the letter sequence in G at positions $n, n + d, \ldots, n + (k - 4)d, n + (k - 3)d + x, n + (k - 2)d + x + y, n + (k - 1)d + x + y + z$. $\delta_h((n, d, k)^{(x,y,z)}, (n', d', k')^{(x,y,z)})$ is defined in the same way as $\delta_h((n, d, k), (n', d', k'))$ is defined and in which f and f' are the Euclidean distances between the unperturbed letters of the two perturbed ELSs respectively. Using the same definitions for μ_h, σ and ω, with perturbed ELSs, we obtain $\Omega^{(x,y,z)}(W, W')$. Note that $\Omega^{(0,0,0)}(W, W') = \Omega(W, W')$.

For (n, d, k) an ELS of W in G, let

$$M(W, W') = \{(x, y, z) \quad | \quad \exists (n, d, k)^{(x,y,z)} \text{ of } W \text{ in } G \text{ and } \exists (n', d', k')^{(x,y,z)} \text{ of } W' \text{ in } G\}$$

and let $m(W, W') = card\{M(W, W')\}$. Note that $m(W, W') \leq 125$. If $(0, 0, 0) \in M(W, W')$ then we define $\upsilon(W, W') = card\{(x, y, z) \in$

$M(W,W')|\Omega^{(x,y,z)}(W,W') \geq \Omega(W,W')\}$. If $m(W,W') \geq 10$ then $c(W,W')$ is defined as $\upsilon(W,W')/m(W,W')$. Note that $c(W,W')$ resembles a normalization of $\Omega(W,W') : 1/125 \leq c(W,W') \leq 1$.

3. The Significance Level of the Compactness Measure

We compute two statistics, ρ_3 and ρ_4, following the notation of [12]. ρ_3 and ρ_4 are defined on the subset, Q, of lists 1 and 2 formed by omitting all appellations that begin with the title "Rabbi". All the personalities in Q have unique appellations whereas this is not the case for the full set of data. We define P_3 as

$$P_3 = \sum_{i=k}^{N} \binom{N}{i} (0.2)^i (0.8)^{N-i} \tag{2}$$

where $k = card\{c(W,W')|c(W,W') \leq 0.2\}$ and $N = card(Q)$. Note that if $c(W,W')$ were independent random variables uniformly distributed on $[0,1]$, then P_3 would be the binomial probability that at least k of the $c(W,W')$ would be less than or equal to 0.2. We define P_4 as $F^N(\prod c(W,W'))$, where $F^N(X) = X(1 - \ln X + (-\ln X)^2/2! + \cdots + (-\ln X)^{N-1}/(N-1)!)$. The $c(W,W')$ are computed for all pairs of W and W' in Q, where W is an appellation of a personality in Q and W' is a name of a community of birth or death for that personality in Q. Note that if the $c(W,W')$ were independent random variables uniform on $[0,1]$ then the probability that $\prod(c(W,W') \leq x$ is equal to $F^N(x)$ (reference Eq.(3.5) in [6]). However, no such assumption is made for either P_3 or P_4; this is merely the motivation for the definitions. To calculate the significance level, $999,999$ pseudo-random permutations, π_i, of the 66 personalities are produced, each permutation thus forming, for each personality in Q, a pseudo-random matching of that personality with the set of communities associated with a (usually different) personality in Q. Each of these permutations, π_i determines statistics $P_3^{\pi_i}$ and $P_4^{\pi_i}$. Then $\rho_3 = (card\{\pi_i|P_3^{\pi_i} \leq P_3\} + 1)/10^6$ is the probability under the null hypothesis that P_3 would rank as low as it is among the $P_3^{\pi_i}$ and similarly for ρ_4. For the communities experiment, the first 9 digits of π were used as the random seed for the pseudo-random number generator.

4. The Linguistic Protocol

The objective of the linguistic protocol (LP) is to derive, for each personality in lists 1 and 2, locations of birth and death, the Jewish community names, and their spellings in an algorithmic unambigous way to the exclusion of linguistic or historical expert consultants. In this way, the LP produces a reproducible data set consisting of accurate locations, place names, and spellings using the ME as a primary and default source of data and the EH as an "expert" instead of a consultant. Spelling rules are in conformity with those described in WRR and its preprint [11].

We present a broad outline of the LP; the exact details and sources can be found in [7]. The LP consists of three parts. Part I determines the geographic location of the place of birth or death. The information is obtained from the ME unless it is not there, or is in conflict with the EH, in which case it is obtained from the EH. Part II determines the Jewish name(s) of the place obtained in part I. These are often different than the secular names. The Jewish names are usually obtained from the EH article on the location, or the index. If no Jewish name is found, then the names obtained from part I are used. Part III determines the spellings of the names obtained. Jewish names are taken as obtained in part II. For other names, the EH index, article on the personality or ME are used. Specific spellings rules are then applied to ensure that the spelling is consistent with the spelling used in the era in which the personality lived and died (as opposed to modern spellings). Spelling rules also introduce valid variations of the spelling as determined by the WRR spelling rules and the EH usage. Finally, a number of prefixes that specifically designate the Jewish community within the secular city are methodically added to the names.

The linguistic protocol and the data generated for the communities experiment using the LP can be found at [7]. In [7] we also include concise indicators associated with each personality, community name, and spelling, which show precisely how the word was obtained using the LP. This facilitates checking the data for accuracy.

5. The Experiment and the Results

The Communities experiment was performed as described above. The list of personalities, their appellations, and their spellings were taken exactly as they appear in lists 1 and 2 in WRR. The list of corresponding Jewish communities of birth and death and their spellings were obtained via application of the LP as described in Section 4 and in [7]. The measure of compactness and the procedure for obtaining the p-level are described in Sections 2 and 3 and are exactly the same as in WRR. Two of the WRR measures, P_3, and P_4, were obtained. Row C in Table 1 gives the associated p-levels for these measures. Row QQ shows the result of adding one new prefix component, קק, to the LP. This addition was suggested by some linguists after the experiment was completed. (This prefix is a commonly used prefix meaning "holy community.") A description of this addition is given in [7], along with the description of the LP.

Thirty control experiments were done in which the letters of each ELS were pseudo-randomly permuted. Identical words in the lists were subjected to the same permutation so as to preserve the dependencies induced by these repeats. The smallest p-level obtained among the 60 p-levels computed was 0.009 for P_4. Row F gives the p-levels obtained by combining the p-levels using Fisher's method [5] on the 30 P_3 and 30 P_4 values. Row A^2 gives the p-levels for the Anderson-Darling statistics [4] on the 30 P_3 and 30 P_4 values. Row D shows the p-levels obtained for a control experiment in which 78 letters (one for each $1,000$ letters) were randomly dropped, and then randomly reinserted in G. Row I shows the p-levels obtained for a control in which successive chapters in G were interchanged. It should be noted that these control texts provide a far better control than, say, a Hebrew translation of "War and Peace" as used in some other studies. This is because these control texts resemble Genesis more closely than any other standard Hebrew text can.

Table 1. Results

	P_3	P_4
C	4×10^{-6}	2×10^{-6}
QQ	4.2×10^{-5}	4.4×10^{-5}
F	0.60	0.51
A^2	0.54	0.48
D	0.20	0.24
I	0.31	0.35

6. Conclusions

We conclude that:

1. The compactness of patterns formed on the surface of a cylinder by ELSs of a priori selected famous Jewish personalities and ELSs of their communities of birth or death is smaller than can be attributed to chance. Specifically, application of the Bonferroni inequality yields a p-level of 4×10^{-6} against the null hypothesis of random distribution of the compactness measures. The same conclusion is obtained with the prefix component קק added to the LP. In this case Bonferroni yields a p-level of 8.4×10^{-5}.

2. When the ELSs or the text is randomized, the WRR procedure produces random p-values uniformly distributed on $[0, 1]$.

3. It is highly likely that the list of appellations and spellings of the personalities was a priori for WRR. For had the data been crafted specifically to produce compact configurations between ELSs of the appellations and ELSs of the dates, then one would not expect those same personality ELSs to form compact configurations with a new data set, the communities, as well.

7. Historical Note

The first communities experiment was completed by the first two authors about 15 years ago. It differed from the experiment described here in the following ways only: (a) there were a handful of errors in application of the LP in the earlier experiment. These were corrected for the current experiment. (b) There was one case of a typographical error in ME which caused an error in the output of the LP (for personality 23). This error was corrected in the earlier list. It was not corrected in the current experiment so that there would be no exceptions to the uniform application of the LP. (c) A probabilistic simulation was used instead of the perturbation method used by WRR in calculating the compactness. (d) Only P_4 was calculated; not P_3. The p-level obtained for this earlier experiment was 5.0×10^{-6}.

More recently, two new communities experiments were performed by a committee [2]. Both produced random results. It is, however, important to note that these experiments had the list of communities and their spellings produced by consultants rather than implementing an algorithmic protocol such as the LP. Indeed, the consultants used to prepare the data as well as some of the committee members who designed the experiments indicate that the data preparation did not always follow the design protocol and that the data contains numerous errors [3]. Thus, the failures of these experiments are not at all relevant to the veracity of the conclusions presented here.

References

[1] *Encyclopedia Hebraica*. Encycopedia Publishing company, Ltd., Jerusalem, 1981.

[2] DP 364 of the center for the study of rationality. http://ratio.huji.ac.il/dp/dp364.pdf, June 2004.

[3] R. Aumann, H. Furstenberg, I. Lapides, and D. Witztum. Analysis of the "Gans" committee report. http://www.Ma.huji.ac.il/ raumann/pdf/dp_365.pdf, pp.6-9.

[4] R. D'Agostino and M. Stephens. *Goodness-of-fit Techniques*, pages 100–101. Marcel-Dekker, Inc.

[5] R. Elston. On Fisher's method of combining p-levels. *Biometrical Journal*, 33:339–345.

[6] W. Feller. *An Introduction to Probability Theory and its Applications*, volume 2. Wiley and Sons, New York, London, Sydney, 1966.

[7] H. Gans, Z. Inbal, and N. Bombach. Patterns of equidistant letter sequence pairs in Genesis. http://www.torahcodes.org/cities/communities_ data.pdf.

[8] M. Margalioth. *Encyclopedia of Great Men in Israel, a Bibliographical Dictionary of Jewish Sages and Scholars from the 9th to the End of the 18th Century*. Joshua Chachik Publishing House Ltd., Tel Aviv, 1961.

[9] B. McKay, D. B. Natan, M. B. Hillel, and G. Kalai. Solving the bible codes puzzle. *Statistical Science*, pages 149–173, May 1999.

[10] D. Witztum. http://www.torahcodes.co.il/ref2.htm.

[11] D. Witztum, E. Rips, and Y. Rosenberg. Equidistant letter sequences in the book of Genesis. Reprint, p13, http://www.torahcodes.org/cities/wrr_preprint_pg_130001.pdf, 1986.

[12] D. Witztum, E. Rips, and Y. Rosenberg. Equidistant letter sequences in the book of Genesis. *Statistical Science*, 9(3):429–438, August 1994.