J Mol Evol (1993) 37:662–664

Journal of Molecular Evolution

Letter to the Editor

Further Results on Error Minimization in the Genetic Code

Nick Goldman

Laboratory of Mathematical Biology, National Institute for Medical Research, The Ridgeway, Mill Hill, London NW7 1AA, England

Received: 23 August 1992 / Revised: 27 May 1993 / Accepted: 19 June 1993

Haig and Hurst (1991) recently examined whether the genetic code is in some respects optimal, or near optimal, with regard to minimizing the effects of errors in amino acid assignments. They considered four attributes of amino acids—namely, polar requirement, hydropathy, molecular volume, and isoelectric point—and found that if they generated random codes only two out of 10,000 were more conservative with respect to polar requirement than the existing (natural) genetic code. For the other attributes, the existing code was not notably superior to the random codes. These results were taken to imply that polar requirement was particularly important during the early evolution of the genetic code.

Haig and Hurst (1991) quantified how conservative a code is by the mean square difference in amino acid characteristic, MS_0, where the mean is taken over all single base mutations in the genetic code (discounting mutations to or from stop codons). Low values of MS_0 imply a more conservative code. The polar requirement for the 20 amino acids was taken from Woese et al. (1966) and was also considered to be relevant to studies of the origins of the genetic code by (e.g.) Di Giulio (1989) and Szathmáry and Zintzaras (1992). Using this characteristic, the natural genetic code has MS_0 = 5.194. Haig and Hurst randomly redistributed the 20 amino acids of the genetic code while maintaining the existing "block structure" of synonymous codons and the positions of the three stop codons. Under this "fixed-block" model, the two "superior" codes found by Haig and Hurst among 10,000 random trials gave MS_0 = 5.167 and MS_0 = 5.189.

It is not feasible to calculate MS_0 for all 20! ($>2 \times 10^{18}$) codes under the fixed-block model: to search for the most conservative code, heuristic algorithms must be used instead. One possibility is the well-known simulated annealing algorithm (Kirkpatrick et al. 1983). Computationally simpler is the "record-to-record travel" (RRT) algorithm of Dueck (1992). In the RRT algorithm, MS_0 is calculated for an initial trial solution. The trial solution is then altered slightly, and MS_0 is recalculated. The new solution is accepted if its score is within a certain distance of the best score attained so far. The process is iterated until the best score has not changed for a long time. Good results are obtained by running the algorithm a number of times, starting each run with a different initial trial solution. The RRT algorithm has performed well in combinatorial optimization problems such as the travelling salesperson problem (Dueck 1992).

I have applied the RRT algorithm to the question of finding the best code under the fixed-block model. Initial trial solutions were selected randomly from the 20! possibilities, and alterations were made by randomly selecting two synonymous codon sets and swapping their associated amino acids. The best solution found is shown in Fig. 1, and has MS_0 = 3.489. This is approximately 4 standard deviations from the mean of Haig and Hurst's 10,000

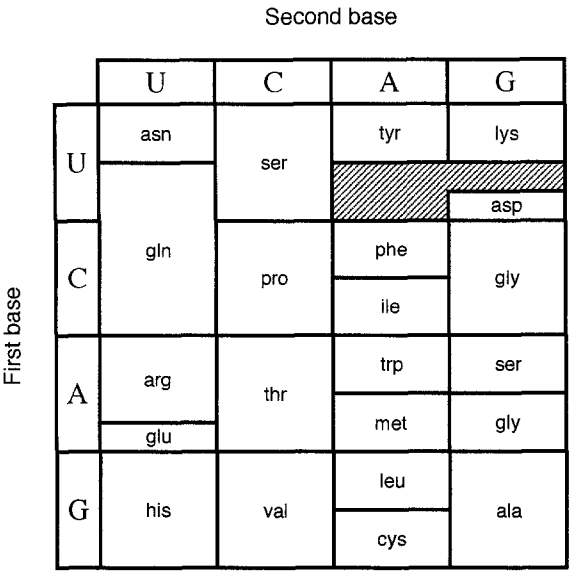

Fig. 1. The most conservative code found by the RRT algorithm under the fixed-block model (MS_0 = 3.489).

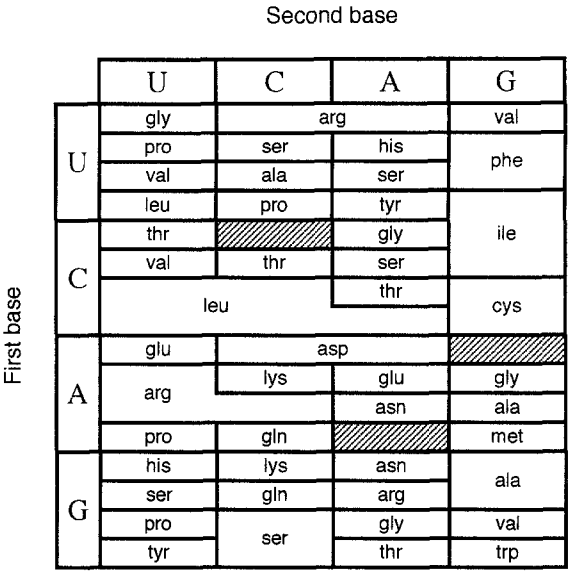

Fig. 2. The most conservative code found by the RRT algorithm under the shuffled-codon model (MS_0 = 4.005).

random codes. Almost every run of the RRT algorithm produced codes more conservative than the natural genetic code. Viewed in the context of an evolutionary optimization problem, the natural genetic code is both (1) far from optimal and (2) easily improved.

In the calculation of MS_0, all single base changes are given equal weight. Scores are invariant to relabeling of the three codon positions, and so the retention of the natural synonymous codon sets in the fixed-block model is in a sense arbitrary. In light of this, it is natural to investigate the effect of varying the block structure of the natural genetic code. To do this, I propose a different model. Instead of varying the assignment of the 20 amino acids to the 20 synonymous codon sets of the natural genetic code, the new model varies the assignment of the 61 amino acids and three stop codons of the natural genetic code to the 64 codons of the three-letter code. In other words, the 64 amino acids and stop codons of the natural genetic code are randomly "shuffled," retaining the same numbers of triplets coding for each (3 stop codons, 2 × phe, 6 × leu, etc.). By doing this, we investigate the effect of the block structure of such codes and remove effects of different numbers of triplets coding for the same amino acids.

Following Haig and Hurst, a large number of random codes were generated under this "shuffled-codon" model; 100,000 random codes gave a mean MS_0 value of 9.37, with standard deviation 0.54. Interestingly, the mean is close to that under the fixed-block model (9.41). Every one of the 100,000 random codes was less conservative than the natural genetic code, whose value MS_0 = 5.194 is more than 7 standard deviations from the mean. This gives clear evidence of the importance of the "block structure" of synonymous codons in the natural genetic code.

Under this model there are $64!/(2!)^9(3!)^2(4!)^8(6!)^3$ ($>10^{65}$) different codes, discounting arbitrary reorderings of equivalent amino acids and relabelings of codon positions. The most conservative code I have found by the RRT algorithm, shown in Fig. 2, has MS_0 = 4.005, approximately 10 standard deviations better than the mean. The improvement over the natural genetic code appears to be due to synonymous codons that differ by single bases having these bases spread more evenly amongst first, second, and third positions. By comparison, the natural genetic code utilizes predominantly third base redundancy. This is shown by the values of MS_1, MS_2, and MS_3, the contributions to MS_0 from changes at first, second, and third codon positions respectively. The code of Fig. 2 has MS_1 = 3.06, MS_2 = 3.67, MS_3 = 5.28—more evenly distributed than the natural genetic code's values of 4.88, 10.56, 0.14, respectively. Under the shuffled-codon model, we again see that the natural genetic code is far from optimal and easily improved.

In conclusion, consideration of the optimal genetic codes (or as near optimal as have been found using heuristic algorithms) under these models provides further evidence that the natural genetic code is conservative with respect to polar requirement. Such results support the hypothesis that codon assignments have evolved to minimize the effects of translation errors. However, the fact that the natural genetic code is far from optimal under both the fixed-block and shuffled-codon models suggests that care must be exercised if the evolution of the genetic code is to be considered in the con-

text of error minimization. If we assume evolution to favor increasingly conservative codes, we must acknowledge that the assignment of amino acids to synonymous codon sets, and the very existence of the observed synonymous codon sets, are being constrained by some as-yet-unmodeled factors which may have significant bearing on Haig and Hurst's (1991) comment that "the translational apparatus would be expected to evolve an inverse relationship between the frequency and severity of an error."

Acknowledgments. I am grateful to members of the Laboratory of Mathematical Biology, National Institute for Medical Research, for helpful discussions on this topic.

References

Di Giulio M (1989) The extension reached by the minimization of the polarity distances during the evolution of the genetic code. J Mol Evol 29:288–293

Dueck G (1992) New optimisation heuristics: the great deluge algorithm and the record-to-record travel. J Comput Phys 104:86–92

Haig D, Hurst LD (1991) A quantitative measure of error minimization in the genetic code. J Mol Evol 33:412–417

Kirkpatrick S, Gelatt CD Jr, Vecchi MP (1983) Optimization by simulated annealing. Science 220:671–680

Szathmáry E, Zintzaras E (1992) A statistical test of hypotheses on the organization and origin of the genetic code. J Mol Evol 35:185–189

Woese CR, Dugre DH, Dugre SA, Kondo M, Saxinger WC (1966) On the fundamental nature and evolution of the genetic code. Cold Spring Harbor Symp Quant Biol 31:723–736