THE ROYAL SOCIETY

Received 11 October 2002
Accepted 27 January 2003
Published online 11 April 2003

Optimal alphabets for an RNA world

Paul P. Gardner[1,3*], Barbara R. Holland[1,3], Vincent Moulton[4], Mike Hendy[1,3] and David Penny[2,3]

[1]*Institute of Fundamental Sciences, and* [2]*Institute of Molecular Biosciences, Massey University, PB 11 222, Palmerston North, New Zealand*
[3]*Allan Wilson Centre for Molecular Ecology and Evolution, Massey University, Auckland, New Zealand*
[4]*Linnaeus Center for Bioinformatics, Uppsala University, Box 598, 751 24 Uppsala, Sweden*

Experiments have shown that the canonical AUCG genetic alphabet is not the only possible nucleotide alphabet. In this work we address the question 'is the canonical alphabet optimal?' We make the assumption that the genetic alphabet was determined in the RNA world. Computational tools are used to infer the RNA secondary structure (shape) from a given RNA sequence, and statistics from RNA shapes are gathered with respect to alphabet size. Then, simulations based upon the replication and selection of fixed-sized RNA populations are used to investigate the effect of alternative alphabets upon RNA's ability to step through a fitness landscape. These results show that for a low copy fidelity the canonical alphabet is fitter than two-, six- and eight-letter alphabets. In higher copy-fidelity experiments, six-letter alphabets outperform the four-letter alphabets, suggesting that the canonical alphabet is indeed a relic of the RNA world.

Keywords: genetic systems; ribozymes; systematic evolution of ligands by exponential amplification (SELEX); RNA world

1. INTRODUCTION

For current models of the origin of modern life an obligatory step is an RNA world (Gesteland *et al.* 1999). This is a point in time when RNA was the predominant biomolecule and served both as the carrier of genetic information and as the primary catalyst for metabolism. These days coding is predominantly carried out by DNA, and metabolic processes by proteins. However, several key roles are still performed by RNAs, leading to the suggestion that some coding and metabolic aspects of current living systems are relicts of the RNA world (Szathmáry 1992; Poole *et al.* 1998; Jeffares *et al.* 1998).

In the RNA world the genotype and the phenotype are expressed in the same molecule. The genotype refers to the sequence of nucleotides within the molecule and the phenotype can be viewed as the specific three-dimensional conformation of the catalytically active functional RNA (fRNA). As a result, the 'fitness' of a ribo-organism can be inferred from the phenotype (Higgs 2000; Joyce 2000). In the case of RNA, the stabilizing forces conferred by the formation of a secondary structure are much greater than those conferred by the rearrangement of secondary structural elements in three-dimensional space (Grüner *et al.* 1996). This fact is supported by the well-documented secondary-structure conservation in fRNAs (Sankoff *et al.* 1978; Hofacker *et al.* 1998; Eddy 1999; Parsch *et al.* 2000). Thus, primary to secondary structure mappings are relevant for studies of evolution in the RNA world. However, an RNA can fold into many near-optimal secondary structures (Zuker 2000; Higgs 2000). In other words, the genotype does not specify a unique phenotype. Part of our study is to find the conditions under which RNA sequences are expected to fold into a small number of stable structures without the assistance of chaperones.

Based on our current knowledge, an RNA world would have been dominated by four coding nucleotides—the two pairs A : U and C : G (although other ribonucleotides may have also been involved as cofactors in catalysis; White 1976). A natural series of questions relevant to understanding the RNA world include the following.

(i) Are two pairs of nucleotides optimal?
(ii) Is there an advantage of a four-nucleotide system over two nucleotides (one pair, either A : U or C : G)?
(iii) If four is better than two, then is six better than four, and eight better than six?

Some researchers have used tools from organic chemistry to investigate the properties of non-canonical RNA systems (here we will refer to the AUCG alphabet as canonical and to the alternatives as non-canonical) such as those with differing nucleotide bases and alternative sugar groups (Rich 1962; Switzer *et al.* 1989; Piccirilli *et al.* 1990; Bain *et al.* 1992; Eschenmoser 1999). Computational experiments have been used to compare non-canonical RNA systems (Szathmáry 1991, 1992; Grüner *et al.* 1996). Szathmáry (1991, 1992) in particular has posed the question 'what is the optimal size for the genetic alphabet?' He concludes that the four-letter genetic alphabet is a 'frozen evolutionary optimum' that was determined in the RNA world. More recently Mac Dónaill (2002) investigated the optimality of the nucleotide alphabet in terms of error minimization using informatic techniques; he concluded that the canonical alphabet is one of the 'better' possibilities. We explore this question further in the context of maps from RNA primary

* Author for correspondence (P.P.Gardner@massey.ac.nz).

structure to RNA secondary structure and from RNA secondary structure to ribo-organism fitness.

In particular, we consider two new approaches to exploring the implications of different alphabet sizes. First, we investigate the average properties of secondary structures derived from both canonical and non-canonical RNAs, and, second, we study how random structures evolve towards some predefined structures for each of the alternative alphabet sizes. From the first investigation we discovered that, while different alphabets lead to secondary structures that have very different properties, beyond some obvious conclusions, it is difficult to determine exactly which of these properties would have been optimal for the RNA world. However, using a 'flow-reactor' simulation (Fontana & Schuster 1998) inspired by systematic evolution of ligands by exponential amplification (SELEX) experiments (Tuerk & Gold 1990), we show that in the RNA world the canonical four-letter (two base pair) alphabet outperforms the non-canonical alphabets under several sets of evolutionary conditions.

2. STATISTICAL MEASURES OF THE RNA SECONDARY STRUCTURE

We study the statistical properties of the 'molecular morphospace' (Schultes *et al.* 1999) of random RNA with respect to alphabet size. We use a modified distribution of VIENNA v. 1.4, in particular the dynamic programming algorithm, RNAFOLD (available from www.tbi.univie.ac.at/~ivo/RNA/). This program uses empirically derived energy values to infer a minimum free-energy secondary structure from a single RNA sequence (Hofacker *et al.* 1994). Note that, while RNAFOLD may not always predict the 'exact' biological RNA secondary structure (Zuker 2000), we are interested in only the average behaviour of different RNA coding regimes, and, therefore, any inaccuracies inherent in this method are not expected to have a significant effect upon our results (Moulton *et al.* 2000*a*). Another important point is that using statistical measures to discriminate between different alphabet sizes is generally 'easier' than discriminating between evolved and random sequences, at least for the canonical alphabet (Rivas & Eddy 2000; Schattner 2002). Hence, for this work, fruitful results may be obtained by studying random sequences.

A feature of RNAFOLD that we exploit throughout this paper is that it allows the prediction of secondary structures for sequences generated from artificial alphabets; ABCD… (where A pairs B, C pairs D, and so on). We restrict our attention to alphabets with two, four, six and eight letters since, as discussed in Szathmáry (1992), there are only $2^3 = 8$ unique hydrogen-donor/acceptor configurations between any two complementary nucleotides and a nucleotide that can be either purine or pyrimidine, yielding 16 unique letters and eight nucleotides. Owing to time and space considerations, we will consider a maximum of only eight letters. Alphabets with an odd number of bases are not considered as this requires two different bases to compete for a complementary site during replication; the base with lower affinity would be lost after just a few generations. There are four energy-parameter options that RNAFOLD uses: the default (0) uses parameters for the canonical alphabet, otherwise it folds sequences generated by an artificial alphabet with (1) G : C, (2) A : U or (3) alternating G : C and A : U energy-parameter assignments for the base pairs AB, CD, ….

Statistics were gathered from 1000 randomly generated sequences of fixed length $N = 120$ for all possible RNAFOLD energy-parameter selections (see figure 1). The statistics shown are: P, the fraction of paired bases within the optimal structure; Q, the Shannon entropy of the base-pairing probability matrix (p_{ij} is calculated using the partition function; McCaskill 1990), defined by

$$Q := \frac{-1}{Q_{\max}} \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} p_{ij} \log_2 p_{ij} \text{ where, } Q_{\max} = \frac{1}{2} N \log_2 N;$$

and F (known locally as the Gardner uniqueness of folding function), that is the Frobenius norm of the base-pairing probability matrix, defined by

$$F := \sqrt{\frac{1}{N}\left(\sum_{i=1}^{N-1} \sum_{j=i+1}^{N} p_{ij}^2\right)}.$$

Q is intended to be a measure of how well defined the predicted secondary structure is—a high Q value indicates uncertainty in the structure assignment, with many alternative structures near to optimal, whereas a low Q value indicates a well-defined assignment. Note from the melting curves for Q that, as the temperature increases Q also initially increases until a threshold is reached when the number of valid structures starts to decrease until the only possible assignment is the completely unfolded structure (see Schultes *et al.* (1999) for further discussion). Unlike the Shannon entropy, F will distinguish between a 'well-folded' stable secondary structure and a completely unfolded molecule. From figure 1 we see that base pairing (P) decreases as the alphabet size increases. For example, the canonical alphabet has an average fraction of paired bases of 0.54 whereas the maximums for the six- and eight-letter alphabets were 0.43 and 0.34, respectively. This can be explained by the fact that there is a higher probability of any given base matching its complement with a smaller alphabet. Similarly, the measure of disorder, Q, is lower indicating a lower degree of structural uncertainty with increasing alphabet size.

Comparing the canonical alphabet and the ABCD alphabet with RNAfold energy parameter 3 we observe that the canonical alphabet has more base pairing, owing to the additional G : U pair that is allowed in the canonical system. As a consequence of this there is also more uncertainty (larger Q values) in the predicted structure.

Each statistic is sensitive to the energy-parameter selection to varying degrees, and is therefore also sensitive to the base composition, which we keep constant (on average). In particular note that a regime containing solely A : U pairing clearly has insufficient pairing potential to maintain RNA secondary structures (from random sequences) of any complexity. From this we note that any pairing regime must contain at least one pair of G : C type to be viable.

Earlier work in this area has shown that a four-letter alphabet would have been a significant improvement upon a two-letter alphabet (Fontana *et al.* 1993; Schuster 1993; Grüner *et al.* 1996). Although ribozymes generated from two-letter alphabets have proven to be adequate enzymes

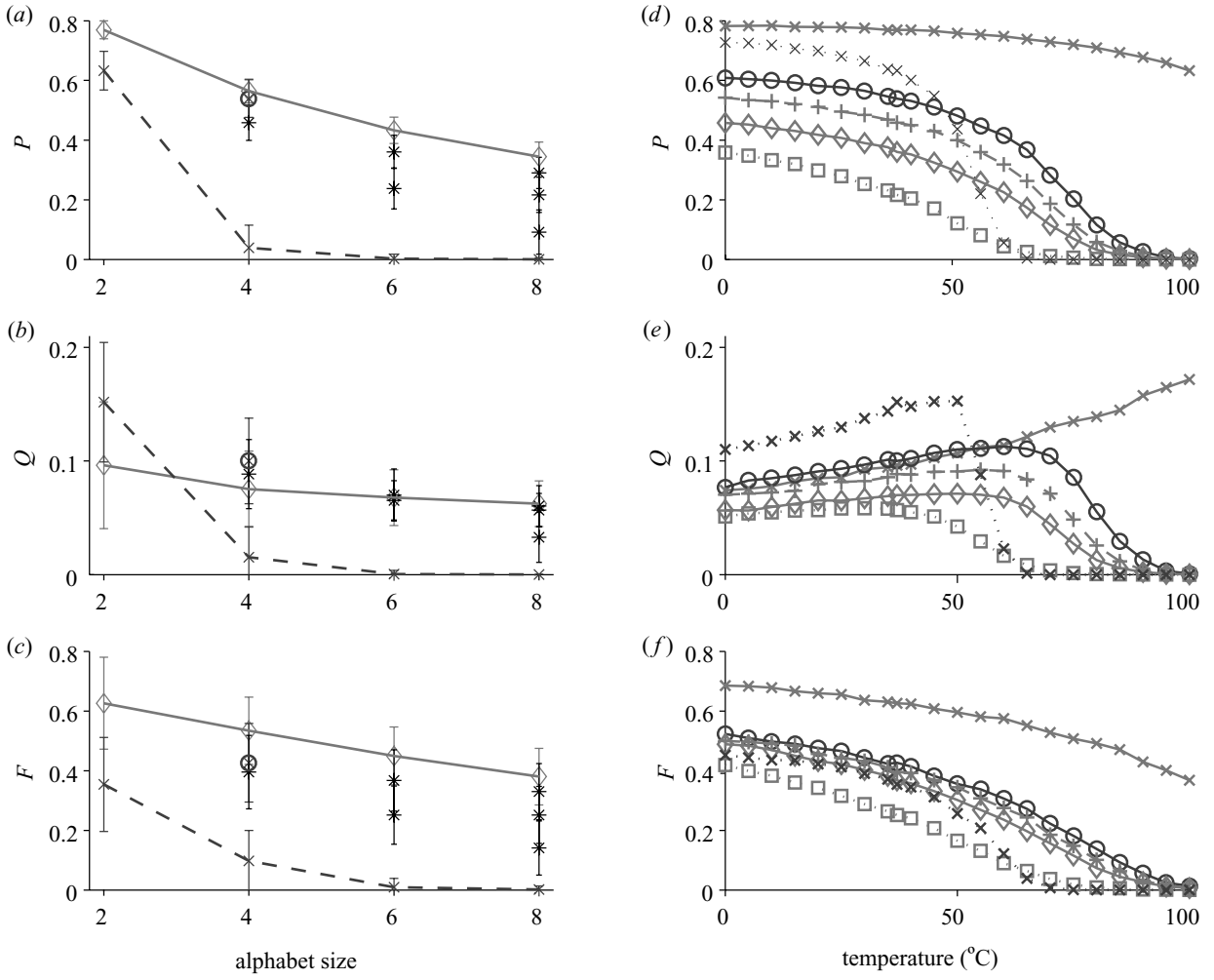

Figure 1. Average RNA secondary structure statistics from 1000 randomly generated sequences with respect to alphabet size (*a–c*) and melting curves (*d–f*). The sequence length is fixed to 120 nucleotides and the temperature is 37 °C (*a–c*). (*a–c*) The solid line connecting the diamonds corresponds to all base-pairs assigned an energy value equivalent to a G : C pair; the dashed line connecting the crosses used only A : U energy parameters. Asterisks, those artificial alphabets where a mixture of energy parameters was used; open circles, the canonical (AUCG) alphabet. (*d–f*) Crosses, points corresponding to the 2-letter alphabets AU (dotted line) and CG (solid line); open circles, the canonical alphabet; plus signs, the 4-letter alphabet with mixed energy parameters (no G : U base pairs); diamonds, the 6-letter alphabet with mixed energy parameters; squares, the 8-letter alphabet with mixed energy parameters.

(Reader & Joyce 2002), the results presented here show that the differences in the statistics between the two- and four-letter alphabets are marked, but the differences between the four- and six-letter alphabets are considerably less so. This suggests that a two-letter alphabet (if one ever existed) would have been rapidly outcompeted by a four-letter one. But does the same argument hold for the four- versus six-letter situation?

3. EVOLVING RNA *IN SILICO*

Having described some statistical properties of random sequences we now consider the impact of alphabet size on the ability of a population of sequences to evolve towards a predefined target structure. We compare the ways in which various non-canonical RNA systems evolve through a fitness landscape. To achieve this we constructed a modified 'flow reactor'. This is a stochastic discrete-time model, with capacity limited to a fixed number of sequences (Fontana & Schuster 1998). This system models SELEX laboratory experiments, where the goal is to artificially evolve RNA aptamers binding specifically to another molecule (Tuerk & Gold 1990). For SELEX experiments the final shape(s) is generally unknown; however, since it is difficult to infer computationally how well an RNA will bind to another molecule, a target structure is defined in advance. Then the probability of survival to the next generation is made a function of the distance to the target.

At generation zero the flow reactor is filled with a pool of randomly generated sequences; successive rounds of amplification with replication error, followed by selection, are used to generate 'evolved' sequences with a corresponding shape that is near some target structure (see figure 2*a*). The probability of selection is a function of the fitness (W), which is dependent upon the distance between the secondary structures of the individual (S_i) and the target (S_{target}). As a distance measure we use the base-pair metric ($d_{BP}(S_{target}, S_i)$), which is a count of the base pairs that two secondary structures (S_{target}, S_i) of equal length do not have in common (see Moulton *et al.* (2000*b*) for a technical description). When compared with

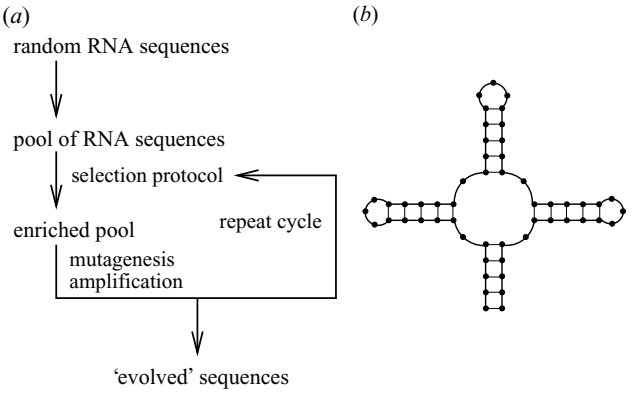

Figure 2. (*a*) An outline of the simulated flow reactor, adapted from Szostak (1993). (*b*) A biological representation of the target structure.

alternative metrics such as the hamming distance and mountain metric (Moulton *et al.* 2000*b*) the base-pair metric showed better discrimination between phenotypes within the evolving population. The metric is scaled to take values bounded by 0 and 1, with 1 being a perfect match to the target structure and 0 being a structure with no base pairs in common with the optimal structure:

$$W = 1 - \frac{d_{\mathrm{BP}}(S_{\mathrm{target}}, S_{\mathrm{i}})}{\max(d_{\mathrm{BP}}(S_{\mathrm{target}}, S_{\mathrm{i}}))}.$$

The selected target, the clover leaf (see figure 2*b*), contained no tetra-loops (which may confer an advantage to the canonical alphabet of Antao & Tinoco (1992)), had a reasonable degree of 'structural complexity' (which would be important for function in the RNA world) and displayed many of the characteristics of a modern ribozyme. Our preliminary studies showed that the empirical behaviour of the flow reactor is largely independent of the energy parameter used (data not shown); this is in contrast to the data in § 2. For all the following experiments we use alternating A : U and C : G energy parameters for the artificial alphabets.

Fontana & Schuster (1998) note that the evolution of this system progresses in leaps towards the target structure, interspersed with periods of no apparent adaptive progress, during which neutral mutations are accrued before the next adaptive step. Here, we are more interested in how possible non-canonical RNA alphabets might perform against each other in an RNA world, so we average over many runs to eliminate the noisy effects of the adaptive leaps.

The function that infers a secondary structure is computationally intensive ($O(N^3)$, where N is the sequence length), and is called many times. To ensure that the computations were completed in a timely fashion we constructed an implementation of the flow reactor that ran on 10 nodes of an Intel-based Beowulf Cluster (code available upon request).

4. COMPUTATIONAL RESULTS

We define the population fitness ($\bar{W}(i)$) as the modal fitness of all the individuals in a population at generation i. Figure 3*a* was generated by taking the mean $\bar{W}(i)$ of 100 runs of the flow reactor for the two-, four-, six- and eight-letter alphabets. The flow reactor was run for 1000 generations and the probability of mutation at each site during replication was 0.01. Inset is a plot of the mean population fitness at generation 1000 as a function of alphabet size showing an optima for the canonical alphabet for this particular parameter set.

The data in figure 3*b* were generated by collecting the mean population fitnesses at generation 1000 for mutation rates ranging from 0 to 0.01. Observe that the four-letter alphabets outperformed the alternatives for high mutation rates. But, as the mutation rate decreases, first the four-letter (ABCD) alphabet then the six-letter alphabet outperform the canonical alphabet. In an RNA world it is unlikely that the mutation rate was low, therefore we can conclude that in this world the canonical RNA alphabet was indeed superior to the alternatives considered here. Only when the *copy fidelity* increased were the four-letter alphabets outperformed by the six-letter alphabets, which is in agreement with the results of Szathmáry (1992). But, by comparing the canonical and non-canonical alphabets where the only difference is that wobble (G : U) base pairs are allowed, we can conclude that allowing wobble pairing for the six-letter alphabet will reduce the advantage of a six-letter alphabet over a four-letter alphabet.

5. DISCUSSION

We have presented a novel approach to investigating the optimal alphabet size in the RNA world. We know that the six- and eight-letter alphabets can cover just as much (if not more) of shape space as the canonical (AUCG) and non-canonical (ABCD) four-letter alphabets, but it would seem from our simulations that the paths from random shapes to our target through sequence space are shorter for the four-letter regimes when the copy fidelity is relatively low. Otherwise the peak shifts to the six-letter alphabet. Although this effect may not be as significant as shown here, owing to the fact that we use the base-pair metric, which penalizes extraneous base pairs, from figure 1 we observe that the four-letter alphabet generally has more base pairs than the larger alphabet sizes. Thus, we conclude that the canonical alphabet was very likely to have been optimal in the RNA world but could indeed be outcompeted (as Szathmáry (1991, 1992) has already suggested) by an alternative six-letter system under a high copy fidelity regimen (although the effects of wobble pairing have not been taken into account here). In addition, copy fidelity decreases with increasing alphabet size (Szathmáry 1992; Mac Dónaill 2002) so it is more realistic to compare high-fidelity two- and four-letter alphabet fitnesses with low-fidelity six- and eight-letter alphabet fitnesses. This would have the effect of increasing the fidelity range where a four-letter alphabet outcompetes a six-letter alphabet.

Ribozymes are usually much larger and more complex than the structure that we use as a target in the flow reactor. Using a more complex structure as the target will improve discrimination between the alphabets. However, experiments to study larger structures will take longer to compute owing to the complexity of the RNA-folding algorithm.

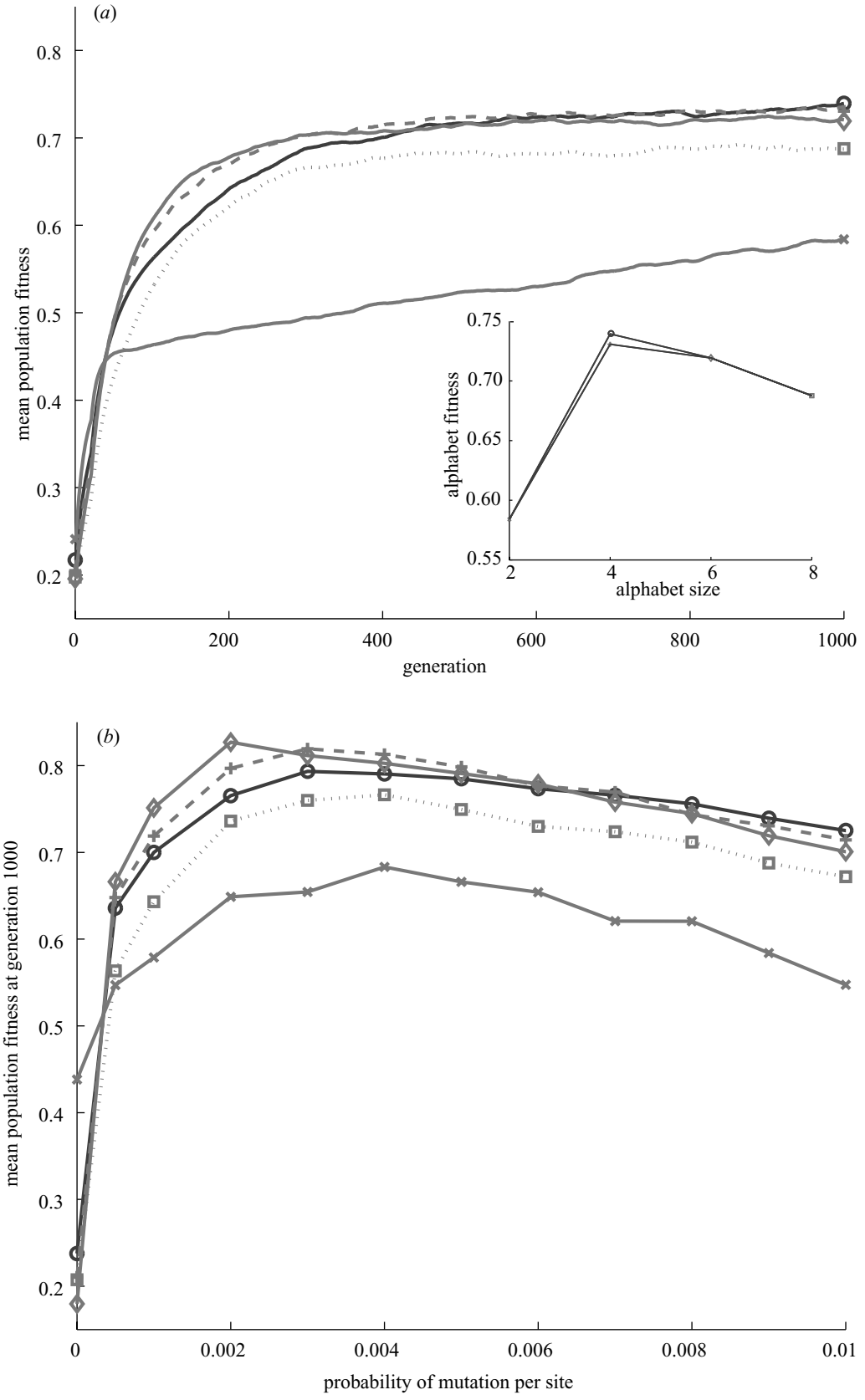

Figure 3. (*a*) Evolving towards a target structure. Results were gathered from the flow reactor (see § 4) for two-, four-, six- and eight-letter artificial genetic alphabets and the canonical AUCG alphabet. The probability of mutation is 0.009 and the population size is maintained at 100. The inset is a plot of the alphabet fitness, which is defined to be mean population fitness at generation 1000 as a function of alphabet size. (*b*) The mean population fitness at generation 1000 as a function of the probability of mutation per site. The results are for two-, four-, six- and eight-letter artificial alphabets (represented by crosses, addition signs, diamonds and squares, respectively) and the canonical AUCG alphabet (circles).

These experiments do not, however, take into account the additional metabolic cost of using longer alphabets. But if we observe four- and six-letter alphabets performing in an almost indistinguishable manner then one can argue that four is optimal owing to the added difficulty of synthesizing an additional base pair (although ribozymes may have been more specific). For the larger 10–16-letter alphabets not included in this study the general trend is an asymptotic approach to a low fitness (data not shown). Additionally, the amount of base pairing with larger alphabets decreases to negligible levels, although the effects of hydrophobic interactions between bases may alter these conclusions (Wu *et al.* 2000).

The authors thank I. Hofacker for implementing the changes in VIENNA necessary for this research, and B. Ryland (bug exterminator). We also thank the administrators of the parallel computing facility used for much of this work (http://sisters.massey.ac.nz). The authors were funded by the New Zealand Marsden Fund and the Swedish Foundation for International Cooperation in Research and Higher Education (STINT), Swedish Research Council (VR).

REFERENCES

Antao, V. P. & Tinoco, I. 1992 Thermodynamic parameters for loop formation in RNA and DNA hairpin tetraloops. *Nucleic Acids Res.* **20**, 819–824.

Bain, J., Switzer, C., Chamberlin, A. & Benner, S. 1992 Ribosome-mediated incorporation of a nonstandard amino-acid into a peptide through expansion of the genetic code. *Nature* **356**, 537–539.

Eddy, S. R. 1999 Noncoding RNA genes. *Curr. Opin. Genet. Dev.* **9**, 695–699.

Eschenmoser, A. 1999 Chemical etiology of nucleic acid structure. *Science* **284**, 2118–2124.

Fontana, W. & Schuster, P. 1998 Continuity in evolution: on the nature of transitions. *Science* **280**, 1451–1455.

Fontana, W., Konings, D., Stadler, P. & Schuster, P. 1993 Statistics of RNA secondary structures. *Biopolymers* **33**, 1389–1404.

Gesteland, R., Cech, T. & Atkins, J. (eds) 1999 *The RNA world*, 2nd edn. Cold Spring Harbor, NY: Cold Spring Harbor Laboratory Press.

Grüner, W., Giegerich, R., Strothmann, D., Reidys, C., Weber, J., Hofacker, I. L., Schuster, P. & Stadler, P. F. 1996 Analysis of RNA sequence structure maps by exhaustive enumeration. *Monatshefte Chem.* **127**, 355–374.

Higgs, P. G. 2000 RNA secondary structure: physical and computational aspects. *Q. Rev. Biophys.* **33**, 199–253.

Hofacker, I. L., Fontana, W., Bonhoeffer, S. & Stadler, P. F. 1994 Fast folding and comparison of RNA secondary structures. *Monatshefte Chem.* **125**, 167–188.

Hofacker, I., Fekete, M., Flamm, C., Huynen, M., Rauscher, S., Stolorz, P. & Stadler, P. 1998 Automatic detection of conserved RNA structure elements in complete RNA virus genomes. *Nucleic Acids Res.* **26**, 3825–3836.

Jeffares, D. C., Poole, A. M. & Penny, D. 1998 Relics from the RNA world. *J. Mol. Evol.* **46**, 18–36.

Joyce, G. F. 2000 Perspectives: RNA structure—ribozyme evolution at the crossroads. *Science* **289**, 401–402.

McCaskill, J. S. 1990 The equilibrium partition function and base pair binding probabilities for RNA secondary structures. *Biopolymers* **29**, 1105–1119.

Mac Dónaill, D. 2002 A parity code interpretation of nucleotide alphabet composition. *Chem. Commun.* **18**, 2062–2063.

Moulton, V., Gardner, P., Pointon, R., Creamer, L., Jameson, G. & Penny, D. 2000*a* RNA folding argues against a hot-start origin of life. *J. Mol. Evol.* **51**, 416–421.

Moulton, V., Zuker, M., Steel, M., Pointon, R. & Penny, D. 2000*b* Metrics on RNA secondary structures. *J. Comput. Biol.* **7**, 277–292.

Parsch, J., Braverman, J. & Stephan, W. 2000 Comparative sequence analysis and patterns of covariation in RNA secondary structures. *Genetics* **154**, 909–921.

Piccirilli, J., Krauch, T., Moroney, S. & Benner, S. 1990 Enzymatic incorporation of a new base pair into DNA and RNA extends the genetic alphabet. *Nature* **343**, 33–37.

Poole, A. M., Jeffares, D. C. & Penny, D. 1998 The path from the RNA world. *J. Mol. Evol.* **46**, 1–17.

Reader, J. & Joyce, G. 2002 A ribozyme composed of only two different nucleotides. *Nature* **420**, 841–844.

Rich, A. 1962 *Horizons in biochemistry*, pp. 103–126. New York: Academic.

Rivas, E. & Eddy, S. R. 2000 Secondary structure alone is generally not statistically significant for the detection of noncoding RNAs. *Bioinformatics* **16**, 583–605.

Sankoff, D., Morin, A. & Cedergren, R. 1978 The evolution of 5S RNA secondary structures. *Can. J. Biochem.* **56**, 440–443.

Schattner, P. 2002 Searching for RNA genes using base composition statistics. *Nucleic Acids Res.* **30**, 2076–2082.

Schultes, E. A., Hraber, P. T. & LaBean, T. H. 1999 Estimating the contributions of selection and self-organization in RNA secondary structure. *J. Mol. Evol.* **49**, 76–83.

Schuster, P. 1993 RNA based evolutionary optimization. *Origins Life Evol. Biosphere* **23**, 373–391.

Switzer, C., Moroney, S. & Benner, S. 1989 Enzymatic incorporation of a new base-pair into DNA and RNA. *J. Am. Chem. Soc.* **111**, 8322–8323.

Szathmáry, E. 1991 Four letters in the genetic alphabet: a frozen evolutionary optimum? *Proc. R. Soc. Lond.* B **245**, 91–99.

Szathmáry, E. 1992 What is the optimal size for the genetic alphabet? *Proc. Natl Acad. Sci. USA* **89**, 2614–2618.

Szostak, J. W. 1993 Ribozymes—evolution *ex vivo*. *Nature* **361**, 119–120.

Tuerk, C. & Gold, L. 1990 Systematic evolution of ligands by exponential enrichment—RNA ligands to bacteriophage-T4 DNA-polymerase. *Science* **249**, 505–510.

White, H. B. 1976 Coenzymes as fossils of an earlier metabolic state. *J. Mol. Evol.* **7**, 101–104.

Wu, Y., Ogawa, A., Berger, M., McMinn, D., Schultz, P. & Romesburg, F. 2000 Efforts toward expansion of the genetic alphabet: optimization of interbase hydrophobic interactions. *J. Am. Chem. Soc.* **122**, 7621–7632.

Zuker, M. 2000 Calculating nucleic acid secondary structure. *Curr. Opin. Struct. Biol.* **10**, 303–310.