Evolution of Biological Information

Thomas D. Schneider *

version = 1.60 of ev.tex 2001 February 9

key words: evolution, information theory, sequence logo.

Nucleic Acids Res, 28(14): 2794-2799, 2000

*National Cancer Institute, Frederick Cancer Research and Development Center, Laboratory of Experimental and Computational Biology, P. O. Box B, Frederick, MD 21702-1201. (301) 846-5581 (-5532 for messages), fax: (301) 846-5598, email: toms@ncifcrf.gov. http://www.lecb.ncifcrf.gov/~toms/

ABSTRACT

How do genetic systems gain information by evolutionary processes? Answering this question precisely requires a robust, quantitative measure of information. Fortunately, fifty years ago Claude Shannon defined information as a decrease in the uncertainty of a receiver. For molecular systems, uncertainty is closely related to entropy and hence has clear connections to the Second Law of Thermodynamics. These aspects of information theory have allowed the development of a straightforward and practical method of measuring information in genetic control systems. Here this method is used to observe information gain in the binding sites for an artificial 'protein' in a computer simulation of evolution. The simulation begins with zero information and, as in naturally occurring genetic systems, the information measured in the fully evolved binding sites is close to that needed to locate the sites in the genome. The transition is rapid, demonstrating that information gain can occur by punctuated equilibrium.

INTRODUCTION

Evolutionary change has been observed in the fossil record, in the field, in the laboratory, and at the molecular level in DNA and protein sequences, but a general method for quantifying the changes has not been agreed upon. In this paper the well-established mathematics of information theory (1, 2, 3) is used to measure the information content of nucleotide binding sites (4, 5, 6, 7, 8, 9, 10, 11) and to follow changes in this measure to gauge the degree of evolution of the binding sites.

For example, human splice acceptor sites contain about 9.4 bits of information on the average (6). This number is called $R_{sequence}$ because it represents a rate (bits per site) computed from the aligned sequences (4). (The equation for $R_{sequence}$ is given in the Results.) The question arises as to why one gets 9.4 bits rather than, say, 52. Is 9.4 a fundamental number? The way to answer this is to compare it to something else. Fortunately, one can use the size of the genome and the number of sites to compute how much information is needed to find the sites. The average distance between acceptor sites is the average size of introns plus exons, or about 812 bases, so the information needed to find the acceptors is $R_{frequency} = \log_2 812 = 9.7$ bits (6). By comparison, $R_{sequence} = 9.4$ bits, so in this and other genetic systems $R_{sequence}$ is close to $R_{frequency}$ (4).

These measurements show that there is a subtle connection between the pattern at binding sites and the size of the genome and number of sites. Relative to the potential for changes at binding sites, the size of the entire genome is approximately fixed over long periods of time. Even if the genome were to double in length (while keeping the number of sites constant), $R_{frequency}$ would only change by 1 bit, so the measure is quite insensitive. Likewise, the number of sites is approximately fixed by the physiological functions that have to be controlled by the recognizer. So $R_{frequency}$ is essentially fixed during long periods of evolution. On the other hand, $R_{sequence}$ can change rapidly and could have any value, as it depends on the details of how the recognizer contacts the nucleic acid binding sites and these numerous small contacts can mutate quickly. So how does $R_{sequence}$ come to equal $R_{frequency}$? It must be that $R_{sequence}$ can start from zero and evolve *up* to $R_{frequency}$. That is, the necessary information should be able to evolve from scratch.

The purpose of this paper is to demonstrate that $R_{sequence}$ can indeed evolve to match $R_{frequency}$ (12). To simulate the biology, suppose we have a population of organisms each with a given length of DNA. That fixes the genome size, as in the biological situation. Then we need to specify a set of locations that a recognizer protein has to bind to. That fixes the number of sites, again as in nature. We need to code the recognizer into the genome so that it can co-evolve with the binding sites. Then we need to apply random mutations and selection for finding the sites and against finding non-sites. Given these conditions, the simulation will match the biology at every point.

Because half of the population always survives each selection round in the evolutionary simulation presented here, the population cannot die out and there is no lethal level of incompetence. While this may not be representative of all biological systems, since extinction and threshold effects do occur, it is representative of the situation in which a functional species can survive without a particular genetic control system but which would do better to gain control *ab initio*. Indeed, any new function must have this property until the species comes to depend on it, at which point it can become essential if the earlier means of survival is lost by atrophy or no longer available. I call such a situation a 'Roman arch' because once such a structure has been constructed on top of scaffolding, the scaffold may be removed, and will disappear from biological systems when it is no longer needed. Roman arches are common in biology, and they are a natural consequence of evolutionary processes.

The fact that the population cannot become extinct could be dispensed with, for example by assigning a probability of death, but it would be inconvenient to lose an entire population after many generations.

A twos complement weight matrix was used to store the recognizer in the genome. At first it may seem that this is insufficient to simulate the complex processes of transcription, translation, protein folding and DNA sequence recognition found in cells. However the success of the simulation, as shown below, demonstrates that the form of the genetic apparatus does not affect the computed information measures. For information theorists and physicists this emergent mesoscopic property (13) will come as no surprise because information theory is extremely general and does not depend on the physical mechanism. It applies equally well to telephone conversations, telegraph signals, music and molecular biology (2).

Given that, when one runs the model one finds that the information at the binding sites ($R_{sequence}$) does indeed evolve to be the amount predicted to be needed to find the sites ($R_{frequency}$). This is the same result as observed in natural binding sites and it strongly supports the hypothesis that these numbers should be close (4).

MATERIALS AND METHODS

Sequence logos were created as described previously (5). Pascal programs **ev**, **evd** and **lister** are available from http://www.lecb.ncifcrf.gov/˜toms/. The evolution movie is at http://www.lecb.ncifcrf.gov/~toms/paper/ev/movie.

RESULTS

To test the hypothesis that $R_{sequence}$ can evolve to match $R_{frequency}$, the evolutionary process was simulated by a simple computer program, **ev**, for which I will describe one evolutionary run. This paper demonstrates that a set of 16 binding sites in a genome size of 256 bases, which would theoretically be expected to have an average of $R_{frequency} = 4$ bits of information per site, can evolve to this value given only these minimal numerical and size constraints. Although many parameter variations are possible, they give similar results as long as extremes are avoided (data not shown).

A small population ($n = 64$) of 'organisms' was created, each of which consisted of $G = 256$ bases of nucleotide sequence chosen *randomly*, with equal probabilities, from an alphabet of 4 characters (a, c, g, t, Fig. 1). At any particular time in the history of a natural population, the size of a genome, G, and the number of required genetic control element binding sites, γ, are determined by previous history and current physiology respectively, so as a parameter for this simulation we chose $\gamma = 16$ and the program arbitrarily chose the site locations, which are fixed for the duration of the run. The information required to locate γ sites in a genome of size G is $R_{frequency} = -\log_2(\gamma/G) = 4$ bits per site, where γ/G is the frequency of sites (4, 14). ⇐Fig 1

A section of the genome is set aside by the program to encode the gene for a sequence recognizing 'protein', represented by a weight matrix (15, 7) consisting of a two-dimensional array of 4 by $L = 6$ integers. These integers are stored in the genome in twos complement notation, which allows for both negative and positive values. (In this notation, the negative of an integer is formed by taking the complement of all bits and adding 1.) By encoding A=00, C=01, G=10, and T=11 in a space of 5 bases, integers from -512 to $+511$ are stored in the genome. Generation of the weight matrix integers from the nucleotide sequence gene corresponds to translation and protein folding in natural systems. The weight matrix can evaluate any L base long sequence. Each base of the sequence selects the corresponding weight from the matrix and these weights are summed. If the sum is larger than a tolerance, also encoded in the genome, the sequence is 'recognized' and this corresponds to a protein binding to DNA (Fig. 1). As mentioned above, the exact form of the recognition mechanism is immaterial because of the generality of information theory.

The weight matrix gene for an organism is translated and then every position of that organism's genome is evaluated by the matrix. The organism can make two kinds of 'mistakes'. The first is for one of the γ binding locations to be missed (representing absence of genetic control) and the second is for one of the $G - \gamma$ non-binding sites to be incorrectly recognized (representing wasteful binding of the recognizer). For simplicity these mistakes are counted as equivalent, since other schemes should give similar final results. The validity of this black/white model of binding sites comes from Shannon's channel capacity theorem, which allows for recognition with as few errors as necessary for survival (1, 16, 7).

The organisms are subjected to rounds of selection and mutation. First, the number of mistakes made by each organism in the population is determined. Then the half of the population making the least mistakes is allowed to replicate by having their genomes replace ('kill') the ones making more mistakes. (To preserve diversity, no replacement takes place if they are equal.) At every generation, each organism is subjected to one random point mutation in which the original base is obtained 1/4 of the time. For comparison, HIV-1 reverse transcriptase makes about one error every 2000-5000 bases incorporated, only 10 fold lower than this simulation (17).

When the program starts, the genomes all contain random sequence, and the information content of the binding sites, $R_{sequence}$, is close to zero. Remarkably, the cyclic mutation and selection process leads to an organism that makes no mistakes in only 704 generations (Fig. 2a). Although the sites can contain a maximum of $2L = 12$ bits, the information content of the binding sites rises during this time until it oscillates around the predicted information content, $R_{frequency} = 4$ bits, with $R_{sequence} = 3.983 \pm 0.399$ bits during the 1000 to 2000 generation interval (Fig. 2b). The expected standard deviation from small sample effects (4) is 0.297 bits, so about 55% of the variance $(0.3^2/0.4^2)$ comes from the digital nature of the sequences. Sequence logos (5) of the binding sites show that distinct patterns appear during selection, and that these then drift (Fig. 3). When selective pressure is removed, the observed pattern atrophies (not shown, but Fig. 1 shows the organism with the fewest mistakes at generation 2000, after atrophy) and the information content drops back to zero (Fig. 2b). The information decays with a half-life of 61 generations. ⇐Fig 2 ⇐Fig 3

The evolutionary steps can be understood by considering an intermediate situation, for example when all organisms are making 8 mistakes. Random mutations in a genome that lead to more mistakes will immediately cause the selective elimination of that organism. On the other hand, if one organism randomly 'discovers' how to make 7 mistakes, it is guaranteed (in this simplistic model) to reproduce every generation, and therefore it exponentially overtakes the population. This roughly-sigmoidal rapid transition corresponds to (and the program was inspired by) the proposal that evolution proceeds by punctuated equilibrium (18, 19), with noisy 'active stasis' clearly visible from generation 705 to 2000 (Fig. 2b, Fig. 3).

An advantage of the **ev** model over previous evolutionary models, such as biomorphs (20), Avida (21), and Tierra (22), is that it starts with a completely random genome, and no further intervention is required. Given that gene duplication is common and that transcription and translation are part of the housekeeping functions of all cells, the program simulates the process of evolution of new binding sites from scratch. The exact mechanisms of translation and locating binding sites are irrelevant.

The information increases can be understood by looking at the equations used to compute the information (12). The information in the binding sites is measured as the decrease in uncertainty from before binding to after binding (4, 14):

$$R_{sequence} = H_{before} - H_{after} \quad \text{(bits per site)}. \tag{1}$$

Before binding the uncertainty is

$$H_{before} = H_g L \tag{2}$$

where L is the site width, $H_g = e(G) - \sum_{b=A}^{T} p(b)\log_2 p(b)$, $e(G)$ is a small sample correction (4) and $p(b)$ is the frequency of base b in the genome of size G. After binding the uncertainty is:

$$H_{after} = \sum_{l=1}^{L}\Big(e(n(l)) - \sum_{b=A}^{T} f(b,l)\log_2 f(b,l)\Big), \tag{3}$$

where $f(b,l)$ is the frequency of base b at position l in the binding sites and $e(n(l))$ is a small sample size correction (4) for the $n(l)$ sequences at l. In both this model and in natural binding sites, random mutations tend to increase both H_{before} and H_{after} since equiprobable distributions maximize the uncertainty and entropy (1). Because there are only 4 symbols (or states), nucleotides can form a closed system and this tendency to increase appears to be a form of the Second Law of Thermodynamics (23, 12), where H is proportional to the entropy for molecular systems (24). Effective closure occurs because selections have little effect on the overall frequencies of bases in the genome, so without external influence H_{before} maximizes at $\sim 2L$ bits per base ($H_g = 1.9995 \pm 0.0058$ bits for the entire simulation). In contrast, by biasing the binding site base frequencies, $f(b,l)$, selection *simultaneously* provides an open process whereby H_{after} can be decreased, increasing the information content of the binding sites according to equation [1] (12).

Microevolution can be measured in haldanes, as standard deviations per generation (25, 26, 27). In this simulation 4.0 ± 0.4 bits evolved at each site in 704 generations, or $4.0/(0.4 \times 704) = 0.014$ haldanes. This is within the range of natural population change, indicating that although selection is strong, the model is reasonable. However, a difficulty with using standard deviations is that they are not additive for independent measures, whereas bits are. A measure suggested by Haldane is the *darwin*, the natural logarithm of change per million years, which has units of nits per time. This is the rate of information transmission originally introduced by Shannon (1). Because a computer simulation does not correlate with time, the haldane and darwin can be combined to give units of bits per generation, in this case 0.006 ± 0.001 bits per generation per site.

DISCUSSION

The results, which show the successful simulation of binding site evolution, can be used to address both scientific and pedagogical issues. $R_{sequence}$ approaches and remains around $R_{frequency}$ (Fig. 2b), supporting the hypothesis that the information content at binding sites will evolve to be close to the information needed to locate those binding sites in the genome, as observed in natural systems (4, 6). That is, one can measure information in genetic systems, the amount observed can be predicted, and the amount measured evolves to the amount predicted. This is useful because when this prediction is not met (4, 28, 29, 6) the anomaly implies the existence of new biological phenomena. Simulations to model such anomalies have not been attempted yet.

Variations of the program could be used to investigate how population size, genome length, number of sites, size of recognition regions, mutation rate, selective pressure, overlapping sites and other factors affect the evolution. Another use of the program may include understanding the sources and effects of skewed genomic composition (4, 7, 30, 31). However, this could be caused by mutation rates, and/or it could be the result of some kind(s) of evolutionary pressure that we don't understand, so how one implements the skew may well affect or bias the results.

The **ev** model quantitatively addresses the question of how life gains information, a valid issue recently raised by creationists (32) (Truman, R. (1999), http://www.trueorigin.org/dawkinfo.htm) but only qualitatively addressed by biologists (33). The mathematical form of uncertainty and entropy ($H = -\sum p \log_2 p$, $\sum p = 1$) implies that neither can be negative ($H \geq 0$), but a decrease in uncertainty or entropy can correspond to information gain, as measured here by $R_{sequence}$ and $R_{frequency}$. The **ev** model shows explicitly how this information gain comes about from mutation and selection, without any other external influence, thereby completely answering the creationists.

The **ev** model can also be used to succinctly address two other creationist arguments. First, the recognizer gene and its binding sites co-evolve, so they become dependent on each other and destructive mutations in either immediately lead to elimination of the organism. This situation fits Behe's (34) definition of 'irreducible complexity' exactly ("a single system composed of several well-matched, interacting parts that contribute to the basic function, wherein the removal of any one of the parts causes the system to effectively cease functioning", page 39), yet the molecular evolution of this 'Roman arch' is straightforward and rapid, in direct contradiction to his thesis. Second, the probability of finding 16 sites averaging 4 bits each in random sequences is $2^{-4\times 16} \cong 5 \times 10^{-20}$ yet the sites evolved from random sequences in only $\sim 10^3$ generations, at an average rate of ~ 1 bit per 11 generations. Because the mutation rate of HIV is only 10 times slower, it could evolve a 4 bit site in 100 generations, about 9 months (35), but it could be much faster because the enormous titer (10^{10} new virions/day/person (17)) provides a larger pool for successful changes. Likewise, at this rate, roughly an entire human genome of $\sim 4 \times 10^9$ bits (assuming an average of 1 bit/base, which is clearly an overestimate) could evolve in a billion years, even without the advantages of large environmentally diverse worldwide populations, sexual recombination and interspecies genetic transfer. However, since this rate is unlikely to be maintained for eukaryotes, these factors are undoubtedly important in accounting for human evolution. So, contrary to probabilistic arguments by Spetner (36, 32), the **ev** program also clearly demonstrates that biological information, measured in the strict Shannon sense, can rapidly appear in genetic control systems subjected to replication, mutation and selection (33).

ACKNOWLEDGMENTS

I thank Denise Rubens, Ilya Lyakhov, Herb Schneider, Natasha Klar, Bruce Shapiro, Richard Dawkins, Hugo Martinez and Karen Lewis for comments on the manuscript, and Frank Schmidt for pointing out that the atrophy should be first order.

REFERENCES

1. Shannon, C. E. (1948) A mathematical theory of communication. *Bell System Tech. J.,* **27**, 379–423, 623–656 http://cm.bell-labs.com/cm/ms/what/shannonday/paper.html.

2. Pierce, J. R. (1980) An Introduction to Information Theory: Symbols, Signals and Noise, Dover Publications, Inc., New York second edition.

3. Gappmair, W. (April, 1999) Claude E. Shannon: The 50th anniversary of information theory. *IEEE Communications Magazine,* **37**(4), 102–105.

4. Schneider, T. D., Stormo, G. D., Gold, L., and Ehrenfeucht, A. (1986) Information content of binding sites on nucleotide sequences. *J. Mol. Biol.,* **188**, 415–431 http://www.lecb.ncifcrf.gov/˜toms/paper/schneider1986/.

5. Schneider, T. D. and Stephens, R. M. (1990) Sequence logos: A new way to display consensus sequences. *Nucleic Acids Res.,* **18**, 6097–6100 http://www.lecb.ncifcrf.gov/˜toms/paper/logopaper/.

6. Stephens, R. M. and Schneider, T. D. (1992) Features of spliceosome evolution and function inferred from an analysis of the information at human splice sites. *J. Mol. Biol.,* **228**, 1124–1136 http://www.lecb.ncifcrf.gov/˜toms/paper/splice/.

7. Schneider, T. D. (1997) Information content of individual genetic sequences. *J. Theor. Biol.,* **189**(4), 427–441 http://www.lecb.ncifcrf.gov/˜toms/paper/ri/.

8. Schneider, T. D. (1997) Sequence walkers: a graphical method to display how binding proteins interact with DNA or RNA sequences. *Nucleic Acids Res.,* **25**, 4408–4415 http://www.lecb.ncifcrf.gov/˜toms/paper/walker/, erratum: NAR 26(4): 1135, 1998.

9. Schneider, T. D. (1996) Reading of DNA sequence logos: Prediction of major groove binding by information theory. *Meth. Enzym.,* **274**, 445–455 http://www.lecb.ncifcrf.gov/˜toms/paper/oxyr/.

10. Shultzaberger, R. K. and Schneider, T. D. (1999) Using sequence logos and information analysis of Lrp DNA binding sites to investigate discrepancies between natural selection and SELEX. *Nucleic Acids Res.,* **27**(3), 882–887 http://www.lecb.ncifcrf.gov/˜toms/paper/lrp/.

11. Zheng, M., Doan, B., Schneider, T. D., and Storz, G. (1999) OxyR and SoxRS regulation of *fur*. *J. Bacteriol.,* **181**, 4639–4643.

12. Schneider, T. D. (1988) Information and entropy of patterns in genetic switches. In G. J. Erickson and C. R. Smith, (ed.), Maximum-Entropy and Bayesian Methods in Science and Engineering, volume **2**, Dordrecht, The Netherlands: Kluwer Academic Publishers pp. 147–154.

13. Laughlin, R. B., Pines, D., Schmalian, J., Stojkovic, B. P., and Wolynes, P. (2000) The middle way. *Proc. Natl. Acad. Sci. USA,* **97**, 32–37.

14. Schneider, T. D. (1994) Sequence logos, machine/channel capacity, Maxwell's demon, and molecular computers: a review of the theory of molecular machines. *Nanotechnology,* **5**, 1–18 http://www.lecb.ncifcrf.gov/~toms/paper/nano2/.

15. Stormo, G. D., Schneider, T. D., Gold, L., and Ehrenfeucht, A. (1982) Use of the 'Perceptron' algorithm to distinguish translational initiation sites in *E. coli. Nucleic Acids Res.,* **10**, 2997–3011.

16. Schneider, T. D. (1991) Theory of molecular machines. I. Channel capacity of molecular machines. *J. Theor. Biol.,* **148**, 83–123 http://www.lecb.ncifcrf.gov/~toms/paper/ccmm/.

17. Loeb, L. A., Essigmann, J. M., Kazazi, F., Zhang, J., Rose, K. D., and Mullins, J. I. (1999) Lethal mutagenesis of HIV with mutagenic nucleoside analogs. *Proc. Natl. Acad. Sci. USA,* **96**, 1492–1497.

18. Gould, S. J. (1977) Is the cambrian explosion a sigmoid fraud?. In Ever Since Darwin, Reflections in Natural History N. Y.: W. W. Norton & Co. pp. 126–133.

19. Gould, S. J. and Eldredge, N. (1993) Punctuated equilibrium comes of age. *Nature,* **366**, 223–227.

20. Dawkins, R. (1986) The Blind Watchmaker, W. W. Norton & Co., New York.

21. Lenski, R. E., Ofria, C., Collier, T. C., and Adami, C. (1999) Genome complexity, robustness and genetic interactions in digital organisms. *Nature,* **400**, 661–664 http://www.krl.caltech.edu/avida/.

22. Ray, T. S. (1994) Evolution, Complexity, Entropy, and Artificial Reality. *Physica D,* **75**, 239–263 http://www.hip.atr.co.jp/~ray/pubs/oji/ojihtml.html.

23. Jaynes, E. T. (1988) The evolution of Carnot's principle. In G. J. Erickson and C. R. Smith, (ed.), Maximum-Entropy and Bayesian Methods in Science and Engineering, volume **1**, Dordrecht, The Netherlands: Kluwer Academic Publishers http://bayes.wustl.edu/etj/articles/ccarnot.ps.gz http://bayes.wustl.edu/etj/articles/ccarnot.pdf pp. 267–281.

24. Schneider, T. D. (1991) Theory of molecular machines. II. Energy dissipation from molecular machines. *J. Theor. Biol.,* **148**, 125–137 http://www.lecb.ncifcrf.gov/~toms/paper/edmm/.

25. Haldane, J. B. S. (1949) Suggestions as to quantitative measurement of rates of evolution. *Evolution,* **3**, 51–56.

26. Hendry, A. P. and Kinnison, M. T. (2000) Perspective: the pace of modern life: measuring rates of contemporary microevolution. *Evolution,* **53**, 1637–1653.

27. Huey, R. B., Gilchrist, G. W., Carlson, M. L., Berrigan, D., and Serra, L. (2000) Rapid evolution of a geographic cline in size in an introduced fly. *Science,* **287**, 308–309.

28. Schneider, T. D. and Stormo, G. D. (1989) Excess information at bacteriophage T7 genomic promoters detected by a random cloning technique. *Nucleic Acids Res.,* **17**, 659–674.

29. Herman, N. D. and Schneider, T. D. (1992) High information conservation implies that at least three proteins bind independently to F plasmid *incD* repeats. *J. Bacteriol.,* **174**, 3558–3560.

30. Stormo, G. D. (1998) Information Content and Free Energy in DNA-Protein Interactions. *J. Theor. Biol.,* **195**, 135–137.

31. Schneider, T. D. (1999) Measuring molecular information. *J. Theor. Biol.,* **201**, 87–92 http://www.lecb.ncifcrf.gov/˜toms/paper/ridebate/.

32. Spetner, L. M. (1998) NOT BY CHANCE! Shattering the Modern Theory of Evolution, Judaica Press, New York.

33. Dawkins, R. (December, 1998) The “Information Challenge”. *The Skeptic,* **18**(4), 21–25 http://www.onthenet.com.au/˜stear/dawkinschallenge.htm and http://www.skeptics.com.au/journal/dawkins1.htm.

34. Behe, M. J. (1996) Darwin’s Black Box: The Biochemical Challenge to Evolution, The Free Press, New York.

35. Perelson, A. S., Neumann, A. U., Markowitz, M., Leonard, J. M., and Ho, D. D. (1996) HIV-1 dynamics in vivo: virion clearance rate, infected cell life-span, and viral generation time. *Science,* **271**, 1582–1586.

36. Spetner, L. M. (1964) Natural selection: an information-transmission mechanism for evolution. *J. Theor. Biol.,* **7**, 412–429.

```
           *         *10       *         *20       *         *30       *         *40
5' t c t t t g c a c g c t a a g t t t g t c a g g a a t t g t a a a c a c c t a a 3'
   --------- A -129    --------- G 450     --------- A 296     --------- G 4
             --------- C -442    --------- T -5      --------- C 251     --------- T 368

           *         *50       *         *60       *         *70       *         *80
5' t c c g t c c a t g a t t t g t c t g a c c t a c a t t g t t g g a c g a g a a 3'
   --------- A -165    --------- G 254     --------- A 369     --------- G -95
             --------- C 334     --------- T -136    --------- C 251     --------- T -480

           *         *90       *         *100      *         *110      *         *120
5' g c t c a t c g g g t a t g c c a g c g g g g c t g g a c g g t c a a t g g c a 3'
   --------- A -396    --------- G -199    --------- A -345    --------- G -304
             --------- C -150    --------- T 294     --------- C -378    --------- T -92

           *         *130      *         *140      *         *150      *         *160
5' t t a c g c g c t g g c a t c t t a a a g a a t g a a c t a g a c t g a a c c g 3'
   --------- Threshold -58                                             (------ ... EvSite-
                   (+++++++++++) EvSite+ 1 evaluation: 171
                                 (-----------) EvSite- 9 evaluation: -443
                                               (-----------) EvSite- 16 evaluation: -1108

           *         *170      *         *180      *         *190      *         *200
5' a c a a c t a g a c g c t t a a t t c a g c a a a a g c g g t g t a a a t a c a 3'
... -----) EvSite- 3 evaluation: -63                                        ( ... EvSite+
        (-----------) EvSite- 2 evaluation: -189         (-----------) EvSite- 13 evaluation: -509
                    (-----------) EvSite- 4 evaluation: -438
                                 (+++++++++++) EvSite+ 14 evaluation: 164

           *         *210      *         *220      *         *230      *         *240
5' g g g g c g a g g a t c g c t a c c c g c a t t c t t c g a t c t c g t g g a g 3'
... +++++++++++) EvSite+ 7 evaluation: 159    (+++++++++++) EvSite+ 5 evaluation: 133
                    (-----------) EvSite- 6 evaluation: -516          (++ ... EvSite+
            (-----------) EvSite- 10 evaluation: -645   (+++++++++++) EvSite+ 15 evaluation: 155

           *         *250      *         *260      *
5' a t c c t t g a t g g t t a t a c t a c a t g c g a 3'
            (+++++++++++) EvSite+ 8 evaluation: 600
... +++++++++) EvSite+ 11 evaluation: 619
                        (-----------) EvSite- 12 evaluation: -140
```

Figure 1: Genetic sequence of a computer organism.

The organism has two parts, a weight matrix gene and a binding site region. The gene for the weight matrix covers bases 1 through 125. It consists of 6 segments 20 bases wide and one tolerance value 5 bases wide. Each segment contains sequence specifying the weights for the four nucleotides. For example, bases 1 to 5 contain tcttt. Translating this to binary gives 1101111111, which is the twos complement number for -129. This is the weight for A in the first position of the matrix. The 16 non-overlapping binding site locations were placed at random in the remaining portion of the genome. Evaluation by the weight matrix is indicated for each site. For example site 1, covering positions 132 to 137, catctt, is evaluated as $-442 + 296 - 136 + 251 + 294 - 92 = 171$. Since this is larger than the threshold (-58), it is 'recognized', and is marked with '+' signs. Evaluations to determine mistakes are for the first 256 positions on the genome. An extra 5 bases are added to the end, but not searched, to allow the sequence logos in Fig. 3 to have complete sequences available at all positions. Mutations are applied to all positions in the genome, so the binding sites and the weight matrix co-evolve. The figure was generated with programs **ev**, **evd** and **lister**.

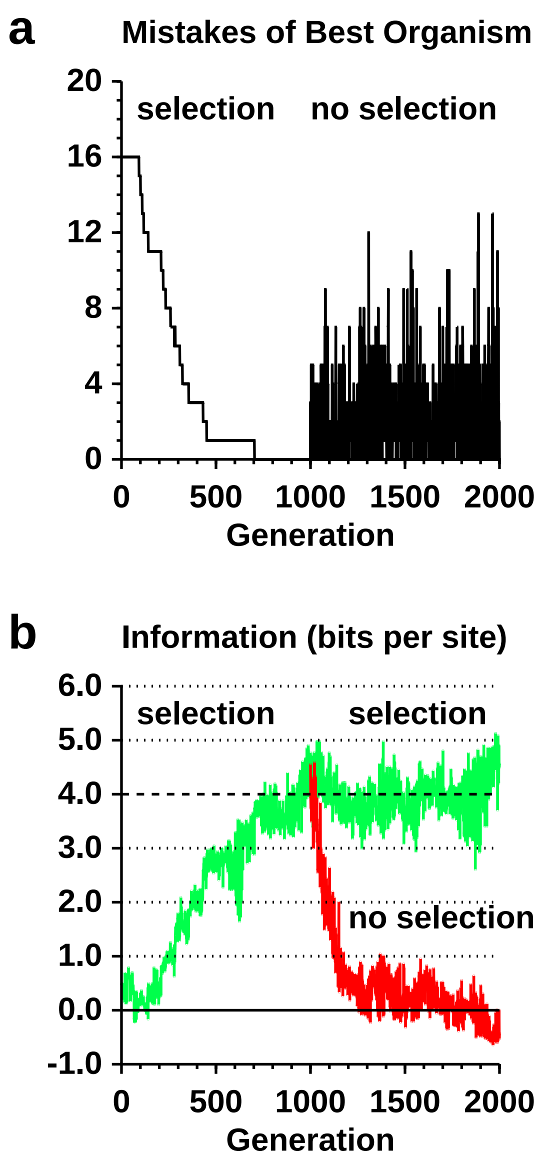

Figure 2: Information gain by natural selection.
a, Number of mistakes made by the organism with the fewest mistakes is plotted against the generation number. At 1000 generations, selection was removed. Because of the initial random number arbitrarily chosen for this particular simulation (0.3), the initial best organism only made mistakes in missing the 16 sites, but this is generally not the case. (Displaying the best organism, which is most likely to survive, is a form of selection that does not affect the simulation.) **b**, The information content at binding sites ($R_{sequence}$) of the organism making the fewest mistakes is plotted against generation number. Selection for organisms making the fewest mistakes was applied from generation 0 to 2000 (top curve, green). The simulation was then reset to the state at 1000 generations and rerun without selection (bottom curve, red). The dashed line shows the information predicted, $R_{frequency} = 4$ bits, given the size of the genome and the number of binding sites.

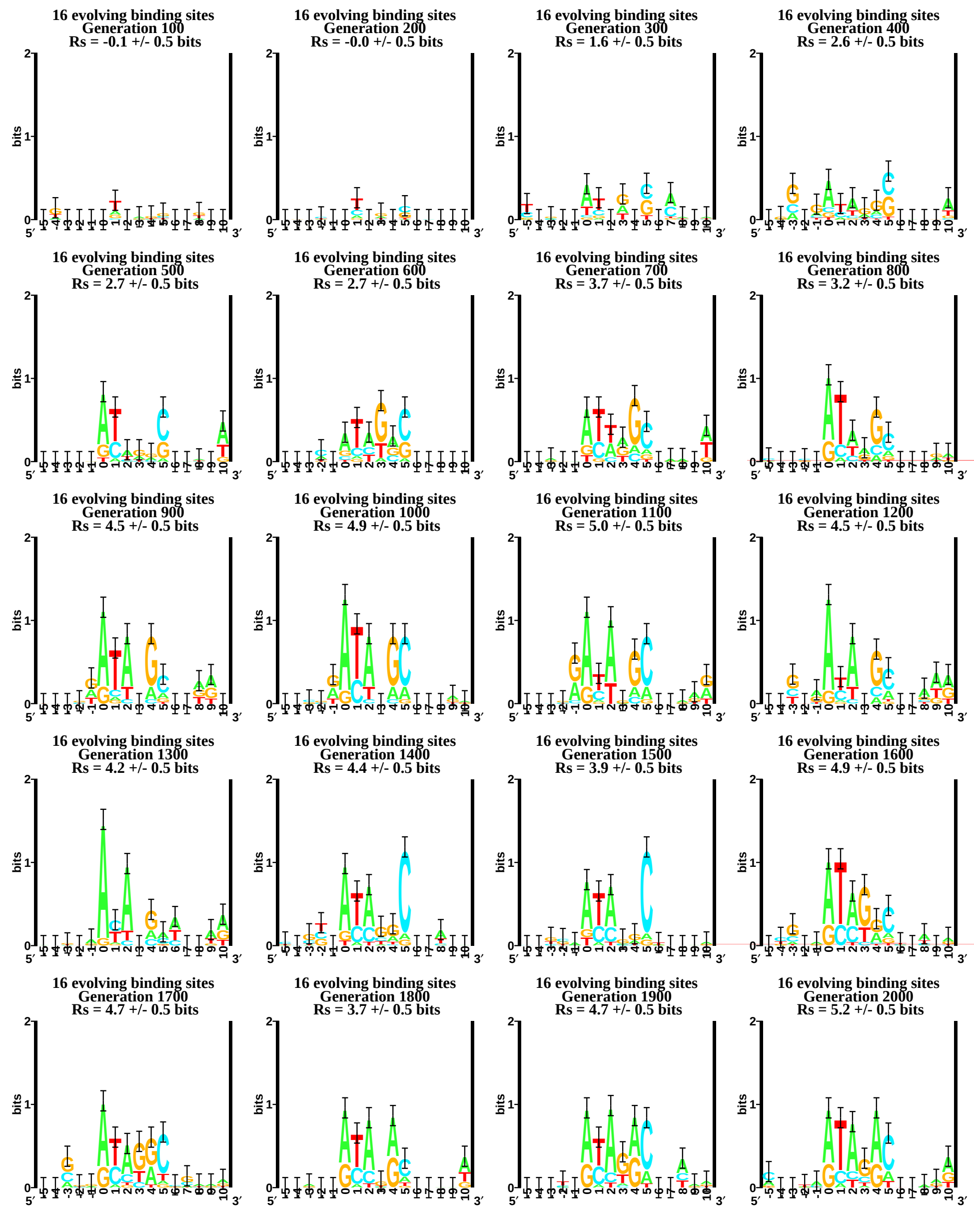

Figure 3: Sequence logos showing evolution of binding sites.

A sequence logo shows the information content at a set of binding sites by a set of stacks of letters (5). The height of each stack is given in bits, and the sum of the heights is the total information content, $R_{sequence}$. Within each stack the relative heights of each letter are proportional to the frequency of that base at that position, $f(b,l)$. Error bars indicate likely variation caused by the small sample size (4), as seen outside the sites, which cover positions 0 to 5. The complete movie is available at http://www.lecb.ncifcrf.gov/˜toms/paper/ev/movie.