

Evolutionary Principles of Genomic Compression

David C. Krakauer

SANTA FE INSTITUTE, HYDE PARK ROAD, SANTA FE, NM 87501 USA.
KRAKAUER@SANTAFE.EDU

The Saturnian stretched out his hand, seized with great dexterity the ship which carried those gentlemen, and placed it in the hollow of his hand without squeezing it too much, for fear of crushing it.... It was not until both Sirian and Saturnian examined the "turds" with microscopes that they realized the amazing truth. When Leeuwenhoek and Hartsoeker first saw, or thought they saw, the minute speck out of which we are formed, they did not make nearly so surprising a discovery. What pleasure Micromegas and the dwarf felt in watching the movements of those little machines, in examining their feats, in following their operations! How they shouted with joy!

Voltaire. Micromegas.

1 The selection for Compression

One of the most remarkable features of life on earth is its diversity. This stands in stark contrast to our typical representation of extraterrestrial life forms, typically conceived of as anatomically extended primates, that are moreover the sole representatives of life within their civilization. In our depiction of Aliens, we are decidedly pre-darwinian. We have imagined a world in which there is but a single species of a single form. Perhaps telescopic distance collapses diversity to such an extent that we are incapable of conceiving of differences. Like Voltaire's gigantic Sirian and Saturnian, who found it difficult to imagine minute humanity, merely homogenous specks - or worse "turds" - capable of reason. Darwin provided us with a theory with which we might understand the origins of diversity at all scales of organization. A theory which I shall argue is particularly suited to the smaller scales of organization. The theory only requires some means of generating diversity, a mechanism for transmitting information, and a phenotype-environment correlation producing selective differences among variants (Lewontin 1970). Those variants best able to survive (high viability), best able to replicate (high fertility) and best able to reproduce, will come to dominate in a population. Darwin's explanation for diversity is summarized in his much cited paragraph on the tangled bank,

" It is interesting to contemplate an entangled bank, clothed with many plants of many kinds, with birds singing on the bushes, with various insects flitting about, and with worms crawling through the damp earth, and to reflect that these elaborately constructed forms, so different from each other, and dependent on each other in so complex a manner, have all been produced by laws acting around us. These laws, taken in the largest sense, being Growth with Reproduction; inheritance which is almost implied by reproduction; Variability from the indirect and direct action of the external conditions of life, and from use and disuse; a Ratio of Increase so high as to lead to a Struggle for Life, and as a consequence to Natural Selection, entailing Divergence of Character and the Extinction of less-improved forms."

A notable feature of this sentence, and of Darwin's theory in general, is that it nowhere provides any mention of the expected range of diversity, or of those biological laws, that might operate to constrain natural diversity. Are there such principles whereby we might predict the diversity of life on earth? Rather than throwing up our hands at the formidable difficulty of this problem, we should consider rephrasing it, all the while ensuring that its character remains intact, but where the degrees of freedom are sufficiently reduced to make the problem tractable. One way in which to do this, is to substitute diversity in life form, with diversity in genome size.

Every naturally occurring, living entity on the planet earth, relies on some form of nucleic acid for inheritance. For all but a few laboratory strains, this means RNA or DNA. Furthermore, all living organisms make use of a triplet code - one in which three nucleotides are required to specify an amino acid - and employ amino acids as the building blocks of proteins. Even upon a more detailed microscopic inspection of the genetic apparatus there are observed common principles of replication, transcription and translation. But when we turn to the

size of the primary RNA or DNA sequence, measured in the total number of nucleotides (the C-value), similarities across species all but disappear. In a virus such as ϕX there are around 0.0000054×10^9 base pairs, in the bacterium *E. coli* there are about 0.004×10^9 nucleotide base pairs. In yeast 0.004×10^9 , in flies 0.18×10^9 and in humans 3.5×10^9 . Lest genome size be mistaken for some measure of phenotypic or evolved complexity, the primitive lungfish possesses on the order of 140×10^9 base pairs and in the common Fritillary 130×10^9 . This is a range of genome sizes spanning over ten orders of magnitude.

The questions that I shall address in this paper all relate to those pressures and constraints favoring small genome sizes, in particular adaptive theories that seek to explain how a given quantity of information can come to be represented by a small message. In other words I seek to discuss mechanisms for compressing genetic information. One thing that we can be certain of, is that all else being equal, larger genomes provide the *opportunity* for more genes and consequently more proteins. More genes and proteins allow for *potentially* greater control over replication, metabolism and over the environment. If as much functional information could be packed into a virus genome with around 10^4 base pairs as into a human genome with 10^9 base pairs, it is likely that there would be far less diversity in genome size.

In this review I shall consider genome size only from from an adaptive compression perspective. There are many alternative hypotheses that seek to explain variation in genome size. These include theories of neutral drift, self replicating parasitic sequences, the contribution of nucleoskeleton structure, and a symbiotic or parasitic mode of life.

Neutral drift theories contend that the selective differences brought about through changes in nucleotide content are so small as to be unimportant to organismal fitness. Increases or decreases in genome size reflect mechanistic biases favoring nucleotide accumulation or nucleotide deletion (Ohno 1976, Charlesworth 1996; Pagel and Johnstone 1992).

Parasitic sequences, such as self splicing introns, are able to propagate themselves horizontally within the host genome. Variation in genome size reflects a mutation selection balance in which parasites strive to increase their numbers, whereas the host seeks to purge unwanted nucleotides (Doolittle and Sapienza 1980, Orgel and Crick 1980). Variation in genome size reflects parasite exposure, parasite tropism, and parasite tolerance.

The total number of nucleotides within a single copy of the sequence of an organisms genome is positively correlated with cell volume (Mirsky and Riss 1951, Cavalier-Smith 1982). Cell volume bears significantly on the rates of cell division, cellular metabolism and development. Once selection has lead to large cell volumes, the bulk properties of DNA rather than the coding properties of its sequence, can play an important role as a support structure or nucleoskeleton (Cavalier-Smith 1982).

Numerous pathogenic and mutualistic intracellular bacterial species possess genomes smaller than their free living relatives. The reduction in genome size is thought to reflect the elimination of those genes encoding peptides that are readily available from the host species (Anderson and Kurland 1998). The causes for genome reduction are hypothesized to be a

reduction in mutation load and or a reduction in the cost of carrying a large genome.

All of these hypotheses are supported by data from various taxonomic groups. None deal explicitly with the selective consequences of compressing information through reduced redundancy, overlapping genes or translational coupling. These issues are discussed in this paper.

We can organize our thoughts on this subject by recognizing three very general principles favoring genomic compression

- the stable propagation of information
- the rapid propagation of information
- the efficient processing of information

Each of these very general biological principles can be coupled to more mathematical concepts: stability to the concept of the error threshold, propagation to strategies for rapid replication, and processing to the efficient exploitation of finite resources. In the following review I shall be adapting ideas from Krakauer (2000), Krakauer and Plotkin (2002), Krakauer and Jansen (2002) and several other works cited throughout the paper.

2 Biological principles of compression

A single stretch of DNA or RNA encodes multiple messages. Most straightforwardly, as non-overlapping linear arrays of genes, encoding proteins by means of a one to one mapping of codon to amino acid. At a higher level, the choice of nucleotide within a chromosome can also reflect selection for chromatin binding motifs or for stabilizing sequences of GC rich DNA. These give rise to the supra-genic structures known as isochores. Within a gene there are further constraints associated with the choice of codon for a given amino acid. There are codes within codes (Trifinov 1989). Selection operates on all of these codes at once, and it is unlikely that they can be concurrently optimized. The genomes we see today are the historical products of each of these countervailing requirements. In the following sections I shall explore how information is compressed in the genetic code and in the translational apparatus, and what implications this has for organismal fitness.

2.1 Redundancy and degeneracy in translation

Each amino acid is associated with one or more codon of three nucleotides of DNA or RNA. Each codon encodes an amino acid through an adaptor molecule carrying an amino acid –

the transfer RNA (tRNA). The tRNA binds to the codon with a complementary anticodon according to the Watson-Crick base pairing rules. These describe the purines adenine and guanine binding with the pyrimidines thymine and cytosine. As there are 64 possible codons and only around 20 amino acids, the mapping from codon to amino acid is many to one giving rise to synonym redundancy in the genetic code. The vast majority of organisms employ the same set of codons to encode amino acids. This led to the code being referred to as universal. More detailed comparative analysis of translation has revealed that there are variant assignment rules, particular in mitochondria, bacteria and protists. The universal code has been demoted to the canonical code. If we examine translation in still greater detail, we find an even larger increase in diversity. This diversity arises in the mapping of codon to anticodon, and by extension, anticodon to amino acid. The genetic anticode is not even canonical.

We can quantify the diversity of the genetic anticode in terms of redundancy and degeneracy. Anticodon redundancy refers to those instances in which a single anticodons process several codons in order to encode an amino acid. Whereas anticodon degeneracy, involves several anticodons binding several codons, to encode a single amino acid. Within the standard code, the total number of tRNAs can vary theoretically between 22, in which case each anticodon recognizes on average 3 codons, to 61 in which case each anticodon binds strictly to its complementary codon (setting aside the 3 stop codons). This leads to two or more tRNAs carrying the same amino acid (isoaccepting tRNAs). Using the above terminology, employing a large number of specific anticodons (61 for example) to encode 20 amino acids is a strategy of degeneracy, as non-identical anticodons must map onto a common set of amino acids. By employing a smaller number of anticodons than codons, the system becomes redundant, as the same anticodon is required to bind to more than one codon.

A redundant strategy allows genomes to carry fewer tRNA molecules and acts as a simple mechanism for error buffering by not distinguishing among similar codons. However, redundancy involves a potential cost in terms of reduced binding specificity and more frequent mismatch errors. Degeneracy provides greater specificity of binding, reducing translation errors by reducing the incidence of near-cognate codon readings. It also has the virtue of greater evolutionary flexibility as the number of elements available for potential modification, is increased. With a degenerate strategy, the set of bound codons can be modified by simply expanding or contracting the set of anticodons. However, a degenerate strategy requires more tRNAs and an attendant increase in the genome size.

2.2 Kinetics of translation and replication

Following Krakauer and Jansen (2002) we assume that there are m different codons, m anticodons and n amino acids. The $m \times 1$ vector \vec{c} specifies the abundances of each of the m codons in the RNA strain. The process of translation involves matching these codons with tRNA anticodons. The abundance of anticodon i in the cell is given by elements v_i of the anticodon vector \vec{v} . The binding rate of anticodon i with codon j is given by element

w_{ij} of the binding matrix \mathbf{W} .

The total rate of binding to codon j is given by $\sum_i^m v_i w_{ij}$, hence the average time it takes for a codon to be matched is given by

$$f_j = \left(\sum_i^m v_i w_{ij} \right)^{-1}. \quad (1)$$

It also follows that codon j is matched with anticodon i with probability

$$u_{ij} = v_i w_{ij} f_j. \quad (2)$$

Because each codon will be matched with an anticodon we have $\sum_{i=1}^m u_{ij} = 1$. The $m \times m$ matrix \mathbf{U} has the probabilities u_{ij} as elements. Each tRNA anticodon is associated with an amino acid. The association between anticodon and amino acid is described by the $n \times m$ binary (i.e. containing only zeros and ones) matrix \mathbf{A} . The elements of \mathbf{A} are denoted a_{ij} . Each column of \mathbf{A} consist of $n - 1$ zeros and only a single one as there is a unique association between tRNA and amino acid. In case there is degeneracy rows of \mathbf{A} can contain many ones. Matrices in which any element $u_{ij} < 1$, have redundancy in the i th anticodon. The abundance in amino acids after translation is given by $\mathbf{AU}\vec{c}$.

The total translation time is given by the sum of the time it takes to initiate the translation process, ϵ (binding of the message to the ribosome), and the total time to match all codons

$$\tau = \epsilon + \sum_j^m (c_j f_j) \quad (3)$$

Assuming that the extension of the protein is the rate limiting step, and not the removal of improperly bound anti-codons, the rate of translation is given by $q = 1/\tau$.

A similar reasoning might be applied to replication of DNA or RNA. For the sake of concision we might assume that replication rate is reduced by a constant amount for each additional codon and tRNA gene carried by the genome. The total number of codons are fixed by functional constraints. The number of anticodons, as we have established above, can vary. The replication time is then a function of total genome length to include $k \sum_i v_i$. The constant k is the average time taken to replicate all codons of a tRNA molecule. In other words replication rate is proportional to,

$$\frac{1}{G + k \sum_i v_i}.$$

Where G is the replication time of all non tRNA-related codons. The fitness of an asexual organism - ignoring for the present rates of deleterious mutation - is a product of its viability (maximum gene expression rate) and replication rates,

$$w = \frac{q}{G + k \sum_i v_i}. \quad (4)$$

It will be observed from equations (1-3) that an increase in redundancy, for a fixed number of anticodons, reduces the rate of translation. Whereas we see in equation (4), that an increase in degeneracy, increases the rate of replication. Maximizing fitness involves a tradeoff between replication and viability. In organisms with small genomes the term $k \sum_i v_i$ in the denominator of (4) will be relatively large in relation to G . Thus we expect there to be greater pressure towards redundancy. In larger genomes the term G will dominate and maximum degeneracy would seem to be desirable. This leads us to the discovery of an important design principle: genomic compression through anticodon redundancy becomes more important the smaller the genome. This is the general pattern that we observe in nature (Krakauer and Jansen 2002)

2.3 Genomic organization and error thresholds

Following the groundbreaking work of Beadle and Tatum (1941) on mutant lineages of *Neurospora*, there dominated the view, that one gene led to the production of one polypeptide. The assumption of co-linearity of genetic message and protein product, played a pivotal role in helping to explain the mechanism of genetic translation (Fox Keller 2000). However an examination of the genomes of viruses, bacteria and protists, show significant violations of this assumption.

By far the majority of viruses and bacteria contain stretches of DNA or RNA in which constituent nucleotides are translated into two or more different polypeptides. Regions of the genome in which there is this one to many translational mapping are termed overlapping reading frames (Normark et al 1983). Overlapping reading frames allow a larger number of proteins to be encoded by a single stretch of nucleotides than would be feasible with one gene one polypeptide. The information for two or more proteins is compressed into the information space typically occupied by a single protein. This raises a number of mechanical and functional questions: (1) How do two or more genes come to occupy overlapping sequences ? (2) What are the implications of overlapping reading frames for the robustness of the underlying genetic message (3) What are the advantages of overlapping reading frames in terms of replication ? (4) Are certain kinds of gene more likely to be found overlapping ?

Before exploring some of these questions as they relate to compression, we need to consider certain evolutionary limits placed on any replicating strand of DNA or RNA, in the absence of overlap. The theory of mutation-selection balance on genetic sequences was first developed by Eigen (1971) to explain the dynamics of RNA replication in a flow reactor. The standard presentation of this concept involves, establishing a sequence space, imposing fitness differentials upon this metric (a fitness landscape), and writing down a dynamical system determining the time evolution of replicators moving through this space. The sequence space is constructed by assuming that there a set of sequences of uniform length comprising N monomeric units each of which can be drawn from a class of size C . The dimension of of the sequence space is N and the total number of unique sequences C^N . We can calculate the distance between any two sequences in this sequence space using the

Hamming distance $d(i, j)$ which tabulates the number of positions in a sequence of length N at which monomers are different. Assuming a wildtype sequence S_w , we can generate Hamming classes by grouping together all of those sequences equidistant from the wildtype, S_d , where $d = d(w, j) \in (1, 2, \dots, N)$. These Hamming classes are important as we shall impose a radial symmetry constraints on our adaptive landscape whereby we assume that all members of a given Hamming class have identical fitness. Moreover, adjacent classes can be reached through a single monomer/base mutation.

During the replication of sequences there are small probabilities of mutational errors. Each site is replicated accurately with a probability q , and hence is mutated with a probability $1 - q$. The mutation probabilities between all sequence pairs (S_i, S_j) of the C^N sequences is given by a mutation matrix,

$$Q = \{Q_{ij}; i, j = 1, 2, \dots, C^N\} \quad (5)$$

where

$$Q_{ij} = q^{N-d(i,j)} \left(\frac{1-q}{C-1} \right)^{d(i,j)} \quad (6)$$

In the standard quasispecies equation it is assumed that the rate constants for replication of sequence S_i are given by a_i , and sequence decay d_i . The differential equation is,

$$\dot{x}_i = (a_i Q_{ii} - d_i - \phi(t))x_i + \sum_{j \neq i} a_j Q_{ji} x_j, \quad i, j = 1, \dots, T. \quad (7)$$

where the total population size (T) is kept constant by matching the productivity of the system with an outflow or ‘flux’ term, $\phi(t) = \sum_{i=1} (a_i - d_i)x_i(t)$. Sequence x_i increases in abundance through replication or through mutation of a sequence S_j into S_i with a probability Q_{ji} . This equation can be substantially simplified by assuming that the decay rates of all sequences are equivalent, and that the back mutation is negligible in comparison with forward mutation: $\sum_{j \neq i} a_j Q_{ji} x_j \ll a_i Q_{ii}$. This gives us a linear system,

$$(AQ - \phi)\mathbf{x} = 0 \quad (8)$$

where the matrix A is the diagonal matrix of replication rates, ϕ the diagonal matrix of outflows, and \mathbf{x} the vector of x_i values. This is now a standard eigenvalue problem, in which the equilibrium distribution of \mathbf{x} are given by the eigenvector associated with the dominant eigen value of AQ .

2.4 Plateau landscapes

To determine the whereabouts and stability of the equilibria we need to establish a geometry of the adaptive landscape. The simplest assumption is that of a single peak landscape, in which the wildtype sequence has replication rate a_w and all other hamming classes a fitness

a_m , where $a_w > a_m$. It is common to assume that $a_w = 1$ and $a_m = 1 - s$, where s is the selective cost of mutation. The selective advantage of the wildtype is $\alpha = 1/a_m$. The equilibrium frequency of the wildtype sequence in the plateau landscape is then given as,

$$\bar{x}_w = \frac{Q_{ww}\alpha - 1}{\alpha - 1} \quad (9)$$

A stability analysis of this equilibrium shows us that this equilibrium remains stable whenever $Q_{ww} > 1/\alpha$, or written differently, when $Q_{ww} > a_m$ or $Q_{ww} > 1 - s$. In order to determine the relationship between stability and sequence length we solve $Q_{ww} = 1/\alpha$ which can be rewritten as $q^N = 1/\alpha$ to give us $N_{crit} = \ln(\alpha)/(1 - q)$. For reasonable values of α ($1 < \alpha < 10^9$), we find that $1 - q \approx 1/N_{crit}$. The error threshold for plateau landscapes is given by a mutation rate equal to the reciprocal of the genome length. From the perspective of genome compression, the error threshold concept illustrates why smaller genomes are more stable than larger genomes – they require lower replication fidelity in order to persist. To ensure that the wildtype is not lost as a result of an increase in genome size, there must be an unrealistically large increases in the fitness of the wildtype. Thus the error threshold imposes a maximum on the amount of information that can be transmitted through a linear genome. An increase in genome length that brings about new adaptive functions, without an increase in replication fidelity, is very unlikely to persist. This has profound implications for the trajectory of the evolution of replicators across plateau landscapes. It states that in order for new gene functions to evolve, a replicator must first have a sufficiently low rate of error to accomodate further contributions to genome length. Stated differently, innovations that increase replication fidelity will occur before innovations that increase replication rate.

2.5 Multiplicative landscapes

A more realistic landscape is the symmetric, multiplicative landscape of the form $a_i = (1 - s)^i$ for which i are the number of mutations and s the selective costs of mutations. There are $N+1$ equilibrium solutions to this system, in which each equilibrium varies in the number of wildtype monomers maintained, from N down to 0 . This system does not strictly have a ‘threshold’ or phase transition at a critical value of the mutation rate, as is observed in the plateau landscape. The abundance of the wildtype sequence gradually decreases for increasing rates of mutation. However, it remains true to say that when the wild type - S_0 is eventually outcompeted, then the population will be driven to extinction. It is convenient to use the notation $\mu = (1 - q)^N$.

The equilibrium frequency the equilibrium frequency distribution in the single peak, multiplicative function landscape is given as,

$$\hat{x}_i = \binom{N}{i} \left(1 - \frac{\mu}{\bar{s}}\right)^i \left(\frac{\mu}{\bar{s}}\right)^{N-i}, \quad (i = 0, 1, \dots, n), \quad (10)$$

in which the mean number of functional sites is $\bar{i} = n(1 - \mu/\bar{s})$, and the population mean fitness is $\bar{w} = (1 - \mu)^n$. As with the plateau landscape, the stability of the equilibria

are calculated by linearising the dynamics. After a rather involved set of calculations the condition for stability of the wildtype is given by ,

$$\mu < s. \tag{11}$$

This result shows that in multiplicative landscapes the error threshold is independent of genome length. It is however misleading, as increasing the genome length leads to very large reductions in the abundance of the wildtype. We can show this by writing down the ratio of of the wildtype to the nearest one-monomer mutant class:

$$\frac{\hat{x}_{N-1}}{\hat{x}_N} = \frac{N\mu/s}{1 - \mu/s} < 1, \tag{12}$$

or

$$N + 1 < \frac{s}{\mu} \tag{13}$$

and hence for large genomes,

$$N_{crit} \approx \frac{s}{\mu} \tag{14}$$

The multiplicative adaptive landscape in the limit of $s \rightarrow 1$, converges to the plateau landscape, and yields the error threshold $N_{crit} \approx \frac{1}{\mu}$. The important difference from the plateau landscape for the evolution of compression is that genome size can increase through genetic innovations that provide adaptive benefits without prior increases in mutation fidelity. In other words, genome size is not as important a factor in genomic evolution across multiplicative fitness landscapes as plateau landscapes.

2.6 Introducing structure: overlapping reading frames

The standard quasispecies model does not distinguish among sites in the genome, and all sequences are of equivalent length. In the simplest plateau landscape, a single mutation at any site, leads to the maximum fitness decrement. In continuously differentiable landscapes, a larger number of mutations are associated with a larger decline in fitness. Genomes with overlapping genes experience different fitness reductions according to the region in which the mutation occurs. In overlapping regions, more than one gene can be damaged, in non-overlapping regions only a single gene can be damaged. Within the overlapping regions there is variation in mutation load according to the direction and phase of overlap. Furthermore, sequences are no longer of uniform length and vary according to their degree of overlap, M . This fact can lead to variation in replication rates. We need to consider each of these properties of overlapping genes in order to derive an appropriate adaptive landscape.

2.6.1 Mutation incidence

Consider a sequence comprising two reading frames each of length N and overlapping reading frame length M . The total genome length is then given by $2N - M$. If mutation rates are homogeneous across the genome, the probability of a single mutation falling within an overlapping region is proportional to,

$$p = \frac{M}{2N - M} \tag{15}$$

and in a non-overlapping region by

$$v = \frac{2(N - M)}{2N - M} = 1 - p \tag{16}$$

The values p and v determine the probable whereabouts of mutations within the genome.

2.6.2 Direction and phase of overlap

Once a mutation arises in an overlapping sequence, the mutation load will depend upon the reading frame within which overlap is observed. This is because the degree to which polypeptides encoded by overlapping genes achieve selective independence, depends upon the direction and phase of overlap. Overlap can increase the mutational load in a genome by reducing the effective redundancy of a sequence. Consider a pair of overlapping genes, in which gene I is read +1 out of phase of gene II. Translation is initiated one nucleotide closer to the 5' end of the genome. The first codon position of the +1-phase (F_{+1}) gene will correspond to the second codon position of the 0-phase gene (F_{+0}), whereas the first codon position of the 0-phase gene corresponds to the third codon position of the +1-phase gene. The two phases of parallel overlap and three of anti-parallel overlap are constructed by shifting each nucleotide along a codon by F_{+1} , F_{+2} , or F_{-0} , F_{-1} , F_{-2} counting mod 3:

```

1 2 3 1 2 3 1 2 3 1 2 3 1 2 3
...1 2 3 1 2 3 1 2 3 1 2 3 1 2 3
.....1 2 3 1 2 3 1 2 3 1 2 3 1 2 3
.....3 2 1 3 2 1 3 2 1 3 2 1 3 2 1
...3 2 1 3 2 1 3 2 1 3 2 1 3 2 1
3 2 1 3 2 1 3 2 1 3 2 1 3 2 1

```

We can provide an estimate of the mean mutational load associated with each of these five overlapping sequence configurations. By examining a table of the canonical genetic code, we determine the probability of a single missense or nonsense mutation in translated polypeptides (non-synonymous mutations) as a result of mutations to the first, second

or third position of a codon. The probability of an amino acid mutation to k polypeptides translated from an overlapping reading frame i nucleotides out of phase is denoted by $\pi_k(F_{*i})$, where $*$ takes the values $+$, $-$. At the first position the probability of a non-synonymous substitution is ≈ 0.65 , at the second position ≈ 0.75 , and at the third position ≈ 0.3 . The corresponding synonymous mutation probabilities are 0.35, 0.25 and 0.7. The average probability of mutation to two polypeptides derived from an overlapping gene is given by the probabilities of non-synonymous mutations at any one of the three codon positions and the probabilities of simultaneous mutations to each of the polypeptides. For a gene in $+1$ or $+2$ phase, the probabilities, denoted by $\pi_2(F_{+1})$ and $\pi_2(F_{+2})$, are given by $\frac{1}{3}(0.65 \times 0.75 + 0.75 \times 0.3 + 0.65 \times 0.3) = 0.30$. Through identical calculations, the probabilities for the remaining configurations are: $\pi_2(F_{-0}) = 0.32$, $\pi_2(F_{-1}) = 0.35$, $\pi_2(F_{-2}) = 0.29$. The probability of a single mutation to a translated polypeptide ($\pi_i(F_{*i})$) is derived from the synonymous mutation probability and the simultaneous non-synonymous mutation probability: $\pi_1(F_{*i}) = 1 - \pi_0(F_{*i}) - \pi_2(F_{*i})$.

2.6.3 Replication rate

Fitness is a product of both viability and fertility or replication rate. In the absence of overlap, fitness is synonymous with viability as all sequences are of uniform length. When the degree of overlap is free to vary, then the fitness function should reflect the replication costs of carrying a larger number of monomers. Replication time increases linearly with increasing sequence length when nucleotides are not rate limiting and there is a single replication initiation site. We allow heterogeneity in genome size to influence replication rate through a weighting function,

$$w(M) = \frac{N}{2N - M}. \quad (17)$$

When overlap is at its maximum for two genes of equal length, $M = N$, the replication rate weighting factor will be equal to 1. When there is no overlap ($M = 0$) the weighting terms will be equal to $\frac{1}{2}$. Assuming a genome comprising two genes of equal length there is a two-fold advantage in replication rate for complete overlap. There is an n -fold advantage for complete overlap in a genome comprising n identical length genes.

2.7 Adaptive landscapes with overlap

In order to specify the adaptive landscape for overlapping genes we shall take into account mutation incidence, the nature of the overlap and the replication rate. We also need to choose the geometry of the landscape. We can either (1) consider the average effect of mutations falling into non-overlapping and overlapping regions of a sequence with a prescribed

degree of overlap or (2) split the sequence population into two sub-populations in which one population describes all those sequences with mutations to non-overlapping regions, and the second population, sequences with mutations to overlapping regions. The first approach is more approximate but allows us to describe all members of the population with a single fitness function. The second approach is more accurate but requires that we describe each overlapping mutant class independently. For plateau landscapes, the second approach is relatively simple, as the dimensions of the system are kept small from grouping together sequences of identical fitness. I shall concentrate on the plateau landscape and only touch on the multiplicative landscape.

2.7.1 Plateau landscapes with overlap

Denote the wildtype sequence x_0 , those genomes with mutations falling within a non-overlapping sequence x_1 and genomes with mutations in overlapping sequences x_2 . Mutations to non-overlapping sequences can be either synonymous or non-synonymous. Mutations falling within overlapping sequences can be synonymous, non-synonymous within a single translated polypeptide or non-synonymous within two translated polypeptides. The per genome per generation mutation rate is given by $\mu = q^{2N-M}$ and we neglect the structure of the mutation matrix Q_{ij} . For a genome with an overlapping reading frame M , the replication dynamics are given by,

$$\dot{x}_0 = (a_0(1 - \mu) - d - \phi(t))x_0 \quad (18)$$

$$\dot{x}_1 = (a_1 - d - \phi(t))x_1 + a_0\mu x_0 v \quad (19)$$

$$\dot{x}_2 = (a_2 - d - \phi(t))x_2 + a_0\mu x_0 p \quad (20)$$

The replication rates a_i for each mutant class with direction and phase of overlap F_{*j} are given by:

$$a_0 = w(M) \quad (21)$$

and

$$a_i = w(M) \frac{1}{i+1} \sum_{k=0}^i \pi_k(F_{*j})(1 - ks), \quad (i = 1, 2) \quad (22)$$

with a flux term,

$$\phi(t) = \sum_{i=0}^2 x_i(t)(a_i - d). \quad (23)$$

The stability of the wildtype is determined by the dominant eigenvalue of the linearized system. From the above definitions, stability is achieved when

$$1 - \mu > \frac{a_0}{a_1} \quad (24)$$

which upon substitution from equation (22) gives,

$$1 - \mu > \frac{2}{\pi_0(F_{*j}) + \pi_1(F_{*j})(1 - s)}. \quad (25)$$

From the LHS of (25) the stability of the wildtype increases as μ decreases which is true for increasing amounts of overlap M . Compressed genomes can tolerate larger error rates, as the total number of mutations per genome is reduced. Stability also depends on the phase and direction of overlap, represented by the values of the $\pi_0(F_{*j})$ and $\pi_1(F_{*j})$ terms. Increases in the probabilities of non-synonymous substitutions increase the stability of the sequence.

An important consequence of compression is a reduction in sequence redundancy. Sequence redundancy can be defined approximately as,

$$r = \frac{\pi_0(F_{*j})}{\sum_{i=0}^2 \pi_i(F_{*j})} \quad (26)$$

which is simply the relative probability of a synonymous substitution in a sequence with overlap phase F_{*j} . High values of redundancy threaten the stability of the wildtype. Low values of redundancy bolster the stability of the wildtype. This is because redundancy reduces the competitive differences among fitness classes in a population and thereby threatens the long term persistence of mutation free genomes. A reduction in redundancy increases the mutation load and thereby more effectively purges mutant genomes from the population (Krakauer and Plotkin 2002, Krakauer and Nowak 1999). Genomic compression can thereby increase stability in two ways: (1) by reducing mutation incidence and (2) by reducing sequence redundancy.

This is the opposite result from the those described in information theory in which redundancy increases the fidelity of information transfer across a noisy channel. The explanation for this important difference derives from the population dynamical transmission of information in evolutionary systems. Here we have multi-access information but where the channel remains undivided and a large number of transmitters compete for their message to be processed by the receiver. The transmission of information is non-cooperative. With small populations, redundancy in the message is once again favored, as it can increase the probability that not all individuals within the population harbor defective genomes (Krakauer and Plotkin 2002).

2.7.2 Multiplicative landscapes with overlap

With multiplicative landscapes we are required to track each mutant class. With two non-overlapping genes of identical length N there are $\binom{2N}{i}$ ways of experiencing i mutations. With an overlap of size M there are $\binom{2N-M}{i-k}$ ways of experiencing $i-k$ mutations to a non-overlapping sequence and $\binom{M}{k}$ ways of experiencing k mutations in an overlapping sequence. To calculate the mean fitness of a genome with a total i mutations we need to consider probable mutations to both overlapping and non-overlapping sequences, and the direction and phase of overlap within these sequences,

$$a_i = w(M) \sum_k \binom{i}{k} p^k v^{i-k} (1 - s\pi_1(F_{*j}) - 2s\pi_2(F_{*j}))^k (1 - s\pi_1(F_{*j}))^{i-k} \quad (27)$$

$$= w(M) v^i (1 - s\pi_1(F_{*j}))^i \left(1 + p \left(\frac{1}{v} + \frac{2s\pi_2(F_{*j})}{vs\pi_1(F_{*j}) - v}\right)\right)^k \quad (28)$$

This function represents the expected fitness after having experienced k out of i mutations to an overlapping sequence and $i-k$ mutations to a non-overlapping sequence. Mutations falling within an overlapping sequence can lead to 0, 1 or 2 amino acid substitutions in a protein with a fitness cost of 0, s or $2s$. Mutations falling within a non-overlapping sequence can lead to 0 or 1 amino acid substitution with a fitness cost 0 or s . Setting $p = 0$, the above equation reduces to the simple multiplicative fitness function for a non-overlapping sequence of monomers.

The analysis of the stability properties of this model are beyond the scope of this review. However a few general properties of this landscape can be seen immediately. Firstly, that overlapping-gene landscapes are steeper than the non-overlapping landscapes, as $1 - s\pi_1(F_{*j}) - 2s\pi_2(F_{*j}) < 1 - s\pi_1(F_{*j})$. Secondly, that increasing the sequence redundancy (r) through changes in the direction and phase of overlap, lead to a reduction in the steepness of the landscape. In other words a reduction in $\sum_{i=0}^2 \pi_i(F_{*j})$. As landscape steepness determines the stability of the wildtype, we see directly how redundancy influences heritability.

2.8 Overlap and limits to evolvability

In the previous section I showed how a compressed sequence encoding more than one protein, is more vulnerable to mutations than a non-overlapping gene. There is another potential fitness constraint on overlapping genes. It is perhaps unrealistic to assume that the same amount of information can be encoded in a sequence of length $2N$ as in a compressed sequence of length N . Standard information theoretical considerations of compression algorithms – used to source encode messages – do so by removing all redundancies. It is assumed therefore that the underlying message remains the same. In evolutionary contexts, we worry

not only about current performance, but potential performance in alternative environments. Genomes, particularly those of bacteria or viruses, need to remain adaptable. Overlapping reading frames impose a limit to adaptability by introducing correlations, described by geneticists as epistasis, into gene functions. Adaptability, or evolvability, are related to mutability in rather complex ways. On the one hand, in rapidly changing environments or in very large population sizes, mutability might promote evolvability by generating a sufficient number of novel variants to spawn an adaptive lineage. In slowly varying environments, or in small populations, redundancy will often be favored as it can promote efficient exploration of sequence space. The choice of penalty for overlapping genes should therefore consider the evolutionary context. If we only consider the case where redundancy would be favored, then the cost of overlap defined in terms of the average degree of mutability, will be negatively correlated with redundancy. Approximately redundancy is the inverse of mutability, which we might define as $R_{*i} = 1 - \pi(F_{*i})$.

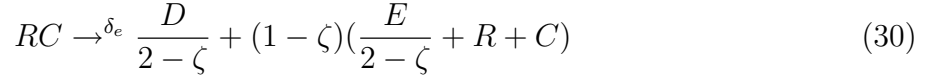
2.9 Efficient processing through translational coupling

Overlapping reading frames can also serve to expedite efficient translation. Rather than thinking about overlap exclusively as a means of increasing rates of replication and minimizing mutation load, we can think about overlap in terms of bringing neighboring genes into contact with translational machinery to ensure some form of coordinated or regulated expression. This is often achieved by fusing the termination codon of one gene with the initiation codon of another; this strategy known as translational coupling is exploited both by bacteria and phages (Oppenheim and Yanofsky 1980, Madison-Antenucci and Steege 1998).

The bacterial *trp* operon is a repressible operon in which tryptophan acts as a corepressor. With low levels of tryptophan there is no repression and the operon is transcribed. With high levels of tryptophan the operon is in the off state and there is no transcription. The *trp* operon consists of a set of structural genes that are expressed as a polycistronic mRNA translated into five enzymes responsible for the synthesis of tryptophan. The genes (*trpE* through *trpA*) are for the five enzymes that catalyze the synthesis of the amino acid tryptophan from chorismic acid. The coding sequence of all but two genes, *trpC* and *trpF* overlap by several nucleotides. The termination codon of *trpE* overlaps with the initiation codon of *trpD* by 29 nucleotides. Premature termination of *trpE* reduces the rate of expression of *trpD*. The putative roles of translational coupling are to (1) reduce the initiation time delay attendant upon tRNA and ribosome binding to the mRNA transcript and (2) modulate the difference in the concentration of *trpE* and *trpD*. Overlap can therefore serve two purposes: decrease the total genome size, and improve the regulation of coordinately expressed genes.

This idea can be clarified by analyzing a simple kinetic model. I shall not model the negative regulation of the *trp* operon, but instead assume that it is constitutively expressed. Suppose that $[R]$ is the concentration of the full polycistronic *trp* operon RNA, and C the joint ribosome tRNA complex. The translation of the mRNA of genes *trpD* and *trpE* into the

polypeptides D and E begin after an initiation delay δ_i , after which the complex [RC] forms, and end after an elongation delay $\delta_e = N/k$ where N is the gene length, k the rate constant of chain elongation. With overlapping genes, and complete translational coupling ($\zeta = 1$), there is translation of the E gene only once D has been translated. Without translational coupling ($\zeta = 0$) both D and E are translated at the same rate. The kinetics of translation can be described by:



The kinetic relations describe how [R] binds to [C] to form the translation complex [RC]. After an elongation delay δ_e the proteins D and E are produced through independent ribosome binding with a probability $1-\zeta$. After a further delay $2\delta_e$, the protein E is produced through translational coupling of the trpD and trpE genes with a probability ζ . The differential equations describing these kinetics are provided below.

$$[\dot{R}] = \frac{1}{2\delta_e}[RC]\zeta + \frac{1}{\delta_e}[RC](1-\zeta) - \frac{1}{\delta_i}[R][C] \quad (32)$$

$$[\dot{RC}] = -\frac{1}{2\delta_e}[RC]\zeta - \frac{1}{\delta_e}[RC](1-\zeta) + \frac{1}{\delta_i}[R][C] \quad (33)$$

$$[\dot{C}] = \frac{1}{2\delta_e}[RC]\zeta + \frac{1}{\delta_e}[RC](1-\zeta) - \frac{1}{\delta_i}[R][C] \quad (34)$$

$$[\dot{D}] = \frac{1}{\delta_e}[RC]\frac{1}{2-\zeta} - g_1[D] \quad (35)$$

$$[\dot{E}] = \frac{1}{2\delta_e}[RC]\zeta\frac{1}{2-\zeta} - (1-\zeta)\frac{1}{\delta_e}[RC]\frac{1}{2-\zeta} - g_2[E] \quad (36)$$

After solving for the equilibrium concentrations of D and E, which we denote as \hat{D} and \hat{E} , we find that we can express \hat{E} in terms of \hat{D} as,

$$\hat{E} = \frac{\hat{D}g_1(2 - \zeta)}{2g_2}. \quad (37)$$

With obligate translational coupling ($\zeta = 1$),

$$\hat{E} = \frac{\hat{D}g_1}{2g_2}. \quad (38)$$

and with independent translation ($\zeta = 0$),

$$\hat{E} = \frac{\hat{D}g_1}{g_2}. \quad (39)$$

Thus translational coupling adds a degree of control over the relative concentrations of coupled proteins by harnessing the translational delay as a mechanism for promoting variation in gene products. Assuming that D and E decay at the same rate, D can be more abundant as a result of a smaller delay in production.

A mechanistic derivation for the delay term (δ_e) was provided in the first section of this paper as,

$$\delta_e = \sum_j^m (c_j f_j) \quad (40)$$

where the $m \times 1$ vector \vec{c} specifies the abundances of each of the m codons in the mRNA, whereas the average time it takes for a codon j to be matched is described by the vector \vec{f} . Thus the degree of redundancy or degeneracy in the translational apparatus, or in the codon composition of the translated genes, further modify the rates of protein production through translational coupling. If the two genes differ in their codon compositions then we write the translational delay terms as $\delta_e^{(E)}$ and $\delta_e^{(D)}$ and the equilibrium concentration of E in terms of the equilibrium concentration of D is given by,

$$\hat{E} = \frac{\hat{D}g_1(\delta_e^{(E)} + \delta_e^{(D)} - \delta_e^{(E)}\zeta)}{g_2(\delta_e^{(D)} + \delta_e^{(E)})}. \quad (41)$$

The introduction of coupling brings with it a variety of mechanisms for regulating protein expression. Thus provided there is some degree of translational coupling ($\zeta > 0$), codon composition, relative gene length, and the decay rates of the respective protein products, all play an important role in equilibrium protein concentrations. Compression of the genome serves to both minimize the nucleotide content and to increase the regulatory possibilities.

2.10 Translation compression in ciliates

In the previous sections I introduced compression as a means of (i) reducing the genome size in order to facilitate rapid replication, and (ii) increase the regulatory efficiency of translated genes. In other words, compression works to reduce the size of heritable messages and to reduce the distance between interpreted messages. The former provides an advantage in gene replication and the latter in gene expression. With the example of translational coupling both forms of compression are realized in the same system. We shall now introduce another example in which translational efficiency and replicative efficiency are correlated. This is the case of gene scrambling in hypotrichous ciliates (Prescott 1997).

Hypotrichous ciliates possess two nucleic acid nuclei: a micronucleus which acts as the germ line, and a macronucleus which acts as a somatic nucleus. From one generation to the next only the micronucleus is transmitted, whereas only the macronucleus is transcribed. This effectively decouples replicatory compression from regulatory compression. Following mating ciliates are diploid with respect to the micronucleus. During the formation of a macronucleus from one of the micronuclei a number of events take place: (1) non-coding regions of DNA present in the micronucleus are excised, (2) Spacer DNA is removed, (3) gene-coding regions of the micronucleus are spliced to form complete genes and (4) DNA encoding genes in the micronucleus is amplified by up to 3 orders of magnitude to increase the gene copy number in the macronucleus. The developmental process from micronucleus to macronucleus thus unscrambles genes, removes all non-coding regions and increases gene numbers. Redundancies are introduced into the macronucleus in order to increase the rate of translation.

From the perspective of genomic compression, hypotrichous ciliates manage to achieve an almost ideal balance: gene numbers are kept low in the replicating micronucleus whereas gene numbers are multiplied in the macronucleus to allow for translation in parallel.

3 Summary

Differences in genome size represent one of the most varied measures of diversity among organisms. Small genomes are often favored in order to (1) promote the stable propagation of information through a reduction in mutation load, (2) promote the rapid propagation of information through a reduction in nucleotide content, and (3) promote an efficient processing of information by minimizing transcriptional and translational delays. All of these pressures favoring smaller genomes are more pronounced in the smallest genomes. This is because in organisms with small genomes, genome size leads to significant kinetic bottlenecks influencing viability. Thus we expect small genomes to get smaller. Because the pressure for compression is less severe on larger genomes, other factors such as nucleoskeleton and drift will dominate. This predicts a bimodal distribution of genome sizes.

Compression can be achieved through greater redundancy in the translational apparatus, through overlapping messages in the DNA or RNA sequence, and through a reduction in the length of translated sequences. Compression can also lead to greater coordination in protein production by coupling the translation of functionally related genes.

As in compression in information theory, compression in biological systems is often the result of the elimination of redundancies. Unlike in information theory where redundancy increases the reliability of messages, the heritable stability of biological messages can be increased by eliminating redundancies. This is a result of increasing the competitive superiority of wildtype sequences. This is a strategy that is only favored in large populations. In small populations, concordant with information theory, message heritability is increased through increases in redundancy.

References

- [1] Beadle, G.W., Tatum, E.L. Genetic control of biochemical reactions in *Neurospora*. Proc. Natl. Acad. Sci. 21: 499-506
- [2] Cavalier-Smith, T. Skeletal DNA and the evolution of genome size. Annu. Rev. Biophys. Bioeng. 11:273-302 (1982)
- [3] Charlesworth, B. The changing sizes of genes. Nature 384: 315-316 (1996)
- [4] Doolittle, W.F. & Sapienza, C. Selfish genes, the phenotype paradigm and genome evolution. Nature 284: 601-603 (1980)
- [5] Eigen, M. 1971. Self-organization of matter and the evolution of biological macromolecules. Naturwissenschaften 58:465-523
- [6] Fox Keller, E. The Century of the Gene. Harvard University Press. Cambridge Mass. (2000)
- [7] Krakauer, D.C, Jansen, V. Red Queen Dynamics and the Evolution of protein translation. (J. theor, Biol. 2002)
- [8] Krakauer, D.C., Plotkin, J. Redundancy, antiredundancy and the robustness of genomes (PNAS 99, 1405-1409 2002)
- [9] Krakauer, D.C. Stability and evolution of overlapping genes. Evolution 54: 731-739 (2000)
- [10] Krakauer, D.C., Nowak, M.A. Evolutionary preservation of redundant duplicated genes. Seminars in Cell and Developmental Biology 10: 555-559 (1999)
- Lewontin, R.C. The units of selection. Ann. Rev. Ecol. Syst. 1:1-18 (1970)

- Madison-Antenucci, S. Steege, D. A. Translation limits synthesis of an assembly-initiating coat protein of filamentous phage λ . *Journal of Bacteriology*. 180: 464-472. (1988)
- [11] Miyata, T., and T. Yasunaga. 1978. Evolution of overlapping genes. *Nature* 272:532-535
- [12] Normark, S., S. Bergstrom, T. Edlund, T. Grundstrom, B. Jaurin, et al. 1983. Overlapping genes. *Ann. Rev. Genet.* 17:499-525
- [13] Ohno, S. So much "junk" DNA in our genome. In *Evolution of genetic systems* (e. H.H. Smith) pp 366-370. Gordon and Breach. New York.
- [14] Oppenheim, D. S., and C. Yanofsky. 1980. Translational coupling during expression of the tryptophan operon of *E.coli*. *Genetics* 95:785-95
- [15] Orgel, L.E., Crick, F.H.C. Selfish DNA: The ultimate parasite. *Nature* 284:604-607 (1980)
- [16] Pagel, M.D, Johnstone, R.A. Variation across species in the size of the nuclear genome supports the junk DNA explanation for the C-value paradox. *Proc. Roy. Soc. Lond. B. Biol. Sci.* 249:119-124. (1992)
- [17] Prescott, D.M. (1997) Origin, evolution, and excision of internal eliminated segments in germline genes of ciliates. *Curr. Opin. Genet. & Dev.* 7:807-813.
- [18] Schumperli, D., K. McKenney, D. A. Sobieski and M. Rosenberg. 1982. Translational coupling at an intercistronic boundary of the *Escherichia coli* galactose operon. *Cell* 30:865-871
- [19] Trifinov, E. N. 1989. The multiple codes of nucleotide sequences. *Bulletin of Mathematical Biology* 51:417-432