Origins of Scaling in Genetic Code

Oliver Obst[1,⋆], Daniel Polani[2], and Mikhail Prokopenko[1]

[1] CSIRO Information and Communications Technology Centre, Locked Bag 17, North Ryde, NSW 1670, Australia
[2] Department of Computer Science, University of Hertfordshire, Hatfield AL10 9AB, United Kingdom
corresponding author: mikhail.prokopenko@csiro.au

Abstract. The principle of least effort in communications has been shown, by Ferrer i Cancho and Solé, to explain emergence of power laws (e.g., Zipf's law) in human languages. This paper applies the principle and the information-theoretic model of Ferrer i Cancho and Solé to genetic coding. The application of the principle is achieved via equating the ambiguity of signals used by "speakers" with codon usage, on the one hand, and the effort of "hearers" with needs of amino acid translation mechanics, on the other hand. The re-interpreted model captures the case of the typical (vertical) gene transfer, and confirms that Zipf's law can be found in the transition between referentially useless systems (i.e., ambiguous genetic coding) and indexical reference systems (i.e., zero-redundancy genetic coding). As with linguistic symbols, arranging genetic codes according to Zipf's law is observed to be the optimal solution for maximising the referential power under the effort constraints. Thus, the model identifies the origins of scaling in genetic coding — via a trade-off between codon usage and needs of amino acid translation. Furthermore, the paper extends Ferrer i Cancho – Solé model to multiple inputs, reaching out toward the case of horizontal gene transfer (HGT) where multiple contributors may share the same genetic coding. Importantly, the extended model also leads to a sharp transition between referentially useless systems (ambiguous HGT) and indexical reference systems (zero-redundancy HGT). Zipf's law is also observed to be the optimal solution in the HGT case.

1 Introduction — coding thresholds

The definition and understanding of the genotype-phenotype relationship continues to be one of the most fundamental problems in biology and artificial life. For example, Woese strongly argues against fundamentalist reductionism and presents the real problem of the gene as "how the genotype-phenotype relationship had come to be" [1], pointing out the likelihood of the "coding threshold". This threshold signifies development of the capacity to represent nucleic acid sequence symbolically in terms of a amino acid sequence, and separates the phase of nucleic acid life from an earlier evolutionary stage. Interestingly, the analysis presented by Woese sheds light not only on this transition, but also on saltations that have occurred at other times, e.g. advents of multicellularity and language. The common feature is "the emergence of higher levels of organization, which bring with them qualitatively new properties,

⋆ Author's list is in alphabetical order.

properties that are describable in reductionist terms but that are neither predictable nor fully explainable therein" [1].

The reason for the increase in complexity can be identified as *communication* within a complex, sophisticated network of interactions: "translationally produced proteins, multicellular organisms, and social structures are each the result of, emerge from, fields of interaction when the latter attain a certain degree of complexity and specificity" [1, 2]. The increase of complexity is also linked to adding new dimensions to the phase space within which the evolution occurs, i.e. expansion of the network of interacting elements that forms the medium within which the new level of organization (entities) comes into existence [1, 2]. An increase of complexity is one of the landmarks of self-organization, typically defined as an increase of order within an open system, without an explicit external control.

As pointed out by Ferrer i Cancho and Solé [3], the emergence of a complex language is one of the fundamental events of human evolution, and several remarkable features suggest the presence of fundamental principles of organization, common to all languages. The best known is Zipf's law, which states that the frequency of a word decays as a (universal) power law of its rank. Furthermore, Ferrer i Cancho and Solé observe that "all known human languages exhibit two fully developed distinguishing traits with regard to animal communication systems: syntax [4] and symbolic reference [5]", and suggest that Zipf's law is a hallmark of symbolic reference [3].

They adopt the view that a communication system is shaped by both constraints of the system and demands of a task: e.g., the system may be constrained by the limitations of a sender ("speaker") trying to encode a message that is easy-to-decode by the receiver ("hearer"). In particular, speakers want to minimise articulatory effort and hence encourage brevity and phonological reduction, while hearers want to minimise the effort of understanding and hence desire explicitness and clarity [3, 6, 7]. For example, the speaker tends to choose ambiguous words (words with a high number of meanings), and this increases the interpretation effort for the hearer. Zipf referred to the lexical trade-off as the *principle of least effort*, leading to a well-known power law: if the words within a sample text are ordered by decreasing frequency, then the frequency of the k-th word, $P(k)$, is given by $P(k) \propto k^{-\alpha}$, with $\alpha \approx 1$.

The main findings of Ferrer i Cancho and Solé are that (i) Zipf's law can be found in the transition between referentially useless systems and indexical reference systems, and (ii) arranging symbols according to Zipf's law is the optimal solution for maximising the referential power under the effort constraints.

Combining terminology of Woese and Ferrer i Cancho and Solé allows us to rephrase these observations as follows: (i) referentially useless systems are separated from indexical reference systems by a coding threshold, and (ii) Zipf's law maximising the referential power under the effort constraints is the optimal solution that is a feature observed at the coding threshold.

In this paper we apply the principle of least effort to genetic coding, by equating, on the one hand, the ambiguity of signals used by "speakers" with codon usage, and, on the other hand, the effort of "hearers" with demands of amino acid translation mechanics. The re-interpreted model confirms that Zipf's law can be found in the transition ("coding threshold") between ambiguous genetic coding (i.e., referentially useless systems) and zero-redundancy genetic coding (indexical reference systems). As with linguistic symbols, arranging genetic codes according to Zipf's law is observed

to be the optimal solution for maximising the referential power under the effort constraints. In other words, the model identifies the origins of scaling in genetic coding — via a trade-off between codon usage and needs of amino acid translation.

This application captures the case of the typical, vertical, gene transfer. We further extend this case to multiple inputs, reaching out toward the case of horizontal gene transfer (HGT) where multiple contributors may share the same genetic coding. We observe that the extended model also leads to a sharp transition between ambiguous HGT and zero-redundancy HGT, and that Zipf's law is observed to be the optimal solution again.

2 Horizontal and Stigmergic Gene Transfer

It is important to realize that during the early phase in cellular evolution the proto-cells can be thought of as conglomerates of substrates, that exchange components with their neighbours freely — horizontally [8]. The notion of vertical descent from one "generation" to the next is not yet well-defined. This means that the descent with variation from one "generation" to the next is not genealogically traceable but is a descent of a cellular community as a whole. Thus, genetic code that appears at the coding threshold is "not only a protocol for encoding amino acid sequences in the genome but also an innovation-sharing protocol" [8], as it used not only as a part of the mechanism for cell replication, but also as a way to encode relevant information about the environment. Different proto-cells may come up with different innovations that make them more fit to the environment, and the "horizontal" exchange of such information may be assisted by an innovation-sharing protocol — a proto-code. With time, the proto-code develops into a universal genetic code.

Such innovation-sharing is perceived to have a price: it implies ambiguous translation where the assignment of codons to amino acids is not unique but spread over related codons and amino acids. [8]. In other words, accepting innovations from neighbours requires that the receiving proto-cell is sufficiently flexible in translating the incoming fragments of the proto-code. Such a flexible translation mechanism, of course, would produce imprecise copies. However, a descent of the whole innovation-sharing community may be traceable: i.e., in a statistical sense, the next "generation" should be correlated with the previous one. While any individual protein is only a highly imprecise translation of the underlying gene, a consensus sequence for the various imprecise translations of that gene would closely approximate an exact translation of it. As noted by Polani et al. [9], the consensus sequence would capture the main information content of the innovation-sharing community.

Moreover, it can be argued that the universality of the code is a generic consequence of early communal evolution mediated by horizontal gene transfer (HGT), and that thus HGT enhances optimality of the code [8]:

> HGT of protein coding regions and HGT of translational components ensures the emergence of clusters of similar codes and compatible translational machineries. Different clusters compete for niches, and because of the benefits of the communal evolution, the only stable solution of the cluster dynamics is universality.

The work of Piraveenan et al. [10] and Polani et al. [9] investigated particular HGT scenarios where certain fragments necessary for cellular evolution begin to play the

role of the proto-code. For example, *stigmeric gene transfer* was considered as an HGT variant. SGT suggests that the proto-code is present in an environmental locality, and is subsequently entrapped by the proto-cells that benefit from such interactions. In other words, there is an indirect exchange of information among the cells via their local environment, which is indicative of stigmergy: proto-cells find matching fragments, use them for coding, modify and evolve their translation machinery, and exchange certain fragments with each other via the local environment. SGT differs from HGT in that the fragments exchanged between two proto-cells may be modified during the transfer process by other cells in the locality.

SGT studies concentrated on the information preservation property of evolution in the vicinity of the "coding threshold", considering a communication channel between a proto-cell and itself at a future time point, and posing a question of the channel capacity constrained by environmental noise. By varying the nature and degree of the noise prevalent in the environment within which such proto-cells exist and evolve, the conditions for self-organization of an efficient coupling between the proto-cell *per se* and its encoding with "proto-symbols" were identified. It was shown that the coupling evolves to preserve (within the entrapped encoding) the information about the proto-cell dynamics. This information is preserved across time within the noisy communication channel. The studies verified that the ability to symbolically encode nucleic acid sequences does not develop when environmental noise φ is outside a specific noise range (an error interval).

In current work we depart from the models developed by Piraveenan et al. [10] and Polani et al. [9], and rather than considering proto-cell dynamics defined via specific dynamical systems (e.g., logistic maps) subjected to environmental noise, we focus on constraints determined by (i) ambiguous codon usage, and (ii) the demands of amino acid translation mechanics. This allows us to abstract away the specifics of the employed dynamical systems [10, 9], and explain the emergence of a coding threshold from another standpoint that takes into account codon usage and amino acid translation. This, in turn, allows us to extend the model to HGT/SGT scenaria with multiple inputs. Both types of models — dynamical systems based [10, 9] and the one presented here — are able to identify an (order) parameter corresponding to the coding threshold: the environmental noise φ [10, 9] or the effort contribution λ [3]. Both types of models formulate objective functions information-theoretically, following the guidelines of information-driven self-organisation [11–13, 10, 14–17]. The main difference from the dynamical systems based models lies, however, in the ability to detect a power law in the codon usage (lexicon) at the threshold.

3 Model

The Ferrer i Cancho and Solé model [3] is based on an information-theoretic approach, where a (single) set of signals S and a set of objects R are used to describe signals between a speaker and a hearer, and the objects of reference the signals are referring to. The relation between S and R are modelled using a binary matrix. As mentioned above, the effort for the speaker is low with a high amount of ambiguity, i.e. if the signal entropy is low. $H_n(S)$ expresses the effort of the speaker as a number between 0 and 1:

$$H_n(S) = -\sum_{i=1}^{n} p(s_i) \log_n p(s_i)$$

The effort for the hearer to decode a particular signal s_i is small if there is little ambiguity, i.e. the probability of a signal s_i referring to one object r_j is high. In [3], this is expressed by the conditional entropy

$$H_m(R|s_i) = -\sum_{j=1}^{m} p(r_j|s_i) \log_m p(r_j|s_i)$$

The effort for the hearer is then dependent on the probability of each signal and the effort to decode it, that is

$$H_m(R|S) = \sum_{i=1}^{n} p(s_i) H_m(R|s_i)$$

A cost function $\Omega(\lambda)$ is introduced to combine effort of speaker and hearer, with $0 \leq \lambda \leq 1$ trading off the effort between speaker and hearer as follows:

$$\Omega(\lambda) = \lambda H_m(R|S) + (1 - \lambda) H_n(S)$$

To consider the effect of different combinations of speaker and hearer effort, different λ from 0 to 1 are used to compute the accuracy of the communication as the mutual information,

$$I_n(S, R) = H_n(S) - H_n(S|R),$$

using matrices evolved, with a simple mutation-based genetic algorithm, for a minimal cost $\Omega(\lambda)$.

3.1 Extension of the model and "readout" modeling

To model codon usage by several neighbours, we extend the original approach that was using one matrix S, to several matrices, which represent different sets of signals S_i for one set of objects R. In the extended model, a separate matrix $\mathbf{A}_i$ is created and evolved to encode different communication channels. The cost $\Omega(\lambda)$ is computed for a matrix $\hat{\mathbf{A}}$ that is obtained by averaging over all individual matrices $\mathbf{A}_i$. During evolution, each $\mathbf{A}_i$ is mutated if the cost of $\hat{\mathbf{A}}$ is higher than the cost of $\hat{\mathbf{A}}$ of the previous generation. Averaging captures SGT between different sources (variables), and is a specific case of "readout", motivated below.

Shannon information between two variables X and X' is determined as an optimum of knowledge extracted from the state of X about the state of X' *under the assumption that both variables and their joint distribution have been identified beforehand*. In our model, however, the transfer of a message fragment from one proto-cell to another does not, in general, enjoy that advantage, because there are multiple candidates for the source of the message. The stigmergic nature of the message transmission in the HGT scenario does not allow for a priori assumptions of who the sender is and who the receiver is. This implies that there might emerge a pressure to "homogenize" the instantiations of sender and the receiver variables in the sense that, where information is to be shared, an instantiation x in the sender variable X evokes to some extent the same instantiation in the receiver variable X'.

To formalize this intuition, we model the sending (and analogously the receiving) proto-cells as mixtures of probabilities, endowed with the "readout" interpretation suggested in [13] which we sketch in the following. Assume a collection of random

variables X_k, indexed by an index $k \in \mathcal{K}$, where all X_k assume values $x_k \in \mathcal{X}$ in the same sampling set $\mathcal{X}$. For a fixed, given $k \in \mathcal{K}$, one can now determine informational quantities involving X_k in the usual fashion; however, if, as in the HGT case, the sender is not known in advance, one can model that uncertainty as a probability distribution over the possible indices $k \in \mathcal{K}$, defining a random variable K with values in $\mathcal{K}$. If nothing else is known about the sender, one can for instance assume an equidistribution on K.

The *readout* of the collection $(X_k)_{k\in\mathcal{K}}$ is then denoted by X_K which models the random selection of one $k \in \mathcal{K}$ according to the distribution of K which selects one of the X_k, followed by a random selection of an instance $x \in \mathcal{X}$ according to the distribution of X_k. Formally, the probability distribution of the readout X_K assuming a value $x \in \mathcal{X}$ is given by

$$\Pr(X_K = x) = \sum_{k\in\mathcal{K}} \Pr(K = k) \cdot \Pr(X_k = x) \tag{1}$$

For a Bayesian network interpretation of the readout, see [13].

4 Results

The average mutual information as a function of λ is shown in Fig. 1 and 2. Figure 1 traces the average mutual information for a single 150x150 matrix (staying within Ferrer i Cancho and Solé model [3], and studying a typical vertical gene transfer), while Figure 2 contrasts the dynamics with the case of HGT, simulated with four 40x40 matrices. In both cases, there are three domains distinguishable in the behavior of $I_n(S, R)$ versus λ.

For small values $\lambda < \lambda^*$, $I_n(S, R)$ grows smoothly, before undergoing a sharp transition in the vicinity $\lambda \approx \lambda^*$. Following Ferrer i Cancho and Solé, we observe that single-signal systems ($L \approx 1/n$) dominate for $\lambda < \lambda^*$: "because every object has at least one signal, one signal stands for all the objects" [3]. Low $I_n(S, R)$ indicates that the system is unable to convey information in this domain (totally ambiguous genetic code). Rich vocabularies (genetic codes with some redundancy), $L \approx 1$, are found after the transition, for $\lambda > \lambda^*$. Full vocabularies (zero-redundancy genetic codes) are attained for very high λ. The maximal value of $I_n(S, R)$ indicates that the associations between signals (codons) and objects (amino-acids) are one-to-one maps, removing any redundancy in the genetic code. In the HGT case, this case is harder to achieve, while the overall tendency is retained.

To investigate transition around $\lambda \approx \lambda^*$, we focus on the lexicon's ranked distribution, and consider the signal's normalised frequency $P(k)$ versus rank k, for different λ. As expected, Ferrer i Cancho and Solé model shows that "Zipf's law is the outcome of the nontrivial arrangement of word-concept associations adopted for complying with hearer and speaker needs" [3]: contrasting the graphs in Fig. 3 reveals the presence of scaling for $\lambda \approx \lambda^*$, and suggests that a phase transition is taking place at $\lambda^* = 0.41$ in the information dynamics of $I_n(S, R)$.

Similar phenomenon is observed for HGT, as shown by Fig. 4. The scaling at $\lambda^* = 0.4$ results in the power law (i.e., $P(k) = 0.2742/k^{1.0412}$, with $R^2 = 0.9474$, consistent with $\alpha \approx 1$ in the power law reported by Ferrer i Cancho and Solé [3]).

Thus, the trade-off between codon usage and needs of amino acid translation in HGT results in a nontrivial but still redundant genetic code.

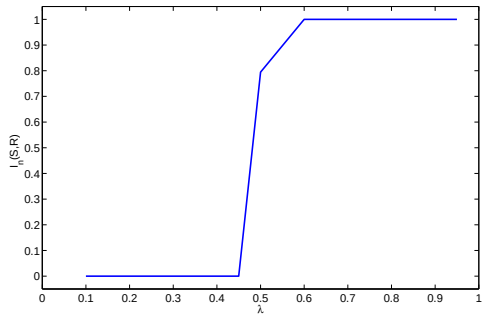

Fig. 1: Single 150x150 matrix. The average mutual information as a function of λ.

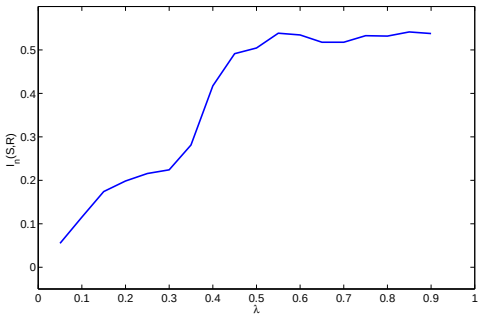

Fig. 2: HGT with four 40x40 matrices. The average mutual information as a function of λ.

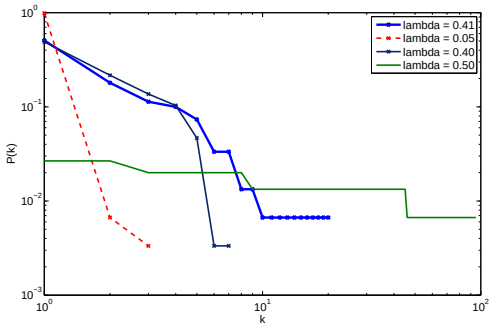

Fig. 3: Single 150x150 matrix. Signal normalised frequency, $P(k)$, versus rank, k, for different lambdas.

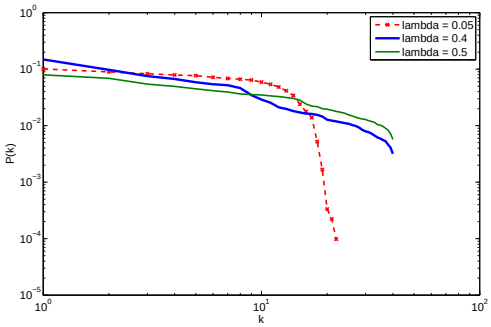

Fig. 4: HGT with four 40x40 matrices. Signal normalised frequency, $P(k)$, versus rank, k, for different lambdas.

5 Conclusions

We applied the principle principle of least effort in communications and the (extended) information-theoretic model of Ferrer i Cancho and Solé to genetic coding. The ambiguity of signals used by "speakers" was equated with codon usage, while the effort of "hearers" provided an analogue for the needs of amino acid translation mechanics. The re-interpreted model captures the case of the typical (vertical) gene transfer, and confirms presence of scaling in the transition between referentially useless systems (i.e., ambiguous genetic coding) and indexical reference systems (i.e., zero-redundancy genetic coding). Arranging genetic codes according to Zipf's law is observed to be the optimal solution for maximising the referential power under the effort constraints. Thus, the model identifies the origins of scaling in genetic coding — via a trade-off between codon usage and needs of amino acid translation. The extended model includes multiple inputs, representing horizontal gene transfer where multiple contributors may share the same genetic coding. The extended model also leads to a sharp transition: between ambiguous HGT and zero-redundancy HGT, and scaling is observed to be the optimal solution in the HGT case as well.

References

1. Woese, C.R.: A new biology for a new century. Microbiology and Molecular Biology Reviews **68**(2) (2004) 173–186
2. Barbieri, M.: The organic codes: an introduction to semantic biology. Cambridge University Press, Cambridge, United Kingdom (2003)
3. Ferrer i Cancho, R., Solé, R.V.: Least effort and the origins of scaling in human language. PNAS **100**(3) (2003) 788–791
4. Chomsky, N.: Language and Mind. Harcourt, Brace, and World, New York (1968)
5. Deacon, T.W.: The Symbolic Species: The Co-evolution of Language and the Brain. Norton & Company, New York (1997)
6. Pinker, S., Bloom, P.: Natural language and natural selection. Behavioral and Brain Sciences **13**(4) (1990) 707–784
7. Köhler, R.: System theoretical linguistics. Theoretical Linguistics **14**(2-3) (1987) 241–257
8. Vetsigian, K., Woese, C., Goldenfeld, N.: Collective evolution and the genetic code. PNAS **103**(28) (July 2006) 10696–10701
9. Polani, D., M. Prokopenko, M., Chadwick, M.: Modelling stigmergic gene transfer. In Bullock, S., Noble, J., Watson, R., Bedau, M.A., eds.: Artificial Life XI - Proceedings of the Eleventh International Conference on the Simulation and Synthesis of Living Systems, MIT Press (2008) 490–497
10. Piraveenan, M., Polani, D., Prokopenko, M.: Emergence of genetic coding: an information-theoretic model. In Almeida e Costa, F., Rocha, L., Costa, E., Harvey, I., Coutinho, A., eds.: Advances in Artificial Life: 9th European Conference on Artificial Life (ECAL-2007), Lisbon, Portugal, September 10-14. Volume 4648 of Lecture Notes in Artificial Intelligence. Springer (2007) 42–52
11. Prokopenko, M., Gerasimov, V., Tanev, I.: Evolving spatiotemporal coordination in a modular robotic system. In Nolfi, S., Baldassarre, G., Calabretta, R., Hallam, J., Marocco, D., Meyer, J.A., Parisi, D., eds.: From Animals to Animats 9: 9th International Conference on the Simulation of Adaptive Behavior (SAB 2006). Volume 4095 of Lecture Notes in Computer Science., Springer (2006) 558–569
12. Polani, D., Nehaniv, C., Martinetz, T., Kim, J.T.: Relevant information in optimized persistence vs. progeny strategies. In Rocha, L., Yaeger, L., Bedau, M., Floreano, D., Goldstone, R., Vespignani, A., eds.: Artificial Life X: Proceedings of The 10th International Conference on the Simulation and Synthesis of Living Systems, Bloomington IN, USA (2006)
13. Klyubin, A., Polani, D., Nehaniv, C.: Representations of space and time in the maximization of information flow in the perception-action loop. Neural Computation **19**(9) (2007) 2387–2432
14. Laughlin, S.B., Anderson, J.C., Carroll, D.C., de Ruyter van Steveninck, R.R.: Coding efficiency and the metabolic cost of sensory and neural information. In Baddeley, R., Hancock, P., Földiák, P., eds.: Information Theory and the Brain. Cambridge University Press (2000) 41–61
15. Bialek, W., de Ruyter van Steveninck, R.R., Tishby, N.: Efficient representation as a design principle for neural coding and computation. In: 2006 IEEE International Symposium on Information Theory, IEEE (2006) 659–663
16. Piraveenan, M., Prokopenko, M., Zomaya, A.Y.: Assortativeness and information in scale-free networks. European Physical Journal B **67** (2009) 291–300
17. Lizier, J.T., Prokopenko, M., Zomaya, A.Y.: Local information transfer as a spatiotemporal filter for complex systems. Physical Review E **77**(2) (2008) 026110