Journal of Theoretical Biology 399 (2016) 22–32

Contents lists available at ScienceDirect

Journal of Theoretical Biology

journal homepage: www.elsevier.com/locate/yjtbi

Simulating evolution of protein complexes through gene duplication and co-option

Loren Haarsma [a,*], Serita Nelesen [b], Ethan VanAndel [b,1], James Lamine [b], Peter VandeHaar [b]

[a] Departments of Physics, Calvin College, Grand Rapids, MI 49546, United States
[b] Departments of Computer Science, Calvin College, Grand Rapids, MI 49546, United States

HIGHLIGHTS

- Digital organisms model evolution of protein complexes with novel functions.
- Greater complexity and interlocking complexity evolve over time.
- Gene duplication, mutation, and co-option are principle mechanisms.
- Many model parameters (e.g. mutation rates, selection, binding) can be varied.

ARTICLE INFO

Article history:
Received 4 September 2015
Received in revised form
29 February 2016
Accepted 17 March 2016
Available online 31 March 2016

Keywords:
Evolution
Complexity
Co-option
Computer model
Digital organism

ABSTRACT

We present a model of the evolution of protein complexes with novel functions through gene duplication, mutation, and co-option. Under a wide variety of input parameters, digital organisms evolve complexes of 2–5 bound proteins which have novel functions but whose component proteins are not independently functional. Evolution of complexes with novel functions happens more quickly as gene duplication rates increase, point mutation rates increase, protein complex functional probability increases, protein complex functional strength increases, and protein family size decreases. Evolution of complexity is inhibited when the metabolic costs of making proteins exceeds the fitness gain of having functional proteins, or when point mutation rates get so large the functional proteins undergo deleterious mutations faster than new functional complexes can evolve.

1. Introduction

Understanding the evolution of complex biological organisms from simpler ancestors is a longstanding challenge. Multi-cellular organisms have complex anatomical features, such as ears, which require multiple parts to be working in order to improve fitness; these features are under the control of many different genes. As Darwin (1859) noted, a step-wise evolutionary history of such complex features can be difficult to deduce, although in some cases like the mammalian middle ear, fossils provide a record of the co-option of existing anatomical features to perform new functions (Crompton and Jenkins, 1979; Allin and Hopson, 1992).

Even single-celled organisms contain biochemical machinery in which many different proteins must bind into complexes in order for their adaptive function to work. It has long been understood that the combination of gene duplication, mutation, and co-option could be an important mechanism for the evolution of protein complexes or complex phenotypic features (Ohno, 1970; Chen et al., 1997; Lynch and Conery, 2000; Wagner, 2001; Chothia et al., 2003; Monteiro and Podlaha, 2009; Shubin et al., 2009; Prochnik et al., 2010). Considerable progress has been made in understanding the evolution of some systems such as eyes (Salvini-Plawen and Mayr, 1977; Goldsmith, 1990; Piatigorsky and Wistow, 1991), the Krebs cycle (Meléndez-Hevia et al., 1996), transcriptional regulatory networks (Gelfand, 2006; Rebeiz et al. 2011), and developmental pathways (Wilkins, 2002; Pires-daSilva and Sommer, 2002).

Our work uses computer modeling to investigate systematically factors that impact the co-option mechanism. Co-option of a protein so that it begins to perform a new, positively selected function could happen either before or after a gene duplication event. Co-option could arise through changes in the gene which encodes the protein, or through changes elsewhere in the genome creating new interactions with an existing protein, or through changes in the environment creating a new function for an existing protein complex. Such events during biological evolution are relatively rare and

* Corresponding author. Tel.: +1 616 526 8517.
E-mail address: lhaarsma@calvin.edu (L. Haarsma).
[1] Present address: Department of Mathematics, University of California Berkeley, Berkeley, CA 94720, United States.

http://dx.doi.org/10.1016/j.jtbi.2016.03.028
0022-5193/

serendipitous, making them difficult to study systematically. Computational modeling can provide insights into these processes and enable a systematic study of parameters which promote or inhibit co-option.

One approach to studying the evolution of novel interactions is network modeling. This has the advantage of allowing rapid calculation of how adding, deleting, or changing interactions between elements changes the behavior of the entire system. At the level of species and ecosystems, network modeling of predator–prey food webs shows increases in ecosystem complexity and speciation events over time under a variety of conditions, including both stable and changing environments (Drossel and McKane, 2000; McKane, 2004; Uchida and Drossel, 2007; Guttenberg and Goldenfeld, 2008; Guill and Drossel, 2008; Braunewell and Bornholdt, 2009). The Evolutionary Constructor modeling tool follows both ecological complexity through population dynamics of interacting species and genomic variation and genomic complexity within species (Lashin et al., 2007, 2012). Protein–protein interactions networks can also be modeled with network modeling (Sun et al., 2014). Solé et al. (2002) demonstrated network modeling of proteome evolution, simulating gene duplication and mutation, resulting in protein–protein interaction maps which reproduce many of the statistical features observed in analysis of real biological proteomes.

A second approach is to model the evolution of populations of digital organisms over fitness landscapes, in which the fitness of the organism is calculated directly from its genome. While digital organisms are greatly simplified and abstracted from biological organisms, studying evolution of digital organisms has several advantages. When digital organisms self-replicate, mutate, and adapt through natural selection, their entire evolutionary history can be recorded. Models by Ohta (1987, 1988a, 1988b) and Walsh (1995) provide insights into the rates at which duplicated genes can spread through a population. Models of this sort which make use of the NK fitness landscapes have been used to study the effects of mutation rate and selection pressure on directed evolution of proteins (Wedge et al., 2009). These models however, do not address the question raised by other researchers (Behe and Snoke, 2004) of how many mutations are typically necessary in order for a protein to acquire a change of function. Recent modeling work on the evolution of protein–protein interactions (Peleg et al., 2014), which calculates protein stability and protein–protein interaction probabilities based on thermodynamic properties of amino acid sequences, begins to address this question.

A third approach is to model evolving digital organisms which interact with an environment. Like real biological organisms, the "fitness" of these digital organisms is not directly calculated from their genomes, but by how well they survive and reproduce in competition with each other in the environment. In the Avida program, digital organisms start with genomes which execute simple logical operations, and evolve novel combinations of operations to execute more complex mathematical functions (Lenski et al., 2003). In the EcoSim program, digital organisms move around an environment, search for food, and avoid predators; through mutations and selections, new species evolve and the ecosystem becomes more complex (Gras et al., 2009; Khater and Gras, 2012; Mashayekhi and Gras, 2014.). Another such digital organism which can show evolution of complexity is Tierra (Ray et al., 1991; Thearling and Ray, 1994; Shao and Ray, 2010). Hintze and Adami (2008) have developed a model of artificial chemistry in which digital organisms gain fitness every time they evolve a novel metabolic reaction. These organisms evolve increases in complexity and information, and their gene–gene interaction networks have several intriguing similarities to real biological protein–protein interaction networks.

Since our goal was to study conditions which promote or inhibit the co-option of proteins into new functions resulting in more complex organisms, we developed a new computation model of this third type which incorporates some of the flexibility of the first two approaches. *Pykaryotes* are digital organisms whose genomes determine how they gather food, move, and make proteins. Proteins and protein complexes can speed up the organism's ability to gather food, and fitness is calculated indirectly from the amount of food gathered. A wide variety of factors are under the experimenter's control, including selection pressure, food distribution, point mutation rates, gene duplication rates, the metabolic costs of large genomes, protein functional probability, protein–protein binding probability, and the number of mutations required for a protein to acquire a new function. All of these play a role in determining the rate at which complexity evolves. Under some conditions, novel protein complexes do not evolve; but over considerable ranges of these variables, organisms gradually evolve novel protein complexes, even to the point where five or more proteins are required in order to perform the food-gathering function, and the removal of any one protein causes the entire complex to stop functioning.

2. Description of the *Pykaryotes* model

The organisms in our model live and can move in an environment with a distribution of several chemicals which serve as food. At any time during the simulation, the state of the organism is determined by a set of characteristics as shown in Table 1.

Each generation, the actions of an organism are dictated by reading its own genome and updating its state accordingly. The genomes are circular, so when the last codon is read it loops back to the beginning. There is one codon type corresponding to each chemical type in the environment, and there are also three special codon types which change the organism's mode of action among gathering, moving, and protein-building.

An overview of the Pykaryotes simulation is show in Fig. 1. After a certain number of codon reads (default = 10,000), each of the organisms is given a fitness score F based on how many units it has stored of each chemical, G_i. For this paper, F is proportional to the sum of the mth root of the number of units of each chemical gathered ($F=\Sigma G_i^{1/m}$). When $m>1$, organisms have greater fitness gathering equal amounts of each type of chemical compared to gathering all of just one chemical. There is also a fitness cost χ proportional to the organism's genome length, to represent the metabolic cost of maintaining a large genome. Each organism then "dies" after reproducing itself into the next generation 0, 1, 2 or more times. The reproduction probability is proportional to $F\varepsilon$, where the selection exponent $\varepsilon>1$ indicates strong selection and $\varepsilon=0$ indicates no selection.

During reproduction, each codon has a probability μ_p of undergoing a point mutation. Each organism also has small probabilities μ_2, μ_c, μ_d, and μ_h, of experiencing genome doubling, copying a portion of its genome, deleting a portion of its genome, and experiencing horizontal transfer from a portion of the genome of another organism. Most point mutations change a codon to a

Table 1
Pykaryote organism state. Each organism has its own values for each datum.

State of an organism
Genome: string of integers which represent codons
Read position: position of next codon to be read/processed
Mode: one of *gather*, *move*, or *build protein*
G_i: Stored amount of each of i chemicals
M_j: Stored amount of each of j protein complexes