Journal of Theoretical Biology 330 (2013) 1–8

Contents lists available at SciVerse ScienceDirect

Journal of Theoretical Biology

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

Sequence and structure space model of protein divergence driven by point mutations

Tomasz Arodź [a,b,*], Przemysław M. Płonka [c]

[a] Department of Computer Science, School of Engineering, Virginia Commonwealth University, 401 W. Main St., Richmond, Virginia 23284, USA
[b] Department of Computer Science, AGH University of Science and Technology, Mickiewicza 30, 30-059 Kraków, Poland
[c] Department of Biophysics, Faculty of Biochemistry, Biophysics and Biotechnology, Jagiellonian University, Gronostajowa 7, 30-387 Kraków, Poland

AUTHOR-HIGHLIGHTS

- A stochastic model of protein evolution in structure and sequence space is proposed.
- It uses real relationship between structure and sequence change.
- It reproduces cluster structure and sequence–structure divergence curve.
- It shows that point mutations alone can lead to the present protein diversity.

ARTICLE INFO

Article history:
Received 27 September 2012
Received in revised form
7 March 2013
Accepted 18 March 2013
Available online 28 March 2013

Keywords:
Protein structure
Protein evolution
Duplication–divergence models
Residue substitutions
Evolution models

ABSTRACT

New folds of protein structures emerge in evolution as a result of insertions, deletions or shuffling of fragments of underlying gene sequences, and from aggregated effects of point mutations. The result of these evolutionary processes is a rich and complex universe of protein sequences and structures, with characteristic features such as heavy-tailed distribution of fold occurrences, and a distinct shape of relationship between sequence identity and structure similarity. Better understanding of how the protein universe evolved to its present form can be achieved by creating models of protein structure evolution. Here we introduce a stochastic model of evolution that involves residue substitutions as the sole source of structure innovation, and is nonetheless able to reproduce the diversity of the protein domains repertoire, its cluster structure with heavy-tailed distribution of family sizes, and presence of the twilight zone populated with remote homologs.

1. Introduction

Understanding evolutionary processes of structure innovation acting on protein domains is instrumental in uncovering the roads that lead to the current diversity of the protein universe. Domains fold independently, and events acting on multi-domain proteins, such as domain insertions, deletions, repetitions or swapping, only join and redistribute folds emerging at the domain level (Orengo and Thornton, 2005). The repertoire of protein domain structures seen in nature has some distinct properties. In genomes of many organisms, the empirical distribution of the number of proteins of a given fold follows a distribution highly skewed towards a selected few folds, resembling a power-law distribution (Huynen and Van Nimwegen, 1998). For prokaryotic organisms, the concordance with power-law is high, while for eukaryotic ones it has been argued that a more localised distribution matches the data better (Abeln and Deane, 2005).

One of the challenges in theoretical biology is to create a model that would approximate the evolution of protein universe, and reproduce its main properties. Some coarse-grained models operate on the level of folds, and are able to reproduce the power-law distribution of frequencies of proteins in folds. These models involve gene duplications, and some form of emergence of new folds (Qian et al., 2001), or, more generally, fold birth, death and fold innovation (Karev et al., 2002, 2003).

A more complex class of models is based on concept from graph theory and, more generally, graphical approach. Graphical approaches have long history in theoretical biology. They are used to analyse interactions and relationships between components of living systems, and to model behaviour and evolution of biological systems. Graphical approach has been used in areas as diverse as enzyme-catalysed reactions (Andraos, 2008; Chou, 1980, 1989;

* Corresponding author at: Department of Computer Science, School of Engineering, Virginia Commonwealth University, 401 W. Main St., Richmond, Virginia 23284, USA. Tel.: +1 804 405 8714; fax: +1 804 828 2771.
E-mail addresses: tarodz@vcu.edu (T. Arodź), przemyslaw.plonka@uj.edu.pl (P.M. Płonka).

0022-5193/$ - see front matter © 2013 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.jtbi.2013.03.015

Chou and Forsén, 1980; Myers and Palmer, 1985; Zhou and Deng, 1984), protein folding kinetics and folding rates (Chou, 1990), enzyme inhibition (Althaus et al., 1993a,c,b; Chou et al., 1994), study of metabolism (Ravasz et al., 2002; Arodź, 2008, 2009; Chou, 2010), protein sequence evolution (Wu et al., 2010), protein–protein interactions (Jansen et al., 2003; Chou et al., 2011; Zhou, 2011), quantitative structure-activity relationships (Wiener, 1947; Randic, 1975; Czech, 2012; Devillers and Balaban, 2000), and vascular architecture (Topa, 2008; Czech et al., 2012). Graphical analysis using cellular automata (Wolfram, 1984; Ermentrout et al., 1993) has been applied to study hepatitis B viral infections and mutations (Xiao et al., 2006a, 2005a), biological sequence representation (Xiao et al., 2005b), discovery of protein attributes (Xiao and Chou, 2007, 2011; Xiao et al., 2006b, 2008, 2011a,b), and tumorigenesis (Kansal et al., 2000; Mallet and De Pillis, 2006).

In the study of protein evolution, one group of graphical approach involves simplified lattice representation of proteins—the structure of protein is modelled as a chain of points on a grid, and relationship between sequence and structure is dictated by a simple form of energy potential. Observations with lattice models confirmed the existence of neutral nets, that is, networks of foldable structures linked by point mutations (Lipman and Wilbur, 1991). Such nets are of different sizes, with a heavy-tailed size distribution, and each is centred on a prototype sequence (Bornberg-Bauer, 1997). Foldable sequences are scattered almost uniformly among all sequences (Bornberg-Bauer, 1997). Some lattice models were analysed in the context of evolutionary dynamics. When a realistic similarity measure was introduced on the lattice-encoded structures (Deeds et al., 2003) within a duplication–divergence model, the heavy-tailed degree distribution characteristic for natural proteins was also recovered. These characteristics remained when a realistic model of divergence was employed (Deeds and Shakhnovich, 2005), in which a sequence is underlying each structure, and divergence is governed by mutations in the sequence, which transfer to structure through a contact-based energy potential. However, the model required high sequence divergence between duplications to recover diversity equivalent to the natural one, leading to conclusion that more rapid mechanisms of sequence divergence, such as recombination, insertion or deletions are necessary to fully reproduce structural diversity. A lattice model that incorporates foldability as a subject of selective pressure operating on the whole phenotype, consisting of several genes, reproduces some main properties of early evolution, the heavy-tailed distribution of protein families, as well as stability and changes in dominant protein structures through evolution (Zeldovich et al., 2007). Also, in the model, stable genomes containing several protein genes emerge once a narrow set of stable protein structures is discovered.

Another group of evolutionary models are inspired by the Protein Domain Universe Graph (Dokholyan et al., 2002), either based purely on structure divergence (Dokholyan et al., 2002), or incorporating an approximate sequence space (Roland and Shakhnovich, 2007). In particular, a model of expanding protein universe (Dokholyan et al., 2002) introduces protein duplication, with a single protein duplicated in each coarse-grained time step. The copy is altered structurally, with uniform distribution of structural distance to the original protein. Also, in each time step, all proteins diverge by a small constant. The model reproduces the power-law distributions of protein cluster sizes, and of number of neighbours of protein domains, but the study was not concerned with the speed of divergence, or the relation of sequence to structure divergence. The Protein Domain Universe Graph was also shown to exhibit power-law distribution of folds individually for proteins from a selected organism (Deeds et al., 2004).

The models above provided many insights into the ability of simple processes to drive evolution of protein structures. However, these models were operating with a simplified view of structural changes—either by focusing on lattice representation, or, in graph models, by relying on an arbitrary structure alteration step. Here, we formulate a different approach to modelling the evolution of protein folds in sequence and structure space. The model relies on availability of information about detailed distribution of structure variation in response to sequence change. We have recently reported the distribution of magnitudes of structure changes brought to protein structures when their sequences undergo a single amino-acid substitution (Arodź and Płonka, 2012). Using this observation, we can link changes in the sequence space to changes in the structure space, and formulate a stochastic model of protein structure evolution.

In the proposed model, each protein has a well-defined structural distance to all other proteins, but its tertiary structure is not explicitly defined. In such a space the distribution of structure change magnitudes can be specified to closely follow the real-world one. As in duplication–divergence models (Dokholyan et al., 2002; Deeds and Shakhnovich, 2005), which relate evolving genome to a birth-and-death process, the simulation starts with a single protein, and expands through duplications and, slightly less frequent, gene deletions, with a form of sequence and structure changes along the way. In our case these are residue substitutions, but the model can be adapted to other sequence modifications as soon as the distributions of their structure effects are measured.

2. Model of protein divergence

The model starts from a single protein with an arbitrary position in both sequence and structure space. A protein is subject to residue substitutions that change both its sequence and its position in the structure space. Also, a protein can be duplicated, or die as a consequence of deletion or silencing of an underlying gene. The interplay of the two forces leads to the expansion of the protein set from a single starting point to a number reaching the size of genomes, while substitutions lead to sequence and structure divergence.

2.1. Probabilities of duplications, deletions and substitutions

The simulated repertoire of proteins evolves in time. The simulation proceeds in 4000 steps, each representing approximately 1 million years. In each time step, duplications, substitutions and deaths of genes encoding amino acid sequences of the simulated proteins are possible. The probabilities of the respective events were chosen to conform to available empirical evolutionary rates. Specifically, each protein can be duplicated with a probability p^t_{dupl} depending on the total number N_t of proteins at time step t, and on the maximal possible number of proteins that can be encoded in the genome, K. The dependency follows a logistic model of growth in an environment with finite carrying capacity, $p^t_{dupl} = p^{max}_{dupl}(1-N_t/K)$. The increase in the number of proteins governed by the logistic model is mitigated by protein deaths, which can result as an effect of deletion or silencing of the underlying gene.

At early stages of the evolution, the number of proteins should increase sharply, but the rise in the number of proteins encoded in the genome should be slower as the genome size grows, as duplications approach equilibrium with protein deaths. We set the equilibrium protein duplication rate at around 0.02 per million years, which fits at the top of the range of empirical observations (Lynch and Conery, 2000). For the equilibrium to occur, the rate of protein deaths should take a similar value. Here, we use a fixed probability p_{del} of protein death equal to 0.02 per million years.