Available online at www.sciencedirect.com
SCIENCE DIRECT®
Physica A 365 (2006) 446–462
PHYSICA A
www.elsevier.com/locate/physa

Content-based network model with duplication and divergence

Yasemin Şengün, Ayşe Erzan*

Department of Physics, Faculty of Sciences and Letters, Istanbul Technical University, Maslak 34469, Istanbul, Turkey

Received 16 October 2005; received in revised form 26 February 2006
Available online 29 March 2006

Abstract

We construct a minimal content-based realization of the duplication and divergence model of genomic networks introduced by Wagner [Proc. Natl. Acad. Sci. 91 (1994) 4387] and investigate the scaling properties of the directed degree distribution and clustering coefficient. We find that the content-based network exhibits crossover between two scaling regimes, with log-periodic oscillations for large degrees. These features are not present in the original gene duplication model, but inherent in the content-based model of Balcan and Erzan. The scaling form of the degree distribution of the content-based model turns out to be robust under duplication and divergence, with some re-adjustment of the scaling exponents, while the out-clustering coefficient goes over from a weak power-law dependence on the degree, to an exponential decay under mutations which include splitting and merging of strings.

Keywords: Networks; Scaling behavior; Genomics

1. Introduction

Biological networks have recently been the subject of many theoretical studies [1–4]. The gene-duplication model introduced by Wagner [5–8] for protein interaction provides a biologically motivated stochastic mechanism for the emergence of a small-world, scale free network [9], with the topological properties depending on just one parameter. This parameter may be adjusted to obtain values of the scaling exponent for the degree distribution, which may be compared with those reported in the literature for the genomic networks. There is a wealth of evidence pointing to the fact that a major portion of the genomic network in many organisms in fact originate from the duplication of transcription factors (TF) and target genes [10–13].

Wagner's mathematical model [5,7] involves nodes, corresponding to genes, and edges, corresponding to genomic interactions, connecting the nodes. Duplication of the nodes results in the inheritance of all of their previous connections. It is postulated that there exists a finite mutation rate which causes the edges to be removed, with some fixed probability. In this minimal model, once a number of interactions are in place, evolution by gene duplication and mutations generically drives the network to a scale free topology. Natural selection can then act on this spontaneously arising scale free network.

*Corresponding author. Tel.: +90 212 2853277; fax: +90 212 2856386.
E-mail address: erzan@itu.edu.tr (A. Erzan).

0378-4371/$ - see front matter © 2006 Elsevier B.V. All rights reserved.
doi:10.1016/j.physa.2006.02.045

The purpose of the present study is to associate the point-like nodes of the Wagner model with binary strings of different lengths, and investigate the behavior of the network under fixed rates of duplication and mutation of these strings. For this, one must (i) postulate a rule for the connectivity of the network, (ii) specify the particular ways in which duplications and mutations are to be affected.

The connectivity rule which we will adopt here was originally proposed by Balcan and Erzan [14], and defined a content-based null model (henceforth, the BE model). In this model, the connectivity is decided by the similarity of the strings associated with pairs of nodes, and gives rise to a spontaneous self-organization of linear codes of sufficient length into a network with a definite scaling behavior [14,15].

The content-based connectivity rule based on string-matching was motivated by the phenomenon of RNA interference (RNAi) [16], where small RNA segments bind messenger RNA (mRNA) via one-to-one Watson–Crick base pairing, and inhibit their translation into proteins. Originally thought to be a just primitive immune mechanism (to prevent the translation of virial RNA), RNAi is being found more and more to be an intrinsic part of so-called "post-transcriptional gene-regulation" [17]. Other biological processes such as conventional gene regulation via transcription factors (proteins coded by some genes) which bind regulatory sequences associated with other genes [10,18,19], or protein–protein interactions [2,8,20–22] which frequently involve sequence homologies [23], also rely on recognition mechanisms which can be thought of as a certain amount of shared information, represented by a common code, between the nodes of a network.

In the present paper we will combine the inputs from the two models, the Wagner model and the BE model, to investigate how the content-based interaction model evolves under duplications and mutations. We will concentrate on the out-degree distribution to characterize the topology of the network. For comparison, we will give an ab initio derivation of the clustering coefficient for the BE model, and investigate how its scaling behavior [24,25] changes under duplications and mutations.

It should be mentioned that there have recently been several studies of models of gene regulatory networks on "artificial genomes"(AG) introduced by Reil [26], based on various alphabets and matching rules. The Wagner model has been simulated with strings mimicking genes, promoter sequences and transcription factors, which evolve under duplication and divergence due to mutations [27–30]. These models employ different degrees of "realism" in their schematization of the interactions involved. In this paper we will stick to the simplest possible rules of perfect matching between binary strings associated with each node, for the establishment of the interaction matrix. Moreover, we will not make any distinctions between regulatory sequences, genes, or TFs, keeping the model to a minimum, so that it may eventually be amenable to analytical treatment.

In Section 2, we outline the original BE model for completeness, and give a summary of its properties, in particular, the degree distribution. In Section 3 we define an out-clustering coefficient, and present analytical results for its scaling behavior as a function of the out-degree, so that we may subsequently compare the effect of duplication and mutations on the clustering coefficient with those on the degree distribution. In Section 4, we introduce the protocols for duplication and mutation, which we have adopted in the present study. In Section 5, we report the results of our simulations. Section 6 provides a discussion with an overview of other content-based models of duplication and divergence.

2. The BE model

The BE content-based network [14] is defined via the following rules: (i) Consider a linear string $\mathscr{C}$ (our AG) of length L, by randomly choosing letters from an alphabet consisting only of the symbols $\{0, 1, 2\}$, according to a predetermined distribution. We will think of this string as a one-dimensional lattice and speak about the "kth site" or the "kth member of the sequence," interchangeably. The character "2" signifies the delimiter between successive words. (ii) Associate the string G_i between the ith and $i+1$st delimiters, with the ith node of a graph, $i = 1, \ldots, N$. (For convenience we assume that a delimiter has been placed at the first and last positions of the artificial genome.) (iii) To form the adjacency matrix w_{ij}, postulate a directed edge from the node i to the node j to exist iff G_i occurs at least once in the string G_j, i.e., $G_i \subseteq G_j$. This "inclusion relation" implies that $w_{ij} \neq w_{ji}$ unless the lengths l_i, l_j of the strings associated with the nodes i,j happen to be equal. Clearly $w_{ij} = 0$ for $l_j < l_i$.