Random distance dependent attachment as a model for neural network generation in the Caenorhabditis elegans

Roi Itzhack\textsuperscript{1,2} and Yoram Louzoun\textsuperscript{1,2,}*  
\textsuperscript{1}Math Department and \textsuperscript{2}Gonda Brain Research Center, Bar Ilan University, Ramat Gan 52900, Israel

ABSTRACT  
Motivation: The topology of the network induced by the neurons connectivity in the Caenorhabditis elegans differs from most common random networks. The neurons positions of the C.elegans have been previously explained as being optimal to induce the required network wiring. We here propose a complementary explanation that the network wiring is the direct result of a local stochastic synapse formation process.

Results: We show that a model based on the physical distance between neurons can explain the C.elegans neural network structure, specifically, we demonstrate that a simple model based on a geometrical synapse formation probability and the inhibition of short coherent cycles can explain the properties of the C.elegans’ neural network. We suggest this model as an initial framework to discuss neural network generation and as a first step toward the development of models for more advanced creatures. In order to measure the circle frequency in the network, a novel graph-theory circle length measurement algorithm is proposed.

Contact: royi.its@gmail.com

Supplementary information: Supplementary data are available at Bioinformatics online.

Received on October 1, 2009; revised on December 10, 2009; accepted on January 12, 2010

1 INTRODUCTION

Neural networks can have up to 100 billion synapses (Kandel and Jessel, 2000). The information defining the position of each synapse is not hard coded into the genome, since the number of genes coding for the properties of the neural network does not exceed a few tens of genes [specifically 93 in the C.elegans (CE)] mainly Nurtans, Ephrins and Semphorins (Schmitz et al., 2007). The position of each synapse must therefore be determined by other means. The synapse positions can be a combination of genetically coded hardwired connections, stochastic processes or learning mechanisms (Aoto et al., 2007; Bear et al., 1987; Cuthbert et al., 2007; Engert and Bonhoeffer, 1997; Federman et al., 2002; Gruart and Delgado-Garcia, 2007; Jelks et al., 2007; Martin et al., 2000; Nikitin, 2007). The relation between the neurons positions and the network wiring has been extensively studied in the context of optimal wiring (Bolz and Castellani, 1997; Buzsaki et al., 2004; Chklovskii et al., 2002; Cruz-Rivera et al., 1998; Gafarov, 2006; Itzkovitz et al., 2008; Kaiser and Hilgetag, 2006; Karbowski, 2001; Kitano and Fukai, 2007; Lee et al., 2008; Milo et al., 2004; Mitchison, 1991; Mizuno et al., 2007; Perez-Escudero and de Polavieja, 2007). The neuron positioning was proposed to be optimized to minimize the total axon and dendrite lengths (Chen et al., 2006; Perez-Escudero and de Polavieja, 2007). The neuron positions were indeed found to reduce the wiring length (compared to a random positioning). However, it is not clear whether the neuron positioning is optimized to minimize the wiring or if the short wiring is simply the result of the preferential attachment of nearby neurons. The only currently existing fully mapped neural network is the one of the CE. As part of the CE project, a full map of the CE cells was produced, and all neurons and synapses were mapped (Achacoso, 1992). The observed CE has 302 neurons and 2170 synapses, which are divided into motor, sensor and intermediate neurons. A wide distribution of synaptic connections characterizes the CE network. Neurons with a high number of synaptic connections are mostly localized on the main nerve ring. The nerve rings, bundles of ∼100 axons encircling the outside of the pharynx, are the principal circumferential tracts in the CE nervous system (White and Rainbow, 1986). Axons from sensory neurons in the anterior ganglia form six labial nerves that enter the nerve ring at its anterior margin. Most axons from neurons in the lateral ganglia enter the nerve ring at its posterior margin either laterally or ventrally via the ventral ganglion. Other neurons, such as sensors and motor neurons, are naturally localized near the edges of the nematode. However, there are also a large number of sensors near the pharynx because the pharynx centralizes much of the sensory activity of the nematode.

In the current work, we propose to study the CE NN generation mechanism using network theory tools. We represent the CE NN neurons as nodes and synapses as edges. Hundreds of network generation mechanisms were proposed in the last few years, some very generic and some specific to a given network (e.g. among many others (Amaral et al., 2000; Barabasi and Albert, 1999; Blank and Solomon, 2000; Jeong et al., 2000; Jose and Bishop, 2003; Louzoun et al., 2006; Michaluk et al., 1999; Nob, 2003; Ramon and Barabasi, 2003; Watts and Strogatz, 1998; Willinger et al., 2002).

The simplest network generation mechanism model was suggested by Erdos and Renyi (1959). In this model, a network is constructed by randomly connecting nodes. For a given Erdos–Renyi Network with $n$ nodes $\{ER(n,p)\}$, the probability of connecting each pair is $p$. This network has a Poisson degree distribution\textsuperscript{1}, and short average geodesic distances between nodes.\textsuperscript{2}

\textsuperscript{1}The degree of a node is the number of other nodes directly connected to it.

\textsuperscript{2}The geodesic distance between two nodes is the minimal number of edges in the path between the two nodes.
Another well known set of models explains Scale Free Networks (SFN). These networks have a power-law degree distribution, where the probability that a random node would have $k$ neighbors is $p(k) = \frac{\gamma - 1}{\gamma} k^{-\gamma}$, with values of $\gamma$ usually between 2 and 3 (Barabasi and Albert, 1999). A more structured model was proposed by Strogatz and Watts (SW) (1998). This model explains the frequently observed combination of short geodesic paths and a high clustering coefficient. The average geodesic path of the network represents the average minimal number of edges from a random node to another one, and the clustering coefficient of the network quantifies approximately the fraction of closed triangles in the network. This model basically transfers a spatial regular lattice into an Erdos–Renyi network. Models incorporating the nodes positions were also proposed such as the gravity model (Burrowes et al., 2004; Chen et al., 2008; Imamishi et al., 2004; Kalveram, 1992; Karpyn and Per, 2007; Lippi Bruni et al., 2008; Piri and Karpyn, 2007; Riviere et al., 2006; Segers et al., 2008; Xiang et al., 2010; Yiotis et al., 2006; Zhang and Jarrett, 1998). In the gravity model, the probability of an edge between a node couple decreases with the distances between the nodes. Specifically, in the context of the CE, Artzy-Randrup (Artzy-Randrup and Ben Tal, 2004) suggested that the small scale motif analysis performed by Milo (Milo et al., 2002) should consider the spatial features of the CE.

Here we propose that the gravity model can explain many of the observed CE neural network properties. In the current context, this model can be interpreted as a synapse formation probability proportional to a power of the physical distance between the soma of the pre-synaptic neuron and the dendrite of the post-synaptic neurons. The properties of this network depend on the spatial positioning of the different neurons.

2 METHODS

2.1 Degree distribution

The degree $k(i)$ of a node $i$ is defined as the number of neighbors of the node $i$. It can be divided into the indegree (degree of incoming edges), outdegree (degree of outgoing edges) and total degree. $P(k)$ is the probability that a randomly selected node has a degree of $k$.

2.2 Distance distribution

The shortest path or geodesic distance $d(i, j)$ is defined as the minimal number of edges from node $i$ to $j$. Note that in a directed graph, the shortest path $d(i, j)$ from $i$ to $j$ is usually not equal to the shortest path $d(j, i)$ between $j$ and $i$. In fact both are longer or equal to the undirected shortest path, since the undirected path often violates the correct edges direction. The average shortest path is defined as the mean of the distance matrix that represents the undirected path often violates the correct edges direction. The average shortest path is defined as the mean of the distance matrix that represents the shortest paths between all connected nodes.

2.3 Clustering coefficient

The clustering coefficient $C(i)$ of a node $i$ is the fraction of its neighbors that are connected one to each other. Given the set of $k(i)$ neighbors of a node $i$, one could in principle have $k(i)(k(i)-1)/2$ undirected edges between them. The number of connected neighbors divided by $k(i)(k(i)-1)/2$ is the C.C. For a large, completely random graph the average CC is $<k^2>/n$, while for a clique (group of nodes which are completely connected to each other) the C.C is 1.

2.4 Betweenness centrality (BC)

Given all the geodesic distance in a network, the betweenness centrality $CB(v)$ of a node $v$ is defined as the fraction of the shortest paths passing through $v$:

$$CB(v) = \sum_{s \neq t \neq v} \frac{\sigma_{st}(v)}{\sigma_{st}}$$

where $\sigma_{st}$ denotes the number of shortest paths from node $s \in V$ to node $t \in V$, and $\sigma_{st}(v)$ is the number of shortest paths from node $s$ to $t$ passing through $v$. We have computed the BC using the highly efficient Brandes algorithm (Brandes, 2001).

2.5 Small-scale motifs

We have characterized the network using the number of small connected subgraph frequency, compared with their expected frequency in a randomized network (Itzhack, 2007; Karpyn et al., 2006; Milo et al., 2002). These 3-4 node sub-graphs are denoted as network motifs. Recently, network motifs have attracted attention as a tool to study networks’ microscopic details, such as the structure of connections between all connected triplets (Milo et al., 2002). We have used the motif of sizes 3 and 4 using a novel and highly efficient algorithm (Itzhack, 2007).

2.6 Circle enumeration

A directed circle is a closed directed path containing no repetitions (i.e. none of the nodes in the circle appears twice). We define a minimal circle as a circle composed of a minimal path from one node to the other and the minimal path back to the original node.

We have developed a rapid algorithm to enumerate all such circles. The naive algorithm would find the minimal geodesic path between all node pairs in a directed graph, $G(V, E)$, using a Breath First Search (BFS) algorithm (Tarre, 2004), and check for each source node whether nodes at the end of the path have a direct edge to the source node $v_{source}$. This algorithm would, however, count each circle multiple times (Supplementary Fig. S1). The only way to count each circle once would be to maintain a list of all circles, and check for those only present once in the list. This naive algorithm is extremely costly, and becomes impractical even for networks of a few hundred nodes.

We developed a practical algorithm that counts each node only once. If a circle of size $r$ is counted $j$ times in the naive algorithm, the circle counter of $s$ increases only by $1/j$, every time this circle is measured, $j$ is the number of nodes in the circle, with a directed distance of $r$ to the node pointing to them in the circle. In other words, if in a circle of length 8 (for example), the distance from a node $v_3$ to its precedent is $<8$, it implies that there is a shortcut in this circle, and one will not count the circles when starting from $v_3$ (Supplementary Fig. S1). A circle of size $r$ would be counted exactly $r$ times only if the direct distance from each node to its precedent is $r-1$. We thus do not use a data structure but only count for each circle how many nodes have a distance of $r-1$ to their precedent.

The total cost of the algorithm is $O(V^2E^2)$, where $E$ denotes the average geodesic path length. The total memory cost is $O(V^2+E)$ (in the worst case), where $V^2$ denotes the number of all the shortest paths between pairs of nodes, and $E$ denotes the average geodesic path length.

2.7 Network analysis toolbox library

We have developed a comprehensive network analysis toolbox library containing graph import and generation tools, graph manipulation routines and network analysis tools to carry out a large number of network measurements, across many applications and disciplines. The toolbox provides a coherent, easily expandable environment, which to the best of our knowledge is the most complete graph analysis package available today. At the time of writing, it implements over 100 analytical and utility methods for graph analysis and manipulation (Muchnik, http://www.levmuchnik.net/Content/Networks/ComplexNetworksPackage.html).
Achacoso, 1992) over a limited spatial region (Achacoso, 1992). The average geodesic distance (4.004 against 4
The CE NN degree distribution also differs from the one expected
The network properties of the CE NN differ from most known
We placed the neurons according to their observed positions, publicly
We used our publicly accessible network analysis toolbox to measure all the
The CE NN is fully mapped and contains 302 neurons and 2170 synaptic
connections between them (Achacoso, 1992; Durbin and Bodmer, 1987; Hall
and Russell, 1991; White and Rainbow, 1986; Wood, 1988). We used the
connectivity data (Chen et al., 2006) that was recently updated from an earlier
version (Achacoso, 1992). The connected part of the network contains 280
neurons and 2170 synaptic connections. Each synaptic connection between
a dendrite and an axon was treated as a single edge and each neuron was
represented as a node. All edges were considered to have an equal weight.

2.8 Translation of the CE neural network to a graph
The CE NN is fully mapped and contains 302 neurons and 2170 synaptic
connections between them (Achacoso, 1992; Durbin and Bodmer, 1987; Hall
and Russell, 1991; White and Rainbow, 1986; Wood, 1988). We used the
connectivity data (Chen et al., 2006) that was recently updated from an earlier
version (Achacoso, 1992). The connected part of the network contains 280
neurons and 2170 synaptic connections. Each synaptic connection between
a dendrite and an axon was treated as a single edge and each neuron was
represented as a node. All edges were considered to have an equal weight.

2.9 Simulation
We placed the neurons according to their observed positions, publicly accessible in the ‘WormAtlas’ project (WormAtlas, 2009). We defined the probability of creating an edge between two neurons as:

\[ p(i \rightarrow j) = \frac{(d_{ij}^c)}{\sum d_{ij}^c} \]

The \( c \) constant was set to produce the proper number of edges in the network (synapses in the neural network), and the value of \( v \) was varied. For each value of \( v \), we check the average properties as measured from 50 simulations. We used our publicly accessible network analysis toolbox to measure all the networks properties (Muchnik).

3 RESULTS
The network properties of the CE NN differ from most known networks. The CE NN stands halfway between an ERN and a scale free network. The CE NN degree distribution is neither a power law (black line in Fig. 1C), nor a Poisson distribution. The CE NN has a large number of nodes with a very low degree on one side and a fat tail on the other side. The CE NN clustering coefficient is high (0.19 versus 0.016 in parallel ERN). Its shortest paths distribution is also longer than the one typical to ERN (Fig. 1A), and it has a peculiar small-scale motif distribution (Supplementary Table S1). The CE NN degree distribution also differs from the one expected in a Strogatz and Watts model (Watts and Strogatz, 1998).

The CE NN evolves through cell division and synapse formation (Bergamasco and Bazzicalupo, 2006; Chaffie et al., 1985; Chang et al., 2006; Chatterjee and Sinha, 2007; Franxs et al., 2006; Kaufman et al., 2006; Klassen and Shen, 2007; Yamamoto and Achacoso, 1992) over a limited spatial region (Achacoso, 1992). Neurons with distant soma have a lower probability of being connected, as can be seen from the observed distribution of synapses and pre-synaptic soma physical distances (Fig. 2). Note that some of the neurons are long and span most of the nematode body affecting the distance distribution. The probability of connecting two neurons decreases with the distance between the soma centers. If only neurons with short span are considered, the decrease is approximately geometrical with a power between 2 and 3, except for a small group of distal connections that will be further discussed.

We here propose an attachment model that follows the observed node position and synapse distance distributions. All neurons were placed in their observed experimental position, and the probability of creating an edge between two neurons was defined as:

\[ p(i \rightarrow j) = \frac{(d_{ij}^c)}{\sum d_{ij}^c} \]

In order to test the validity of our model, we tested whether the model reproduces the generic features of the observed CE NN. The only free parameter in the simulation is \( \alpha \). We run multiple simulations with different values of \( \alpha \) and for each such value, we did 50 simulations. Most of the properties of the LS are similar to the CE NN for \( \alpha \approx 2.5 \) (Fig. 3 and Table 1). The average geodesic distance (4.004 against 4.01 \pm 0.1), the full shortest path distribution and the average connectivity (244.1 \pm 6.67

against 239.4) are similar between the LS and the CE NN, and very different from the ERN or SFN. The connectivity of node 1 is defined as the number of nodes in its connected component. The only feature that is different between the LS and the CE NN is the clustering coefficient, which is 2 SD away from the LS results. The SD of each
The upper panels represent the average shortest path and the average connectivity with other nodes, while the horizontal lines represent the observed value in the real CE NN. The lower panels represent the absolute difference between the nodes degree distribution of the LS and the CE NN. The absolute difference of the betweenness centrality distribution as a function of \( \alpha \). The difference between the two distributions is computed as \( \sum_{i} \left| \pi_{i} - \overline{\pi}_{i} \right| \) using constant bins. The vertical line represents \( \alpha = 2.5 \) in all plots. As one can see the fit for this value is optimal (the error is minimal) (see also Table 1).

Table 1. All network properties

<table>
<thead>
<tr>
<th>Network</th>
<th>Clustering coefficient</th>
<th>Average distance</th>
<th>Average connectivity</th>
<th>Bidirectional links</th>
</tr>
</thead>
<tbody>
<tr>
<td>CE NN</td>
<td>0.19</td>
<td>4.004</td>
<td>239.4</td>
<td>197</td>
</tr>
<tr>
<td>LS</td>
<td>0.09±0.04</td>
<td>4.01±0.1</td>
<td>244.1±6.67</td>
<td>171±15.07</td>
</tr>
<tr>
<td>ERN</td>
<td>0.02±0.01</td>
<td>3.04±0.01</td>
<td>279±0.8</td>
<td>28.86±0.95</td>
</tr>
<tr>
<td>SFN</td>
<td>0.12±0.01</td>
<td>3.07±0.04</td>
<td>265.2±2.1</td>
<td>12.6±0.54</td>
</tr>
</tbody>
</table>

The results from all the networks that we measured are summarized. The real CE NN compared with the average results of 50 LS network (\( \alpha = 2.5 \)), 50 ERN, and 50 SFN. The rows represent the networks and the columns the properties. The results of the CE NN are very similar to the spatial simulation. The only difference between the CE NN and the LS demonstrates a very similar behavior for all \( \alpha \). For very low values of \( \alpha \), the CE NN is also very different from the ERN (Supplementary Table S1). However, the CE NN and the LS demonstrate a very similar behavior for all \( \alpha \), with the exception of the three nodes simple directed circle motif (motif number 9). This motif appears half as often as expected in the CE NN, while in the simulated network it appears twice as often. This motif is simply a coherent 3 nodes circle. The specific difference in a single motif hints that directed circles are selected against. In order to systematically test this possibility, we developed a new measure—the minimal circle length distribution. A minimal circle is defined as a closed couple of shortest paths between two neighbors. Such circles in general tend to be short, since large circles containing a short circle are ignored (Supplementary Fig. S1). The circle distribution of the CE NN has a long tale with a maximal length of 11 neurons, and many 8–11 neuron long circles (25% of the circles). This is in contrast with an average circle length of <3 in the ERN (Fig. 4). It is also much longer than predicted by the LS. These differences suggest a functional mechanism limiting directed short circles. Such a mechanism is expected if neural networks require long circles of information flow.

To further validate this conclusion, we changed our model to mimic destruction of synapses participating in three and four node circles. First, 500 extra edges were added to the network. These edges were then removed selectively from 3 and 4 node circles. We denote this model as LS1. In LS1 the main features of the network did not change significantly (Supplementary Table S2). The motif frequency now fully fits the observed CE NN, and the number of circles of length 2 to 5 is similar to the real CE NN (Fig. 4, Supplementary Table S3).

One can thus summarize that the CE NN network structure shows that the CENN can be formed using a simple spatial stochastic model accompanied by a mechanism limiting short circles of information flow.

The negative selection of short circles explains part of the increase in the long circle number, but not all of it. The discrepancy in the long

---

\*Groups of nodes which are completely connected to each others.
The CE NN is currently the only fully mapped network. Beyond its here, the long circles are a hard coded feature of the CE NN. The same ganglia and form pre-defined consistent physical pathways. From the simulation. Moreover, all nodes in long circles belong to with a large number of edges is much longer in reality than expected (of the neurons in the CE NN (Chen et al., 2006) and the ganglion to which they belong. The Euclidian length (the sum of the distances (between the centers of the soma) of experimentally measured circles with a large number of edges is much longer in reality than expected from the simulation. Moreover, all nodes in long circles belong to the same ganglia and form pre-defined consistent physical pathways. One can thus conclude that in contrast to all other features studied here, the long circles are a hard coded feature of the CE NN.

4 DISCUSSION

The CE NN is currently the only fully mapped network. Beyond its obvious physical parameters (size, number of neurons, number of synapses and the physical position of the neurons), it has several well-defined properties, such as its peculiar degree distribution, clustering coefficient, centrality and shortest distance and circle distribution. This fully mapped network allows us to test if the network structure is predetermined and rigid, or is based on a simple stochastic process. We have tested the simplest stochastic process where the probability of creating a synapse decreases geometrically with the physical distance. This model contains a single parameter and can explain the vast majority of the network properties. The main difference between the proposed model and the experimentally measured CE NN is the low number of coherent circles in the observed network. This may be due to a secondary process that limits useless very short information cycles and increasing the number and length of long circles. Such an inhibitory process is reasonable, taking into account that most axon guidance and neural adhesion genes are inhibitory. Thus, the network generation is a mixture of spatial and functional limitations. The proposed model suggests that the network properties can be explained by the need to maintain global connectivity and a local structure. Such a balance leads to the creation of local nerve rings with a high internal clustering on the one hand and the creation of long range axons on the other hand. At the mathematical level, these two requirements limit the values of α to a very narrow range and explain most of the network properties.

A complementary view of neural networks is that the neurons properties have evolved over long periods to minimize the wiring length. Neuron positions may have evolved to optimize some functionality, for example, through the aggregation of neurons in rings. However, the observed minimal wiring length cannot serve as a proof for such a mechanism. The proper positioning of neurons based on their wiring properties can be an artifact of the network generation mechanism. We have applied an optimal wiring algorithm (Chen et al., 2006) to the network produced by the simulation and obtained a average error of 0.17 ± 0.03 in the node positions (random positions would give 0.35 ± 0.01). A similar error was obtained when the CE NN was used. The analysis was performed assuming a single synapse between each pair of connected neurons.

While the nematode’s NN is much simpler than that of primates and advanced vertebrates, its generation rule may hint to the way much more complex networks are generated. One may extrapolate from these results two basic building elements: (i) a local stochastic creation mechanism combined with functional limitations on short coherent circles, (ii) local structure and long range connectivity need to be maintained. Such a model can be the prototype for the way the structure of large complex networks can be theoretically coded by a very small number of genes.

Conflict of Interest: none declared.

REFERENCES


