Pattern Discovery in Breast Cancer Specific Protein Interaction Network

Xiaogang Wu, PhD[1, 2], Scott H. Harrison, PhD[1, 2], Jake Yue Chen, PhD[1, 2, 3, *]
[1]School of Informatics, Indiana University, Indianapolis, IN;
[2]Indiana Center for Systems Biology and Personalized Medicine, Indianapolis, IN;
[3]Department of Computer and Information Science, Purdue University, Indianapolis, IN
***Email: jakechen@iupui.edu**

Abstract

The interest in indentifying novel biomarkers for early stage breast cancer (BRCA) detection has become grown significantly in recent years. From a view of network biology, one of the emerging themes today is to re-characterize a protein's biological functions in its molecular network. Although many methods have been presented, including network-based gene ranking for molecular biomarker discovery, and graph clustering for functional module discovery, it is still hard to find systems-level properties hidden in disease specific molecular networks. We reconstructed BRCA-related protein interaction network by using BRCA-associated genes/proteins as seeds, and expanding them in an integrated protein interaction database. We further developed a computational framework based on Ant Colony Optimization to rank network nodes. The task of ranking nodes is represented as the problem of finding optimal density distributions of "ant colonies" on all nodes of the network. Our results revealed some interesting systems-level pattern in BRCA-related protein interaction network.

Introduction

The interest in indentifying novel biomarkers for early stage breast cancer (*BRCA*) detection has become grown significantly in recent years[1-3]. Known BRCA susceptibility genes, e.g. P53, BRCA1, BRCA2, ERBB2 and PTEN, only account for 15-20% of the familial risk[4]. Identification of these genes[5], while extremely precious, is only a first step to understand BRCA progression. From a view of network biology[6], these genes never function in isolation[7], one study re-characterized them in a molecular interaction network for BRCA, and identified HMMR as a new susceptibility locus[3]. Another study integrated protein interaction network and gene expression data to improve the prediction of BRCA metastasis[8]. These works suggest that protein interaction networks, although noisy and incomplete, can serve as a molecule-level conceptual roadmap to guide future network biomarkers studies[9].

On the other hand, it is found that both biological shape[10, 11] and physiological signals[12, 13] have chaotic and/or fractal characteristics[14], which indicate that many biological systems and networks could be analyzed effectively by applying *nonlinear dynamical approaches* involving chaos, fractal, bifurcation, pattern formation and complex systems[15]. For these studies, the concept of dynamical biomarkers was firstly introduced on a speech by A.L. Goldberger in 2006[16], which can be seen as an initiation of using nonlinear dynamical properties as biomarkers, although this concept has not extended to the area of molecular networks.

Based on the relationship between features of complex networks (e.g. scale-free) and nonlinear dynamical properties (e.g. fractals)[17], systems-level biomarkers (*sys-biomarkers*), as an innovative concept shown in Figure 1, derive from the marriage of network biomarkers and dynamical biomarkers. Although many methods have been presented in network biology, including network-based gene ranking for molecular biomarker discovery[18], and graph clustering for functional module discovery[19], it is still hard to find sys-biomarkers hidden in disease specific molecular networks.

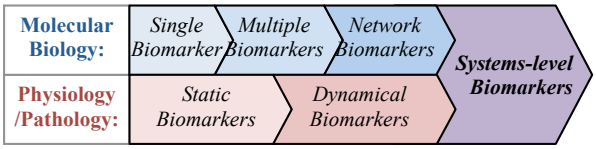


Figure 1. Evolvement of concepts on diagnostic biomarkers.

Starting with the initial motivation of systems biology[20], we reconstructed BRCA-related protein interaction network by taking BRCA-associated genes/proteins as seeds, using the nearest-neighbor expansion method[21], and expanding them in an integrated protein interaction database.

Our method allows BRCA experts to merge their prior knowledge on the BRCA-associated genes/proteins into a manually curated list (protein seeds), which could be obtained from the OMIM™ database (Online Mendelian Inheritance in Man™).

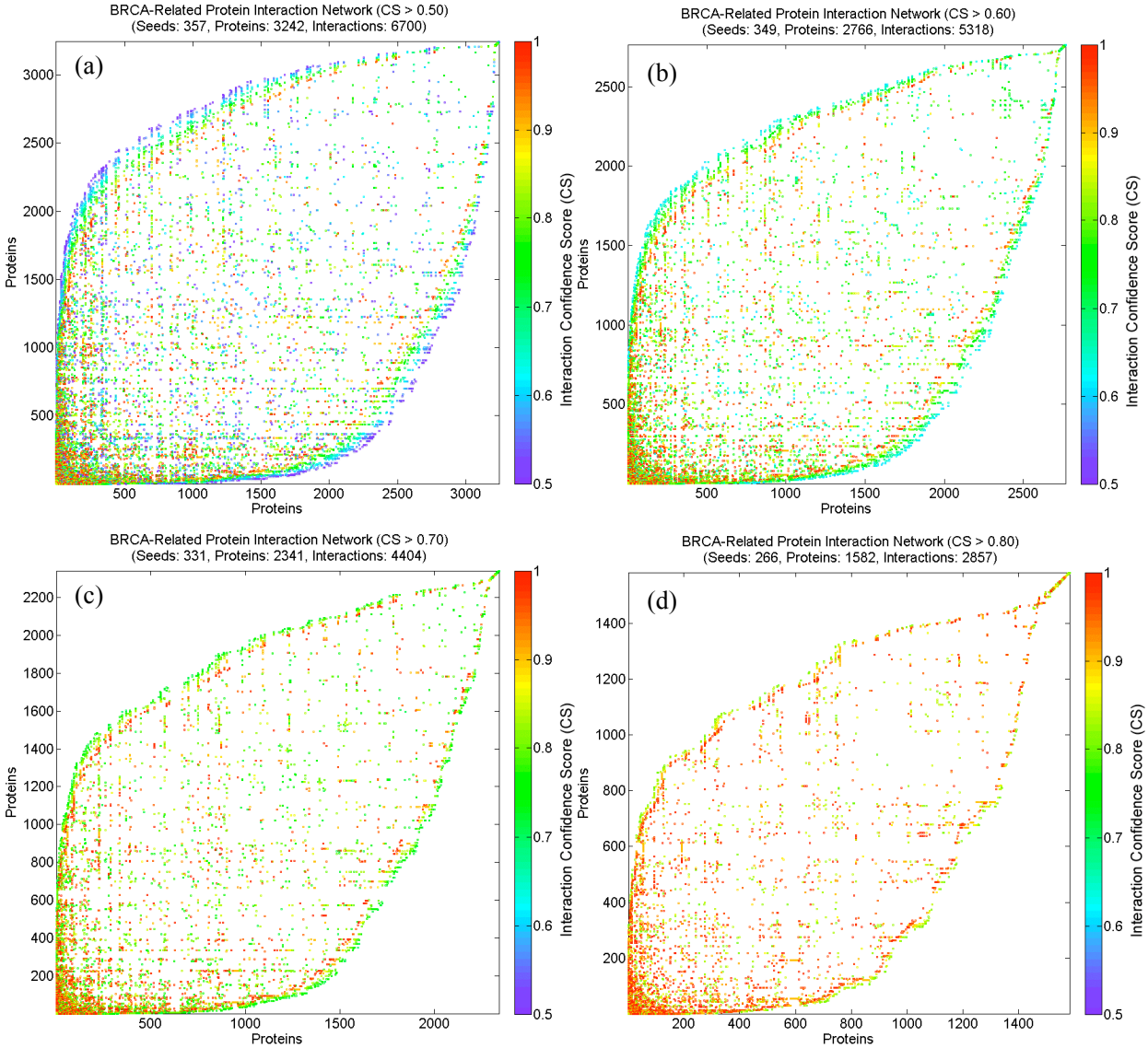
Here, we use the latest high-quality subsets of protein interaction data integrated into the Human Annotated

and Predicted Protein Interaction (HAPPI, http://bio.informatics.iupui.edu/HAPPI) database. In this database, all protein interactions are weighted, with a confidence score (SC) encoding prior knowledge of experimental and literature evidence supporting each protein interaction.

We further developed a computational framework based on Ant Colony Optimization (ACO) [22] to rank network nodes. The task of ranking nodes is represented as the problem of finding optimal density distributions of "ant colonies" on all nodes of the network. Our results revealed some interesting systems-level pattern in BRCA-related protein interaction network.

Results

In our experiments, we firstly constructed an BRCA-related protein interaction network as described above. Using the ACO ranking algorithm, the ranking results of the weighted BRCA-related protein interaction network are shown in Figure 2, which show the ranked adjacency matrix according to the final density distribution. The top 20 proteins from the ranking result shown in Figure 2(f) are highlighted in Figure 3, which shows a high-quality BRCA-related protein interaction network when taking interaction confidence scores CS > 0.99. Node degree distributions plotted in Figure 4 for each BRCA-related protein interaction network taking different CS thresholds, are all very close to power-law distribution, which implies scale-free features.



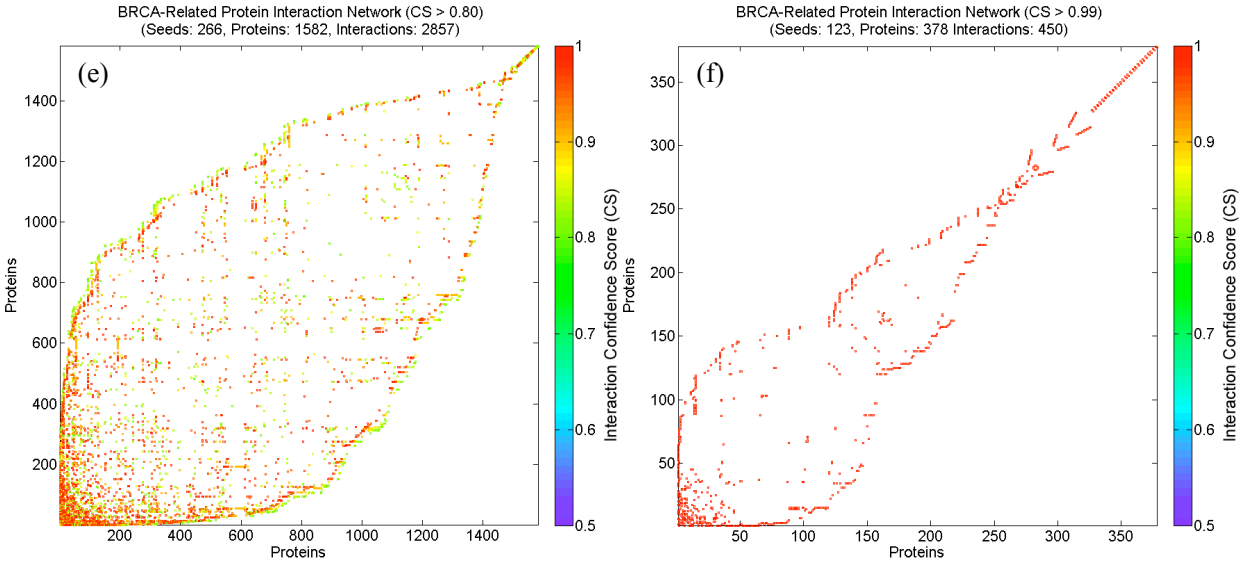


Figure 2. Node ranking of the weighted BRCA-related protein interaction network. (a) CS > 0.50; (b) CS > 0.60; (c) CS > 0.70; (d) CS > 0.80; (d) CS > 0.90; (d) CS > 0.99.

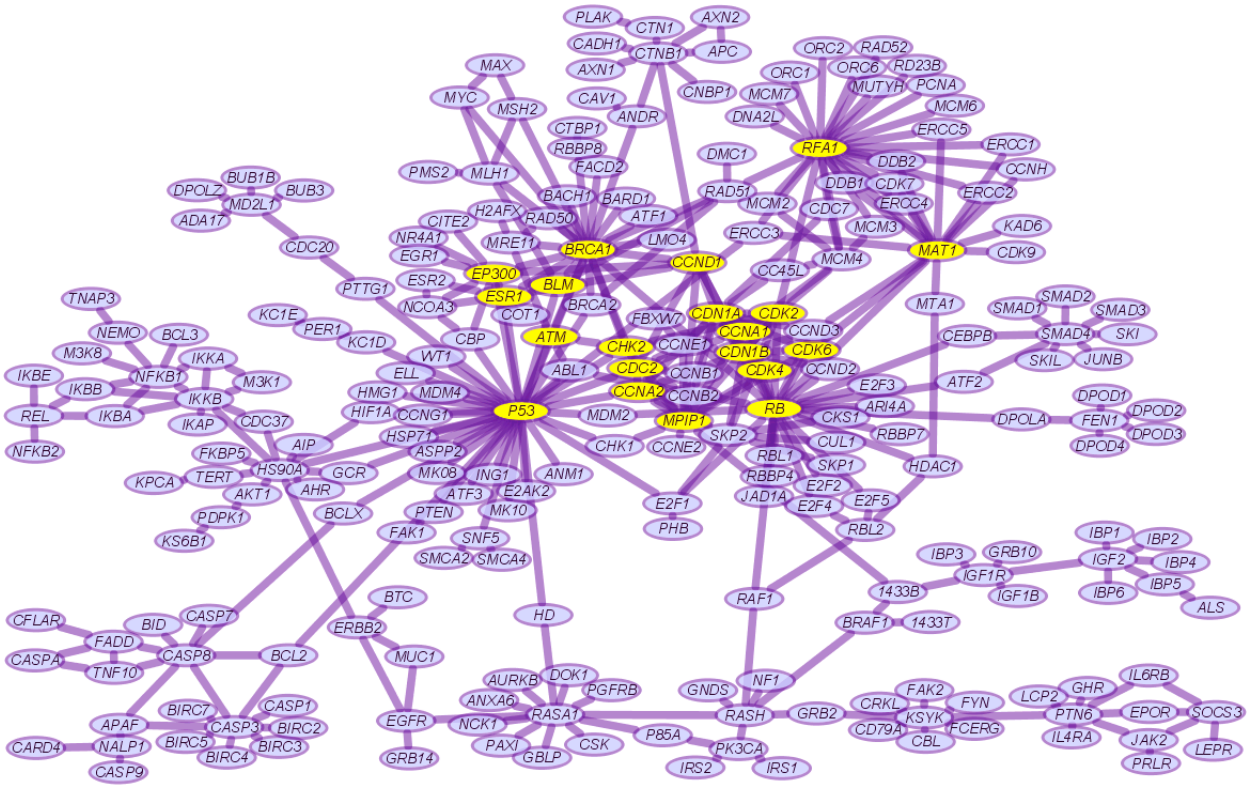


Figure 3. A visual layout of the BRCA protein interaction network (CS > 0.99). Top 20 proteins from the ranking result shown in Figure 2(f) are highlighted.

Discussion

ACO is a dynamic process effective in solving optimization problems such as those of phylogenetic analyses in biology[15]. Here, we represent the task of finding network relevant nodes as an ant colony optimization problem, in which simulated ants (*s-ant*) roam all possible network paths iteratively. By designing various strategies of s-ants for each step taken to walk in a network, the iteration process can be manipulated to get the density distribution of s-ants crowding on each node. According to this

density distribution, the adjacency matrix of the network with ranked nodes is shown as a map in order to reveal the system-level features of the network. Experiments on an BRCA-relevant protein interaction network demonstrated that this method finds the key nodes in the network, and also reveals a fractal feature of the scale-free network through a quick-populating strategy of colonization. Analyses for both unweighted and weighted protein interaction networks based on this framework are given to exhibit the feasibility and flexibility of our method. Comparisons with previous works on BRCA-related protein interaction networks show the reliability of ACO.

Conclusion

Proteins ranked from an BRCA network using our method not only show system-level fractal characteristics but are also useful for subsequent translational biomedical discoveries of gene/protein-disease associations. The highly-ranked proteins from the case study for BRCA could be prioritized for "drug target candidates" and, with additional validation, for "disease biomarker candidates", where proteins may be differentially and specifically expressed in tissues/biofluids based on an associated condition of health or disease. We found that ACO-adapted framework to be robust in identifying fractal-like organization with or without confidence weightings of network connections. Our results revealed fractal features not previously reported in disease-specific molecular interaction networks. Our results are comparable but seem more sensitive than a previous study[11], suggesting convergence of different algorithmic approaches in revealing the same network characteristics of BRCA-related proteins. Proteins in this disease-specific network could have dramatically different characteristics than in the global network. For example, as labeled in Figure 2, CDK5 is a major BRCA-related protein and a "mini-hub" in the BRCA protein interaction network, but it is not a major hub in the global networks based on having a node degree of only 22 in the HAPPI database[12]. If we accept that fractal features reflect a high level of "orderness" eventually interpretable in biology, the results of our study and methodology could point to a brand-new direction of finding and ranking proteins and genes systematically for all human diseases with public data available to bioinformatics researchers today.

Methods

In this framework, node ranking is seen as an optimization problem, which is why the concept of an "ant colony" can be utilized. ACO is mostly like a multi-agent system, but each s-ant (also can be seen as an agent in the system) will mark its path in ACO in a manner comparable to the natural situation where a real ant will leave a pheromone on its track. The pheromone on the ground will stimulate other ants to work together and the whole ant colony will become more cooperative, in a phenomenon of self-organizing communication called *stigmergy*[13]. This characteristic of self-organization leads to an emergence of a complex system, and we propose to leverage this characteristic into solving the problem of complex biological networks by using it as a basis for complex systems modeling. In our developed methodology, s-ants roam all possible network paths iteratively, and marks signed by the s-ants act to accelerate the optimization process. By designing various strategies $\mathbf{F}_i$ of s-ants for each step taken to walk in a network, the iteration process can be manipulated to get the density distribution s_i of s-ants crowding on each node, as shown in Eq. (1). According to this density distribution, the ranked adjacency matrix of the network will be shown as a map to reveal the system-level feature of the network.

$$\begin{aligned} &s_{i+1} = \mathbf{F}_i(\mathbf{M}_i) \times s_i, s_i \in R^n, \mathbf{M}_i \in R^{n \times n}, \\ &\mathbf{F}_i \in R^{n \times n} \to R^{n \times n}, i = 0,1,...,N-1 \end{aligned} \quad (1)$$

Here $\mathbf{M}_i$ is determined by both the network features under analysis (including topology and weighted information) and the marks signed by s-ants. The initial column vector can be evaluated as $s_0 = (1/n, 1/n,\ldots, 1/n)^{\mathrm{T}}$ to describe the equivalence of each node in the network. The final density distribution s_N will determine the rank of each node. Moreover, marks signed from outside will easily switch this scheme from an unsupervised mode into a supervised one.

In a simple case of the proposed scheme, s-ants never sign a mark on the network, and $\mathbf{M}_i$ is only determined by the network, which means it is invariable. Eq. (1) can be reduced as:

$$\begin{aligned} s_{i+1} &= \mathbf{F}_i(\mathbf{M}_i) \times s_i = \mathbf{F}_i(\mathbf{M}) \times s_i \\ &= \mathbf{F}_i \cdots \mathbf{F}_1 \cdot \mathbf{F}_0(\mathbf{M}) \times s_0 \end{aligned} \quad (2)$$

For further simplification, s-ants can be modeled by the constraint of maintaining a constant walking strategy, and Eq. (2) can be reduced as:

$$s_{i+1} = \mathbf{F}_i(\mathbf{M}) \times s_i = \mathbf{F}^{(i)}(\mathbf{M}) \times s_0 = \mathbf{M}^i \times s_0 \quad (3)$$

Here $\mathbf{M}$ becomes the state transition probability matrix about the network. From Eq. (3), we observe there to be a typical Markov Chain. Let $\mathbf{P}$ denote the adjacency matrix of the network (in spite of directed versus undirected or unweighted versus weighted). In

the event where s-ants fail to populate, **M** can be obtained by Eq. (4).

$$\mathbf{P} = \{p_{i,j}\}, \mathbf{M} = \{m_{i,j}\},$$
$$m_{i,j} = \frac{p_{i,j}}{1 + \sum_{j} p_{i,j}}, i, j = 1,2,...,n \quad (4)$$

We established by proof that the final density distribution s_N has a convergent limit as described by Eq. (5).

$$\lim_{N \to \infty} s_N = S = \{S^{(i)}\},$$
$$s^{(i)} = \frac{1 + \sum_{j} p_{i,j}}{n + \sum_{i} \sum_{j} p_{i,j}}, i, j = 1,2,...,n \quad (5)$$

If s-ants populate quickly, **M** can be simply evaluated as **M** = **P**. In this situation however, a convergent property of this algorithm cannot be assured for all kinds of networks. In our experiments, it seems to be related with a scale-free feature.

References

1. Hartwell, L., Mankoff, D., Paulovich, A., Ramsey, S. & Swisher, E. Cancer biomarkers: a systems approach. *Nature Biotechnology* **24**, 905-908 (2006).
2. Hinestrosa, M.C. et al. Shaping the future of biomarker research. *Nature Reviews Cancer* **7** (2007).
3. Pujana, M.A. et al. Network modeling links breast cancer susceptibility and centrosome dysfunction. *Nature Genetics* **39**, 1338-1349 (2007).
4. Balmain, A., Gray, J. & Ponder, B. The genetics and genomics of cancer. *Nature Genetics* **33**, 238-244 (2003).
5. Easton, D.F. et al. Genome-wide association study identifies novel breast cancer susceptibility loci. *Nature* **447**, 1087-1095 (2007).
6. Barabasi, A.L. & Oltvai, Z.N. Network biology: understanding the cell's functional organization. *Nature Reviews Genetics* **5**, 101-113 (2004).
7. Goymer, P. Cancer genetics: Networks uncover new cancer susceptibility suspect. *Nature Reviews Genetics* **8**, 823-823 (2007).
8. Chuang, H.Y., Lee, E., Liu, Y.T., Lee, D. & Ideker, T. Network-based classification of breast cancer metastasis. *Molecular Systems Biology* **3**, 140-149 (2007).
9. McCarthy, N. Tumour profiling: Networking, protein style. *Nature Reviews Cancer* **7**, 892 - 893 (2007).
10. Tatsumi, J., Yamauchi, A. & Kono, Y. Fractal Analysis of Plant Root Systems. *Annals of Botany* **64**, 499 (1989).
11. Palmer, M.W. Fractal geometry: a tool for describing spatial patterns of plant communities. *Plant Ecology* **75**, 91-102 (1988).
12. Goldberger, A.L. et al. Fractal dynamics in physiology: Alterations with disease and aging. *Proceedings of the National Academy of Sciences* **99**, 2466-2472 (2002).
13. Costa, M., Goldberger, A.L. & Peng, C.K. Broken Asymmetry of the Human Heartbeat: Loss of Time Irreversibility in Aging and Disease. *Physical Review Letters* **95**, 198102-198105 (2005).
14. Peitgen, H.O., Jugens, H. & Saupe, D. Chaos and Fractals: New Frontiers of Science. (Springer, 2004).
15. Amaral, L.A.N. et al. Emergence of Complex Dynamics in a Simple Model of Signaling Networks. *Proceedings of the National Academy of Sciences of the United States of America* **101**, 15551-15555 (2004).
16. Goldberger, A.L., G.B. Moody, and C.K. Peng, Techniques, applications and future directions, Heart Rate Viability 2006, April 20-23, 2006.
17. Goh, K.I., Salvi, G., Kahng, B. & Kim, D. Skeleton and Fractal Scaling in Complex Networks. *Physical Review Letters* **96**, 18701-18704 (2006).
18. Morrison, J.L., Breitling, R., Higham, D.J. & Gilbert, D.R. GeneRank: Using search engine technology for the analysis of microarray experiments. *BMC Bioinformatics* **6**, 233 (2005).
19. Bar-Joseph, Z., Gifford, D.K. & Jaakkola, T.S. Fast optimal leaf ordering for hierarchical clustering. *Bioinformatics* **17**, S22-S29 (2001).
20. Ideker, T. Systems biology 101: What you need to know. *Nature Biotechnology* **22**, 473-475 (2004).
21. Chen, J.Y., Shen, C. & Sivachenko, A.Y. Mining Alzheimer disease relevant proteins from integrated protein interactome data. *Biocomputing 2007-Proceedings of the Pacific Symposium* **11**, 367-378 (2006).
22. Dorigo, M., Bonabeau, E. & Theraulaz, G. Ant algorithms and stigmergy. *FUTURE GENER COMPUT SYST* **16**, 851-871 (2000).