Agent-Based Simulation Tools
in Computational Epidemiology

Padmavathi Patlolla, Vandana Gunupudi, Armin R. Mikler, and Roy T. Jacob

Department of Computer Science and Engineering
University of North Texas, USA 76203

Abstract. An agent-based approach is evaluated for its applicability as
a new modeling technology in the emerging area of Computational Epi-
demiology, a research domain that attempts to synergistically unite the
fields of Computer Science and Epidemiology. A primary concern of epi-
demiologists is investigating the spread of infectious diseases. Computer
Scientists can provide powerful tools for epidemiologists to study such
diseases. The existing simulation approaches available to epidemiologists
are fast becoming obsolete, with data being stored in newer formats
like GIS formats. There is an urgent need for developing computation-
ally powerful, user-friendly tools that can be used by epidemiologists to
study the dynamics of disease spread. We present a survey of the state-
of-the-art in agent-based modeling and discuss the unique features of
our chosen technique. Our agent-based approach effectively models the
dynamics of the spread of infectious diseases in spatially-delineated en-
vironments by using agents to model the interaction between people and
pathogens. We present preliminary results of modeling an actual tuber-
culosis disease outbreak in a local shelter. This model is an important
step in the development of user-friendly tools for epidemiologists.

1 Introduction

Computational epidemiology is a developing research domain that attempts to
unite the disparate fields of computer science and epidemiology. Epidemiologists
are primarily concerned with investigating disease outbreaks and risk assessment
in spatially delineated environments, investigating vaccination strategies to con-
trol the spread. Thorough understanding of the dynamics of disease transmission
is key to predicting the spread of a disease and controlling it. Epidemiologists
employ various statistical methods to analyze the data relating to a disease, but
there are no specific tools that they can use to study a disease, its spread, the
spread of the infective agent and other factors. With the emergence of new in-
fectious diseases like Lyme disease, Hepatitis-C, West Nile Virus, and HIV, the
need to develop tools that epidemiologists need to study these diseases becomes
apparent. As more strains of diseases like tuberculosis and pneumonia become
resistant to antibiotics, it becomes imperative to track the progression/mutation
of these strains. The method of field trials, currently available to epidemiologists,
is either prohibitively expensive or unethical. Applying mathematical models and

using computers to process the copious amounts of data related to a particular disease outbreak can aid in understanding the dynamics of the disease spread to a certain extent. But, even today, in spite of the availability of computational resources to process the huge data involved, epidemiologists face tough challenges in trying to understand the results obtained by applying the mathematical models and computer programs. Epidemiologists are not trained to understand the intricacies of mathematical theories or the subtleties of computer programs. This is where computer scientists can step in, by harnessing the power of powerful computing resources now available and developing user-friendly tools that epidemiologists can use.

Epidemiologists are often faced with the challenge of dealing with data that are sparse, widely distributed, and incomplete (often due to confidentiality and other constraints). This may result in conflicting information that confound or disguise the evidence, leading to wrong conclusions. Today, the role of epidemiologists has become even more pronounced as the significance of Public Health has been recognized. To meet the increasing demands, the field of Epidemiology is in need of specific computational tools that would enable the professionals to respond promptly and accurately to control and contain disease outbreaks. Increased globalization, highly mobile populations, and possible exposure to infectious diseases pose new public health threats. It is vital to develop new tools that take advantage of today's communication and computing infrastructures. Computational models for the simulation of global disease dynamics are required to facilitate adequate what-if analyses. This necessitates adapting fundamental Computer Science concepts to the specific problems in Epidemiology.

One of the primary challenges that Computational Epidemiologists face today is trying to understand the spread of a disease globally. The results obtained from the processing of the available data are difficult to analyze because of the sheer size of the population involved. In such scenarios, providing visualization tools would allow the scientists to process available data and draw relevant conclusions. For example, consider the critical issue of limiting the spread of an infectious disease in a particular area (city or town, for example). The data immediately available to epidemiologists may include information about the extent of disease spread, the areas the epidemic has spread to and the number of vaccine doses available for immunization. If the number of doses is limited, a decision must be made about vaccination strategies, *i.e.* what is the optimal way to vaccinate in order to effectively limit the spread of the disease. The goal is to provide epidemiologists simulation tools that will allow them to analyze the effect of various immunization strategies to derive an optimal strategy for immunization with the available resources. For example, they may arrive at the conclusion that ring vaccination, i.e. vaccinating in the area of 1 mile radius around the infected area, will contain the spread of the disease.

In 2003, an outbreak of SARS in a province of China spread quickly to geographically-remote parts of the world like Toronto. When such a disease outbreak occurs, epidemiologists must study the outbreak to predict its spread. At this time, there are no general-purpose computational tools available to epidemi-

ologists to model disease outbreaks. Outbreaks like SARS in 2003 illustrate the need for developing tools that epidemiologists can use to model an outbreak quickly and predict its spread successfully. These include mathematical or computational modeling tools to effectively monitor the spread of a disease, track mutations of different strains, and implement effective vaccination strategies, and investigate the spread of diseases.

Computational Epidemiology addresses the broader aspects of epidemiology, primarily disease tracking, analysis and surveillance. An example of the synergistic collaboration of computing power and biology is the genome research project, which involved mapping the human genome. Similarly, we can use high performance computing and data visualization techniques to develop tools to simulate disease outbreaks and aid in the investigative process allowing them to respond promptly and effectively contain the spread of diseases. The availability of data from geographic information systems (GIS), new visualization techniques (like virtual reality) and high-performance computing paradigms, such as cluster and grid computing, will greatly contribute to the development of tools that facilitate the work of today's epidemiologists.

2   Outbreak Models

Mathematical models are important tools that can be combined with analytical tools to model disease outbreaks, study mutations of viruses, help in developing effective immunization and vaccination strategies. When combined with other analytical methods, they can become powerful tools for epidemiologists. Mathematical models can predict future outbreaks, present risk analysis, compare alternatives and methods, and even help prepare effective response strategies for bioterrorism attacks. In order to provide epidemiologists with user-friendly tool, a synergy is required between computer scientists and epidemiologists to help in the utilization of currently available tools and development of new ones. It is imperative that both computer scientists and epidemiologists cooperate in order to develop effective tools.

Dynamical modeling and statistical methods have been used in epidemiology for many years, but their role is changing now due to increasing size of the associated data. There has been little effort until now to harness the power of computational resources and apply them to existing approaches. Other modeling paradigms, such as agent-based modeling and stochastic cellular automata, can now be applied to epidemiology since we can use available data sets in the form of GIS data and large computer databases. The SARS outbreak was a global outbreak, whereby the disease that originated in a remote part of China spread rapidly to other parts of the world like Toronto. Global outbreaks of diseases are dependent on a number of factors like demography, geography and culture of a region, socioeconomic factors, and travel patterns. Local disease outbreaks, on the other hand, are outbreaks of diseases in spatially delineated areas like factories, homeless shelters. The spread of the disease is thus dependent on airflow rates, heating and cooling, the architectural properties of the delineated space, and social and spatial interactions.

3   Computing Paradigms

Different computational paradigms have been used to model the behavior of biological phenomena like disease outbreaks. Depending on the nature of the outbreak, different paradigms are required. For example, the same computing paradigm cannot be used to model both global and local outbreaks. Stochastic Cellular Automata is a novel paradigm that can successfully model global disease outbreaks, taking into account the factors responsible for the outbreak. Modeling global outbreaks is a particularly challenging task for which stochastic cellular automata are a useful modeling tool. Each region consists of cells, where each cell is influenced by its neighbors. The state of each cell is dependent on the state of its neighbors. This paradigm is useful for modeling vaccination strategies such as ring vanninations. It can also be used to model other vaccination strategies.

Whether it is global or local, modeling a disease outbreak is particularly challenging. The sheer complexity of the problem overwhelms traditional analytical tools. A particularly promising paradigm for modeling local outbreaks is that of agent-based modeling, which involves assigning an agent to each object in the environment that we want to model. We can exploit the features of agent-based modeling to address problems that can be intractable using traditional models. We discuss agent-based modeling in the following section.

3.1   Agent-Based Modeling

Agent-based models are simple but allow modeling of complex phenomena. An agent can be defined simply as an entity that acts on behalf of others, while displaying some form of autonomous behavior in choosing its actions. Using this simple definition of an agent, we can build very complex systems. By assigning agents to the entities in a system, we can run controlled experiments that allow us to change some parameters of the system while keeping the others constant. In this manner, an entire history of the system can be developed. Agent-based modeling is particularly useful for modeling local disease outbreaks and can be combined with other techniques and, by harnessing computational resources, used to effectively simulate local disease outbreaks.

The functionality in an agent-based system is implemented through the interaction of agents. Assigning agents to different interacting entities in the outbreak will allow is to model the spatial and social interactions in spatially delineated environments. The infected people as well as the infectious bacilli or viruses can be agents and purposeful movement can be incorporated to simulate the disease in a realistic manner. Purposeful movement implies that the agents can move of their own volition, simulating the movement of the objects (people and bacilli) in the actual disease.

Incorporating purposeful movement is not a simple task. We need first to investigate the type of environment and the nature of the entities we want to model, study their desires and model them as a function of some parameters. When the desire function reaches a threshold, agents tend to change their behavior, such as move in a different direction or perform a certain action. For

example, if we are trying to model humans with desires to smoke and to drink, we associate with each an agent, with variables representing their present level of desire to smoke or drink. When its desire level hits a threshold, an agent will perform an action, such as moving toward the smoking or toward the drinking fountain.

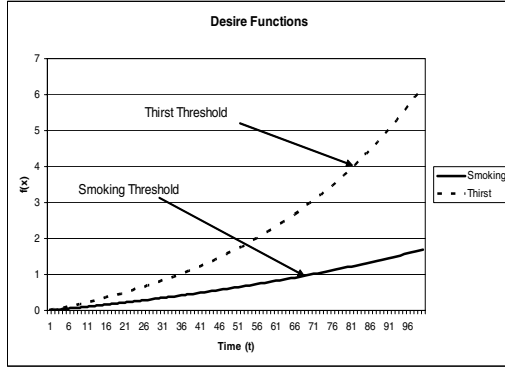Figure 1 shows the desire levels as a function of time and threshold values.

Fig. 1. Desire Function

Now when the desire level reaches the threshold, the agent, which is in the area A in Figure 1 tries to move towards the area D, here it can quench its thirst. In order to move toward the destination D from the source A, we can design a Source/Destination routing table which will help the agent move in the right direction and reach the destination with least effort.
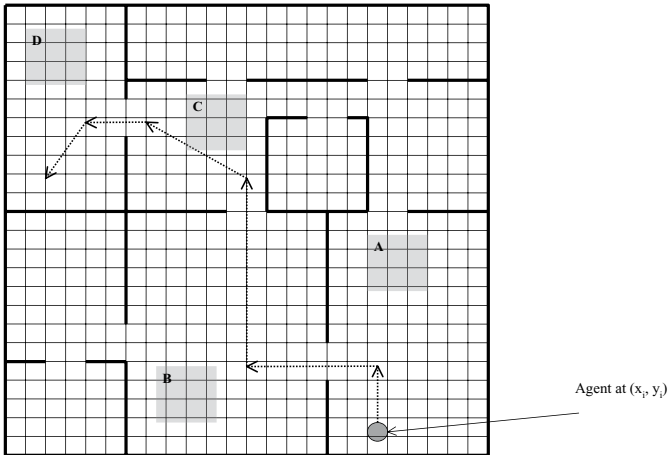

Fig. 2. Grid Environment

S/D	A	B	C	D	...
A	-	B	C	B	...
B	A	-	C	C	...
C	A	B	-	D	...
D	C	C	C	-	...
...	...	...	...	...	...

Fig. 3. Source/Destination Routing Table

3.2 State-of-the-Art in Agent-Based Modeling

Agent-based modeling has been used in the biological domain extensively, and various modeling tools are available. The following section presents a brief overview of the current state-of-the-art in agent-based modeling. This section outlines some of the agent-based computational tools that are currently available.

Swarm. Developed at the Santa Fe institute, Swarm [13] provides a set of libraries that can be used to model complex systems. As the name suggests, the basic component of agent organization is a "swarm", a collection of agents with a schedule of events over the collection. The swarm represents the entire model, i.e, the agents as well as the evolution of these agents over a time period. Swarm is a very powerful and flexible agent platform. However, it is very domain-specific in that extensive knowledge of the Java programming language is required. Consequently, this tool cannot be used directly by many epidemiologists.

Ascape [14], which is modeled on Swarm, is a platform that allows users to develop complex models. It has been used to develop the well-known Sugarscape simulation. Easier to learn and use than Swarm and providing many user-friendly tools. it is also implemented using Java. Some knowledge of the language is required to be able to use it effectively.

RePast [15] is a framework for creating agent-based simulations. Developed by the University of Chicago's Social Science Research Group, it requires knowledge of Java. It has built-in GIS, Excel import/export facility, and support to produce charts and graphs and allows objects to be moved by the mouse. It aims to model

the agents using recursive social constructions and to replay the simulations with altered assumptions.

StarLogo [12] is a multi-agent simulation tool developed at the MIT Media laboratory. It is specifically aimed toward providing support to build decentralised multi-agent simulations. Starlogo is an extension of the programming language Logo, which allows us to define the movement of an agent called turtle on the computer screen by giving it commands. Starlogo extends this idea and provides support to create and control thousands of turtles in parallel, allowing them to move around a world defined by the user. Starlogo provides support to program thousands of patches that make up the turtles world. This allows a purposeful movement for the turtles in which they can sense their surroundings and move by choice. Starlogo is very easy to learn and provides a graphical user interface that allows epidemiologists with no prior programming experience to use it easily. It provides support for plotting graphs, allows the user to define slide bars, buttons and monitors allowing the user to control the simulation, monitoring the various parameters of the simulation and observing how they change with time.

3.3   What's New

Even though agent-based modeling tools useful to epidemiologists exist today, the unique features of epidemiology require the development of new tools. Data from various sources and in different formats need to be input into these models, highlighting the need for developing tools to convert existing data into uniform formats. Also, data are most commonly available in GIS format, but agent-based tools are not able to directly read data from these sources. Therefore, either existing tools need to be modified to read GIS data or new tools must be designed that read the data in GIS format and output the simulation results.

4   TB Outbreak in a Homeless Shelter

Tuberculosis (TB), an extremely infectious disease, is of particular concern when people interact in spatially-delineated environments like factories and homeless shelters. It was a leading cause of death in the 19th century. It spreads through the air, with the bacilli having a settling rate of 3 feet/hour. The tuberculosis bacilli reside in the lungs, so the immune system responds quickly to the infection. Most infected individuals never develop TB, *i.e.* they never become infectious. Most exposed individuals remain "latent-for-life," but around 10 percent of those exposed become infectious. The reasons that TB persists in the United States are co-infection with HIV/AIDS, lack of public knowledge about transmission and treatment, and immigration, which accounts for a large number of new cases. We propose to model the dynamics of an outbreak that occurred in 2000 to model the spread of infection from patient zero to other infected individuals. Using agent-based modeling, we can simulate the outbreak and study the

transmission patterns unique to this setting. We model the layout of the homeless shelter and the interacting entities. The bacilli and the individuals are modeled as agents. Each individual is given a color indicating its level of infection.

4.1 Data Collection

Data was collected through interviews during targeted surveillance screenings of homeless people who use the shelter. The facility opens daily at 4 p.m. and people line-up outside the shelter, waiting to enter it. These individuals are tested at least once a week, with non-regular people being tested more often than the people who sleep at the shelter regularly. The data has been sanitized in compliance with HIPAA regulations, whereby all identifying information has been removed. This data has been provided to us by local health authorities in GIS format and is incorporated into the agent-based model. For the homeless shelter, the data used in the model includes the following information in each case:

- Date tested (relative to t_0)
- Status of tuberculosis
- Location in the facility
- Length of time spent in the facility

4.2 Agent-Based Simulation of the Outbreak

We have modeled this disease outbreak using StarLogo, an agent-based modeling tool that uses "turtles" and "patches" to model the interaction in the environment. Figure 4 shows a screen shot from the simulator giving the layout of the homeless shelter.

The shelter contains mats and beds, with the mats shown in light-grey and the beds shown in dark-grey (see Figure 4). The occupants of the beds are regular inhabitants of the homeless shelter who pay a nominal rent in return for being guaranteed a bed. The people who sleep on mats are usually short-term occupants who sleep in the shelter sporadically. There are separate sleeping areas for men and women, with different sleeping areas for people over 50. The simulation shows the placement of the beds and mats in the different sleeping areas. The upper left-side section shows the beds occupied by men that sleep at the shelter regularly. The lower left-side shows the placement of the mats occupied by men who sporadically sleep at the shelter. The upper right corner shows the beds used by the men that are over fifty years of age. Finally, the lower right area shows the women's sleeping area that has both beds and mats.

The model shows that the beds are spaced further apart from each other than the mats, which illustrates the difference in the infection rate in these areas. The small compartments that are between the men's and the women's area are the restrooms and the lower area without mats and beds is the smoking area. People ocassionally wake up during the night to smoke, congregating in that
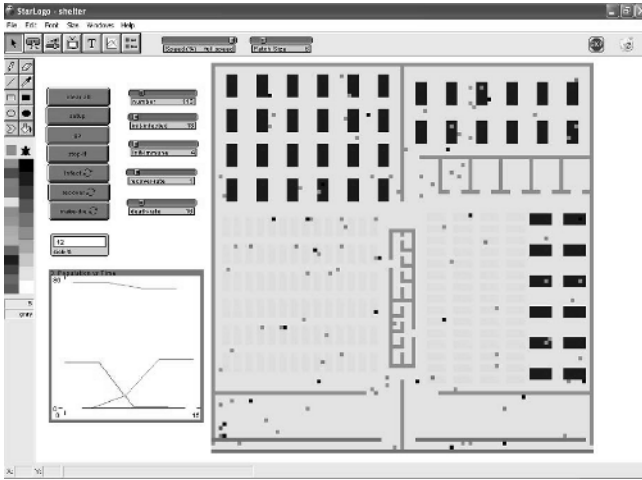
Fig. 4. StarLogo Screen Shot

area. Incidentally, people enter the shelter through the smoking area. The figure shows the queues near the smoking area from which people enter the shelter.

We use different colors (not shown in Figure 4) to represent people in various stages of infection and resistance to tuberculosis. For example, green dots are used to denote healthy people, people who are immune (already immunized) are represented using black, and red is used to represent the infected people. The people, represented as agents, move about randomly in the homeless shelter. Whenever they approach a bed or mat, they may choose to sleep on it for a random amount of time. The amount of time an agent rests on a bed or a mat can be varied in order to simulate the different behaviors of the people. For example, there is more interaction among people during the daytime hours than in the nighttime hours. Accordingly, the agents linger on the beds or mats for a longer time when we want to simulate the nighttime behavior. In our model, the days and nights continue in a cycle until we stop the simulation.

We also simulate the behavior of smokers by allowing the agents to randomly move about the smoking area. The risk of infection among smokers is different from that of non-smokers due to the complexities of the transmission of the tuberculosis infection. Since smoke lingers in the air for a while, the bacilli may survive in the smoke-filled air for a longer time than in recirculated air, so we use different settling rates for the tuberculosis bacilli in recirculated air than for smoke-filled air. In a future paper, we plan to take on the difficult task of modeling the air-flow in the homeless shelter, in order to predict the spread of the infection.

The model shows the spread of infection from the infected people to the healthy people who are not immune. Whenever a healthy person encounters an infected person, he may get infected with probability p. Infected people can take medication, recovering to become immune to future infections. The infected peo-

ple can also die in which case they disappear from the population. The functional parameters of the simulation can be controlled using the slider bars shown in Figure 4. In this model we have one slider to indicate the total number of people in the homeless shelter before we start to run the simulation, the number of initially infected people, number of initially immune people, the death rate and the recovery rate. The model allows us to drag on the sliders, increasing and decreasing the functional parameters to engage in a what-if analysis and arrive at useful conclusions about the transmission of infection among the population.

The movement of the people, the spread of infection among individuals, and the rate with which they recover or die, can be examined visually from the model. At any time, the precise number of people that fall into the infected, recovered (immune), healthy, and dead classes can be seen from the graph plot at the lower left side of the screen shot. Also at any time, the percentage of infected (sick) people is displayed in a small monitor box.

The simulation can be controlled using control buttons that can be switched on or off during the simulation. We have one button each to control the spread of infection, process of recovery, and process of death. We can at any time switch these off or on as needed to analyze the effect of a factor on the spread of infection.

The infection can spread through direct contact between agents while they are moving around in the homeless shelter or by exposure to bacilli that are in the air. The simulation can model these aspects to show that the smoking areas and the restrooms are the areas with high risk of TB infection. The model can be extended to include the other architectural details of the shelter such as the elevation. Ventilation which was better on the women's side than the men's side, may prove to be the reason why the men's side shows high risk of infection.

4.3   Analysis

Primary results show that we are able to model the outbreak in a spatially delineated environment successfully by incorporating "purposeful movement" in the agent entities. This simulation can be extended to incorporate other parameters such as the effect of smoking in the spread of this infection.

The dynamics of the spread were studied using the graphical output generated by the simulation and the behavior of the spread is exactly that explained by the classical SIR model. Also, the threshold levels of the SIR model could be related and observed by using the graphical outputs. There were certain points where the infection rapidly spread, and other points where it slowed. These were related to the threshold levels calculated using the SIR model. Also, we could see that few areas in the shelter had high risks of TB infection. This was due to the differences in air flow, the distance of separation between the beds, and the proximity to the smoking area and restrooms where people usually gather.

5   Conclusion

We have presented a survey of the state-of-the-art in modeling tools for computational epidemiology. We looked at different computational paradigms and

how they can be used to model disease outbreaks. We proposed simulating the outbreak of TB in a local homeless shelter and presented preliminary results.

In order for effective tools to be developed for epidemiologists by computer scientists, a synergistic union of the disparate fields as well as existing computational paradigms is required. Every year millions of dollars are invested by US in research towards finding ways to improve public health and lower the risk of epidemics and disease spread. Recently, attention has also shifted towards bio-terrorist attacks. In such cases, it becomes imperative to understand the social network and its behavior in order to understand the spread of the disease. Once sufficient knowledge about the social groups of the network is established, multi-agent simulations help in modeling and analyzing the risk of disease spread through socially connected groups. In order to do so, we first have to understand the characteristics of the social group, the disease, and the environmental factors that affect the spread of the disease.

The tools discussed above can successfully model social behavior, the interactions, the air flow, the air suspensions, the atmospheric conditions favorable or unfavorable for spread of any disease, the characteristics of the infection spreading agents and their behavior and responses to different prevailing conditions. This helps in analyzing the risk of propagation, the spread of propagation, and the medium of propagation, and helps in mitigating the risk of such spreads or, in some cases, totally eradicating the disease through immunization strategies or other measures. Being able to do so would take the epidemiologists a step further in the way of analyzing disease outbreaks and the spread of epidemics. The results of the simulations and the associated graphs can help to improve understanding of the dynamics of transmission and to take better steps towards the prevention and control of disease spread.

References

1. Gimblett, H. Randy, "Integrating Geographic Information Systems and Agent-based Modeling Techniques", Santa Fe Institute, Oxford University Press, 2002.
2. Carley K. M., & Schreiber, C., "Information Technology and Knowledge Distribution in C3I Teams", 2002 Command and Control Research and Technology Symposium, Monterey, CA, 2002.
3. Buckeridge, DL, Graham, J, O'Connor, MJ, Choy,MK, Tu, SW, Musen, MA., "Knowledge-Based Bioterrorism Surveillance", American Medical Informatics Association Symposium, SanAntonio, TX, November 9-13, 2002.
4. Barrat, Rod, "Atmospheric dispersion modeling: an introduction to practical applications", Earthscan,2001.
5. Carley, K. M. and Hill, V., "Structural Change and Learning Within Organizations", in A. Lomi and E. R. Larsen (eds), *Dynamics of Organizations: Computational Modeling and Organization Theories.* Menlo Park, CA: MIT Press/AAAI, 2001.
6. Lawson, Andrew B., *Statistical Methods in Spatial Epidemiology*, John Wiley & Sons, 2001.
7. Meselson, Matthew, "Note Regarding Source Strength", *ASA Newsletter*, article 01-6a (www.asanltr.com).

8. Carley, K.M., & Gasser, L., "Computational organization theory", in G.Weiss (ed.), *Multiagent Systems: A Modern Approach to Distributed Artificial Intelligence*, MIT Press, 1999.
9. Turner, D. Bruce, "Workbook of Atmospheric Dispersion Estimates: An Introduction to Dispersion Modelling", Lewis Publishers, 1994.
10. Carley, K.M., "A theory of group stability", *American Sociological Review 56(3)*, pp. 331-354, 1991.
11. Miller RA, Pople HE, Myers JD, "Interist-I, An Experimental Computer-based Diagnostic Consultant for General Internal Medicine", N Engl J Med 307, pp. 468-76, 1982.
12. Documentation available from: http://education.mit.edu/starlogo/, Last Accessed: May 3, 2004.
13. Documentation available from: http://swarm.org, Last Accessed: May 3, 2004.
14. Documentation available from: http://www.brook.edu/es/dynamics/models/ascape/, Last Accessed: May 3, 2004.
15. Documentation available from: http://repast.sourceforge.net, Last Accessed: May 3, 2004.