GenePath: From Mutations to Genetic Networks and Back

Peter Juvan[1], Janez Demsar[1], Gad Shaulsky[2] and Blaz Zupan[1,2,*]

[1]Faculty of Computer and Information Science, University of Ljubljana, Ljubljana, Slovenia, [2]Department of Molecular and Human Genetics, Baylor College of Medicine, Houston, USA

[*]To whom correspondence should be addressed. Tel: +386 1 4768 402; Fax: +386 1 4768 386; Email: blaz.zupan@fri.uni-lj.si

Abstract

GenePath is a web-based application for analysis of mutant-based experiments and synthesis of genetic networks. Here we introduce GenePath and describe a number of new approaches, including conflict resolution, handling cyclic pathways, confidence level assignment, what-if analysis and new experiment proposal. We illustrate the key concepts using data from a study of adhesion genes in *D. discoideum*, and show that GenePath discover genetic interactions that were ignored in the original publication. GenePath is available at http://www.genepath.org/genepath2.

Introduction

Discovery of genetic networks is a major goal in functional genomics and bioinformatics. Exploring all plausible connections within a genetic pathway is a formidable task that can be greatly aided by computation. To support that task we have developed a computational method based on principles of epistasis analysis (1). The method and the related program, GenePath, use logical patterns to infer relations between genes from mutant-based experiments, and implement them for automated construction of genetic networks (2,3). The system was prototyped in Prolog and later extended with a web-based interface (4). The implementation was limited in the type of data it could accept and could only analyze linear pathways. GenePath now offers mechanisms that help the researchers analyze cyclic pathways, assign confidence levels to data, resolve conflicts (through explanation or confidence levels), perform exploratory analysis and plan experiments. GenePath is implemented as a stand-alone web application with an intuitive user interface.

Methods

Data input and genetic network inference

GenePath accepts experimental data on mutant-based morphological and transcriptional phenotypes. Figure 1 shows data from a study of adhesion genes and their role in intercellular communication during *D. discoideum* development (5). The data include morphological (Figure 1c) and transcriptional (Figure 1d) changes following knockout or overexpression of the genes *lagC*, *lagD* and *comC*.

GenePath first infers regulatory relations between genes from the experimental data. Its logic is essentially identical to that of epistasis analysis (1,2), where patterns are encoded in the form ‘IF a certain combination of data exists, THEN a certain relationship between a gene and a biological process is hypothesized’ (2,3). The relations between pairs of genes are then assembled into a network. The network inferred from our example data (Figure 1b) reveals that *comC* both inhibits and excites *lagC* and *lagD*, which excite each other in a cyclic relation and exhibit both positive and negative influences on development. Detailed description of the inference of relations and network construction is provided in the supplement (*GenePath: Inference of Relations and Networks* at NAR Online).

Explanation and conflict resolution

GenePath traces every relation back to the relevant data, and provides a textual explanation of the reasoning. Such explanation is particularly useful for resolving conflicts. For instance, GenePath found that *comC* both inhibits and excites *lagC*, whereas the original publication reported only a negative influence (5). The evidence shown in Figure 1e reveals that *comC* excites *lagC* because a knockout mutation in either gene suppresses the ability to form fruiting bodies, and the streaming pattern of the double mutation is more similar to $lagC^{-}$ than it is to $comC^{-}$ (E2, E4 and E9 in Figure 1e). This relation was overlooked by the experimentalists (5), illustrating the important role GenePath can play in data analysis thanks to its formalized, systematic search for all possible relations. This task is hard for manual consideration of small data sets, such as the above, and is nearly impossible for larger data sets.

Handling cycles

Biological systems often utilize auto-regulatory mechanisms in the form of positive and negative feedback loops. In real life, these loops have a temporal component that is usually missing from genetic analyses. As a result, a genetic network cannot define clear input and output points from resulting cycles and exhibits logical conflicts in negative feedback loops. The presence of cycles represents an algorithmic problem for the integration of relations into a network. For example, consider the cyclic pathway between genes *lagC* and *lagD* in Figure 1b. One cannot determine whether *lagC* influences development directly or through *lagD* from the experimental results. GenePath overcomes this problem by inserting the genes involved in cycles into a single node, thus constructing an acyclic network. Figure 1f shows how GenePath visualizes the contracted nodes (genes in a bounding box) and calls the biologist's attention to the presence of a potential feedback loop. The biologist then decides how to continue.

Confidence levels

GenePath allows researchers to translate their subjective belief in experimental methods and published results into internally consistent confidence levels. It also assigns default confidence levels that are related to the number and type of mutations (Figure 1c-d). GenePath treats confidences as probabilities; although they model subjective beliefs, they still conform to the calculus of the probability theory. Confidence levels of the inferred relations are computed from the confidences assigned to the experimental data (see supplemental material *GenePath: Analysis Methods* for details).

GenePath reports confidence levels of relations above edges in the networks (Figure 1b). Interestingly, the relations reported in (5) received higher confidence levels than those subsequently discovered by GenePath. If one would consider only the most confident relations, the network derived by GenePath and the published network (5) would be the same. Confidence levels provide grounds for an approach to automated resolution of conflicts and thus represent a significant step towards formalizing the process of automatic construction of genetic networks from mutant data.

What-if analysis

The what-if analysis is a powerful tool for interactive exploration of experimental results. It allows the user to test the consequences of ignoring a set of experiments, changing the outcome of a selected experiment, or adding hypothetical experiments. The optimized execution code minimizes the time required to process and display changes in the data. This feature provides an on-the-fly environment for exploratory analysis and hypothesis testing.

In our example, the experimentalists produced two types of data, morphological and transcriptional (5). We utilized the what-if analysis to test the consequences of ignoring the transcriptional data. As a result (Figure 1g), the cyclic relation between *lagC* and *lagD* was lost and so was the inhibitory effect of *comC*. The confidence levels were reduced across the board as well.

Experiment proposal

Experimental proposal may help the geneticist plan the next step. Consider the network in Figure 1g, and suppose we suspect that *comC* excites *lagD*. Which mutations should be generated and what outcome would support the hypothesis? Among the numerous possibilities, what experiments would benefit the most from the existing observations and reagents?

GenePath reverses the reasoning used to infer relations (4) in order to find what experiments are needed to test missing or low-confidence relation. For the above example, GenePath proposed 26 experiments that would test the relation between *comC* and *lagD*, and ranked them according to an estimated laboratory cost (see supplement *GenePath: Analysis Methods*). The two highest-rated scenarios based on morphological phenotypes are presented in Figure 1h. They both introduce one new experiment (displayed in red), a knockout of *comC* in either $lagD^-$ or $lagD^+$ background. If either $comC^-lagD^-$ cells are unable to aggregate, or $comC^-lagD^+$ cells are able to form fruiting bodies, a single experiment would be sufficient to support the hypothesis that *comC* excites *lagD*. The experimentalist can change the cost and effort estimates to fit individual laboratory circumstances, thus optimizing these two critical parameters and increasing efficiency.

Interface

GenePath is a web-based application that runs on a dedicated server and is accessed through a web browser. The interface consists of a navigation menu and related parts showing information on the current project, experimental data and inferred network (Figure 1).

GenePath handles each problem as a project that consists of a list of genes, phenotypes, genetic experiments and prior knowledge. After a new project is created or an existing project is loaded, a navigation menu appears at the top of the browser window (Figure 1a). The buttons in the first row allow the user to manage data entry. The second row handles data analysis and the third row navigates between open projects. An integrated notebook can be used for additional background information on a particular project, comments about the data, intermediate results of the exploratory data analysis, or comments about the final results. GenePath maintains all of the data within a session that runs on the server. GenePath projects, including data and figures from the notebook, can be saved in an XML format on the local computer.

GenePath is implemented in Microsoft Visual Basic .NET using the ASP.NET technology. It runs on a Microsoft Windows platform with support of Internet Information Services and .NET Framework. Implementation details are provided in the supplement (*GenePath: Web Interface and Application History*).

Conclusion

GenePath can assist biologists in the systematic exploration of mutant data, in identifying and testing new relations, and in documenting and communicating genetic data. For brevity, the example in Figure 1 included only three genes, but GenePath performs just as well on much

larger data sets. Thanks to its interface and textual explanation, GenePath may also help in teaching the concepts of genetic data analysis.

Examples available on line

A number of ready-to-run examples are included on the GenePath web page, including gene network studies on *D. discoideum* (transition from growth to development, spore formation, and intercellular communication) and *C. elegans* (programmed cell death and dauer larva formation).

Availability and supplementary information

GenePath is available on line at http://www.genepath.org/genepath2. The code and the executable may be obtained under GPL license upon request. The supplementary information available at NAR Online is split into three parts and includes *GenePath: Web Interface and Application History*, *GenePath: Inference of Relations and Networks*, and *GenePath: Analysis Methods*.

Acknowledgements

This work was supported in part by a grant from the Slovene Ministry of Education, Science and Sports and by a grant from the National Institute of Child Health and Human Development, P01 HD39691.

References

1. Avery,L. and Wasserman,S. (1992) Ordering gene function: the interpretation of epistasis in regulatory hierarchies. *Trends Genet.*, **8**, 312-316.
2. Zupan,B., Demsar,J., Bratko,I., Juvan,P., Halter,J.A., Kuspa,A. and Shaulsky,G. (2003) GenePath: a system for automated construction of genetic networks from mutant data. *Bioinformatics*, **19**, 383-389.
3. Demsar,J., Zupan,B., Bratko,I., Kuspa,A., Halter,J.A., Beck,R.J., Shaulsky,G. (2001) GenePath: a computer program for genetic pathway discovery from mutant data. Proc. 10th World congress on medical informatics, London, UK. In Patel,L., Rogers,R., Haux,R. (ed.), *Medinfo 2001*, IOS Press, Amsterdam, 10, Pt. 2, pp. 956-959.
4. Zupan,B., Bratko,I., Demsar,J., Juvan,P., Curk,T., Borstnik,U., Beck,J.R., Halter,J., Kuspa,A., Shaulsky,G. (2003) GenePath: a system for inference of genetic networks and proposal of genetic experiments. *Artif. Intell. Med.*, **29**, 107-130.
5. Kibler,K., Svetz,J., Nguyen,T.L., Shaw,C. and Shaulsky,G. (2003) A cell-adhesion pathway regulates intercellular communication during Dictyostelium development. *Dev. Biol.*, **264**, 506-521.

Figure legends

Figure 1: Elements of the GenePath user interface. **(a-d)** An example of analysis of intercellular communication in *D. discoideum*, showing a window with a navigation menu, an inferred network (green edges for excitation, red for inhibition; confidence levels shown above), a table with morphological phenotypes ("+" and "–" indicate gene activation (overexpression) and inactivation (knockout), respectively) and transcriptional phenotypes ("0" indicates wild type expression, "+" and "–" indicate higher and lower than wild type expression, respectively). **(e)** A report on evidence for a positive influence of *comC* on *lagC*. **(f)** Compact representation of a genetic network with a positive feedback cycle between *lagC* and *lagD*. **(g)** A genetic network constructed from morphological data, ignoring the transcriptional data. **(h)** The two highest-rated experimental sets that would support the relation "*comC* excites *lagD*". Experiments displayed in red were proposed by GenePath to supplement those already in the experimental set (black).

Figures

Figure 1