

Towards Automated Explanation of Gene-Gene Relationships

Wacław Kuśnierczyk,¹ Astrid Lægreid,² Agnar Aamodt¹

Keywords: microarrays, gene-gene relationships, knowledge-intensive problem solving, iterative search, public databases and tools.

1 Introduction.

During the recent decades research in molecular biology experienced several paradigm shifts that changed the researchers' approach to solving particular problems in the field. The invention of microarray technology in the last decade of the previous millennium can certainly be seen as one of those paradigm shifts [3]. However, although there appear first reports showing efforts to combine various resources of genomic data instead of investigating just one source (e.g., [2, 4]), the understanding and interpretation of the results—the key issue in any attempt to a discovery—is still entirely left to the human.

Microarray data represent a reasonable source of new hypotheses on gene function and between-gene relationships. In order to fully understand and explain these hypotheses, biological background information from many resources has to be explored and combined.

We propose a novel method intended to aid a researcher in understanding hypothetical relationships between genes, e.g., genes not previously known to be related. In order to justify a tentative link between genes, the sequences, promoter regions, protein structure, function and other properties may have to be investigated for these and possibly other related genes. Although a manual search for information that would link two genes is theoretically possible, in practice it may be a very tedious task. Our approach is an attempt to design and implement a high-level wrapper for existing databases and tools, providing an automated process of forming relevant human-readable explanations. The proposed solution draws from the achievements of research in artificial intelligence—knowledge representation and knowledge-intensive reasoning, non-deductive inference mechanisms, and machine-learning [1].

2 Methods.

The proposed system is an intelligent interface between the user on one side, and remote databases and publicly available tools on the other side, enhanced by background knowledge in molecular biology. Its modular architecture is illustrated in Fig 1. A typical question that can arise from a microarray experiment and may be asked to the system is of the form *How are genes g_1 and g_2 related?* or *What might be the causal relationship between genes g_1 and g_2 ?* The exact syntax of the query depends on the actual implementation of the query interface module (QI).

The query, translated into the internal representation language, is interpreted by the core reasoner (CR), which utilizes general domain knowledge (GDK) to construct an explanation chain that links the two investigated genes. The GDK is modelled as a multi-relational semantic network, a kind of ontology, where each concept and relation are represented as a

¹Department of Information and Computer Science, Norwegian University of Science and Technology, Sem Sælandsv. 7, 7491 Trondheim, Norway. E-mail: {waku,agnar}@idi.ntnu.no

²Department of Cancer Research and Molecular Medicine, Norwegian University of Science and Technology, Olav Kyrresg. 3, 7489 Trondheim, Norway. E-mail: astrid.lagreid@medisin.ntnu.no

distinct object. The CR contains inheritance and propagation methods for reasoning over a combined set of relations (e.g. *causes*, *has-function*, *has-structure*, *has-subclass*, *has-part*, *triggers*, *inhibits*, etc.), assigning a specific ‘explanatory strength’ to each relation. Data of different types, related to the queried genes, are retrieved from databases (DB) and matched with the help of available tools (T). The connection between the system and various public resources goes through dedicated interfaces responsible for contacting a resource suitable for a particular task and in a way specified by the resource’s query interface. After the concepts have been instantiated with specific data coming from the query and from databases, the system attempts to construct an explanation by searching for one or more paths in the semantic network that would connect the two genes. The retrieval of information needed to instantiate the concepts is repeated iteratively until a pathway is found, or specific search-limiting criteria are met.

An explanation is output through the explanation module (EI), and the user may or may not accept it, thus giving feedback for the search process. The explanation may be refined at a later time, and the result may be retained in the case-base (CB) for further reference and to boost a search similar to a previously completed one. An explanation is of the form *Gene g_1 regulates gene g_x which in turn produces a protein that may interfere with the action of gene g_2* , though typically it would be much more complex.

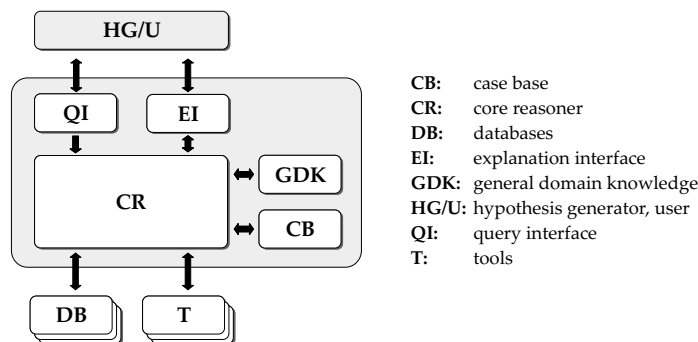


Figure 1: The architecture of an automated system for explanation of gene-gene relationships.

Note on the implementation The system is currently in the design phase. The specific modules will be implemented concurrently and a functional version of the system is planned to be released by the end of year 2005.

References

- [1] Aamodt, A. 1991. *A knowledge intensive, integrated approach to problem solving and sustained learning*. PhD thesis, Norwegian University of Science and Technology.
- [2] Bar-Joseph, Z. et al. 2003. Computational discovery of gene modules and regulatory networks. *Nature Biotechnology* 21(11):1337–1342.
- [3] Brown, P.O. and Botstein, D. 1999. Exploring the new world of the genome with DNA microarrays. *Nature Genetics* 21:33–37.
- [4] Köhler, J. and Schulze-Kremer, S. 2002. The Semantic Metadatabase (SEMEDA): Ontology based integration of federated molecular biological data sources. *In Silico Biology* 2.