

# Knowledge-Intensive Case-Based Support for Automated Explanation of Biological Phenomena

Wacław Kuśnierczyk<sup>1</sup>, Agnar Aamodt<sup>1</sup>, and Astrid Lægreid<sup>2</sup>

<sup>1</sup> Department of Information and Computer Science,  
Norwegian University of Science and Technology (NTNU),  
Sem Sælands v. 7–9, N-7491 Trondheim, Norway  
{waku, agnar}@idi.ntnu.no

<sup>2</sup> Department of Cancer Research and Molecular Medicine,  
Norwegian University of Science and Technology (NTNU),  
Olav Kyrres gt. 3, N-7489 Trondheim, Norway  
astrid.lagreid@ntnu.no

**Abstract.** The rapid growth of data stored in molecular biology-related databases has stimulated the development of integrative tools for retrieval and presentation of the data in the form of, e.g., biological association networks. We argue that a general and specific knowledge-based approach may provide substantial support for automated reconstruction of networks which otherwise tend to be large and, eventually, unreadable. This knowledge-based approach introduces a novel strategy with the potential to greatly enhance the explanatory power of automatically generated biological association networks. We discuss the motivation for a study of the process an expert employs while building a network, and suggest that a series of expert sessions be used as a case library for future reference. An example of a biological problem and the shape of its solution is described, and the types of knowledge involved are discussed.

## 1 Introduction

During the recent years research in molecular biology and bioinformatics has experienced several paradigm shifts that have changed the researchers' approach to investigating particular problems within these fields. Rapid, both quantitative and qualitative, growth of data stored in molecular biology-related databases — as for January 2005, there are 719 such publicly available resources [5] — is partly the result of application of new advanced high-throughput technologies such as gene expression microarrays. In turn, it has stimulated parallel development of a number of methods and tools for improved search and retrieval, analysis and presentation of the stored data.

However, automated production of massive amounts of data has been far easier than integrated and automated analysis of them. The analysis is complicated by the multitude of data types and the concepts they represent, and by

the multitude of quality measures, database schemas, interfaces and communication protocols etc. Bioinformatics has recently merged with other disciplines to the highly integrative field of Systems Biology [12, 17] where one of the major challenges is to unify or combine various methods (tools) and various sources of data (databases) in order to provide a basis for models that enable understanding biological systems.

Much effort has recently been invested into building systems that would integrate diverse biological data [16]; databases that store data of different types and web interfaces presenting heterogeneous search results from such databases (or from multiple distributed resources) have been around for years (e.g., ENSEMBL [3], GENECARDS [6]).

Biological association networks (BANs) are examples of how pieces of information can be organized and presented as a connected structure so that they provide explanations for biological phenomena in which there are numerous interactions and dependencies (the network's edges) between entities such as genes and their products, biological molecules etc (the network's nodes). There exist tools that facilitate building such networks (e.g., [9, 10, 15, 18]); these tools provide automated or semiautomated support on the level of data retrieval, matching, filtering, and presentation. However, although simple networks that include relatively few nodes and interactions may be meaningful to the researcher, larger networks built in this way are usually much more complex and eventually become unreadable as they grow, and thus require further processing.

As the Nobel Prize laureate Sydney Brenner says, "*The great challenge in biological research today is how to turn data into knowledge. I have met people who think data is knowledge but these people are then striving for a means of turning knowledge into understanding.*" [4] We observe that biologists employ a heuristic, context-dependent reasoning to extract meaning from large, complex networks retrieved by a network building tool (NBT). Such reasoning is not supported by the existing tools; it is thus entirely left to the human and forms an implicit part of the network building process.

We argue that a knowledge-based approach including both general (domain model) and specific (episodic experience, cases) knowledge may provide substantial support for the reasoning part of building a BAN. Employing such an approach would presumably result in networks that both are more readable and have higher explanatory power in a particular context, and at the same time make the process feasible for larger sets of nodes. In our work we focus on how a molecular biologist constructs a network to explain the relationships between a collection of genes. It is important to stress that it is the process of building a network (*how* it is being constructed) and not the network itself (*what* is being constructed) that is discussed here. We do not propose novel explanations for particular biological phenomena; we propose a novel approach to automated construction of such explanations from existing (experimentally confirmed, curated by experts) pieces of information. In this paper, our goal is to motivate a new line of research at the crossroads of bioinformatics and artificial intelligence

(AI); we do not present any testable implementation of the idea — it would necessarily be overly simplistic and inaccurate at this stage.

The rest of the article is organized as follows. In Section 2 we present the goal and motivation of our study. Next, in Section 3, we examine the problem of explaining the role genes play in biological phenomena, and in Section 4 we discuss the process a biologist employs to derive a solution and justify the need for a knowledge-intensive case-based approach to automatize this process. Finally, Section 5 wraps up the conclusions and presents directions for future work.

## 2 The Goal and Motivation

Our goal is to automate the process of finding explanations for molecular level biological phenomena. To get an insight into how and to what degree one could achieve this, we study how experts in molecular biology use the available tools and resources to build such explanations, how this process is guided by their knowledge of the domain, and how this knowledge can be used to automate the process.

We focus on how biologists state their problems, what are the solutions they produce, and how the solutions are derived. Most psychologists and epistemologists agree that people, when solving problems, combine generalized knowledge (e.g. causal models, heuristic rules) with episodic knowledge (past episodes, experienced cases). We study how biological expert knowledge can be collected in a non-generalized form, i.e. as cases, and reused in combination with existing models of general domain knowledge, in a way that provides an efficient assistant in the expert’s explanatory process.

Each problem-search-solution<sup>3</sup> session with an expert is an experience that (partially) explains how the expert reasons and thus contributes to an (implicit) ‘model’ of the expert’s reasoning. A history of such problem solving sessions can be stored as past cases that, whenever a new problem has to be solved, serve for purposes of analogical reasoning,<sup>4</sup> with possible reuse of past solutions or inference patterns that led to them. Nonetheless we need to stress that this case-based way of reasoning<sup>5</sup> is chosen because it is difficult to model the expert’s performance otherwise, and *not* to mimic the expert’s reasoning, which is likely to involve a much more complex integration of reasoning processes..

To our best knowledge, there have been no attempts yet to automatize this part (i.e., reasoning) of the explanatory process. So far, most post-genomic research related to genetic networks has been focused on the discovery of new (previously unknown) relations between genes, e.g., Boolean, Bayesian and other

---

<sup>3</sup> This expression does not imply that the workflow is linear; it is, in fact, iterative and more complex.

<sup>4</sup> Understood as in e.g., [19], not as a reasoning by *transdomain* matching.

<sup>5</sup> We consider the terms *analogical reasoning* and *case-based reasoning* synonymous in this context, though in general they are not.

statistical approaches to the analysis of microarray data, application of a number of machine learning techniques. Various tools facilitate retrieval, text-mining, filtering, matching and presentation of biological information; none of them can, however, produce an extensive explanation of the function or interaction of genes without explicit manual control and processing by an expert.<sup>6</sup> If we think of the available information represented by a (immense) graph<sup>7</sup> with nodes corresponding to pieces of data and edges corresponding to relations between these, an explanation of a particular biological phenomenon is a subgraph of the whole graph; to find this explanation one needs to search the space of all possible subgraphs.

The existing NBTs, e.g., PATHWAYASSIST [15], realize this search (implicitly) by employing fixed constraints on how a subgraph can be build from an initial set of nodes, rather than using a specification of the desired solution. This means that the user has to explicitly set the conditions for extending a network, such as the types of nodes and relations to be used, the maximal length of an indirect connection between two nodes etc. each time an extension is desired. Successful production of such networks involves a substantial amount of contextual, implicit knowledge and mindful reasoning. Although the process of locating, downloading, matching and composing pieces of information into a network, based on user-definable parameters, is now well supported in NBTs, the decision on how to reduce or extend the network, and how to estimate the explanatory power<sup>8</sup> of the already built network, cannot be made by the machine and the decisive reasoning lies entirely on the user's side.

Manual construction of explanations for biological phenomena, although possible, is often a very tedious task. It involves searching or browsing vast resources containing various types of data, retrieving, evaluating and extracting information from free text scientific publications.<sup>9</sup> What makes the process slow is not, however, the machine-side part of the retrieval of information from available resources, which is fairly quick with current hardware, database management systems and advanced query engines; the slow-down is due to the human-side reasoning that has to be performed in order to evaluate and filter correct, relevant and applicable information from the search results, to link the chosen pieces of information, and to make decisions on further queries.

---

<sup>6</sup> Except for the (trivial) case when a precompiled metabolic, signal transduction or other pathway can be retrieved from a specialized database, and is both relevant and satisfactory as an explanation.

<sup>7</sup> A graph-like representation is appealing for us here for several reasons: it is more readable for a user inexperienced in other representation formalisms, and it is the backbone of the CBR system CREEK we intend to use.

<sup>8</sup> By this we mean the *plausibility* of the network in the particular context, not its biological *validity*, since the graph which all such networks are subgraphs of is assumed to be biologically valid.

<sup>9</sup> In this article we somewhat blurr the difference between data and information, as it is of no major importance to distinguish them here; it should, however, be kept in mind that in general the words *data*, *information* and *knowledge* refer to related, though not equivalent concepts.

The primary function of NBTs is to disengage the user from manually searching multiple databases and to prevent her from ‘drowning in a sea of data’ [17]. Yet if, due to the continuously growing amount of publicly available data, paralleled by increasing information-retrieval capabilities of tools supporting genome-wide (and, so to say, available data-wide) studies, the recall of information relevant to a query increases substantially and the retrieved networks tend to be unreasonably large, then in fact the result is to the user hardly more than just another form of data that has to be manually inspected with considerable effort and in considerable time. To keep it manageable, manual processing has to be (at least partially) automatized.

### 3 The Biological Problem

The class of problems in the domain of molecular biology that we consider here is explaining functions of genes in various biological events. A biologist wishes to understand the role a gene plays in a physiological or pathological state; the relationships between this and other genes in a disease; or the interactions between gene products (proteins) and other biological entities such as signalling molecules, transcription factors etc.

A problem is thus defined by a set of genes, questions concerning their roles and relationships, and contextual information used to guide and constrain the search and explanation process. The problem is solved in a human-machine interaction setting; unfortunately, the context is mostly implicit. It is used by the expert in her (external to the machine) reasoning process, and only occasionally it fragmentarily enters the query as additional constraints.

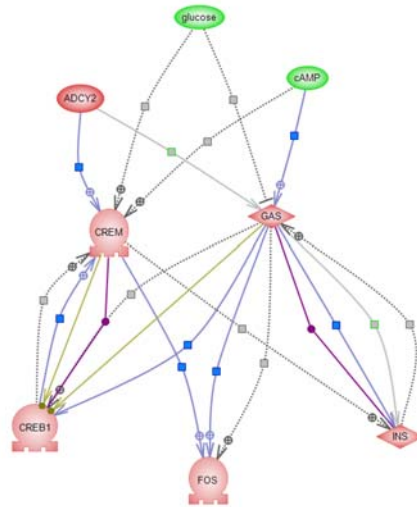
Nevertheless, it is the information included in the context that actually allows the biologist to produce a meaningful explanation<sup>10</sup> from the retrieved data. Without the context the search result, although related in one or another way to the query, may keep the expected explanation hidden among uninteresting, though correct, pieces of information. By making the context explicit and machine-understandable, we can assign a part of the reasoning task to the machine, and thus acquire the desired speedup.

The idea of an explanation is illustrated in Fig. 1: it presents a (simplified) network-like structure built to explain some of the (much more complex in reality) relationships between the genes GAS (gastrin) and CREM (cAMP responsive element modulator); the initial problem is the role of CREM in gastrin-induced gastric cancer.

An explanation in this sense includes only the model in its final form, which reflects the gained understanding or the message to be communicated; it is, in a sense, a solution to the underlying problem without any explanation or justification of how the solution was reached. This interpretation of an explanation should not be confused with that of a reference to the *process of deriving* an

---

<sup>10</sup> It is important to note that an explanation is meaningful within a particular context, and may be completely meaningless in another situation, despite the same biological validity in both cases.



**Fig. 1.** A simplified explanation of relationships between the gastric hormone gastrin (GAS) and the transcription factor cAMP-responsive element modulator (CREM). Relations involving the enzyme adenylyl cyclase (ADCY2), the second messenger cAMP as well as the nutrient glucose are shown together with relations to the transcription factors cAMP responsive element binding protein (CREB1), c-Fos (FOS) and to the hormone insulin (INS). Different node shapes and line types denote different classes of biological entities (hormones, transcription factors, small molecules) and their relationships (induction, inhibition, signalling), respectively. Labels denote names of the involved molecules.

explanatory model, used further in this text and closer to the understanding of this term in AI. To distinguish these distinct meanings, in the following we refer to explanations in the sense of an expert’s final model of a phenomenon as *solutions*, and the term *explanation* is used to refer to an explanatory process, whether an explication of the expert’s reasoning to the machine, a justification of the machine’s performance to the expert, or the internal rationalization of the machine’s inference to itself. To be able to reuse past experience in building new solutions, we need to record the details of the reasoning process as well.

## 4 The Explanation Process

In an attempt to find a solution to a question, a biologist starts with specifying the initial information — the question itself plus some explicit context. The requirements for the final solution and the constraints on the process of deriving it are usually implicit and become more clear during the process; the problem is thus initially heavily underconstrained, has a large number of solutions that are

‘correct’ with respect to the domain model, and can be recognized as plausible or not only by being criticized by the expert.

#### 4.1 Stepwise Construction of a Solution

The initial model is iteratively modified until it meets the (explicit only at that event) expectations. A detailed study of the tasks, methods and processes involved in solving the problem is ongoing work in our group; here we restrain the description to the most important elements.

At each step of the process the expert analyzes the current state of the model of a tentative solution and performs the following actions:

- *inspects* the model to assess whether it can be accepted as a final solution to the problem; the criteria may include the size of the model, coverage of particular entities or entities of a particular type etc.;
- *extends* the model by adding new nodes or relationships; the nodes may represent properties of the already included genes, or new genes, and the relationships may link already included nodes, or may require new nodes to be added;
- *constrains* the model by removing some of the nodes or relationships.

If the inspection of the model leads to its rejection as a final solution, the model is modified by employing actions of the other two types.<sup>11</sup> An extension is achieved by retrieval, evaluation and filtering of information related to selected nodes in the model. The retrieved information may be incorrect (due to misinterpretation of free text by a text-mining tool, an error in a database etc.), it may be correct but irrelevant (not related to the problem), it may be relevant but of little importance; since detection of such situations often requires extensive knowledge and intense reasoning, they may be noticed only at a later time. Thus nodes previously accepted may have to be removed afterwards; nodes may be removed also when the the model grows too large and there are other nodes of superior relevance; the process of building a solution is not necessarily monotonic.

#### 4.2 Supporting Knowledge

The knowledge an expert employs and that we have to consider comes from several related and partially overlapping domains that together form what we call *expertise*, in this case expertise in explaining biological phenomena: the domain of molecular biology (biomedicine), that of managing and analyzing biological

---

<sup>11</sup> Obviously, it is also possible that a model fully explaining the problem cannot be produced; in fact, such a situation may be anticipated and even desirable if the researcher aims at finding ‘holes’ in the existing knowledge that might be filled by her wet-lab experiments.

information (bioinformatics), and that of a biologist's reasoning processes (that is, an operational 'model' of the researcher).<sup>12</sup>

The first covers an extensively, though not completely, understood domain; this knowledge can be relatively easily acquired and, in fact, has already been partially modelled in, e.g., the Gene Ontology (GO, [2]) and the Kyoto Encyclopedia of Genes and Genomes (KEGG, [11]) projects.

The domain of bioinformatics is fairly well understood (since it describes man-made artifacts such as databases, tools and algorithms), but has not been well documented in a comprehensive model so far. Examples of models from this domain are the Transparent Access to Multiple Bioinformatics Information Sources (TAMBIS, [7]) and the Semantic Metadatabase (SEMEDA, [14]) projects.

Human reasoning has been extensively studied within the field of psychology of problem solving. There are many models of how humans solve particular classes of problems, and some of them served as inspiration for machine problem solving, e.g., Model-Based Reasoning (MBR), Rule-Based Reasoning (RBR) and Case-Based Reasoning (CBR). These models are, however, too general to be easily instantiated and applied to our problem; they include many details irrelevant to our domain, and do not explain well the particular situation of answering questions in molecular biology. Moreover, the same question may be answered dissimilarly by two biologists, exposing their different background, interest and other sorts of bias. Finally, the knowledge of how to build an explanation is of the operational type rather than of the declarative. Therefore, although in the two former cases we intend to partly build new and partly reuse and merge existing domain models (a non-trivial task, anyway), in the last case the choice is to store and reuse particular experiences for future reference rather than generalize an *a priori* model of user preferences — that is, use CBR. The knowledge content will come from recordings of expert problem solving sessions in our research group.

### 4.3 Representation and Reasoning

At the symbol level the knowledge will be represented within the frame-based, semantic network-like formalism CREEKL [1]. The great benefit of this approach is that both general domain (model) and specific episodic (cases) knowledge can be represented within the same, consistent language; in fact, CREEKL is in principle capable of representing virtually any kind of knowledge — deep and shallow, general and specific, declarative and operational. Other advantages and disadvantages of CREEKL, both with respect to this task and more generally, will be discussed elsewhere.

CREEKL as a knowledge representation language is a part of a complete framework for knowledge intensive, integrated problem solving and sustained learning, partially implemented in the system CREEK (mostly the CBR part of

---

<sup>12</sup> Other, more loosely related, domains may possibly be employed in cross-domain analogical reasoning; these are not discussed in this text.

the framework). Our system is intended as an instantiation of this theoretical framework.

A substantial effort has to be put into investigation of how the existing implementation must be extended to meet the specific requirements of the domain. In particular, the case base has to be (re)designed so that a case records a problem solving session as fully as possible, but at the same time allows efficient storage, indexing, retrieval and reuse. A case must contain the initial definition of the problem, the solution, and a trace of its derivation. A stratified, or nested, structure is desirable; each step within an explanation process is actually a (sub)case with its own problem description (the initial state of the model), its solution (the actions performed) and the outcome (the new state of the model after modifications). The problem of splitting cases into reusable subparts (called *snippets*) has already been recognized and addressed in a number of studies (e.g., [13, 8]).

## 5 Summary and Future Work

In this paper we discussed the need for a knowledge intensive support for an automated system that would provide answers to biological questions such as the role of genes in a particular disease. We anticipate that in near future such ‘smart’ tools will be indispensable to enable a molecular biologist to extract meaning from the ever growing mass of publicly available data. To achieve our goals, we need to further specify the methods and processes the expert employs in her reasoning; an implementation of these will instantiate the general model of knowledge intensive problem solving developed in our group.

In the next step we will implement a simple, non-sophisticated tool for building biological association networks that will enable us to automatically construct a library of network-building cases. These cases will be then used to experiment with providing various levels of support for the expert: from simple case retrieval, through semi-automated suggestion of subsequent solution modifications to fully automated construction of final solutions.

## Acknowledgments

This work has been financed from a graduate study grant provided by the Norwegian University of Science and Technology (NTNU). The figures in this article are loosely inspired by a discussion with Ola Ween researching at NTNU the role the gene CREM plays in gastric cancer; they were produced with the PATHWAYASSIST software powered by Ariadne Genomics, Inc., distributed by Stratagene, Inc, designed to support building pathways, exchanging data, accessing public repositories and extracting facts from publication abstracts. The authors thank anonymous reviewers for their comments on an earlier version of the article.

## References

1. Aamodt, A., (1991). A Knowledge Intensive, Integrated Approach to Problem Solving and Sustained Learning. Norwegian University of Science and Technology, Dept. of Information and Computer Science, PhD Dissertation.
2. Ashburner, M. et al, (2000). Gene Ontology: tool for the unification of biology. *Nature Genetics*, **25**:25–29.
3. Birney, E. et al. (2004). An Overview of Ensembl. *Genome Research*, **14**: 925–928.
4. Brenner, S. (2002). Ontology Recapitulates Philology. *The Scientist*, **16**(6):12.
5. Galperin, M. Y. (2005). The Molecular Biology Database Collection: 2005 Update. *Nucleic Acids Research*, **33**:D5–D24.
6. Gene Cards, [www.genecards.com](http://www.genecards.com)
7. Goble, C. A., (2001). Transparent access to multiple bioinformatics information sources. *IBM Systems Journal*, **40**(2).
8. Grimnes, M. J. F. (1998). ImageCreek: A knowledge level approach to case-based image representation. Norwegian University of Science and Technology, Dept. of Information and Computer Science, PhD Dissertation.
9. Hoffmann, R., Valencia, A. (2004). A gene network for navigating the literature. *Nature Genetics*, **36**(7):664.
10. Jenssen, T.K., Lægreid, A., Komorowski, J., and Hovig, E. (2001). A literature network of human genes for high-throughput analysis of gene expression. *Nature Genetics*, **28**(1):21-8.
11. Kanehisa M., Goto S., Kawashima S., Okuno Y., and Hattori M., (2004). The KEGG resource for deciphering the genome. *Nucleic Acids Res.*, **1**.
12. Kitano, H. (2002). Systems biology: a brief overview. *Science*, **295**(5560):1662–4.
13. Kolodner, J. L., and Simpson, R. L. (1989). The MEDIATOR: Analysis of an early case-based problem solver. *Cognitive Science* **13**(4): 507–549.
14. Köhler, J., Lange, M., Hofestädt, R., and Schulze-Kremer, S., (2000). Logical and Semantic Database Integration. *Proc. of the IEEE Symposium Bioinformatics and Biomedical Engineering*, ed. Young, D. C.
15. Nikitin, A., Egorov, S., Daraselia, N., and Mazo I., (2003). Pathway studio—the analysis and navigation of molecular networks. *Bioinformatics*, **19**.
16. Palsson, B. (2000). The challenges of in silico biology. *Nature Biotechnology*, **18**.
17. Roos, D. S. (2001). Bioinformatics—trying to swim in a sea of data. *Science*, **291**(5507):1260–1261.
18. Sohler, F., Hanisch, D., and Zimmer, R. (2004). New methods for joint analysis of biological networks and expression data. *Bioinformatics*, **20**(10):1517–21.
19. Sowa, J. F., and Majumdar, A. K. (2003). Analogical Reasoning. *Conceptual Structures for Knowledge Creation and Communication*, LNAI 2746, eds. Aldo, A., Lex, W., and Ganter, B.