

# Towards Automated Explanation of Gene-Gene Relationships

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

<sup>1</sup>Department of Information and Computer Science and <sup>2</sup>Department of Cancer Research and Molecular Medicine, Norwegian University of Science and Technology, Trondheim, Norway

## The Goal

An automated system that will assist a biologist in explaining actual and hypothetical relationships between sets of genes using background knowledge and various types and sources of data

## Overview

Although there are reports showing attempts to combine different sources of genomic data (e.g., [2,3]), the interpretation of the retrieved results is entirely left to the human. We propose a novel method that, given a query about a relationship between some genes of interest, would use background knowledge, past experience and available data to build a detailed explanation of the relationship, if possible.

## Methods

An implementation of the system will be based on the architecture of Creek [1], a system for knowledge intensive reasoning and sustained learning.

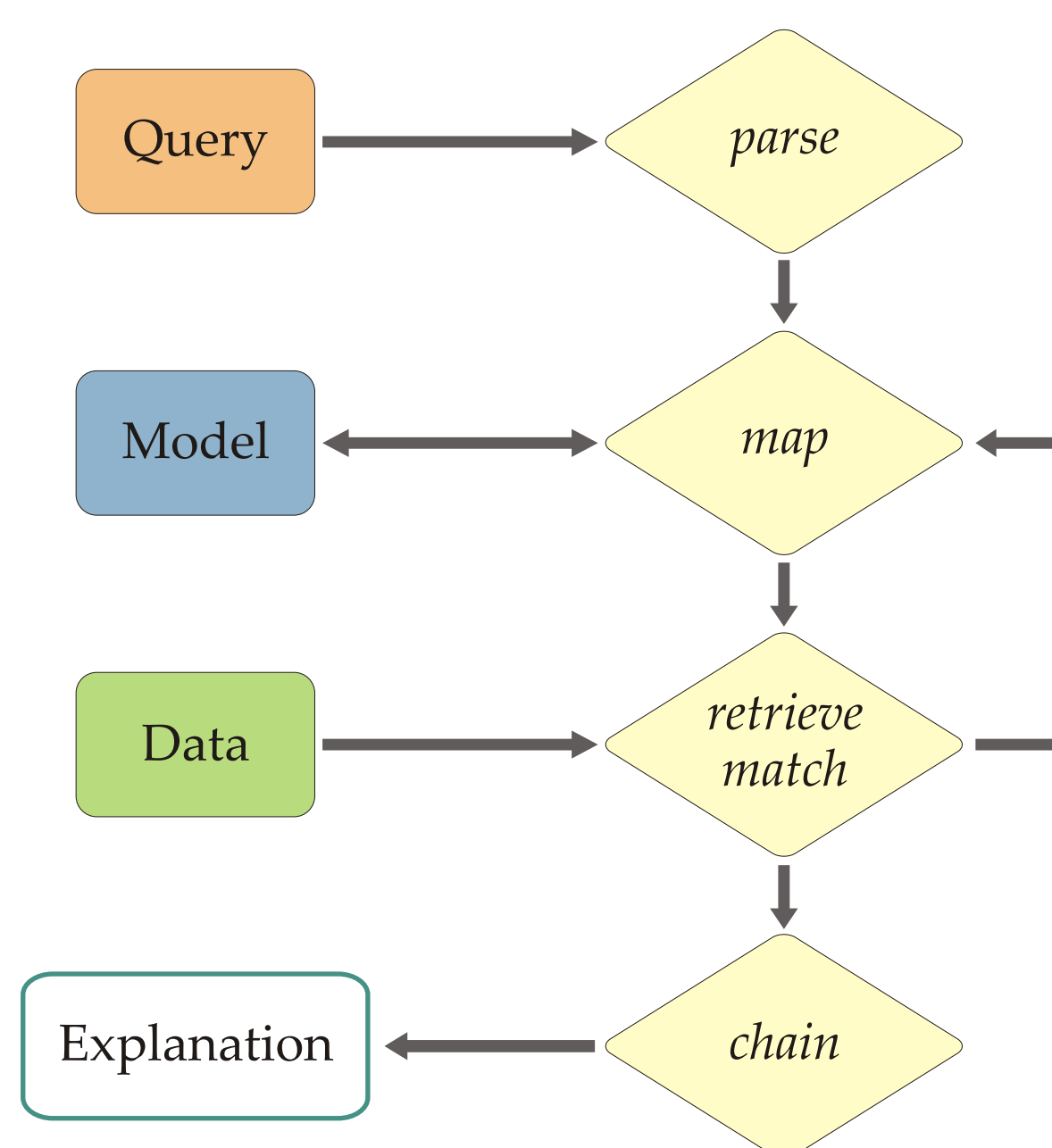


Fig. 1. A schematic representation of the iterative process of answering a query (forming an explanation).

1. A query is parsed and represented in an internal language, and the objects within the query are mapped onto the conceptual structures of the background knowledge, and matched to past cases;
2. The knowledge is modelled as a semantic network, where concepts are linked by relationships to other concepts and relationships;
3. The data necessary in an iteration is retrieved from a particular database via a dedicated interface (a centralized datawarehouse-based solution is also considered), matched using various tools, and mapped onto the knowledge structures;
4. The elements of an explanation path are evaluated with respect to an explanatory strength measure, selected and chained, and an explanation is produced.

## A (Trivial) Example

What is the similarity between gastrin (GAS) and cholecystokinin (CCK) in the context of function?

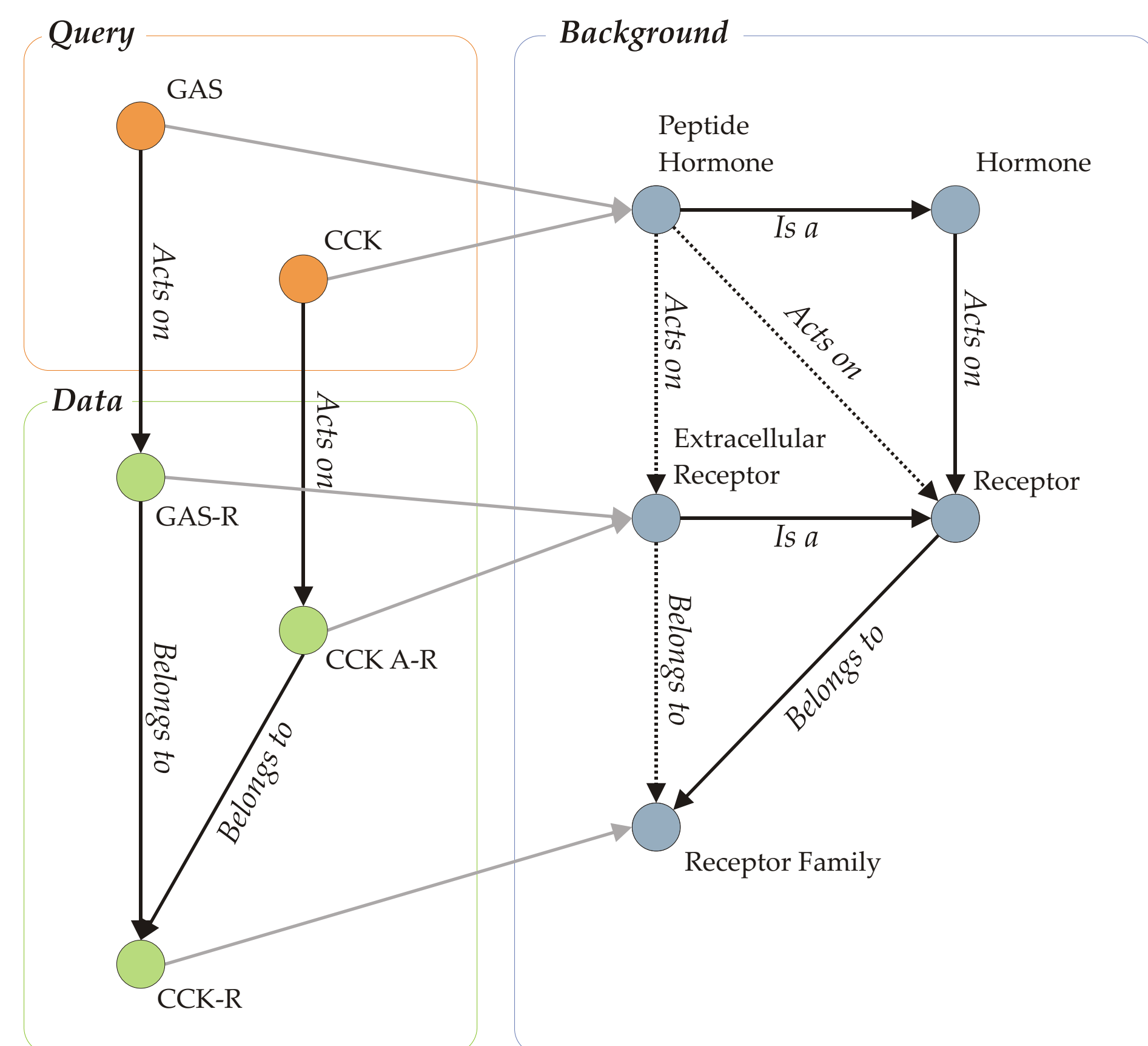


Fig. 2 A simplified network illustrating one possible inference pattern for the GAS-CCK problem. Gray arrows denote mapping of case-specific data onto the domain knowledge model. Dotted arrows denote inferred relationships.

Both GAS and CCK are peptide hormones acting on their respective extracellular receptors GAS-R and CCKA-R, and both receptors belong to the same receptor family CCK-R.

The system will be used to answer more complex queries related to our studies (e.g., [4]).

## Future Work

- design/adapt/modify ontologies (top and domain knowledge model levels), and implement within the Creek framework
- design/adapt inference methods, and implement within the Creek framework
- implement methods for learning from past cases, to boost solving of queries similar to those already solved
- design interfaces to publicly available databases and tools, and implement as plugins to Creek

## Reference

1. A. Aamodt, 1991. *A Knowledge Intensive, Integrated Approach to Problem Solving and Sustained Learning*. PhD Thesis, Norwegian University of Science and Technology.
2. Z. Bar-Joseph et al., 2003. Computational Discovery of Gene Modules and Regulatory Networks. *Nature Biotechnology* 21(11):1337-1342.
3. J. Köhler and S. Schulze-Kremer, 2002. The Semantic Metadatabase (SEMEDA): Ontology Based Integration of Federated Molecular Biology Data Sources. *In Silico Biology* 2.
4. F. Yadetie et al., 2003. Liver Gene Expression in Rats in Response to the Peroxisome Proliferator-Activated Receptor- Agonist Ciprofibrate. *Physiological Genomics*, 15:9-19.

## Contact information

Wacław Kuśnierczyk, Dept. of Information and Computer Science, NTNU, Sem Sælands vei 7-9, 7491 Trondheim, Norway. Email: waku@idi.ntnu.no, phone: +47 73591875, fax: +47 73594466