A Multi-Agent System-driven AI Planning Approach to Biological Pathway Discovery

Salim Khan and **Keith Decker**
Computer and Information Sciences
University of Delaware
Newark, DE 19716, USA.
{skhan,decker}@cis.udel.edu

William Gillis and **Carl Schmidt**
Animal and Food Sciences
University of Delaware
Newark, DE 19716, USA.
{kerouac,schmidtc}@udel.edu

Abstract

As genomic and proteomic data is collected from high-throughput methods on a daily basis, subcellular components are identified and their *in vitro* behavior is characterized. However, much less is known of their *in vivo* activity because of the complex subcellular milieu they operate within. A component's milieu is determined by the biological pathways it participates in, and hence, the mechanisms by which it is regulated. We believe AI planning technology provides a modeling formalism for the task of biological pathway discovery, such that hypothetical pathways can be generated, queried and qualitatively simulated. The task of signal transduction pathway discovery is re-cast as a planning problem, one in which the initial and final states are known and cellular processes captured as abstract operators that modify the cellular environment. Thus, a valid plan that transforms the initial state into a goal state is a hypothetical pathway that prescribes the order of signaling events that must occur to effect the goal state. The planner is driven by data that is stored within a knowledge base and retrieved from heterogeneous sources (including gene expression, protein-protein interaction and literature mining) by a multi-agent information gathering system. We demonstrate the combined technology by translating the well-known EGF pathway into the planning formalism and deploying the Fast-Forward planner to reconstruct the pathway directly from the knowledge base.

Keywords

Applications of planning and scheduling; Planning and scheduling with complex domain models

Introduction

The advent of high-throughput methods have revolutionized the field of genomics, creating an information explosion in their wake. The daily accretion of experimental data from sequencing projects, gene array experiments and other high-volume data pipelines, has prompted calls (P.D.Karp 2001) to encode scientific theories into symbolic form so that inference engines may be employed to generate promising hypotheses and flag anomalous results. In this paper, we demonstrate that AI planning technology, built atop a multi-agent system (MAS) driven knowledge base (KB) platform, provides just such a formalism when applied to the task of biological pathway discovery. We describe our experiences in applying planning techniques to generate hypothetical pathways that can be tested, queried and qualitatively simulated. While our focus is on the generation of signal transduction pathways, the discussion is relevant to the discovery of other types of biological pathways, such as those of metabolism and gene regulation.

Living systems exhibit great robustness and versatility in adjusting their intracellular molecular machinery to changes in the external environment. The cellular processes by which cells detect, convert and internally transmit information regarding the external environment are collectively referred to as signal transduction (ST) pathways. The study of ST pathways is vital to our understanding of many diseases, including cancer, diabetes and neural disorders, as they commonly result from the malfunctioning of signaling components. While the malfunction of a single entity might be tolerated, the combined effect of multiple components malfunctioning can be substantial (Weng, Bhalla, & Iyengar 1999). For these reasons, it is important that we study subcellular molecular interactions in the context of the signaling pathways that they participate in.

Though individual ST components have been identified, and their *in vitro* behavior characterized, ST pathways have proven difficult to study *in vivo* due to their inherent complexity. Much like the complex systems studied in the mathematical and physical sciences (Weng, Bhalla, & Iyengar 1999), the complexity of ST pathways arises not only in the number and connectedness of its components, but also from the diversity of the message forms and the translating interfaces required. The complexity is exacerbated by the degree of conditional branching, nesting and looping present within these pathways. The pathway components also exhibit physical properties of dynamic assembly, translocation between intracellular compartments, and eventually, degradation. Finally, ST pathways exhibit a greater degree of inter-pathway crosstalk as compared to other biological pathways.

Before we map the ST pathways domain into the planning framework, we first provide an overview of the domain and list other computational approaches to the ST pathway discovery problem. To guide the comparison between these

approaches and ours, we list criteria deemed necessary for a successful model of the ST pathways domain.

Signal Transduction Pathways Domain

A signal transduction pathway can be decomposed into the following steps (Lodish *et al.* 1999):

1. A signaling molecule arrives from outside the cell.
2. A receptor on the surface of the cell interacts with the signaling molecule.
3. The receptor interacts with intracellular pathway components, setting off a cascade of protein interactions within the cell.
4. The signal arrives at its destination and elicits a functional response, *e.g.*, gene transcription.

A signaling molecule may take many forms — proteins, steroids, peptides, etc. The signal can originate from different sources —- nearby cells (*paracrine signaling)*, hormones released into the bloodstream (*endocrine signaling*) and from the receiving cell itself (*autocrine signaling*). Regardless of the signal origin, we limit ourselves to only the intracellular portion of ST pathways.

There exists a finite set of mechanisms by which a signal is transferred within a cell, all of which can be described using basic chemical reaction schemas (Voet & Voet 1995). In this paper, we consider reactions of the following three types:

1. *Complex formation*. Two bio-molecules with sufficient kinetic energy coalesce to from a complex.
$$A + B \rightarrow AB$$
2. *Complex decomposition*. An unstable complex may decompose into its constituent bio-molecules.
$$AB \rightarrow A + B$$
3. *New product formation*. If an unstable complex can overcome the activation barrier of the reaction, then a new product P results, perhaps accompanied with the creation of some byproducts.
$$AB \rightarrow P$$

While a reaction may be described overall as

$$A \rightarrow P$$

it may actually proceeds through intermediate steps

$$A \rightarrow I_1 \rightarrow I_2 \rightarrow P$$

that can be mapped to one of the three reaction types described above.

A fundamental tenet of biology is that the shape of a biomolecule imputes its function. The primary mode of signal transduction is through structural changes of the participating pathway components, which are mostly proteins. Proteins can be regarded as long chains of amino acids, coiled into a distinct conformation, *i.e.*, 3-dimensional shape.

A protein can be decomposed into *domains*, *motifs* and *active sites*, which can roughly be generalized as different structural units on the protein's surface that are associated with some functional activity. We collectively refer to these structures as *docking sites*, because their peg-to-hole conformation allows proteins to "dock" with one another, forming protein complexes. A protein can affect the conformation of one or several other proteins by binding with those proteins to form a transient complex, thereby *activating* or *inhibiting* the activities of those proteins.

A conformational change can also be effected by *post-translational modifications*. Post-translational modifications refer to the addition of chemical groups to the amino acid side chains and/or the terminal amino and carboxyl groups of a protein. Common post-translational modifications include phosphorylation, ubiquitination and acetylation. Post-translational phosphorylation, a key modification involved in signal transduction within mammalian cells, is the addition of one or more phosphate groups at specific locations (*i.e.*, amino acids) on a protein. A *kinase* protein catalyzes the phosphorylation reaction by which adenosine triphospate (ATP) transfers a phosphate group to a protein and is converted to adenosine di-phosphate (ADP) in the process. Phosphorylation occurs as a two-step enzymatic process

$$E + S + ATP \rightleftharpoons ES \rightarrow E + S^* + ADP$$

where E is the kinase (enzyme) whose target substrate S is phosphorylated (indicated by S^*) by first forming the intermediate enzyme-substrate complex. In general, ample ATP is assumed to exist within the cell, and is dropped altogether from the above schema.

A *phosphatase* protein catalyzes the inverse reaction of dephosphorylation, *i.e.*, the removal of a phosphate group

$$E + S^* \rightleftharpoons ES^* \rightarrow E + S$$

where E is the phosphatase (enzyme) which dephosphorylates the activated substrate S^*.

Signal transduction is heavily dependent on the action of kinases and phosphatases, as most of the intracellular portion of signaling pathways are cascades of protein phosphorylations and dephosphorylations. Each step leads to the activation or inhibition of events further downstream, or feeds back on upstream events.

While the processes described so far are reversible (*e.g.*, phosphorylation and dephosphorylation), signal transduction can also occur via irreversible processes such as *proteolytic cleavage*. A proteolytic cleavage is a reaction catalyzed by *proteases* that remove amino acid residues from a protein, to reveal an active form of the protein.

There are many other types of signaling events that we do not describe here. Most, however, can be viewed as variations of the reactions described herein.

Criteria for Successful Qualitative Modeling

Many computer science techniques have been applied to the task of biological pathway discovery. In this paper, we limit ourselves to qualitative models that capture the causal relationships that propagate the signal within a pathway. Qualitative models can be compared along the following dimensions (Peleg, Yeh, & Altman 2002; Regev, Silverman, & Shapiro 2000):

- *Representation of static-structural, dynamic and functional views*. A static-structural view is a representation of the many molecular species, such as protein complexes, biopolymers and chemicals, that participate within ST pathways. A dynamic model is one that captures the temporal ordering (control flow) of processes within a pathway. In particular, the model should support sequential, parallel, conditional and iterative processes. A functional view of a signaling system describes the functionality of every component within that system.

- *Hierarchical representation of biological processes*. A qualitative model that can represent high-level physiological processes and connect them to molecular-level functions is preferred, for four reasons: firstly, the representation of high-level biological processes in the context of low-level molecular interactions provides a context to evaluate the effect of perturbing specific interactions along a ST pathway. Secondly, a hierarchical visualization is easier to comprehend and study from a human perspective. Thirdly, a hierarchical representation is more likely to scale with an increase in the problem size. Finally, the data collected is inherently heterogeneous as a result of components within the same pathway being studied to different levels of biochemical detail.

- *Verification of system properties and behavior*. A qualitative formalism must allow for the checking of the model in order to corroborate the results generated. Preferably, the formalism can be mathematically or logically verified so that the system properties are guaranteed not to be violated.

- *Capable of inference*. A useful model allows for the qualitative simulation of the ST pathway. The reasoning mechanism must be able to perform, for instance, the reachability analysis of a certain state from some set of initial conditions.

- *Inclusion of ontological information*. A model that incorporates a biological concept model is very powerful. Firstly, ontologies provide consistent definitions ad interpretations of biological concepts, and enables the software re-use of this knowledge consistently (T.R.Gruber 1995). Ontologies can further extend the logical inferencing capabilities by generalization and explanation (S.Schulze-Kremer 1998).

- *Amenable to computational analysis*. The verification of system properties must be computationally possible, and feasible. The scope of the ST pathways domain makes it imperative that we find a computational solution to the problem of pathway discovery.

- *Abstract schemata for processes*. A qualitative model must be able to provide generalizable definitions of the processes occurring within ST pathways. By doing so, specific instances of these processes can be automatically generated, thus eliminating the need to hand-code every reaction occurrence. Also, an abstract schemata can be applied to infer new instances of ST processes.

Other Approaches

We summarize some of the ongoing efforts at qualitative modeling and simulation of the ST pathways domain.

Ecocyc (P.D.Karp 2000) is a bio-ontology that represents a model functional for metabolic reactions in *E. coli* using a frame-based formalism. This framework has been extended to ST pathways as well. Pathways are represented as ordered lists of reactions with branch points. An expert system is employed to trace the pathway between any two system states (P.R.Romero & P.Karp 2001). However, the reactions need to be fully described, *i.e.*, common processes cannot be schematically defined and instantiated when required. Also, Ecocyc does not capture dynamic properties, including parallelism and temporal constraints.

System dynamics and structure exist within human organizations, and the methods applied to improve their efficacy can be translated to the realm of biological systems as well. **Statecharts** (D.Harel & E.Gery 1997) are state machines with hierarchy, orthogonality/parallelism, and broadcast communication capabilities. They can represent concurrent behavior by splitting the ST pathways domain into its constituent components. As a formalism, statecharts focus on capturing system dynamics. They do not, however, capture the functional roles played by the components. Statecharts also do not allow for the formal verification of the model.

Another graph-based technique is that of **Petri Nets** (J.L.Peterson 1981). A Petri Net is a formal model used to model concurrent systems. It is represented by a directed, bipartite graph in which vertices are either places or transitions, signifying conditions and processes, respectively. Tokens are placed on vertices to indicate that those conditions are true. When all the arcs to a transition are tokenized, the transition is said to be enabled. High-level Petri Nets include extensions that allow temporal and hierarchical input. Petri Nets are mathematically well understood, thereby allowing a formal analysis of system properties. But Petri Nets do provide a static-structural view of the components within a ST pathway. For example, they cannot reason about the different molecular species represented, or the functional roles played by each species. As with statecharts, the state description and state transition must be fully described. This only supports pathway simulation and not the discovery of novel hypothetical pathways.

Pi-calculus (Regev, Shapiro, & Shapiro 2001), a process algebra originally developed for specifying mobile communication systems, provides a model that is mathematically well-defined and computationally amenable. It therefore allows a formal verification of the ST pathway properties being modeled. A component within a pathway is represented as a *computational process* that interacts with each other component via communication through *channels*. A multi-domain protein is modeled as a set of computational processes, one per domain, that communicate over private channels. As with Petri Nets, pi-calculus is unable to capture the static-structural view of the concept ontology. Furthermore, as the complexity of the signaling network increases, the messages passed increases quadratically, threatening the scalability of this approach. The pi-calculus system is pro-

gressing towards a stochastic and rate-based simulating platform.

Maude (Eker *et al.* 2002) is a symbolic abstraction algebraic structure which implements rewriting logic to capture the ST pathways domain. It is capable of state and concurrent computation. As a well-defined mathematical formalism, it can perform inference and is amenable to computational analysis. Typing and sub-typing of objects within a model provides a type hierarchy and thus, a detailed static view. However, Maude is unable to incorporate the ontology of biological functionality.

More detailed models, using hybrid approaches that combine two or more techniques, have also been attempted, *e.g.*, workflow/Petri Net model (Peleg, Yeh, & Altman 2002).

Encoding ST Pathways as a Planning Domain

We demonstrate the correspondence between planning and ST pathway discovery by reconstructing the epidermal growth factor (EGF) pathway. The EGF pathway, a member of the well-characterized group of receptor tyrosine kinase - mitogenic activated protein kinase (RTK-MAPK) pathways, has been modeled by other computational approaches (Regev, Shapiro, & Shapiro 2001; Eker *et al.* 2002), and thus provides a good basis for comparison.

The Planning Problem

A classical planning problem is characterized (Weld 1999) by (1) an *initial state*, (2) one or more *goal states* and (3) a set of *operator schemata* that comprise the domain theory. A **plan** is defined as a sequence of *actions* (*i.e.*, instantiated operators) that can be applied to the initial state, thereby transforming it into a goal state.

The pathway discovery problem can similarly be stated as the task of finding the ordered sequence of subcellular processes which, when applied to a subset of cellular components (present in some initial configuration), elicits a specific cellular response. Formally, the task of signal transduction pathway discovery is re-cast as a planning problem $\mathcal{P} = (\mathcal{O}, \mathcal{I}, \mathcal{G})$ where

- *Initial state*. The initial state $\mathcal{I}$ is the conjunction of the initial configuration of the pathway components present within the subcellular milieu. Every protein is initialized to some state, as are its docking sites, if present.
- *Operator schemata*. Signal transduction proceeds in mammalian cells via a finite set of subcellular processes. Every process can be broadly defined in terms of the modifications it makes to some subset of objects within the domain. Thus, every process can be converted to an operator schema which can be instantiated when the schema preconditions are met within the subcellular milieu. The set $\mathcal{O}$ is the union of these process schemata.
- *Goal state*. The response elicited from the signal target is described by the goal state, $\mathcal{G}$.

Overview of EGF Pathway

The EGF pathway can be summarized thus:

1. An epidermal growth factor (EGF) signaling molecule binds to two EGF receptors on the surface of a cell. The EGF receptor can be thought to consist of three domains, extracellular, transmembrane and intracellular, one per each cell compartment that it spans. The EGF molecule binds with the extracellular domain.
2. Each of the two bound EGF receptors phosphorylates the other's intracellular domain at specific locations.
3. An adapter protein Grb2, present within the cytosolic compartment, binds to the phosphorylated intracellular domain of the EGF receptor with its SH2 domain and undergoes a conformational change that results in the opening of its SH3 domain.
4. The guanyl-nucleotide exchange factor (GEF) activity of the Sos protein then activates the Ras molecule.
5. Activated Ras recruits the Raf protein kinase to the cell membrane where Raf is phosphorylated. A kinase cascade follows.
6. Activated Raf binds to and phosphorylates the MEK protein at two locations, giving rise to single and double phosphorylated MEK species.
7. Double phosphorylated MEK phosphorylates the MAP kinase, which is the ERK protein. ERK can be single and double phoshphorylated as well.
8. Activated ERK translocates to the nucleus, where it triggers *de novo* gene expression.

It must be noted that the above description, although fairly detailed, still represents a high-level abstraction of the EGF pathway. In reality, each step can occur via multiple reactions, and can be regulated by components from other pathways.

Mapping the EGF Pathway to Planning Formalism

Before we describe the EGF pathway using the planning formalism, we identify the objects within the domain and their interal state representations. Also, we examine the invariants the ST pathways domain — compartmentalization and docking site state representation. The domain is specified in terms of the Planning Domain Definition Language (PDDL) (D.McDermott *et al.* 1998).

Objects The ST pathways domain is primarily composed of proteins and genes (DNA segments). A protein can be further decomposed in terms of its sequence features, namely, active sites, motifs and domains. Thus, the types allowed within the domain are specified by `gene`, `protein`, `dom`, `motif` and `site`, respectively.

Within the EGF pathway, for example, the EGF molecule is a `protein` and the EGF receptor (EGFR) is a protein with three domains - intracellular, extracellular and membrane. This is specified by

```
(:objects
    EGF EGFR - protein
    EGFR-intra EGFR-extra EGFR-mem - dom
    ...)
```

Compartmentalization We simplify our domain by subdividing the cell into four distinct compartments - *extracellular, membrane, cytoplasmic solution (cytosol)* and *nucleus*. For the purposes of this paper, we trace the signal propagation through these compartments only. The cellular localization of any object is given by the `in` predicate. The smallest unit of localization is the docking site. If docking site information is unavailable then the protein's localization is provided.

The EGF receptor spans three domains. Thus, each domain is sequestered from proteins in other compartments. The domain localization for EGFR's intracellular domain is given by

```
(in EGFR-intra cytosol)
```

The EGF molecule is present in the extracellular domain. It is treated as a single protein whose structure serves as one functional domain. Thus, its protein localization is described by

```
(in EGF extracellular)
```

Docking Site State Representation As mentioned previously, docking sites exhibit functional activity by helping cellular components to "dock" or bind with each another to create complexes. Domains characteristically bind with a specific set of other domains, leading researchers to infer protein-protein complexes based on the proteins' constituent domains (M.Deng *et al.* 2002). Such domain-domain interactions are possible when both domains, on either protein, are *open* for docking.

Thus, when the extracellular domain of the EGF receptor is not yet bound to EGF, *i.e.* it is open, it is represented as

```
(domain-state EGFR-extra open)
```

Initial State The initial state is the conjunction of all the literals that describe the initial configuration of the components within the pathway. Each component is described in terms of its constituent docking sites, their cellular localization, and the initial state representation for every docking site.

Goal State Like the initial state, the goal state lists the conditions desired at the end of the signaling pathway. Usually within the ST pathways domain, the goal state is a targeted cellular response activated by the signaling events that comprise the pathway.

Abstract Operators Biological processes can be defined in terms of the modifications they effect upon components within the domain. As was previously stated, the two most important processes within signal transduction are *binding*, which results in protein complex formation, and *protein modification*, such as phosphorylation, which results in the activation and consequent propagation of the signal.

Abstract operators are written at two levels of abstraction reflecting the differences in the biochemical detail of the data collected.

Protein complex formation: The two abstraction levels that we model within the planning formalism are

1. *protein-level*: Certain data sources, such as gene expression data and two-hybrid systems, provide activation information at the protein level. The constituent docking site information of the proteins is implicit within the analysis. In our formalism, a bind interaction at the protein level is modeled explicitly by the `protein-bind` operator.

```
(:action protein-bind
 :parameters (?comp - compartment
    ?x ?y - protein)

 :precondition (and (in ?x ?comp)
    (in ?y ?comp)
    (can-ppi ?x ?y)) ;;; in PPI list

 :effect (protein-bound ?x ?y))
```

From the preconditions, the action requires the two proteins to be in the same compartment. The predicate `(can-ppi ?x ?y)` admits only a restricted list of protein-protein interactions. This list is compiled from within the KB, and can be populated directly from protein-protein interaction data, or inferred from gene array experiments, co-occurrence within the literature, etc.

2. *domain- or docking site-level*: As more detailed characterization of proteins accumulates, interactions can be modeled at the level of docking sites. A domain-level bind interaction is modeled by the `domain-bind` operator.

```
(:action domain-bind
 :parameters (?comp - compartment
    ?x ?y - protein
    ?xdom ?ydom - dom)

 :precondition (and (has-domain ?x ?xdom)
    (has-domain ?y ?ydom)
    (in ?xdom ?comp)
    (in ?ydom ?comp)
    (domain-state ?xdom open)
    (domain-state ?ydom open)
    (can-ddi ?xdom ?ydom)) ;;; in DDI list

 :effect (and (domain-bound ?xdom ?ydom)
    (protein-bound ?x ?y)
    (not (domain-state ?xdom open))
    (not (domain-state ?ydom open)))))
```

The preconditions of the `domain-bind` action require that for two proteins ?x and ?y to bind via their interacting domains ?xdom and ?ydom respectively, the two domains must be "open" for binding. The predicate `(can-ddi ?x ?y)` restricts the domain-domain bind interactions to a restricted list that is compiled within the KB from different data sources.

Binding between proteins characterized at different levels of biochemical detail is allowed only if the knowledge model allows it explicitly, *i.e.*, they must meet the specifications of either the `protein-bind` or the `domain-bind` action schemata. For example, the interaction between the EGF molecule and the extracellular domain of the EGF receptor is obtained because, in addition to being a protein, the EGF molecule is modeled as a single domain, with a corresponding entry `(can-ddi EGF EGF-extra)` in the knowledge base.

Protein Modification - Phosphorylation: Protein modifications can be modeled at the protein-level as well the domain-

level. For example, the protein-level phosphorylation specification

```
(:action protein-phosphorylation
 :parameters (?comp - compartment
    ?x ?y - protein)

 :precondition (and (in ?x ?comp)
    (in ?y ?comp)
    (can-phosphorylate ?x ?y)
    (not (has-phosphate ?y)))

 :effect (and (has-phosphate ?y)))
```

ignores many docking site-level details, such as the number of phosphorylation sites, their locations, etc. The only criteria are that ?x must be able to phosphorylate ?y and ?y is not previously phosphorylated. The predicate `(can-phosphorylate ?x ?y)` is derived from a *phosphorylation list* which is compiled within the KB.

Modeling the Internal Circuitry of Proteins By a protein's *internal circuitry* is meant the protein-specific conformational changes that can occur when activated or inhibited by some other protein. We chose to model a protein's internal circuitry as a grounded action with conditional effects. The conditions are the changes in the environment that initiate the conformational changes, while the effects are the corresponding conformational changes. We guarantee the internal circuitry action is called every time a molecule participates in a reaction by setting a flag that must be reset before that molecule can be re-used again, *i.e.*, participate in a new reaction.

Take the example of the Grb2 protein. When its SH2 domain is phosporylated by the EGF receptor, its conformation changes, revealing its SH3 domains, which are now ready to bind.

Multi-Agent System Platform for Planning

The AI planning inference engine described above, can only generate new and promising hypotheses if it is applied to new data that can improve and/or modify existing plans. The volume of data collected requires that the acquisition and translation of data into the planning formalism be automated. Additionally, the data must be encoded in a representation such that all the biological concepts and relationships involved are clearly represented and transparently manipulable by the planning system.

For these reasons, we have incorporated the planning engine within a multi-agent system that is also responsible for gathering data from multiple sources and populating a knowledge base. The task of pathway discovery can then be framed from the information present within the knowledge base.

Multi-Agent System for Information Gathering and Inference

DECAF Our MAS for biological network simulation was built using DECAF (Distributed, Environment Centered Agent Framework) (Graham & Decker 2000) which is Java-based toolkit for creating multi-agent systems. We chose DECAF because it offers several advantages. Firstly, several agent-building tools are available within DECAF, allowing for the rapid prototyping of the agent system. Secondly, the internal architecture of a DECAF agent is built much like an operating system. The modular design and robustness of each service allows us to track the control flow within the agent, thereby reducing the time taken for testing and debugging.

DECAF Support for Information Gathering DECAF supports building information gathering systems by providing useful middle agents and a shell for quickly building information extraction agents for wrapping web sites and external programs. DECAF has previously been applied to create an information gathering MAS for the purpose of genomic annotation (K.Decker *et al.* 2002).

Agent name servers, matchmakers, brokers, and other middle-agents support the creation of open systems where elements may come and go over time. Dynamic information change is supported by reusable Information Extraction Agent behaviors that include the ability to push data values to the user, or to set up persistent queries that pull data from providers only when the answer changes significantly.

DECAF Support for Pathway Discovery

In addition to the domain-independent agents mentioned above, we have developed agents specifically for the task of pathway discovery. The **Planner** agent is created to wrap the planning system. Pathway discovery queries are posed to it by the **Manager Agent** which can receive these queries from the **User Query Agent**. It is conceivable that in the future, the Manager Agent will be able to pose pathway discovery problems independently, perhaps motivated to resolve inconsistencies within the KB, or in order to extend existing pathways.

Information Extraction Agents are used to extract information from multiple data sources, such as gene expression, literature mining, etc.

The ontology/knowledge base is wrapped by the Knowledge Base (KB) Agents that can respond to queries issued by the Manager. The KB agents can also talk directly with the Information Extraction agents to upload the data collected.

ST Pathways Ontology And Knowledge Base

We have created a signal transduction ontology, which guides the design of the knowledge base used to store information collected from multiple sources. A major problem associated with collecting data from sources as diverse as literature mining and gene expression is that the information is not only heterogeneous, but is also produced at different levels of biochemical detail. The ST ontology alleviates this problem by providing a consistent set of definitions and interpretations of biological concepts that spans the diversity of the data collected. The ontology is broadly divided (see Figure 2) into the following orthogonal aspects:

- *Signaling Event*: Captures all the processes that participate in ST pathways and help transfer the signal. Includes

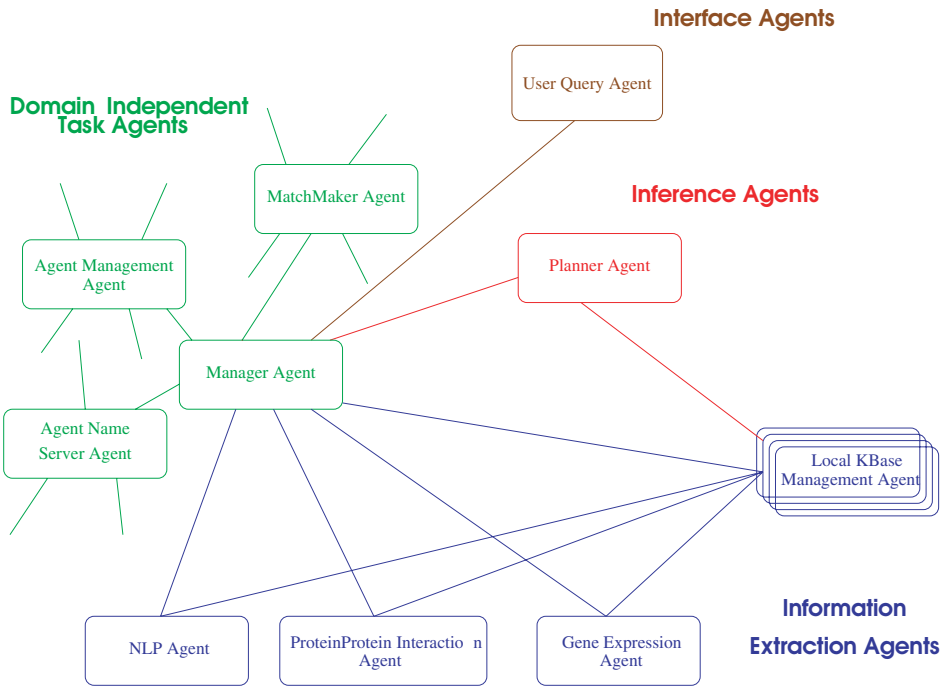

Figure 1: The Planning Multi-Agent System

binding, phosphorylation, etc.

- *Chemicals*: Provides a class hierarchy for the biomolecules found to participate in the ST domain.
- *Chemical Attributes*: Conceptual decomposition of chemicals into functional units that help propagate the signal despite lacking independent physical existence. Includes protein constituents such as domain, motif, and active sites.
- *Cellular Compartmentalization*: Provides a controlled vocabulary for the distinct sub-compartments into which the cell has been divided.

We have created our ontology with the aid of Protege (N.Noy, R.Fergerson, & M.Musen 2002). Our ontology incorporates concepts from Gene Ontology (Consortium 2000) and other bio-KBs (P.D.Karp 2000; Fukuda & Takagi 2001; F.Schacherer *et al.* 2001) as well.

Implementation

The EGF pathway, at the level of granularity modeled, consists of 11 domains distributed over 9 proteins. The signal is propagated through 3 compartments — extracellular, cytosol and membrane.

We chose the Fast Forward planner (J.Hoffman & B.Nebel 2001) to apply to the ST pathways domain. The ST domain theory currently consist of a total of 14 operator schemas. The planner was able to "discover" the EGF pathway in seconds (see planner output, Figure 3).

To test the generalizability of these operators, we applied them without modification to map the Fas Ligand (FasL) pathway. We were successful in doing so. To test the planner against larger data sets, we added artificial objects to the domain. There was an exponential blow-up in the time taken to find the EGF pathway in the increased presence of these distracting elements (see Figure 5).

Discussion

By reformulating the task of signal transduction pathway discovery into a planning problem, we can apply methods that are amenable to inference as well as qualitative, and to a lesser extent, quantitative, simulation.

Evaluation of Planning as Modeling Formalism

We revisit the criteria deemed necessary for a successful qualitative model, and demonstrate that planning can be a successful modeling formalism for ST pathways.

Planning provides static-structural, dynamic and functional views of the ST pathways domain. A planning domain is defined in terms of its objects; the ST pathways domain is described in terms of subcellular objects and their properties, which together represents the *static-structural view* of the domain. A plan is a representation of the *dynamic* properties of the pathway, *i.e.*, the flow of control through atomic processes or actions. A *functional view* is obtained by the very act of instantiating operators: each action (*i.e.*, a biological process), is bound to a set of inputs (preconditions) and outputs (effects).

As previously mentioned, we are interested in applying planning to conduct a *reachability* analysis, which is essentially a qualitative approach. Thus, we eschew the

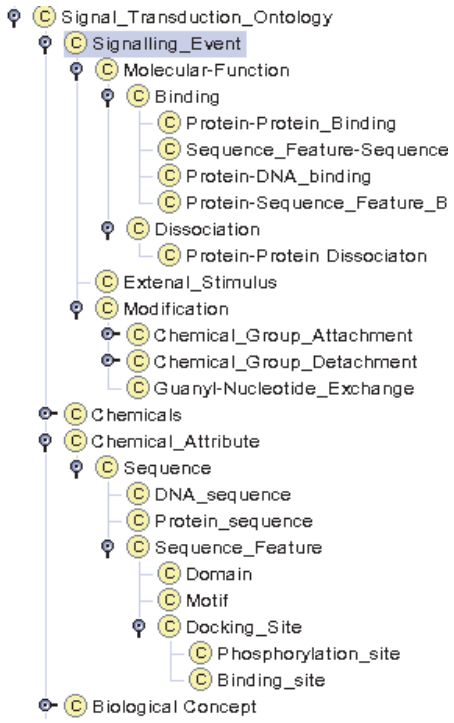

Figure 2: A snapshot of our Signal Transduction Ontology

molecular dynamics of reactions, such as reaction rates, etc. and believe that other approaches of applying differential equation modeling and simulation would be better suited for such a task.

Planning includes a biological ontology of the ST pathways domain. Although the planner does not explicitly require the ontology, we base the planning formalism on our ontological specification. The predicates that describe a state are transcribed from slots within ontology class definitions. The *Chemicals* and *Chemical Attributes* ontologies are used to classify objects within the ST pathways domain.

Planning allows a hierarchical view of the ST pathways domain. Although Fast-Forward, a STRIPS-style planner, is unable to capture multi-level abstractions of the ST pathways domain, Hierarchical Task Network (HTN) planners can be applied to encode activities at different levels. The non-primitive tasks are associated with biological processes while the primitive tasks correspond to their molecular-level functional counterparts. Our initial experiences of applying HTN planning to biological pathway discovery, are detailed further below.

Planning is mathematically-based and allows verification of system properties. Properties of planners, and planning in general, have been well investigated (T.Bylander 1994; Weld 1999). Planning progresses by executing an applicable operator in the current state. Thus the system properties (*i.e.*, state description) is known before and after the action

```
STEP 0: DOMAINBIND EXTRACELLULAR EGF EGFR EGF EGFEXTRA
STEP 1: PROTEINBIND MEMBRANE EGFR EGFR
STEP 2: PROTEINPHOSPHORYLATION CYTOSOL EGFR EGFR
STEP 3: ACTIVESITEBIND CYTOSOL EGFR GRB2 EGFREXTRA SH2
STEP 4: DOMAINBIND GRB2 SOS SH3 SH3
STEP 5: PROTEINBIND CYTOSOL SOS RAS
STEP 6: PROTEINPHOSPHORYLATION CYTOSOL RAS RAF
STEP 7: PROTEINPHOSPHORYLATION CYTOSOL RAF MEK
STEP 8: PROTEIN PHOSPHORYLATION CYTOSOL MEK ERK
```

Figure 3: FF output plan corresponding to the EGF pathway

is executed. At any point in the plan, we can list the predicates that are true; their conjunction represents the current state description.

Planning provides important inference capabilities. The planner can be used to verify if some state is reachable from some initial condition. We can also conduct perturbation analyses of two types: 1) modify the properties of one or more objects in the domain to check if a pathway can be constructed and 2) modify the schema definitions of operators to check if new pathways can be discovered. From such analyses, we can generate *in silico* hypotheses regarding the outcome of knockout and/or mutation experiments.

Planning representation is intuitive to biologists. Planning operators are written in propositional logic, which is very simple to read and understand. The ground predicates are properties of objects in the domain, and are easily specified.

Challenges Encountered

The extraction of knowledge and its subsequent representation into the planning formalism was the hardest challenge we faced. While encoding the EGF pathway was relatively straightforward, extending our analysis to larger pathways involved in apoptosis (cell suicide) was cumbersome, as the data had to be manually gathered. Currently, we are in the process of integrating the NLP and Gene Expression Agents. With the aid of these and other tools we plan to include as part of the MAS, we hope to be able to tackle more information-dense pathways.

The second problem was that of scalability. While planning competitions have used plan length as the difficulty metric, successful planning within the ST domain means that a valid plan must generated after sifting through the huge volume of information gathered. In our experience, we found that the number of predicates describing the state was a better measure of difficulty. From our testing, we found that time taken to find a plan rose exponentially with increases in the problem size.

A problem that we have faced when applying Fast Forward is the inability to make use of the object hierarchy within the ST ontology. The ability to infer that a *kinase* is a protein with a phosphorylating capability, for example, is very useful when specifying the domain. Likewise, incorporating a hierarchical notion of operator schemas would simplify the task of translating between the ontology and the

planning formalism. An example of this type would be the *phosphoprotein binding* operator which is a bind operator where one of the proteins is phosphorylated.

The use of axioms to infer relationships at every stage of the plan can be helpful when encoding the ST pathways domain. For example, in order to simplify the representation, we avoided the creation of a protein complex type, deciding instead, to imply complex formation via the the `protein-bound` predicate. However, when multi-protein complexes are created, the situation becomes unwieldy if we wish to enforce the transitivity of the protein-bound predicate. An axiom that compounds individual proteins into a single complex following a call to the bind operator would resolve this issue.

Given the surplus of data, it is conceivable that many hypothetical pathways can be generated for some pair of initial and final states. However, the functional conservation found in nature suggests that most of these pathways will not have real-world counterparts. By ascribing a confidence measure to a pathway that is based on the data that supports it, we can sift the numerous pathways to select the ones of interest. Incorporating the calculation of a confidence measure or some other probabilistic notion would be a vital addition to the system.

Intracellular reactions can take anywhere from milliseconds to minutes to occur. This vast range in reaction velocity is an important aspect of the ST domain, because even though a pair of reactions might be hypothesized by the planner, their relative reaction rates might prevent them from being in physical proximity to react at any given time. A potential solution to this modeling shortcoming is the use of durative actions.

Finally, Fast-Forward and other STRIPS-style planners are unable to plan at different levels of abstraction. This stymies the visualization and human comprehension of the plan/pathway, besides being more resource-intensive when applied to larger pathways.

HTN Planning

Individual ST pathways can be treated as parameterized modules (Endy & Brent 2001). Larger pathways can be built by combining these modules, while minimizing cross-talk. HTN planning provides one such solution, where tasks representing independent pathways are combined within a network to form a larger pathway. HTNs can also be used to encode the ST pathway at multiple levels of abstraction. In addition to providing powerful and specialized tools to reason about temporal and resource utilization (Wilkins & desJardins 2001), hierarchical representations can help with scalability, as is seen in many practical planning systems (K.Erol, D.Nau, & J.Hendler 1994). We employed O-Plan (Currie & Tate 1991), to represent the EGF pathway from before. To test the scalability, we generated artificial objects with a fair amount of detail (*e.g.*, proteins were created with at least 2 domains each). As the problem size (measured in terms of the number of pathway components present) increased, O-Plan scaled better than FF in our preliminary trials. However, HTN planning has shortcomings of its own. HTN planners require a task structure in order to generate a plan (see Figure 4). This hand-coding amounts to pre-knowledge of the pathway in some detail. Unfortunately, since not all pathways are as well characterized as the EGF pathway, framing a HTN is not always possible. Secondly, by creating a pre-defined task structure, we are limiting the solution to conform to a known framework. This design pre-empts the discovery of novel pathways and unexpected interconnections.

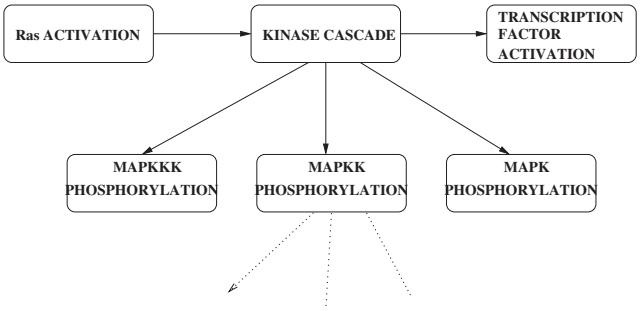

Figure 4: HTN representation of the EGF pathway.

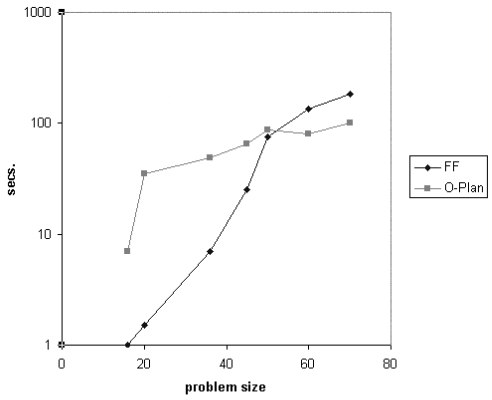

Figure 5: Run time curves for FF and O-Plan. The problem size is the number of objects present at the initial state within the domain. The first point on both curves are the run times for the FasL pathway. The second point is the run time for the EGF pathway. The third is the time taken to discover the EGF pathway in the presence of FasL components at runtime. Data points beyond the third point were created for artificial data. Note that the time is in logarithmic scale.

As suggested by others (Wilkins & desJardins 2001), we believe that the combination of STRIPS-style and HTN planning techniques will be required to generate successful plans within the large and complicated domain of signal transduction pathways.

References

Consortium, T. G. O. 2000. Gene ontology: tool for the unification of biology. *Nature Genetics* 25:25–29.

Currie, K., and Tate, A. 1991. O-plan - the open planning architecture. *Artificial Intelligence* 52:49–86.

D.Harel, and E.Gery. 1997. Executable object modelling with statecharts. *IEEE Computer* 30:31–42.

D.McDermott; M.Ghallab; A.Howe; C.Knoblock; A.Ram; M.Veloso; D.Weld; and D.Wilkins. 1998. The planning domain definition language.

Eker, S.; Knapp, M.; Laderoute, K.; and Lincoln, P. 2002. Pathway logic: Symbolic analysis of biological computing. In *Pacific Symposium on Biocomputing*, 400–412.

Endy, D., and Brent, R. 2001. Modelling cellular behaviour. *Nature* 409:391–395.

F.Schacherer; Choi, C.; Gotze, U.; Krull, M.; Pistor, S.; and Wingender, E. 2001. *Bioinformatics* 17(11):1053–1057.

Fukuda, K., and Takagi, T. 2001. *Bioinformatics* 17(9):829–837.

Graham, J., and Decker, K. 2000. Towards a distributed, environment-centered agent framework. *Intelligent Agents VI* LNAI-1757:290–304.

J.Hoffman, and B.Nebel. 2001. The ff planning system: Fast plan generation through heuristic search. *Journal of AI Research* 14:253–302.

J.L.Peterson. 1981. *Petri Net Theory and the Modelling of Systems*. Englewood Cliffs, NJ: Prentice Hall.

K.Decker; S.Khan; C.Schmidt; G.Situ; R.Makkena; and D.Michaud. 2002. Biomas: A multi-agent system for genomic annotation. *Intl. J. of Coop. Info. Sys.* 11:265–292.

K.Erol; D.Nau; and J.Hendler. 1994. Htn planning: complexity and expressivity. In *AAAI*, 1123–1128.

Lodish, H.; Berk, A.; Zipursky, L.; Baltimore, D.; and Darnell, J. 1999. *Molecular Cell Biology*. New York: W.H. Freeman and Company.

M.Deng; S.Mehta; F.Sun; and T.Chen. 2002. Inferring domain-domain interaction from protein-protein interactions. In *Proceedings of the Sixth Annual Intl. Conf. on Comp. Biol.*, 117–126.

N.Noy; R.Fergerson; and M.Musen. 2002. The knowledge model of protege-2000:combining interoperability and flexibility. *Second Intl. Conf. on Knowledge Engg. and Knowledge Models* 265–292.

P.D.Karp. 2000. An ontology for biological function based on molecular interactions. *Bioinformatics* 16:269–285.

P.D.Karp. 2001. Pathway databases: A case study in computational symbolic theories. *Science* 293:2040–2044.

Peleg, M.; Yeh, I.; and Altman, R. 2002. Modelling biological processes using workflow and petri net models. *Bioinfo. J.* 18(6):825–837.

P.R.Romero, and P.Karp. 2001. Nutrient-related analysis of pathway/genome databases. In *Proceedings of the Sixth Pacific Symposium on Biocomputing*, 471–482.

Regev, A.; Shapiro, W.; and Shapiro, E. 2001. Representation and simulation of biochemical processes using the π-calculus process algebra. In *Pacific Symposium on Biocomputing*, 459–470.

Regev, A.; Silverman, W.; and Shapiro, E. 2000. Representing biomolecular processes with computer process algebra: π-calculus programs of signal transduction pathways. *Technical Report, Weizmann Institute*.

S.Schulze-Kremer. 1998. Ontologies for molecular biology. In *Proceedings of the Third Pacific Symposium on Biocomputing*, 693–704.

T.Bylander. 1994. The computational complexity of propositional strips planning. *Atificial Intelligence* 69:165–204.

T.R.Gruber. 1995. Towards principles for the design of ontologies using knowledge sharing. *Internal Journal of Human-Computer Studies* 43.

Voet, D., and Voet, J. 1995. *Biochemistry*. New York: John Wiley and Sons.

Weld, D. 1999. Recent advances in ai planning. *AI Magazine*.

Weng, G.; Bhalla, U.; and Iyengar, R. 1999. Complexity in biological signal systems. *Science* 284:92–96.

Wilkins, D. E., and desJardins, M. 2001. *AI Magazine* 22(9).