Extending a Multi-Agent System for Genomic Annotation

Keith Decker, Salim Khan, Carl Schmidt, and Dennis Michaud

Computer and Information Sciences Department
University of Delaware, Newark, DE 19716
{decker}@cis.udel.edu

Abstract. The explosive growth in genomic (and soon, expression and proteomic) data, exemplified by the Human Genome Project, is a fertile domain for the application of multi-agent information gathering technolgies. Furthermore, hundreds of smaller-profile, yet still economically important organisms are being studied that require the efficient and inexpensive automated analysis tools that multi-agent approaches can provide. In this paper we give a progress report on the use of the DECAF multi-agent toolkit to build reusable information gathering systems for bioinformatics. We will briefly summarize why bioinformatics is a classic application for information gathering, how DECAF supports it, and recent extensions underway to support new analysis paths for genomic information.

1 Introduction

Massive amounts of raw data are currently being generated by biologists while sequencing organisms. Most of this raw data must be analyzed through the piecemeal application of various computer programs and hand-searches of various public web databases. Typically both the raw data and any valuable derived knowledge will remain generally unavailable except in published natural language texts such as journal articles. However, it is important to note that a tremendous amount of genetic material is *similar* from organism to organism, even when they are as outwardly different as a yeast, fruit fly, mouse, or human being. This means that if a biologist studying the yeast can figure out what a certain gene does—its *function*—that other biologists can at least guess that similar genes in other organisms play similar roles. Thus huge databases are being populated with sequence data and functional annotations [3]. All new sequences are routinely compared to known sequences for clues as to their functions.

Furthermore, the *methods* by which biologists hypothesize function other than by similarity become other important clues to the identification of gene function in new organism sequences. The "function" of a gene can refer to one or all of concepts such as the molecular/biochemical function of a gene product, how this is used in some larger biological process, and also where in the cell the gene product does its work [27]. For example, biologists can get clues to gene function by:

– looking for certain patterns of amino acids (called *motifs* or larger *domains*) that indicate likely molecular/biochemical activity

– looking for certain patterns of amino acids (especially at the ends of a protein) that indicate where the gene product will be used
– looking for information about similar gene products (proteins), rather than just similar genes

A large amount of work in bioinformatics over the past ten years has gone into developing algorithms (pattern matching, statistical, and/or heuristic/knowledge-based) to support the work of hypothesizing gene function. Many of these are available to biologists in various implementations, and now many are available over the web. Meta-sites combine many published algorithms, and sites specialize in information about particular topics such as protein motifs.

From a computer science perspective, several problems have arisen, as we have described elsewhere [7]. To summarize, what we have is a large set of heterogeneous and dynamically changing databases, all of which have information to bring to bear on the biological problem of determining genomic function. We have biologists producing thousands of possible genes, for which functions must be hypothesized. For the case of all but the largest and well-funded sequencing projects, this must be done by hand by a single researcher and their students.

Multi-agent information gathering systems have a lot to contribute to these efforts. Several features make a multi-agent approach to this problem particularly attractive: information is available from many distinct locations; information content is heterogeneous; information content is constantly changing; much of the annotation work for each gene can be done independently; biologists wish to both make their findings widely available, yet retain control over the data; new types of analysis and sources of data are appearing constantly.

We have used DECAF, a multi-agent system toolkit based on RETSINA [26, 11, 8]: and TAEMS [10, 28], to construct a prototype multi-agent system for automated annotation and database storage of sequencing data for herpesviruses [7]. The resulting system eliminates tedious and always out-of-date hand analyses, makes the data and annotations available for other researchers (or agent systems), and provides a level of query processing beyond even some high-profile web sites.

Since that initial system, we have used the distributed, open nature of our multi-agent solution to expand the system in several ways that will make it useful for biologists studying more organisms, and in different ways. This paper will briefly describe our approach to information gathering, based on our work on RETSINA; the DECAF toolkit; our inital annotation system; and our new extensions for functional annotation, EST processing, and metabolic pathway reasoning.

2 The RETSINA model of information gathering

We view information gathering as a catch-all phrase indicating information retrieval, filtering, integration, analysis, and display. In particular, information gathering is done in domains where information (unique, redundant, or partially redundant) is available at many different locations and is constantly being changed or updated, with even new information sources appearing over time.

We promote the use of the RETSINA multi-agent organization for building information gathering systems. The RETSINA approach consists of three general classes of agents[26, 11]:

- Information Extraction Agents, which interact directly with external data sources, i.e. wrapping sensors, databases, web pages.
- Task Agents, which interact only with other agents to handle the bulk of the information processing tasks. These include both domain-dependent agents that take care of filtering, integration, and analysis; also domain-independent "middle agents" that take care of matchmaking, service brokering, and complex query planning.
- Interface Agents, that interact directly with the end user.

A surprisingly large number of these agents are all or mostly reusable [8], which contributes to faster prototyping of information gathering systems. Contributing to this in turn is the reliance on query processing as the basic, common agent action. In particular, the multi-agent implementation of such a system extends traditional database systems in handling dynamic Information (data or derived information that changes over time), creating open systems (data or derived information sources come and go over time), and even in achieving secondary user utility (users don't just expect an answer, but they often have expectations about the time it will take to get that answer or how many resources (e.g. money) to spend to achieve an answer of some characterization (quality, certainty, etc.)).

DECAF (described in the next section) provides an implementation of these middle agents, reusable agent classes, and other tools for building multi-agent information gathering systems.

3 DECAF

DECAF (Distributed, Environment-Centered Agent Framework) is a Java-based toolkit for creating multi-agent systems [16]. In particular, several tools have been developed specifically for prototyping information gathering systems. Also, the internal architecture of each DECAF agent has been designed much like an operating system—as a set of services for the "intelligent" (resource-efficient, adaptively-scheduled, soft real-time, objective-persistent) execution of agent actions. DECAF consists of a set of well defined control modules (initialization, dispatching, planning, scheduling, and execution, each in a separate, concurrent thread) that work in concert to control an agent's life cycle. There is one core task structure representation that is shared between all of the control modules. This has meant that even non-reusable domain-dependent agents can be developed more quickly than by the API approach where the programmer has to, in effect, create and orchestrate the agent's architecture as well as its domain-oriented agent actions. This section will first discuss the internal architecture of a generic DECAF agent, and then discuss the tools (such as middle agents, system debugging aids, and the information extraction agent shell) we have built to implement multi-agent information gathering systems. The overall internal architecture of DECAF is shown in Figure 1. These modules run **concurrently**, each in thier own thread.

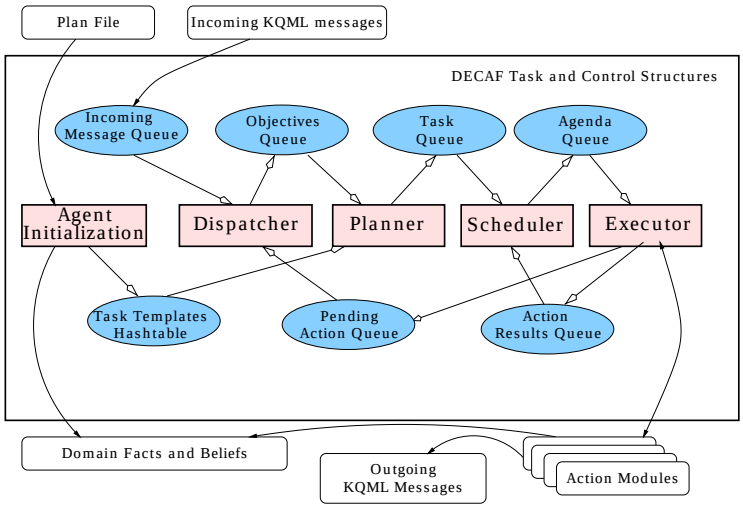

Fig. 1. DECAF Architecture Overview

Agent Initialization: The agent initialization module will read a *plan file* that describes the agent's capabilities as a specially-annotated HTN (Hierarchical Task Network). Each task reduction specified in the plan file will be added to the *Task Templates Hash table* (plan library).

Dispatcher: waits for an incoming KQML or FIPA message. If the message is part of ann ongoing conversation, the dispatcher will find the corresponding action in the *Pending Action Queue* and set up the tasks to continue the agent action. Otherwise, the message may indicate that it is part of a new conversation. If so a new *objective* is created (similar to the BDI "desires" concept[24]) and placed on the *Objectives Queue* for the Planner. An agent typically has many active objectives, not all of which may be achievable.

Planner: The Planner monitors the Objectives Queue and matches new goals to an existing task template as stored in the Plan Library. A copy of the instantiated plan, in the form of an HTN corresponding to that goal, is placed in the *Task Queue* area, along with a unique identifier and any provisions that were passed to the agent via the incoming message. If a subsequent message comes in requesting the same goal be accomplished, then another instantiation of the same plan template will be placed in the task networks with a new unique identifier. The Task Queue at any given moment will contain the instantiated plans/task structures (including all actions and subgoals) that should be completed in response to an incoming request.

Scheduler: The *Scheduler* waits until the Task Queue is non-empty. The purpose of the Scheduler is to determine which actions *can* be executed now, which *should* be executed now, and in what order they should be executed. This determination is currently based on whether all of the provisions for a particular module are available. It is possible to add significant reasoning ability to the scheduling module [15, 28, 17].

This effort involves annotating the task structure with performance and scheduling information to allow the scheduler to select an "optimal" path for task completion.

Executor: The *Executor* is set into operation when the Agenda Queue is non-empty. Once an action is placed on the queue the Executor immediately places the task into execution. One of two things can occur at this point: The action can complete normally (Note that "normal" completion may be returning an error or any other outcome) and the result is placed on the *Action Result Queue*. The framework waits for results and then distributes the result to downstream actions that may be waiting in the Task Queue. Once this is accomplished the Executor examines the Agenda queue to see if there is further work to be done. The Executor module will start each task in its own separate thread improving throughput and assisting the achievement of the real-time deadlines. Alternatively, an action may fail and not return, in which case the framework will indicate failure of the task to the requester.

3.1 DECAF Support for Info Gathering

DECAF provides core internal architectural support for secondary user utility. Thus DECAF plans can include alternatives, and these alternatives can be chosen dynamically at runtime depending on user constraints on answer timeliness or other resource constraints. DECAF also supports building information gathering systems by providing useful middle agents and a shell for quickly building information extraction agents for wrapping web sites. The **Agent Name Server (ANS)** ("white pages") is an essential component for agent communication. It works in a fashion similar to DNS (Domain Name Service) by resolving agent names to host and port addresses. The **Matchmaker** serves as a "yellow pages" to assist agents in finding services needed for task completion. The **Broker** agent acts as a kind of "middle manager" to assist an agent with collections of services. The broker can now provide a larger service than any single provider can, and often manage a large group of agents more effectively [9]. A **Proxy** agent allows web page Java applets to communicate with DECAF agents that are not located on the same server as the applet. The **Agent Management Agent (AMA)** allows MAS designers a look at the entire running set of agents spread out across the internet that share a single agent name server. This allows designers to query the status of individual agents and watch or record message passing traffic.

Information Extraction Agent Shell The main functions of an information extraction agent (IEA) are [8]: Fulfilling requests from external sources in response to a *one shot query* (e.g. "What is the price of IBM?"). Monitoring external sources for *periodic* information (e.g. "Give me the price of IBM every 30 minutes."). Monitoring sources for patterns, called *information monitoring* requests (e.g. "Notify me if the price of IBM goes below $50."). These functions can be written in a general way so that the code can be shared for agents in any domain.

Since our IEA operates on the Web, the information gathered is from external information sources. The agent uses a set of *wrappers* and the wrapper induction algorithm STALKER [22], to extract relevant information from the web pages after being shown several marked-up examples. When the information is gathered it is stored in the local IEA "infobase" using Java wrappers on a PARKA [18] knowledgebase. This makes

new IEA's fairly easy to create, and forces the difficult parts of this problem back on to KB ontology creation, rather than the production of tools to wrap web pages and dynamically answer queries. Currently, there are some proposals for XML-based page annotations which, if adopted, will make site wrapping easier syntactically (but still, does not solve the ontology problem—but see projects such as OIL).

4 A DECAF Multi-Agent System for Genomic Analysis

These tools can be put to use to create a prototype multi-agent system for various types of genomic analysis. In the prototype, we have chosen to simplify the query subsystem by materializing all annotations locally, thus removing the need for sophisticated query planning (e.g. [21]). This is a reasonable simplification since most of our work is with viruses and bacteria that have fairly small genomes (around 100 genes for a herpesvirus and around 30 herpesviruses) or with larger organisms (e.g. chickens) for which we are constructing a consensus database explicitly.

Figure 2 shows an overview of the system as four overlapping multi-aent organizations. The first, *Basic Sequence Annotation*, is charged with integrating remote gene sequence annotations from various sources with the gene sequences at the Local KnowledgeBase Management Agent (LKBMA). The second, *Query*, allows complex queries on the LKBMAs via a web interface. The third, *Functional Annotation* is responsible for collecting infomormation needed to make an informed guess as to the function of a gene, specifically using the three-part Gene Ontology [27]. The fourth organization, *EST Processing* enables the analysis of expressed sequence tags (ESTs) to produce gene sequences that can be annotated by the other organizations.

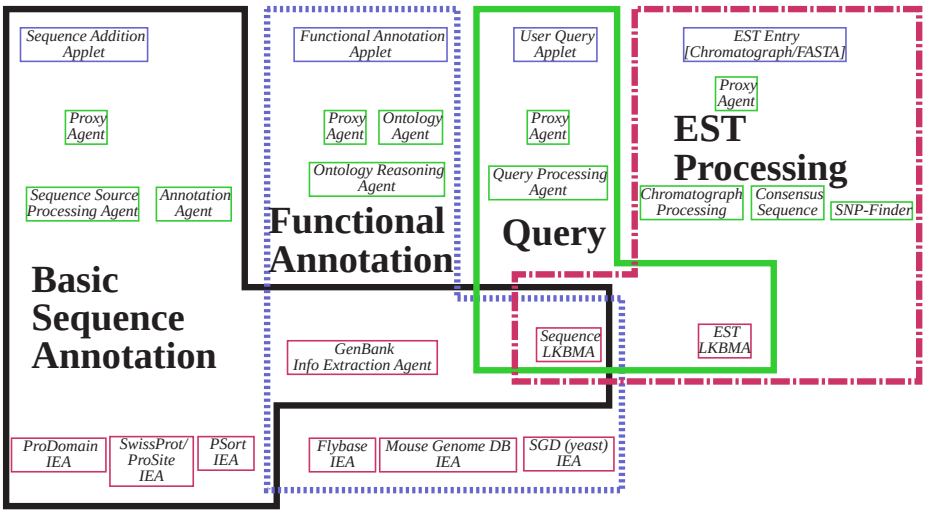

Fig. 2. Overview of DECAF Multi-Agent System for Genomic Analysis

An important feature to note is that we are focussing on annotation and anaysis services that are not organism specific. In this way, the resulting system can be used to build and query knowledgebases fro several different organisms. The original subsystems (basic annotation and the simple query system) were built to annotate the newly sequenced herpesvirus of turkey (the bird), and then to compare it to the other known sequenced herpesviruses. Work is just beginning to build a new knowledgebase from chicken ESTs, and to extend the depth of the herpesvirus KB for Epstein-Barr Virus (human herpesvirus 4) which has clinical significance for pediatric organ transplant patients.

4.1 Basic Sequence Annotation and Query Processing

Figure 3 shows the interaction details for the basic sequence annotation and query subsystems. We will describe the agents by their RETSINA classification.

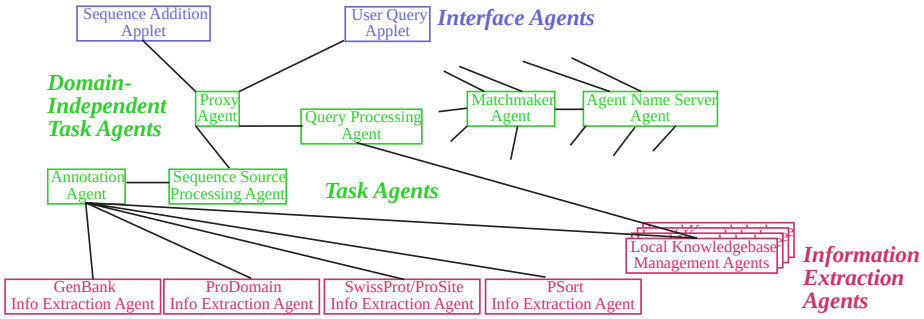

Fig. 3. Basic Annotation and Query Agent Organizations

Information Extraction Agents. Currently 4 agents based on the IEA shell wrap public web sites. The Genbank wrapper primarily supplies “BLAST” services: given the sequence of a herpesvirus gene, what are the most similar genes known in the world (called “homologs”)? The answer here can give the biologist a clue as to the possible function of a gene, and for any gene that the biologist does not know the function of, a change in the answer to this query might be significant. The SwissProt wrapper primary provides protein motif pattern searches. If we view a protein as a one-dimensional string of amino acids, then a motif is a regular expression matching part of the string that may indicate a particular kind of function for the protein (i.e. a prenylation motif indicates a place where the protein may be modified after translation by the addition of another group of molecules) The PSort wrapper accesses a knowledge-based system for estimating the likely sub-cellular location that a sequence’s encoded protein will be used. The ProDomain wrapper allows access to other information about the encoded protein; a protien domain is similar to a motif but larger. As we move to new organisms, many more resources could be wrapped at this level (almost all biologists have a “favorite” here).

The local knowledgebase management agent (KBMA) is a slightly different member of this class because unlike most IEAs it actually stores data via agent messages rather than only querying external data sources. It is here that the annotations of the genetic information are materialized, and from which most queries are answered. Each KBMA is updated with raw sequencing data indirectly from a user sequence addition interface that is then automatically annotated under the control of an annotation task agent. KBMAs can be "owned" by different parties, and queried seperately or together. In this way, researchers with limited computer knowledge can create shareable annotated sequence databases using the existing wrappers and other analysis tools as they are developed, without having to necessarily download and install them themselves. Using a PARKA-DB knowledgebase allows efficient, modern relational data storage on the back end and query as well as limited KB inferencing [18].

Task Agents. There are two domain task agents; the rest are generic middle agents described earlier. The Annotation Agent directs exactly what information should be annotated for each sequence. It is responsible for storing the raw sequence data, making queries to the various wrapped web sites, storing those annotations, and also indicating the provenance of the data (meta-information regarding where an annotation came from). The Sequence Source Processing Agent takes almost raw sequence data in ASN.1 format as output by typical sequence estimation programs or stored in Genbank. The main function of this agent is to test this input for internal consistency.

Interface Agents. There are two interface applets that communicate via the proxy agent with other agents in the system. One is oriented towards adding new sequences to a local knowlegebase (secured by a password) and the other allows anyone to query the complete annotated KB (or even multiple KBs). The interface hardly scratches the surface of the queries that are actually possible, but a big problem is that most biologists are not comfortable with complex query languages. Indeed, the simple interface that allows simple conjunctive and disjunctive queries over dynamic menus of annotations (constructed by the applet at runtime from the actual local KB) is quite advanced as compared to most of the existing public sites that allow textual keyword searches only.

4.2 Functional Annotation

Figure 4 shows the Functional Annotation subsystem. This subsystem is responsible for assisting the biologist in the difficult problem of making functional annotations of each gene. Unfortunately, many of the millions of genes sequenced so far have fairly haphazard (from a computer scientist's perspective) functional annotation: simply free natural language descriptions. Recently, a fairly large group representing at least some of the primary organism databases have created a consortium dedicated to creating a gene ontology for annotating gene function in three basic areas: the biological process in which a gene plays a part, the molecular function of the gene product, and the cellular localization [27]. The subsystem described here supports the use of this ontology by biologists as sequences are added to the system, eventually leading to even more powerful analysis of the resulting KBs.

Information Extraction Agents. Besides the gene sequence LKBMA and the GenBank IEA, we are wrapping three new organism-specific gene sequnce DBs, for Drosophila (fruit fly), Mus (Mouse), and Saccrynomaeces cervasie (yeast). Each of these organisms

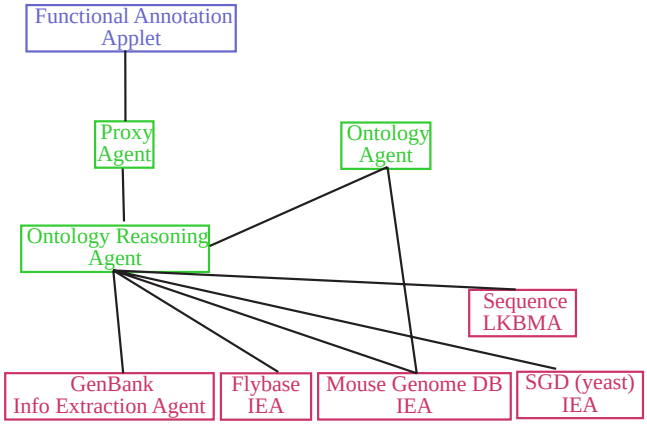

Fig. 4. Functional Annotation Agent Organization

is part of the Gene Ontology (GO) consortium, and has spent considerable time in making the proper functional annotation. Each of these agents, then, finds GO-annotated, close homologs of the unannoted gene and proposes the annotation of the homologs for the annotation of the new gene.

Task Agents. There are two new task agents, one is a domain-independent ontolgy agent using the FIPA ontology agent specification as a starting point. The ontology agent contains both the GO ontologies and several mappings from other symbologies (i.e. SwissProt terms) to GO terms. In fact, the Mouse IEA uses the Ontology agent to map some non-GO terms for certain records to GO terms. Although not indicated on the figure, some of the other organism DB IEA agents must map from GO ontology descriptive strings to the actual unique GO ID. The other service provided by the ontolgy agent (and not explicitly mentioned in the experimental FIPA Ontology Agent specification) is for the ontology reasoning agent to ask how to terms are related in an ontology. The Ontology Reasoning Agent uses this query to build a minimum spanning tree (in each of the three GO ontologies) between all the terms returned in all the homologies from all of the GO organism databases. This inforamation can then be used to propose a likely annotation, and to display all of the information graphically for the biologist via the interface agent.

Interface Agents. The functional interface agent/applet consists of two columnar panes: on the left, the top pane displays the gene being annotated, and the bottom displays the general homolgies from GenBank with their natural language annotations. On the right, three panes display the subtrees from the three GO ontologies (biological process, molecular function, cellular location) marked in color with the homologs from the three organism databases.

4.3 EST Processing

One way to broaded the applicability of the system is to accept more kinds of basic input data to the annotation process. For example, we could broaden the reach of the sys-

tem by starting with ESTs (Expressed Sequence Tags) instead of complete sequences. Agents could wrap the standard software for creating sequences from this data, at which point the existing system can be used. The use of ESTs is part of a relatively inexpensive approach to sequencing where instead of directly sequencing DNA we instead sequence the genes being expressed in the cell (mRNA) using a method that produces many short sequences that partially overlap. The sequences are actually produced by attaching luminous markers (a different color for each of the four nucleotides A, T, C, and G) and then reading the resulting spectrographs. By finding the overlaps in the short sequences, we can eventually reconstruct the entire sequence of each expressed gene. Essentially, this is a "shotgun" approach that relies on statistics and the sheer number of experiments to eventually produce complete sequences. Figure 5 shows a multi-agent subsystem for automating the processing of ESTs to produce consensus gene sequences.

As a side effect of this processing, information is produced that can be used to find Single Nucleotide Polymorphisms (SNPs). SNPs indicate a change of one nucleotide (A,T,C,G) in a single gene between different individuals (often, conserved across strains or subspecies). These markers are very important for identification even if they do not have functional effects.

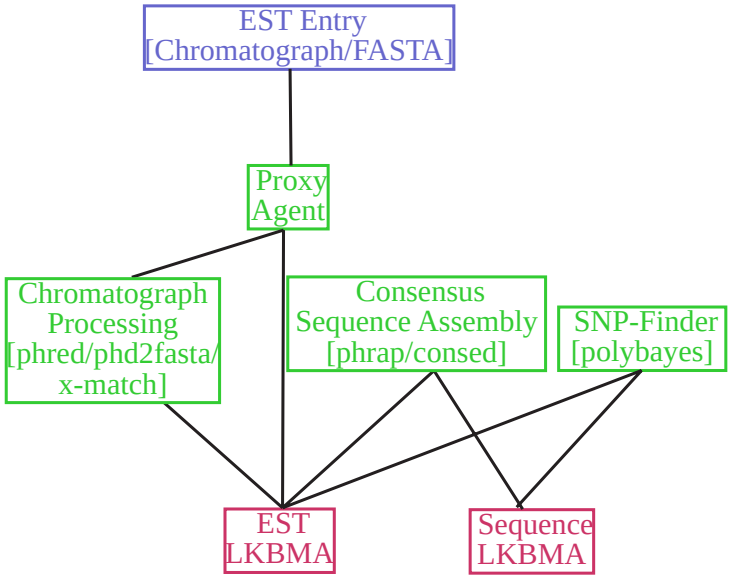

Fig. 5. EST Processing Agent Organization

Information Extraction Agents. The process of consensus sequence building and SNP identification does not require any external information, so the only IEAs are the LKBMAs. Up until now, there has only been one LKBMA, responsible for the gene sequences and annotations. EST processing adds a second LKBMA responsible for storing the ESTS themselves and the associated information discussed below. Primarily,

this is because (especially early on in a sequencing projoect) there will be thousands of ESTs that do not overlap to form contiguous sequences, and that ESTs may be added and processed almost daily.

Task Agents. There are three new domain-level task agents. The first deals with processing chromatographs. Essentially the chromatograph is a set of signals that indicate the relative strengths of the wavelengths associated with each luminous nucleotide tag. Several standard Unix analysis programs exist to process this data, essentially "calling" the best nucleotide for each position. The chromatograph processing agent wraps three analysis programs: Phred, which "calls" the chromatograph and also separatly produces an uncertainty score for each nucleotide in the sequence; phd2fasta which converts this output into a standard (FASTA) format; and x-match which removes a part of the sequence that is a byproduct of the sequencing method, and not actually part of the organism sequence. The consensus sequence assembly agent uses two more programs (Phrap and consed) on all the ESTs found so far to produce a set of candidate genes by appropriatley splicing together the short EST sequences. This produces a set of candidate genes that can then be added to the gene sequence LKBMA and from which the various annotation processes described earlier may commence. Finally, a SNP-finder agent operates the PolyBayes program which uses the EST and Sequence KBs and the uncertainty scores produced by Phred to nominate possible single nucleotide polymorphisms. Each of the wrapped programs (especially Phred, Phrap, and PolyBayes) has a large number of parameters to control. Currently we set these in consultation with our biologist parteners, but as we get more experience we plan to experiment with various automated learning mechanisms for parameter adjustment.

Interface Agents. There is only one simple interface agent, to allow particpants to enter data in the system. Preferrably, this is chromatograph data from the PCR sequencers, because the original chromatograph allows Phred to calculate the uncertainty associated with each nucleotide call. However, FASTA-format (simple "ATCG…" named strings) ESTs called from the orignal chromatographs can be accomodated. These can be used to build consensus sequences, but not for finding SNPs.

5 Gene Expression Processing

A new kind of genomic data is now being produced, that may swamp even the amount of sequenceing data. This is so-called *gene expression* data, and indicates quantitatively how much a gene product is expressed in some location, under some conditions, at some point in time. We are developing an multi-agent system that uses available online genomic and metabolic pathway knowledge to extend gene expression analysis. By incorporating known relationships between genes, knowledge-based analysis of experimental expression data is significantly improved over purely statistical methods. Although this system has not yet been integrated into the existing agent community, eventually relevant genomic information will be made available to the system through the existing GenBank and SwissProt IEAs. Metabolic pathways of interest to the investigator are identified through a KEGG (Kyoto Encyclopedia of Genes and Genomes) database wrapper. Analysis of the gene expression data is performed through an agent that executes SAS, a statistical package that includes clustering and PCA analysis meth-

ods. Results are to be presented to the user through webpages hyperlinked to relevant database entries.

Gene expression studies, which identify the set of active genes within a particular state, are beginning to explore the dynamic nature of intracellular behavior [13]. As cells respond to environmental or physiological changes of state, individual genes are induced and repressed and the corresponding messenger RNA (mRNA) and protein levels are modified, all according to a complicated but predetermined reaction network. However, interpreting these complex underlying networks from gene expression experiments is a difficult process. Current techniques for gene expression analysis have primarily focused on the use of clustering algorithms, which group genes of similar expression patterns together [14]. Based on the results of such analysis, researchers have identified conserved control regions (promoters) upstream of co-expressed genes that appear to be responsible for the similar expression behavior [7]. However, experimental gene expression data can be very noisy and the complicated pathways within organisms can generate coincidental expression patterns, which can significantly limit the benefits of standard cluster analysis. In order to separate gene co-regulation patterns from co-expression, the gene expression processing organization was developed to gather available pathway-level information in order to presort the expression data into functional categories. Thus, clustering of the reduced data set is much more likely to find genes that are actually regulated together. The system also promises to be useful in discovering regulatory connections between different pathways. A diagram outlining the agents within the system is shown in Figure 6. Although the system currently only uses the KEGG database to identify pertinent metabolic pathways, other sites are available and shall be incorporated in the future. One advantage of using the KEGG database is that its gene/enzyme entries are organized by the EC (Enzyme Commission) ontology, and so are easily mapped to gene names specific to the organism of interest. The gene expression database agent for this system currently queries a local copy of the publicly available S. cerevisiae (yeast) whole cell expression data and allows the user to access data related to diauxic shift, cell cycle, and sporulation. Once the experimental data has been reduced by the system to only those genes within the metabolic pathways of interest, the SAS statistical package is then used to cluster the remaining data to identify what are hopefully closely regulated genes within the organism. The co-regulation of genes is confirmed though the identification of conserved promoter motifs within each gene's genetic code. Final results of an analysis run will be presented to the user in graphic form showing clustered expression patterns along with hyperlinks to relevant on-line database entries for further exploration. Currently, we have automated access to the local database, the cluster analysis, and presentation to demonstrate the biological utility of the approach. Integration with the rest of the systems described here is planned.

6 Related Work

There has been significant work on general algorithms for query planning, selective materialization, and the optimization of these from the AI perspective, for example TSIMMIS [5], Information Manifold [20], Infosleuth [23], HERMES [1], SIMS [2],

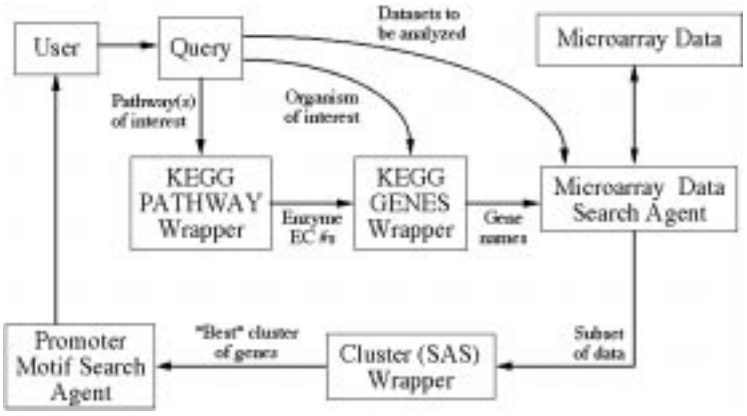

Fig. 6. Overview of gene expression processing organization

etc., and of course on applying agents as the way to embody these algorithms [21, 26, 11, 19].

In Biology, compared to the work being done to create the raw data, all the work on how to organize and retrieve it is relatively small. Most of the work in computer science directed to biological data has been in the area of heterogeneous databases, focusing on the semi-structured nature of much of the data that makes it very difficult to store usefully in commercial relational databases [6]. Some work has begun in applying the work on wrappers and mediators to biological databases, for example TAMBIS [25]. These systems differ from ours in that they are pure implementations of wrapper/mediator technology that are centralized, do not allow for dynamic changes in sources, support persistent queries, or consider secondary user utility in the form of time or other resource limitations.

Agent technology has been making some inroads in the area. The word "agent" with the popular connotation of a single computer program to do a user's bidding is found in the promotional material for Doubletwist (`www.doubletwist.com`). Here, an "agent" stands for a persistent query (e.g. "tell me if a new homolog is found in your database for the following sequence"). There is no collaboration or communication between agents.

We know of two truely multi-agent projects in this domain. First, InfoSleuth has been used to annotate livestock genetic samples [12]. The flow of information is very similar to our system. However, the system is not set up for noticing changes in the public databases, for integrating new data sources on the fly, or for consideration of secondary user utility. Secondly, the GeneWeaver project [4] is another true multi-agent system for annotation of genomes. GeneWeaver has as a primary design critereon the observation that the source data is always changing, and so annotations need to be constatly updated. They also express the idea that new sources or analysis tools should be easy to integrate into the system, which plays to the open systems requirement, although they do not describe details. The primary differences are the way in which an open system is achieved (it is not clear that they use agent-level matchmaking, but rather possibly

CORBA specifications) and that GeneWeaver is not based on a shared architecture that supports reasoning about secondary user utility. In comparision to the DECAF implementation, GeneWeaver uses CORBA/RMI rather than TCP/IP communication, and a simplified KQML-like language called BAL.

7 Discussion

The system described here is operational and normally available on the web at `http://udgenome.ags.udel.edu/herpes/`. This is a *real* working prototype, and so the interface is strongly oriented to biologists only. In general, computational support for the *processes* that biologists use in analyzing data is primitive (Perl scripts) or non-existant. In less than 10 min, we were able to annotate the HVT-1 sequence, as well as store it in a queryable and web-publishable form. This impressed the biologists we work with, compared to manual annotation and flat ASCII files. Furthermore, we have recently added approximately 25 other publicly available herpesvirus sequences (e.g. several strains of Human herpesvirus, African swine fever virus, etc.). The resulting knowledgebase almost immediately resulted in queries by our local biologists that indicated possible interesting relationships that may result in future biological work. This summer we will begin testing with viral biologists from other universities.

Other things about the system which have excited our biologist co-workers are the relative ease by which we can add new types of annotation or analysis information, and the fact that the system can be used to build similar systems for other organisms, such as the chicken. For example, the use of open system concepts such as a matchmaker allow the annotation agent to access and use new annotation services that were not available when it was initially written. Secondary user utility will become useful for the biologist when faced with making a simple office query vs. checking results before publication.

The underlying DECAF system has been evaluated in several ways, especially with respect to the use of parallel computational resources by a single agent (all of the DECAF components and all of the executable actions are run in parallel threads), and the efficacy of the DRU scheduler which efficiently solves a restricted subset of the design-to-criteria scheduling problem [17]. Running the gene annotation system as a truly multi-agent system results in true speedups, although most of the time is currently spent in remote database access (see Table 1). Parallel hardware for each agent will be useful for some of the more locally computationally intensive tasks involving EST processing.

BLASTP	PSort	Motif	Distr. BLASTP	Distr. PSort	Distrib. Motif
1994	1296	1833	775	346	809

Table 1. Average processing times on uniprocessor or distributed hardware

8 Conclusions and Future Work

In this paper we have discussed the very real problem of making some use of the tremendous amounts of genetic sequence information that are being produced. While there is much information publically available over the web, accessing such information is different for each source and the results can only be used by a single researcher. Furthermore, the contents of these primary sources are changing all the time, and new sources and techniques for analysis are constantly being developed.

We cast this sequence annotation problem as a general information gathering problem, and proposed the use of multi-agent sytems for implementation. Beyond the basic heterogeneous database problem that this problem represents, an MAS solution gives us mechanisms for dealing with changing data, the dynamic appearance of new sources, minding secondary utility characteristics for users, and of course the obvious distributed processing achievments of parallel development, concurrent processing, and the possibility for handling certain security or other organizational concerns (where part of the agent organization can mirror the human organization).

We currently are offering the system publically on the web, with the known herpesvirus sequences. A second system based on chicken ESTs should be available by the end of summer 2001. We intend to broaden the annotation coverage and add more complex analyses. An example would be the estimation of the physical location of the gene as well as its function. Because biologists have long recorded certain QTLs (Quantitative Trait Loci) that indicate that a certain *physical region* is responsible for a trait (such as chickens with resistance to a certain disease), being able to see what genes are physically located in the QTL region is a strong indicator as to thier high-level genetic function.

In general, we have not yet designed an interface that allows biologists to take full advantage of the materialized data —they are uncomfortable with complex query languages. We believe that it may be possible to build a graphical interface to allow a biologist, after some training, to create a commonly needed analysis query and to then save this for use in the future by that scientist, or others sharing the agent namespace.

Finally, the next major subsystem will be agents to link and analyze gene expression data (which will in turn interoperate with the metabolic pathway analysis systems described above). This data needs to be linked with sequence and function data, to allow more powerful analysis. For example, linked to QTL data, this allows us to ask questions such as “what chemicals might prevent club root disease in cabbage?”.

9 Acknowledgments

This work could not have happened without the work of Wei Chen, Hongru Cui, Foster McGeary, Mohamed Mostagir, and Furong Zhen on the functional annotation subsystem. This work was supported by the National Science Foundation under grants IIS-9812764, IIS-9733004 and BDI-0092336.

References

1. S. Adali and V.S. Subrahmanian. Amalgamating knowledge bases, III: Distributed mediators. *International Journal of Intelligent Cooperative Information Systems*, 1994.
2. Y. Arens and C.A. Knoblock. Intelligent caching: Selecting, representing, and reusing data in an information server. In *Proc. 3rd Intl. Conf. on Information and Knowledge Management*, 1994.
3. D.A. Benson and et al. Genbank. *Nucleic Acids Res.*, 28:15–18, 2000. `http://www.ncbi.nlm.nih.gov`.
4. K. Bryson, M. Luck, M. Joy, and D.T. Jones. Applying agents to bioinformatics in geneweaver. In *Proceedings of the Fourth International Workshop on Collaborative Information Agents*, 2000.
5. S. Chawathe, H. Garcia-Molina, J. Hammer, K. Ireland, Y. Papakonstantinou, J. Ullman, and J. Widom. The TSIMMIS project: integration of heterogeneous information sources. In *Proceedings of the Tenth Anniversary Meeting of the Information Processing Society of Japan*, December 1994.
6. S. B. Davidson and et al. Biokleisli:a digital library for biomedical researchers. *Intnl. J. on Digital Libraries*, 1(1):36–53, 1997.
7. K. Decker, X. Zheng, and C. Schmidt. A multi-agent system for automated genomic annotation. In *Proceedings of the 5th Intl. Conf. on Autonomous Agents*, Montreal, 2001.
8. K. S. Decker, A. Pannu, K. Sycara, and M. Williamson. Designing behaviors for information agents. In *Proceedings of the 1st Intl. Conf. on Autonomous Agents*, pages 404–413, Marina del Rey, February 1997.
9. K. S. Decker, K. Sycara, and M. Williamson. Middle-agents for the internet. In *Proceedings of the Fifteenth International Joint Conference on Artificial Intelligence*, pages 578–583, Nagoya, Japan, August 1997.
10. Keith S. Decker and Victor R. Lesser. Quantitative modeling of complex computational task environments. In *Proceedings of the Eleventh National Conference on Artificial Intelligence*, pages 217–224, Washington, July 1993.
11. Keith S. Decker and Katia Sycara. Intelligent adaptive information agents. *Journal of Intelligent Information Systems*, 9(3):239–260, 1997.
12. L. Deschaine, R. Brice, and M. Nodine. Use of infosleuth to coordinate information acquisition, tracking, and analysis in complex applications. Technical Report MCC-INSL–008-00, MCC, 2000.
13. J. L. DiRisi, V.R. Iyer, and P.O. Brown. Exploring the metabolic and genetic control of gene expression on a genomic scale. *Science*, 278:680–686, 1997.
14. M.B. Eisen, P.T. Spellman, P.O. Brown, and D. Botstein. Cluster analysis and display of genome-wide expression patterns. *Proc. Nat. Acad. Sci.*
15. Alan Garvey and Victor Lesser. Design-to-time real-time scheduling. *IEEE Transactions on Systems, Man, and Cybernetics*, 23(6):1491–1502, 1993.
16. J. Graham and K.S. Decker. Towards a distributed, environment-centered agent framework. In N.R. Jennings and Y. Lesperance, editors, *Intelligent Agents VI*, LNAI-1757, pages 290–304. Springer Verlag, 2000.
17. John Graham. *Real-time Scheduling in Multi-agent Systems*. PhD thesis, University of Delaware, 2001.
18. J. Hendler and Merwyn Taylor Kilian Stoffel. Advances in high performance knowledge representation. Technical Report CS-TR-3672, University of Maryland Institute for Advanced Computer Studies, 1996. Also cross-referenced as UMIACS-TR-96-56.
19. L. Kerschberg. Knowledge rovers: cooperative intelligent agent support for enterprise information architectures,. In P. Kandzia and M. Klusch, editors, *Cooperative Information Agents*, LNAI-1202. Springer-Verlag, 1997.

20. T. Kirk, A. Levy, J. Sagiv, and D. Srivastav. The information manifold. Technical report, AT&T Bell Labs, 1995.
21. C.A. Knoblock, Y. Arens, and C. Hsu. Cooperating agents for information retrieval. In *Proc. 2nd Intl. Conf. on Cooperative Information Systems*. Univ. of Toronto Press, 1994.
22. I. Muslea, S. Minton, and C. Knobloch. Stalker: Learning expectation rules for simistructured web-based information sources. In *Papers from the 1998 Workshop on AI and Information Gathering*, 1998. also Technical Report ws-98-14, University of Southern California.
23. M. Nodine and A. Unruh. Facilitating open communication in agent systems: the infosleuth infrastructure. In M. Singh, A. Rao, and M. Wooldridge, editors, *Intelligent Agents IV*, pages 281–295. Springer-Verlag, 1998.
24. A.S. Rao and M.P. Georgeff. BDI agents: From theory to practice. In *Proceedings of the First International Conference on Multi-Agent Systems*, pages 312–319, San Francisco, June 1995. AAAI Press.
25. R. Stevens and et al. Tambis: Transparent access to multiple bioinformatics information sources. *Bioinformatics*, 16(2):184–185, 2000.
26. K. Sycara, K. S. Decker, A. Pannu, M. Williamson, and D. Zeng. Distributed intelligent agents. *IEEE Expert*, 11(6):36–46, December 1996.
27. The Gene Ontology Consortium. Gene ontolgy: tool for the unification of biology. *Nature Genetics*, 25(1):25–29, May 2000.
28. T. Wagner, A. Garvey, and V. Lesser. Complex goal criteria and its application in design-to-criteria scheduling. In *Proceedings of the Fourteenth National Conference on Artificial Intelligence*, Providence, July 1997.