PROCESS INTEGRATION FOR BIO-MANUFACTURING GRID

ZHIQI SHEN*[1], HUI MIEN LEE[1], CHUNYAN MIAO[2], MEENA SAKHARKAR[3], ROBERT GAY[1] AND TIN WEE TAN[4]

[1]School of EEE, Nanyang Technological University, Singapore 639798;
[2]School of Computer Engineering, Nanyang Technological University, Singapore 639798;
[3]School of Mechanical and Aerospace Engineering, Nanyang Technological University, Singapore 639798 and [4]Department of Biochemistry, Yong Loo Lin School of Medicine, National University of Singapore, Singapore 117597
Email: Email: [1]{zqshen, eklgay, leeh0015}@ntu.edu.sg; [2]ascymiao@ntu.edu.sg; [3]mmeena@ntu.edu.sg and [4]tinwee@bic.nus.edu.sg

Recently there has been a great demand for extending high throughput life science research to bio-manufacturing. However, there is a gap from life science research to bio-manufacturing. Most existing bio-workflow tools/grid computing systems only provide an isolated solution to help bio-scientist to orchestrate bio-R&D operations such as bio-database query, bio-computation and analysis for biological problems in specific verticals. They are static and lack the ability to adapt in a dynamic changing environment. Manufacturing of biological materials such as diagnostics, therapeutics and prophylactics, or bio-manufacturing, typically involves many bioprocesses, each of which requires a set of bio-workflows to be choreographed. This is currently achieved by manually defining and managing the workflows for bioprocesses through different workflow tools. This paper proposes a novel goal-oriented approach to modeling bioprocesses, choreographing bio-workflows from different workflow tools, and to integrating agents, web services and workflows for automated execution. It demonstrates how a multi-agent system on a grid infrastructure can be further derived to adapt and automate complex bio-manufacturing workflow/processes in a dynamic changing environment. In this way, database access in a large data grids, high performance computing in a computational grid, and remote device control in a manufacturing grid, can be coupled to a supply chain management system to form a broad scale bio-manufacturing grid that streamlines the entire value chain from R&D, productisation and design of biological products.

1. INTRODUCTION

Recent outbreaks of highly contagious diseases worldwide highlight the urgent need to extend high throughput life science research to bio-manufacturing which aims to connect the current broad base of life science research to manufacturing areas and

translate the technological know-how and research output into designs and subsequently manufactured products in a timely and rapid manner, and if possible, in quasi-real time.

Collaborative life science R&D activities revolve around data integration to computation integration and to virtual bio-labs with complex laboratory information management systems (LIMS). Bioscientists often need to perform experiments using shared data resources distributed worldwide. In addition to raw data resources that are shared on the web, many different bio-tools/applications that operate on it have also been developed, most of them with restricted functionality and targeted at performing highly specific tasks. A bio-R&D activity is performed through a set of bioprocesses. Collaborative life science R&D not only requires the sharing of bio-data resources but also the integration of bio-operations using various bio-tools. Therefore, it is both challenging and vital to integrate bio-tools/applications in various bioprocesses.

With the emergence of service-oriented technology, it is expected that any entity in interconnecting bio-applications/tools will be viewed as a service, whether it is a bio-instrument device data acquisition operation, or a bio-application/tool for signal processing or analysis of acquired data or a bio-database transaction. Web/grid services are fast emerging as enabling technologies for seamless bio-application integration. Workflows that orchestrate the bio-operations using different tools are poised to automate various bioprocesses to increase efficiency and productivity. As a result, many attempts for using web service technology and workflow management systems to tackle the above issues have been made [1-4].

However, most of existing bio-workflow tools/systems only provide an isolated solution to help bio-scientist to orchestrate bio-R&D operations such as bio-database query, bio-computation and analysis for specific bio-problems of limited scope. They are static and lack the ability to adapt in a dynamic changing environment. In the case of bio-manufacturing, many bioprocesses are involved, each of which requires a set of bio-workflows to be choreographed. This is currently achieved by manually defining and managing the workflows for bioprocesses through different workflow tools. They are also often not well interconnected with the research and development, or the discovery and design process.

Most of current research on integration of the available bio-services is based on web service architecture. For example, Taverna [1] (part of MyGrid project) is a leading research project of the UK government's e-Science programme. The Taverna software is a workflow workbench that provides a language and software tools to facilitate easy use of workflow and distributed services within the e-Science community.

With the growing number of bioinformatics resources such as computational tools and information repositories being made available as Web services, the Taverna project aims to provide a modeling tool with graphic user interface for designing and constructing bioinformatics workflows on top of bio-services over the web. Taverna provides user-friendly interfaces for bio-scientists to select and compose web services in a sequential order to form a workflow. Each operation in the defined workflow is able to invoke a specific bio-web service. Integrated bio-services for a specific problem are realized by running the corresponding workflows.

Some other research groups also focus on bioinformatics workflows on top of the available bio-services. They provide similar tools that allow users to compose and execute workflows, such as Pegasys [2], Wildfire [3] and BioWBI [4]. One of the common limitations in these is that most of them do not support dynamic and adaptive workflows. Moreover, none of the existing workflow systems supports the integration of workflows defined by different workflow tools.

The "Integrated Bio-laboratory Manufacturing and Services System" was a national project of Singapore in 2005, aimed at investigating how an "Integrated Workflow Infrastructure" for offering manufacturing and services involving the integration of bioinstrumentation, measurement systems and related databases as web services, could be developed. Concurrently, since 1997, the Bioinformatics Centre of the National University of Singapore and subsequently, its spin-off company, KOOPrime, a Singapore based company, has developed a suite of products and solutions, KOOPlatform, which addresses the needs of life sciences processes, specifically in the area of genomics and proteomics research [5]. This suite is known as "Workflows for Life Sciences" originally developed for a GlaxoWellcome-funded natural products drug discovery research centre's IT operations.

Although attempts have been made to offer descriptions by manually categorizing the services and sharing related workflows, these workflow management systems only provide a partial solution to the service integration. Existing Bio-workflow systems such as Pegasys, Taverna, Wildfire, BioWBI and KOOPlatform have limited user-centric modeling, abstracting, reasoning and automation capabilities. They are useful for composing partial, low-level processes, but lack the ability to integrate and automate a complete pipeline from R&D to manufacturing.

To meet the above challenges, we present Goal Net [6], a goal-oriented approach for modeling, integrating and automating bioprocesses as well as bio-workflows that are adaptive to dynamic changing environments. It also

demonstrates how an agent-oriented system can be further derived to automate complex bio-manufacturing processes in a service-oriented environment.

The activities carried out according to a requirement are usually organized in groups of inter-related activities called processes that can be seen as a set of operations, rules and constraints specifying the steps that must be taken, and conditions that must be satisfied, in order to accomplish a given goal. This new methodology will lead to the design of an integrated workflow infrastructure to straddle across manufacturing and services involving integration of bioinstrumentation, measurement systems, related databases, software and web-based services.

2. IMPLEMENTATION

We have developed a prototype of bio-manufacturing system using the approach presented in this paper. In this system, we use Taverna and KOOPlatform as the bio-workflow systems. Web services that wrap different bio-services are orchestrated through the two workflow systems respectively. Goal Net is used to choreograph the workflows for modeling different bioprocesses (Figure 1).

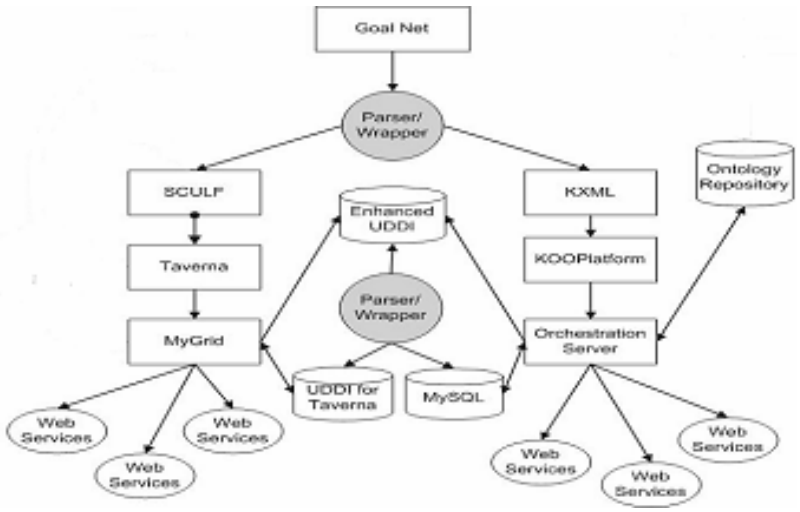

Figure 1. Architecture of the prototype system.

In this system, Taverna and KOOPlatform provide two sets of workflows. The extended UDDI provides common place for the services provided by the two systems. Goal Net provides a process integration platform for modeling bioprocesses and automating the bioprocess execution. The orchestration of existing bio-services not only needs a consistent definition of terminologies of bioprocesses but also the semantic linkage among various bioprocesses. The ontology repository is constructed in the system to store the defined consistent terminologies and concepts, and the semantics between the concepts.

The multi-agent development environment (MADE) that we have developed allows easy insertion of additional task libraries into the framework. The task libraries are the entry point where different functionalities can be added to MADE. To enable agents created by MADE to invoke web services and workflows in the two workflow systems, we have extended the MADE by adding two web service invocation components, Taverna integration component and KOOPlatform integration component.

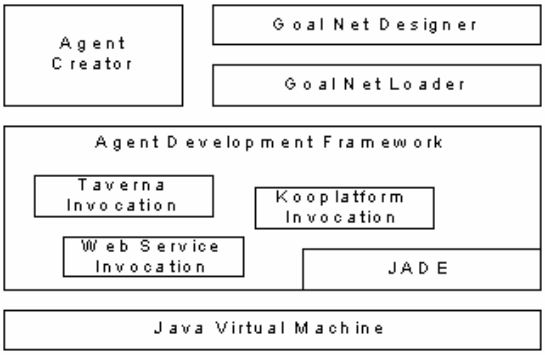

Figure 2. Structure of the extended MADE.

The Web service invocation component provides the API calls to the existing AXIS Web service tool [7] provided by Apache. AXIS is chosen because it is the latest Web service tool that provides good features with reasonable performance. By providing the API calls to the Taverna and KOOPlatform workflow systems, we have proven that the extended framework is able to act as a coordination framework for multiple atomic Web services and other existing Web services composition workflow models.

There are two types of storage methods for storing Goal Nets: XML-syntax description file and database storage. These two methods can be used to keep the

goals, arcs, transitions, attributes and their interconnection information for dynamic Goal Net loading. As shown in Figure 3, there is a File/DB access library to access Goal Net storage system. The user can specify the rules and configuration for running agents through the user rule configuration file. This user rule configuration file will keep the information that agents require in decision making using action selection and goal selection inferences. JADE platform acts as agent creator which agents can be dynamically generated during run-time. Agents' deployment and un-deployment can be done through JADE services too.

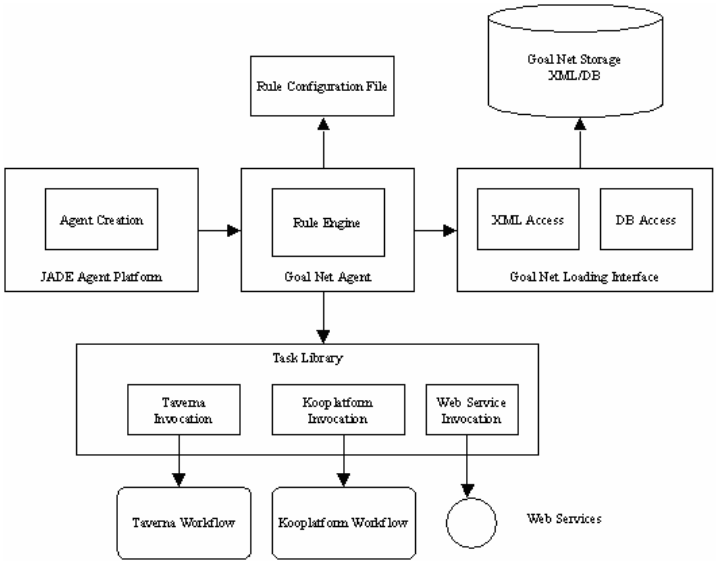

Figure 3. Agent-oriented architecture of the bioprocess execution system.

The action selection and goal selection mechanisms in the Goal Net model have been combined to become a rule engine. The Goal Net rule engine (Figure 3) makes decisions in goal selection and action selection based on the conditions defined in the user rule configuration file. Users need to specify the desired conditions and variables in the rule configuration file. Besides, users need to register tasks and the mapping of rules for the goal selection and action selection mechanisms to the rule engine. Finally web services and workflows can be invoked through the task libraries, that is, namely the web service invocation component, Taverna workflow invocation component and KOOPlatform workflow invocation component.

A typical scenario demonstrated by Taverna is that it provides a workflow to compare two genes X and Y. To illustrate our method, we designed a goal net to represent a process in which a ***compareXandY*** workflow will be invoked according to the user designed goal net. We then created an agent using the extended MADE and loaded the goal net to the agent. Figure 4 shows the running result of an invocation to a Taverna workflow from the agent.

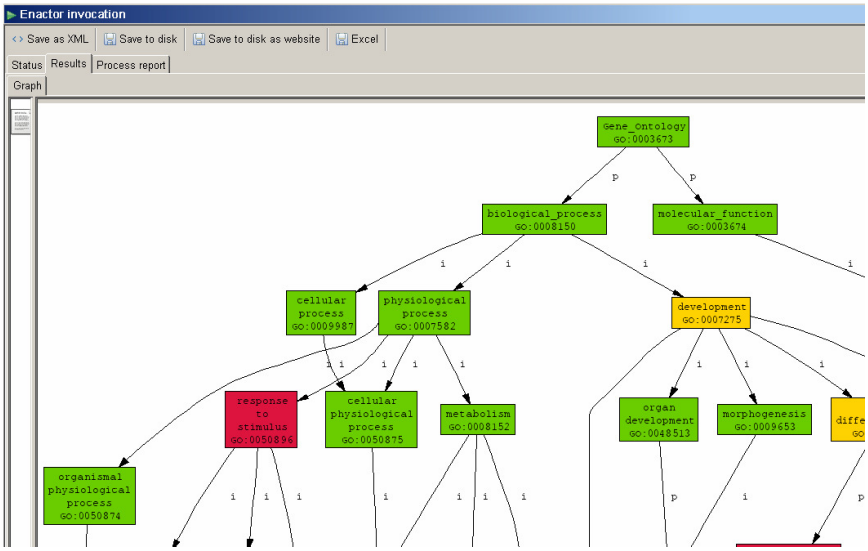

Figure 4. Goal Net agents can invoke Taverna workflows. The run results of the agent in Taverna are shown here.

Originally a user needed to prepare the data and invoke the workflows manually through the Taverna workbench, a GUI tool for Taverna workflow operations. With the system we have developed, previously designed and configured bioprocess can be stored in a database in the form of a goal net. In this way, a user only needs to create an agent and load the specific goal net to get the expected results and bypass the manual invocation. Furthermore the goal nets can be reused with different data for different requirements.

3. BIOPROCESS MODELING

A bioprocess is a specific ordering of activities with clearly identified inputs and outputs that achieve a certain goal. For example, a high level bio-manufacturing

process takes a sample of an unknown infectious disease such as SARS or bird flu and produces a diagnostic material such as a DNA chip for rapid development of diagnostics and eventually for design of RNAi (miRNA/siRNA) therapeutics and DNA/peptide vaccines. The activities involved in a bioprocess can be bio-workflow executions, web service invocations, or other bio-application executions. The activities of a bioprocess and the order of the activities may be different in different situations. Currently most researchers in life science still manually manage the activities of a bioprocess, according to the current situation and based on their expertise, to adapt to the dynamic environment. In this paper, we adopted a goal-oriented approach to modeling bioprocesses by which life science researchers are able to transform their expertise to the process models. Then we build an agent oriented system to automate the process execution based on the process models.

A bioprocess model facilitates the alignment of bioprocess specifications with the technical framework that IT development needs. The challenges for modeling a bioprocess include:

1. The bioprocess model should capture relevant information consistently and thoroughly so that both life science researchers and the IT developers can understand the process requirements that are captured in the model.
2. The bioprocess model should capture alternatives and exceptions to standard in addition to normal operations.
3. The bioprocess model should be easily executed towards execution automation.

4. THE GOAL NET

In this paper, Goal Net is used to model the bio-manufacturing process. Goal Net is a composite goal hierarchy which is composed of goals and transitions. Round rectangles are used to represent the goals that agent need to go through in order to achieve its final goal. The transitions, represented by arc and rectangle, connect one goal to another specifying the relationship between the goals that it joins. Each transition must have at least one input goal and one output goal. Each transition is associated with a task list which defines the possible tasks that an agent may be required to perform in order to transit from the input goal to the output goal. Figure 5 shows a simple goal net.

Goal Nets can represent four types of basic temporal relationships between goals: sequence, concurrency, choice and synchronization. Sequence relationship represents a direct sequential relationship between one input goal and one output

goal; concurrency relationship means one goal has more than one next goals, and all its next goals can be achieved simultaneously; choice relationship specifies a selective connection from one goal to other goals; synchronization relationship specifies a synchronization point from different input goals to a single next goal. With different combinations of the basic temporal relations, Goal Net supports a wide range of complicated temporal relations among goals. This is one of the major differences between Goal Net and other goal modeling methods.

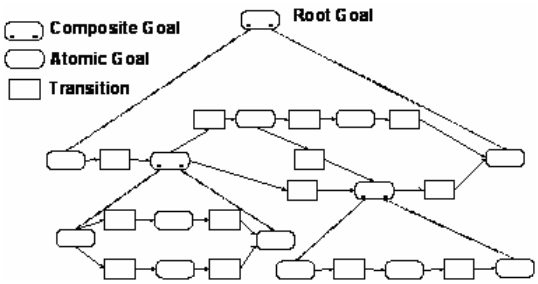

Figure 5. A simple Goal Net with two types of goals in Goal Nets, atomic goal and composite goal. An atomic goal accommodates a single goal which could not be split anymore; a composite goal may be split into sub-goals (either composite or atomic) connected via transitions.

In Goal Net, a composite goal needs to be decomposed into sub-goals. Here we do not intend to present a complete bio-manufacturing process model. Rather, we want to illustrate how a process can be mapped or modeled using Goal Net. A goal is a desired state that an agent intends to reach. Goal Net is an agent goal model. A goal net can be executed by an agent. We have proposed and developed an agent development framework based on Goal Net so that the created agent will refer to a goal net as its goal model to infer and guide its behaviors. When an agent is created, the agent has no goal and is running in an idle status, that is, it is an agent body. A goal net that represents a bioprocess is then loaded to the agent as its brain. Now the agent will start goal pursuit based on the goal net from the initial state to achieve the final goal of the bioprocess.

5. AUTOMATED BIOPROCESS EXECUTION

From an unknown infectious agent such as a deadly virus, to the elucidation of its complete genome, from its genome to the complete analysis and design of specific diagnostic DNA reagents, from the designed candidate diagnostics to the

manufactured biochemical products, to the testing of these products against clinical samples, to the fine tuning and optimization of the diagnostic material such as a DNA chip – each of these steps can be individually semi-automated for high throughput today. Yet no-one has attempted to connect these disparate and distributed steps into a complete workflow chain of manufacturing and design steps end-to-end according to the best of our knowledge. Each step in the process can be modeled and choreographed on a software platform to achieve a specific high-level goal. The steps are then all represented digitally as services over a grid to be made available to and callable by geographically distributed life scientists through integrated workflow orchestration system as services. Together with these services, the relevant resources comprising human operator, machinery, equipment, laboratory instrumentation, materials and computers will also be made online over the grid. These steps will capture all the processes spanning over the entire value chain together with their relationships from business to operational to manufacturing. They are captured modularly and at different granularity. Furthermore, these steps are all semantically aligned and they are mapped ontologically and semantically for process integration and compatibility. Most of the steps are generic enough to be reconfigurable in different workflows.

In a bioprocess model represented by Goal Net, each step is to pursue a goal. An invocation of workflow system or a web service indicates a transition from one achieved goal to the next goal. Workflows and web services are invoked as tasks of the transitions in a goal net. In this way, the individual workflows and web services are integrated for bioprocesses by using Goal Net. The execution of a goal net represents the automated execution of the bioprocess represented by the goal net.

6. AGENT DEVELOPMENT ENVIRONMENT AND EXECUTION PLATFORM

Figure 2 shows the multi-agent development environment (MADE) and the execution platform we have developed. In this figure, Goal Net Designer is a tool to design bioprocesses using Goal Net. Agent Creator is built on top of the agent development framework [8] to provide an agent development environment.

Goal Net Loader is an interface for users to load a goal net to a created agent. Users can use it to assign different goal nets to different agents according to the users' requirements. In addition, MADE has been enhanced by incorporating the popular agent development environment JADE [9], which is in compliance with industry standard FIPA [10], supports standard agent communication mechanism and provides multi-agent running platform. Through the integration with JADE,

MADE can provide goal model development environment and at the same time, supports standard agent communication mechanism and agent running platform.

7. RESULTS AND DISCUSSION

With Goal Net, the composition of each bioprocess is designed in order to achieve a specific goal. A bioprocess can be decomposed into a hierarchy of sub-processes and activities. These sub-processes and activities are then assigned to different recognized workflows during run-time. Hence we see here a combination of various processes taking place at different locations of the virtual organizations in order to achieve the global goal of the high level process. The problem of the supervision or coordination of such a process at its various levels of decomposition is critical and especially so in this context, where the process and activities are not limited to a single organization, but to a set of autonomous, distributed, and heterogeneous nodes that need to cooperate. With Goal Net, the supervision and coordination are automatically derived during the process decomposition phase.

The advantages of using Goal Net include:

1. GoalNet is a novel goal oriented process modeling tool which can decompose a complex process (goal) into executable sub-processes (sub-goals) for achieving a common goal. The temporal relationships between processes (goals) are modeled. This is the key difference between Goal Net and other goal-oriented models.
2. GoalNet has reasoning capability. An agent running a goal net can reason the next goal to pursue and the next task to take for achieving the selected goal based on the current situation. Therefore the agent can compose the low level workflows to form a complete pipeline in a dynamic changing environment.
3. GoalNet is also a multi-agent modeling tool by which a multi-agent system can be derived from the process modeling for automating the processes execution.

Goal Net therefore provides a rich set of relationships and goal/action selection mechanisms to achieve a dynamic and highly autonomous process integration model. This approach of process integration is viable, in which:

1. The interactions between bioprocesses, bio-workflows and web services are represented as a Goal Net.

2. A bio-workflow or web service operation is represented by a transition task and a goal shows that a particular goal is reached after the execution of the transition tasks.
3. Combinations of different relationships between goals, sub-goals, and transitions can be used to represent complex bioprocess logic.
4. The dynamic bioprocess flow is achieved by defining action selection and goal selection mechanisms.

The goal-oriented bioprocess modeling proposed in this paper can handle atomic Web services and other existing web service compositions such as the workflows defined in Taverna and KOOPlatform. This is easily achieved through calling the APIs provided by the external composition models as transition tasks of Goal Nets.

With such a composite goal hierarchy and various temporal relations within the hierarchy, a complex system can be recursively decomposed into sub-goals and sub-goal-nets. In such a manner, a system can be easily modeled and simplified. For example, Figure 6 shows a bio-manufacturing process. Figure 7 shows the goal net which models the process.

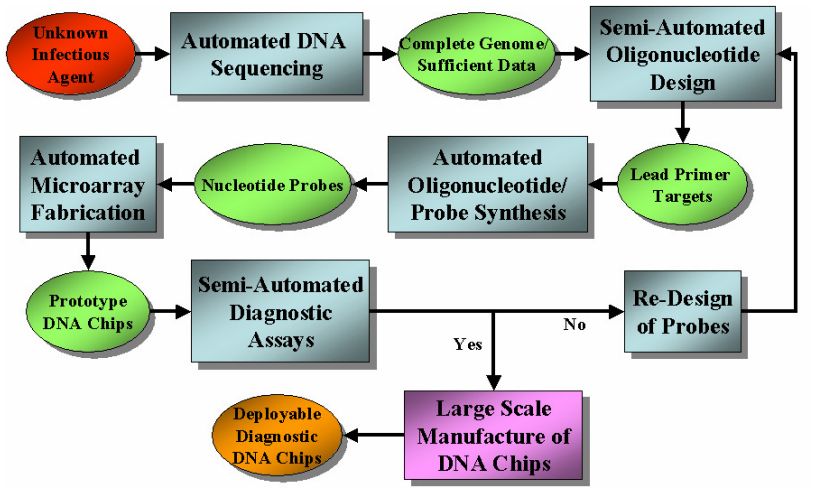

Figure 6. A bio-manufacturing workflow process with an underlying bio-manufacturing grid can be used to design, prototype and scale-up the manufacture of DNA chips against initially unknown infectious agents.

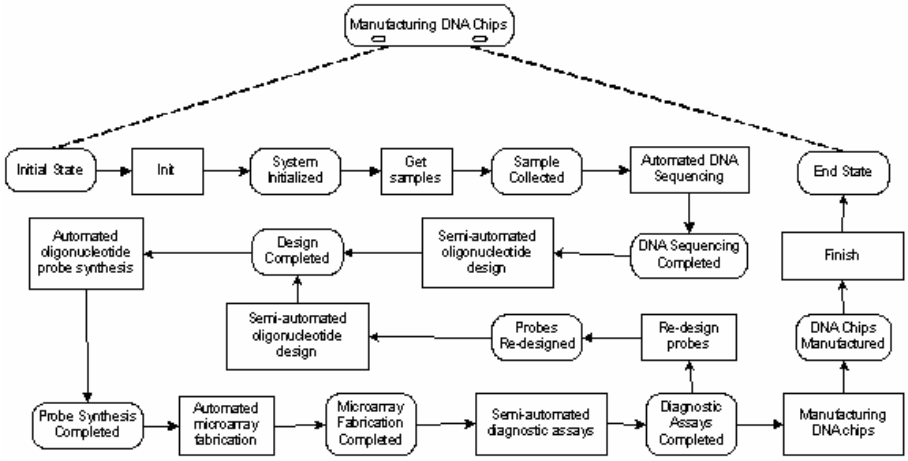

Figure 7. The corresponding Goal Net that models the bio-manufacturing workflow process in Figure 6.

8. CONCLUSION

To be able to integrate a complete pipeline from R&D to the manufacturing of diagnostic kits is not only an important advancement in terms of R&D value generated, but also vital in both economic and social context as it will allow us to build up a platform that can effectively respond against new outbreaks of infectious disease or bioterrorist attacks.

In this paper, we present a goal-oriented approach for bioprocess modeling and integration. A multi-agent platform for integrating bio-workflows and automating the bioprocess execution has been presented. The results generated by the developed prototype system shows that our goal to integrate various existing workflows and to automate the process execution using proposed approach has been achieved.

In fact, the core of biopharmaceutical manufacturing of PCR diagnostics, RNAi (RNA interference) agents, peptide vaccines etc, are processes which assemble linear polymers of biochemical monomers from a linear sequence of genetic information, mimicking the way each living cell does it. Automated DNA/RNA sequencers, oligonucleotide synthesizer machines, peptide synthesizers etc are readily available bio-instruments today with a plethora of service vendors on remote locations. Each instrument can be called to produce specific sequences by sending to the machines a text file generated from database searches and bioinformatics computation. Each system relies on standard sets of reagents and buffer solutions, which constitute the supply chain manufacturing and management system which,

when integrated with these systems, will allow high throughput, automated or semi-automated manufacturing to take place.

Software Availability and Requirements

The source code and the executable tool are available at the site http://www.ntu.edu.sg/home/zqshen/imss-bio. The system requires Taverna 1.0, KOOPlatform 4.0, Java Runtime Environment 1.5 or higher.

ACKNOWLEDGEMENTS

We thank Kuay Chongthong and Xiong Luying for having developed the prototype system during their study for Masters degrees. We also thank Lim Teck Sin from KOOPrime Pte Ltd for his support during our research.

REFERENCES

[1] Oinn T, Addis MJ, Ferris J, Marvin DJ, Greenwood M, Carver T, Wipat A and Li P. Taverna, lessons in creating a workflow environment for the life sciences. *In* GGF10; Berlin, Germany. (2004)

[2] Shah SP, He DYM, Sawkins JN, Druce JC, Quon G, Lett D, Zheng GXY, Xu T, and Ouellette BFF. Pegasys: software for executing and integrating analyses of biological sequences. BMC Bioinformatics, **5**:40. (2004)

[3] Tang F, Chua CL, Ho LY, Lim YP, Issac P and Krishnan A. Wildfire: distributed, Grid-enabled workflow construction and execution. BMC Bioinformatics **6**:69. (2005)

[4] Leo P, Marinelli C, Pappadà G, Scioscia G and Zanchetta L. BioWBI: an Integrated Tool for building and executing Bioinformatic Analysis Workflows. In BITS2004; Mar 26-27 2004; Padova, Italy. (2004)

[5] KOOPlatform [http://www.kooprime.com]

[6] Shen ZQ, Gay R, Miao CY and Tao XH. Goal Oriented Modeling for Intelligent Software Agents. *In* IEEE/WIC/ACM International Conference on Intelligent Agent Technology (IAT'04); September 20 - 24, 2004, Beijing, China. (2004)

[7] WebServices - Axis [http://ws.apache.org/axis/]

[8] Shen ZQ, Gay R, Miao CY and Tao XY. Goal Autonomous Agent Architecture. *In* 28th Annual International Computer Software and Applications Conference (COMPSAC'04), September 28 - 30, 2004, Hong Kong, China. (2004)

[9] Bellifemine F, Poggi A and Rimassa G. JADE: a FIPA2000 compliant agent development environment. *In* 5th International Conference on Autonomous Agents, Montreal, Quebec, Canada. pp 216 - 217. (2001)

[10] FIPA Agent Management Specification [http://www.fipa.org/specs/fipa00023/]