A Programming Environment for Global Activity-based Application

Emanuela Merelli
Dip. Matematica e Informatica
Università di Camerino

joint work with

Flavio Corradini
Dip. di Informatica
Università di L’Aquila

Leonardo Mariani
Dip. di Informatica, Sistemi e Comunicazione
Università di Milano Bicocca

WOA03
Villasimius
Outline

• Activity-based Applications in the Bioinformatics Domain
• Workflow-based Activity Coordination
• The Proposed Approach
  • 3-layered Software Architecture for the Agent-based Middleware
  • User Level to Agent Level Wf Compiler
  • Application-agents and Wrapper-agents cooperative environment
• Future Activities and Conclusion
Activity-Based Application in Bioinformatics Domain

**Problem:** To find the crystallographic structure of the 10 proteins more similar to that of a given genetic sequence, e.g., `atggag ... tga`.

**Objective:** To use several Bioinformatics Software Tools available on Internet in order to find the wanted result.

1. **1st Activity**
   - Select the 10 proteins more similar to the given sequence `atggag ... tga`
   - by using BLASTn in GenBank at NCBI

2. **2nd Activity**
   - Search for the PDB ID (crystallographic structure identifier) of each selected proteins,
     - by using BLASTp in SWISS-PROT at EMBL-EBI
     - by retrieving from PubMed via Entrez Retrieval System at NCBI, abstracts containing PDB-ID information

3. **3rd Activity**
   - Search for the Crystallographic Structure of any selected PDB ID
     - find 3-D biological macromolecular structure in Protein DataBank repository

4. **4th Activity**

**Aim:** the automatic execution of the activities by coordinating several Bioinformatics tools, freeing the Bioscientist from the need to know details on data and to continuous interact with remote sites.

By defining an integrated environment supporting activity coordination data and services integration
large scale distributed management system in order to automatically execute an experiment.
Tools suitcase

- **Wf Management**
  - **Wf Editor** - modelling and definition
  - **Wf Checker** - analyzing consistency of the model
  - **Wf Compiler** - translating model to executable code
  - **User interface**: Web-based GUI, Console, ...

- **System Management**
  - Middleware platform deployment tool
  - Account management tool
  - System’s Diagnosis tool
  - System maintenance tool
  - Traceability tool
  - ...

Workflow-based Activity Coordination

Select similar sequences

1st Activity

select PDB structures from sequences

CONIF: PDB

Sequence=sequence
BlastVersion="Blastp"

2nd Activity

Find crystallographic structures from PDB Structures

CONIF: PDB

array1=structure
array2=structure

Find crystallographic structures from articles

CONIF: PDB

Articles=articles

3rd Activity

Select articles from sequences

CONIF: ARTICLES

Sequence=sequence

4th Activity

Union

content=structures

FTP: Upload Result

CONIF: FTP

SIMILARITY
Tool=Blast!
SeqRepository="GenBank"

FTP: File://www.unicam.it
UserName='unicam'
Password='unicam'

PDB
PDBRepository="SIMS/PRT"
ARTICLES
ArticleRepository="PubMed"
The proposed approach
From Data to Knowledge and vice versa

- data + algorithms
- Different data format
- XML elements
- meta-data

ontologies (human concepts) + workflow

Information + coordination

data + algorithms
### System’s software architecture

<table>
<thead>
<tr>
<th>User Application Workflow</th>
<th>User Layer</th>
</tr>
</thead>
<tbody>
<tr>
<td>Workflow Management</td>
<td></td>
</tr>
<tr>
<td>Application Agents</td>
<td>System Layer</td>
</tr>
<tr>
<td>Application Agents Management</td>
<td></td>
</tr>
<tr>
<td>Service Agents</td>
<td>Run-Time Layer</td>
</tr>
<tr>
<td>Core Level</td>
<td></td>
</tr>
</tbody>
</table>
A general system’s architecture
**Agent-based System Architecture**

**User Layer**
- User Application Workflow
- Workflow Mng
- Long-transaction

**System Layer**
- User Agents
- Knw Mng Agent Service
- Wrapper Agent
- Retrieval Service
- Temporary Data Repository

**Run-Time Layer**
- Remote data format
  - EMBL
  - FASTA
  - ASN.1
  - GenBank
  - RDB
  - HTML
  - XML
  - TXT
  - …

**Remote data format**
PEGAA: The Agent-based Middleware
Activity Database (ADb)

MaxSequences: Integer
SequenceType: seqType
newSequence: String

Select similar sequences

CONGIF: SIMILARITY

Sequences: String[]
BlastVersion: String

select PDB structures from sequences

CONGIF: PDB

PDBlist : PDBstruct[]

Find crystallographic structures from PDB structures

CONGIF: PDB

Articles: File[]

Find crystallographic structures from articles

CONGIF: PDB

File[]

CONGIF: ARTICLES

Articles: File[]

CONGIF: ARTICLES

Cstructure[]

Cstructure[]

array1: []
array2: []

Union

[]

CONGIF: FTP

FTPUploadResult

content: ANY

FTPUploadResult

content: ANY

FTP

FTPSite: String
UserName: String
Passwords: String

SIMILARITY
Tool: String
SeqRepositories: String[]

FTP
FTPSite: String
UserName: String
Passwords: String

POB
PDBRepositories: String[]

ARTICLES
ArticleRepositories: String[]
User Level Workflow

GLOBAL VARIABLES:
- sequence: String
- PDBs: PDBarray
- Structures1, Structures2, Structure3
- articles: File

MaxSequence = "10"
SequenceType = "mRNA"
nextSequence = "ME"...

CONIF: SIMILARITY

Select similar sequences

CONIF: PDB

Sequence = sequence
BlastVersion = "Blastp"

Select PDB structures from sequences

PDBe = PDBs

CONIF: ARTICLES

Sequence = sequence
Select articles from sequences

PDBlax = PDBs

Find crystallographic structures from PDB structures

Artikles = articles

Find crystallographic structures from articles

array1 = Structures1
array2 = Structures2
content = Structures1

Union

FTP Upload Result

SIMILARITY Tool: "Blast"
SeqRepository = "GenBank"

FTP Site: "www.unica.it"
Username = "unicam", Password = "unicam"

PDBe PDBRepository = "SISMS PDB"

ARTICLES ArticleRepository = "PubMed"
User level to Agent level Compiler

The compiler generates a pool of cooperating agents from the workflow specification.

Two steps compilation:
• Step 1: from user level Wf to agent level Wf
  • User Level Activity Database (ULAD)
    • contains the mapping from each user activity to an agent level Wf specification

• Step 2: agent synthesis
  • Database of Skeletons (DoS)
    • contains the “empty” implementation of an agent role (skeleton)
  • Agent-Level Activity Implementation Database (ALAID)
    • contains the implementation of each agent level activity

Context-awareness
The compiler takes advantage of information about the state of the global environment during the agent synthesis.
Compiler Step 1: Agent level workflow

Agent1 VARIABLES
- sequence: String []
- PDBs: PDB structures
- structures: structures

ULAD

select similar sequences

Agent1

Place = "GenBank"

MoveTo

Tool = "Blast"
CurrentRepository = "GenBank"
MaxSequences = "10"
Sequence Type = "mRNA"
new Sequence = "EP", "DSD"

Get sequence by similarity

sequence

Create Agent

sequence

Agent2 VARIABLES
- sequence: String []
- articles: File
- structures: structures

select articles from sequences

Agent2

Place = "PDB"

MoveTo

Tool = "Blast"
CurrentRepository = "SWISS-PROT"

Get sequences by similarity

PDBs

Place = "PDB"

MoveTo

sequences = sequence

Get articles containing sequences

articles
Compiler Step 2: executable application agents
Application-agents (WE) and Wrapper-agents (WA) cooperative environment
Wrapper-Agent: general scenario

- WA
  - QueryString: ……
  - XML
  - AIXO
  - HTML Web Page

- WA
  - ProgramOption: ……
  - XML
  - AIXO
  - Flat Files from Command Line Program

- WA
  - XML
  - AIXO
  - RDBMS
  - SELECT…. FROM… WHERE …..

Wrapper-based data integration: general scenario
Bioinformatics Tools

Tool 1:
Environment: NCBI (WebSite): html format
Data: GenBank (DB): proprietary format
Tool: BLASTn (Algorithm): Takes nucleotides sequences in FASTA format
Output: GenBank Format

Tool 2:
Environment: EMBL-EBI (WebSite): html format
Data: Swiss-Prot (DB): proprietary format
Tool: BLASTp (Algorithm): Takes protein sequences in FASTA format
Output: SwissProt format

Tool 3:
Environment: NCBI (WebSite): html format
Data: PubMed: ANS.1 format
Tool: Entrez Retrieval System: Takes protein name
Output: XML format

Tool 4:
Environment: Protein DataBank web site
Data: PDB (DB): proprietary format
Tool: FASTA (Algorithm)
Output: PDB format (3 dimensional)
AIXO - NCBI-Wrapper

Retrieval articles about P53 protein

Filter and Map + XSLT = XML

XML Trasl. + XML GRAMMAR = XML

TEXT

<...>
<entry>
  <ID name="P53_HUMAN" type="STANDARD" molecule="PRT" length="393"/>
  <AC value="P04637"/> <AC value="Q16848"/> <AC value="Q9UBI2"/>
  <DT day="13" month="AUG" year="1987" rel="05"/>
</entry>
BioAgent & O2I

The agent sends an XML request to AIXO specifying the wrapper to be selected:

- AIXOAgentService
- BioAgent
- Swiss_prot_abstract
- medline_abstract

The Swiss-Prot document is converted into an XML document. The XML output is obtained by an XSLT transformation.

http://www.bioagent.net
Future Activities and Conclusion

We are

- developing the first prototype of the compiler to allow the automatic generation of user-agents
- enriching the set of tools supporting workflow specification
- evaluating our approach in different application domains (Supply Chain Mng, Product quality Testing, ...)

We plan to use JADE as agent middleware

Projects

- Oncology Over Internet - O2I (MIUR Cofin.),
- Sistemi COoperativi Multiagente - SICOM (CIPE-Regione Marche),
- Components traceability for product testing - FarMAS (Gruppo Loccioni),
- Sahara (MIUR 40%)

Collaborations with industries

- AEA, Guzzini, Merloni, SIGMA SpA
P53 Crystallographic Structure
DNA and nucleotide sequence

atggaggagccgcagtcagatcctagctgcagcccccctctgagtcaggaaacattttca
M E E P Q S D P S V E P P L S Q E T F S
gacctatggaaaactacttcttgtgaaaaaacacgttctgtcccccttgccgtccccagcaatg
D L W K L L P E N N V L S P L P S Q A M
gatgatttgatgtgtccccggacgcagatattgaacaatgttcaactgaagacccaggttca
D D L M L S P D D I E Q W F T E D P G P
gatgaagctcccagaatgcaccagaggctgtctccccggctggcccccttgaccagcagctctct
D E A P R M E A A P R V A P G P A A P
acaccggcggccctctgacaccagcccccccctctggccccctgtcatcttcctgtcccttccag
T P A A P A P A P S W P L S S S V P S Q
aaaacctaccaggccagctacggtttccgtctgggcttctgattctgggacagccaag
K T Y Q G S Y G F R L G F L H S G T A K
tctgtgacttgcacgtactcccccctgcacccacactgtgctttgcattctggtgaccagcagccag
S V T C T Y S P A L N K M F C Q L A K T
tgccccgtgcagctgtgggtgtgattccacaccccccgccccgccacccccgctgcgccccatg
C P V Q L W V D S T P P P G T R V R A M
gccatctacaagcagtcacagcactgacggagggtgttgaggcgctgccccaccatgag