An agent-based layered middleware as tool integration

Flavio Corradini
University of L'Aquila
ITALY

Leonardo Mariani
University of Milano
ITALY

Emanuela Merelli
University of Camerino
ITALY

Helsinki
FSE/ESEC 2003
Tool Integration Workshop
Outline

- The Tool Integration problem in the Bioinformatics Domain
- The Workflow-based Task Coordination (High Level Tool Integration)
- The Wrapper-based Data Integration (Low level Tool Integration)
- The Proposed Approach:
  An Agent-based Middleware for Tool Integration
- Preliminary Results
- Future Activities and Conclusions
The Tool Integration problem in Bioinformatics Domain

Problem: To find the crystallographic structure of the 10 proteins more similar to a new genetic sequence, e.g. X=MEEP ... DSD,

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

1. Select the 10 proteins more similar to the X=MEEP ... DSD sequence
   • by using BLASTn in GenBank at NCBI

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

Aim: To integrate the four Bioinformatics tools freeing the Bioscientist from the need to continous interact with remote sites.

For Tool Integration we mean
1) Supporting Tasks Coordination
2) Allowing Data Integration

in order to automatically execute an experiment
Wrapper-based System: general scenario

QueryString: ……

ProgramOption: …..

SELECT.... FROM... WHERE .....
Wrapper-based System: Bioinformatics Tools

**Tool 1:**
- Environment: NCBI (WebSite): html format
- Data: GenBank (DB): proprietary format
- Tool: BLASTn (Algorithm): Takes nucleotides sequences in FASTA format, GenBank Accession numbers or GI numbers and compares them against the NCBI nucleotide databases
- 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, GenBank Accession numbers or GI numbers and compares them against the NCBI protein databases.
- Output: FASTA Format

**Tool 3:**
- Environment: NCBI (WebSite): html format
- Data: PubMed & MEDLINE: ANS.1 format
- Tool: Entrez Retrieval System
- Output: XML

**Tool 4:**
- Environment: Protein DataBank web site
- Data: PDB(DB): proprietary format
- Tool: FASTA (Algorithm):
- Output: FASTA Format
Wrapper-based System: Retrieval MedLine articles about P53 proteine

Filter and Map

XML Trasl.

Access

<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>
Wrapper-based System: the software architecture (AIXO)

- **DataSource**: HTTP, RDBMS, Command Line program, …
- **ResourceToXMLReader**: HTML, FlatFile, …
The Tool Integration Problem in Activity-Based Applications

**Problem:** To integrate and coordinate multiple software tools for retrieving and integrating heterogeneous, distributed and frequently redundant data.

**Objective:** To integrate and coordinate several software tools in order to provide a uniform way and an high level of abstraction for users.

**Aim:** To define an *integrated environment* freeing the user from the need to know details on data repository and to coordinate the intermediate steps of an experiment (tasks).

**Proposed Approach:** To define an application as a workflow of tasks; to coordinate the execution of cooperative tasks by using software agent tools.
## 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>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Application Agents</th>
<th>System Layer</th>
</tr>
</thead>
<tbody>
<tr>
<td>Application Agents Management</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Service Agents</th>
<th>Run-Time Layer</th>
</tr>
</thead>
<tbody>
<tr>
<td>Core Level</td>
<td></td>
</tr>
</tbody>
</table>
A general system’s architecture

User Layer
- User Application Workflow
- Web Services
- Knowledge Base

System Layer
- Workflow Mng
- High Level Integration Module
- Low Level Integration Module
- Temporary Data Repository
- Retrieval Service

Run-Time Layer

Remote Place

where Tools are available
Agent-based System Architecture

User Layer
- User Application Workflow
- Workflow Mng
- Long-transaction
- Short-transaction
- Web Services

System Layer
- User Agents
- Knw Mng Agent Service
- Retriev Service
- Wrapper Agent

Run-Time Layer
- Knowledge Base
- Temporary Data Repository
- Remote data format
- EMBL
- FASTA
- ASN.1
- GenBank
- RDB
- HTML
- XML
- TXT
...
From Data to Knowledge and vice versa

- data + algorithms
- Information + coordination
- Different data format
- XML elements
- meta-data
- ontologies (human concepts) + workflow

tool Integration
The Proposed Approach: an Agent-based Middleware
Preliminary Results: User-agent as high level Tool Integration
Preliminary Results: Wrapper-agent as low level Tool Integration
BioAgent
Future Activities and Conclusions

For different application domains (i.e supply chain, components traceability for testing…) we plan to:

- Develop wrapper agents
- Design and develop the knowledge database to manage software tools
- Develop the compiler to allow the automatic generation of user-agents
- Evaluate the possibility to include mobility to user-agents in order to minimize the data transfer during tasks execution.

We conclude saying that software tool integration for real applications, as those in Bioinformatics domain, is a very difficult task due to both heterogeneity of data format and wide variety of tools which continuously evolve.
NCBI – main databases

NCBI (National Center for Biotechnology Information) is a database resource funded by the US National Institutes of Health. NCBI creates and maintains databases, conducts research in computational biology, develops software for analyzing genome data, and distributes biomedical information. NCBI is a valuable resource for researchers and scientists in the field of biotechnology and molecular biology.
NCBI - Entrez

Entrez is a retrieval system for searching several linked databases. It provides access to:

- PubMed: biomedical literature
- Nucleotide: sequence database (GenBank)
- Protein: sequence database
- Structure: three-dimensional macromolecular structures
- Genome: complete genome assemblies
- Books: bookshelf online books
- Domains: conserved domains (CDD)
- 3D Domain: domains from Entrez Structure
- GEO: Gene Expression Omnibus
- OBO: Ontology for Biomedical Investigations
- OBO Summary: curated OBO data sets
- Journal: journals in Medline
- MeSH: medical subject headings
- NCBI Tools: NCBI Web site search
- OMIM: Online Mendelian Inheritance in Man
- PMD: full-text digital archive of life sciences journal literature
- PopSet: population study databases
- SNP: single nucleotide polymorphisms
- Taxonomy: organisms in GenBank
- UniGene: gene-oriented clusters of transcript sequences
- Unigene: markers and mapping data

Pre-computed similarity searches are available for most databases, which produces a list of related items, version, structure matches, as well as related articles.
NCBI - BLAST

Basic Overview

- What is BLAST?
- What is the Best BLAST for me?
- How can I access the BLAST services at NCBI?
- What is the difference between BLAST and Basic BLAST Search?

Q: What is BLAST?

BLAST (Basic Local Alignment Search Tool) is a set of similarity search programs designed to explore all of the available sequence databases regardless of whether the query is protein or DNA. The BLAST programs have been designed for speed, with minimal sacrifice of sensitivity to detect sequence relationships. The scores assigned in a BLAST search have a well defined statistical interpretation, making real matches easier to distinguish from random background hits. BLAST uses a heuristic algorithm that seeks local as opposed to global alignments and is therefore able to detect relationships among sequences that share only isolated regions of similarity (Altschul et al. 1990). For a better understanding of BLAST, you can refer to the BLAST Course, which explains the basics of the BLAST algorithm. There is also a description of BLAST services located here. Also for details on BLAST and theory of similarity search, see the References section.

Q: What is the Best BLAST for me?

There are many different types of BLAST available from the main BLAST website. Choosing the right one depends on the type of sequence you are searching with (i.e., gene, nucleotide, protein), and the desired database. In order to help guide you to the right service for your search please consult the BLAST Program Selection Guide.

This guide provides information on what database and BLAST service (megabLAST, translation BLAST etc.) is ideal for your search. There is also additional information describing how the tools work and why they are best for certain types of searches.

Q: How can I access the NCBI's BLAST services?
NCBI – ASN.1

ASN.1, or Abstract Syntax Notation One, is an international Standard Organization (ISO) data representation format used to achieve interoperability between platforms. NCBI uses ASN.1 for the storage and retrieval of data such as nucleotide and protein sequences, structures, genomes, and MEDLINE records. It permits computers and software systems of all types to reliably exchange both the data structure and content.

The documentation for the NCBI Software Development Toolkit contains information about the ASN.1 specification. The toolkit documentation is quite large (over 300 pages) and can be downloaded as the MS Word file "toolkit.doc" from the NCBI FTP site. Please see the "readme" file in that directory for more information.

The NCBI Software Development Toolkit (known as the "NCBI Toolbox") is a set of software and data exchange specifications used by NCBI to produce portable, modular software for molecular biology. The software in the Toolbox is primarily designed to read ASN.1 format records. It is available to the public in the toolbox/ncbi_tools directory of NCBI's ftp site, and can be used in its own right or as a foundation for building tools with similar properties. The readme files in the toolbox and toolbox/ncbi_tools directories of the FTP site contain more information about the toolbox and ASN.1.

The following publications might also be of interest:
NCBI - fomats
• PDB-ID (P53) = 1TSR

• www.rcsb.org (}
DNA and nucleotide sequence

```
DNA sequence:
atggaggagccg cagtcagatcctagcgtcgagccccctctgagtcaggaaacattttca
M E E P
Q S D P S V E P P L S Q E T F S
gacctatggaaactacttcctgaaaacaacgttctgtcccccttgccgtcccaagcaatg
D L W K L L P E N N V L S P L P S Q A M
gatgatttgatgctgtccccggacgatattgaacaatggttcactgaagacccagggtca
D D L M L S P D D I E Q W F T E D P G P
gatgaagctcccagaatgccagaggctgctccccgcgtggcccctggaccagcagctcct
D E A P R M P E A A P R V A P G P A A P
cacccggcggcccctgcaccagccccctcctggcccctgtcatcttctgtcccttcccag
T P A A P A P A P S W P L S S S V P S Q
aaaacctaccagggcagctacggtttccgtctgggcttcttgcattctgggacagccaag
K T Y Q G S Y G F R L G F L H S G T A K
tctgtgacttgcacgtactcccctgccctcaacaagatgttttgccaactggccaagacc
S V T C T Y S P A L N K M F C Q L A K T
tgccctgtgcagctgtgggttgattccacacccccgcccggcacccgcgtccgcgccatg
C P V Q L W V D S T P P P G T R V R A M
gccatctacaagcagtcacagcacatgacggaggttgtgaggcgctgcccccaccatgag
A I Y K Q S Q H M T E V V R R C P H H E
cgctgctcagatagcgatggtctggcccctcctcagcatcttatccgagtggaaggaaat
R C S D S D G L A P P Q H L I R V E G N
ttgcgtgtggagtatttggatgacagaaacacttttcgacatagtgtggtggtgccctat
L R V E Y L D D R N T F R H S V V P Y
```
Significant References

Y. Papakonstantinou, H. Garcia-Molina & J. Widom ’95

OEM: Object Exchange Across Heterogeneous Information Sources

S. Bergamaschi … ’00

Momis: Mediator environment for Multiple Information Sources

G. Cabri, L. Leonardi & F. Zambonelli ’00

MARS: A Programmable coordination Architecture for Mobile Agents

…

AIXO: Any Input XML Output, a generalized wrapper

E. Bartocci, L. Mariani & E. Merelli ’03

PEGAA: A Programming Environment for Global Activity-based Applications

F. Corradini, L. Mariani & E. Merelli ‘03