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 EBI
2. *Search for the PDB ID (crystallographic structure identifier) for 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

1st Tool

2nd Tool

3rd Tool

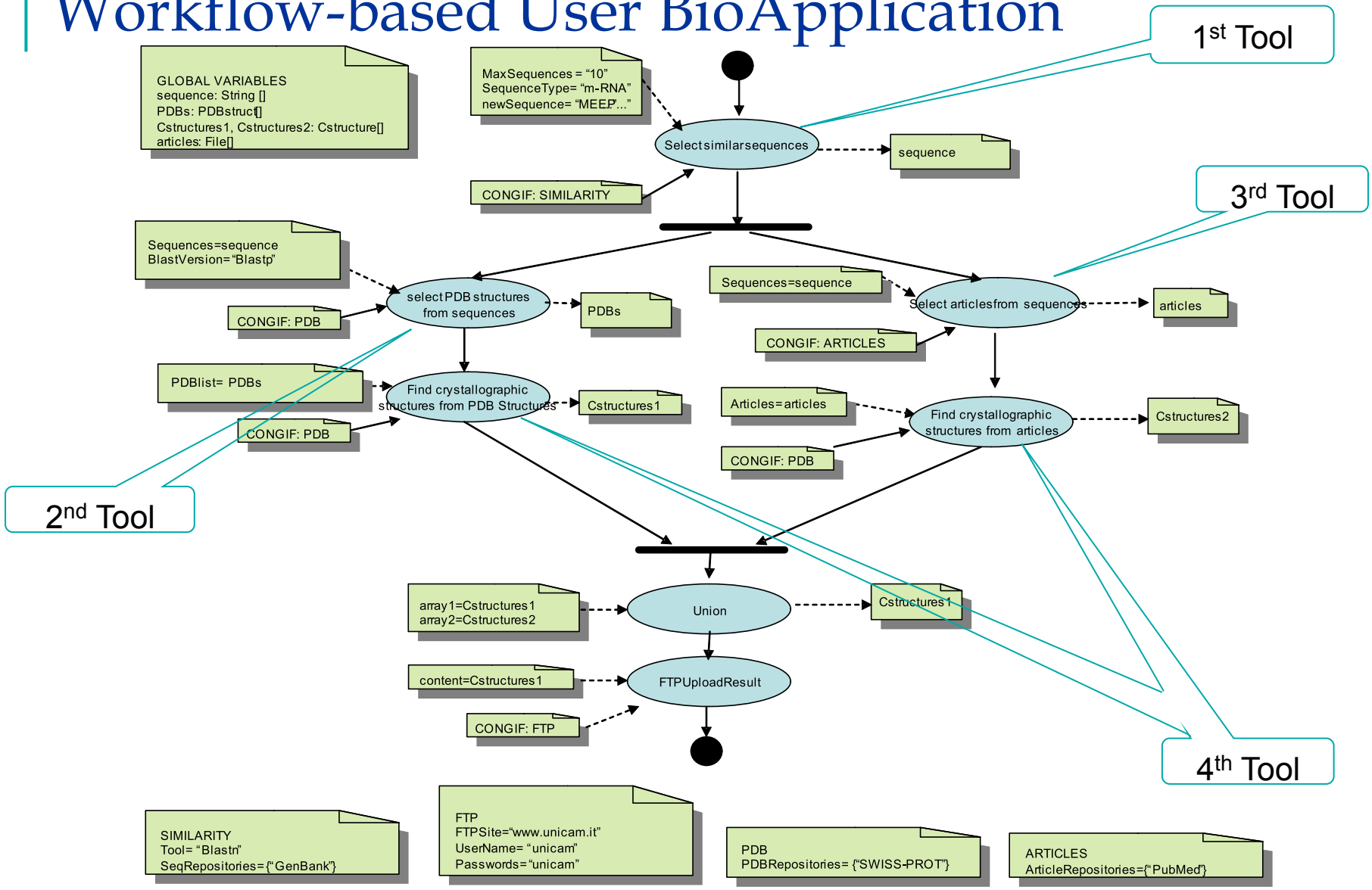
4th Tool

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

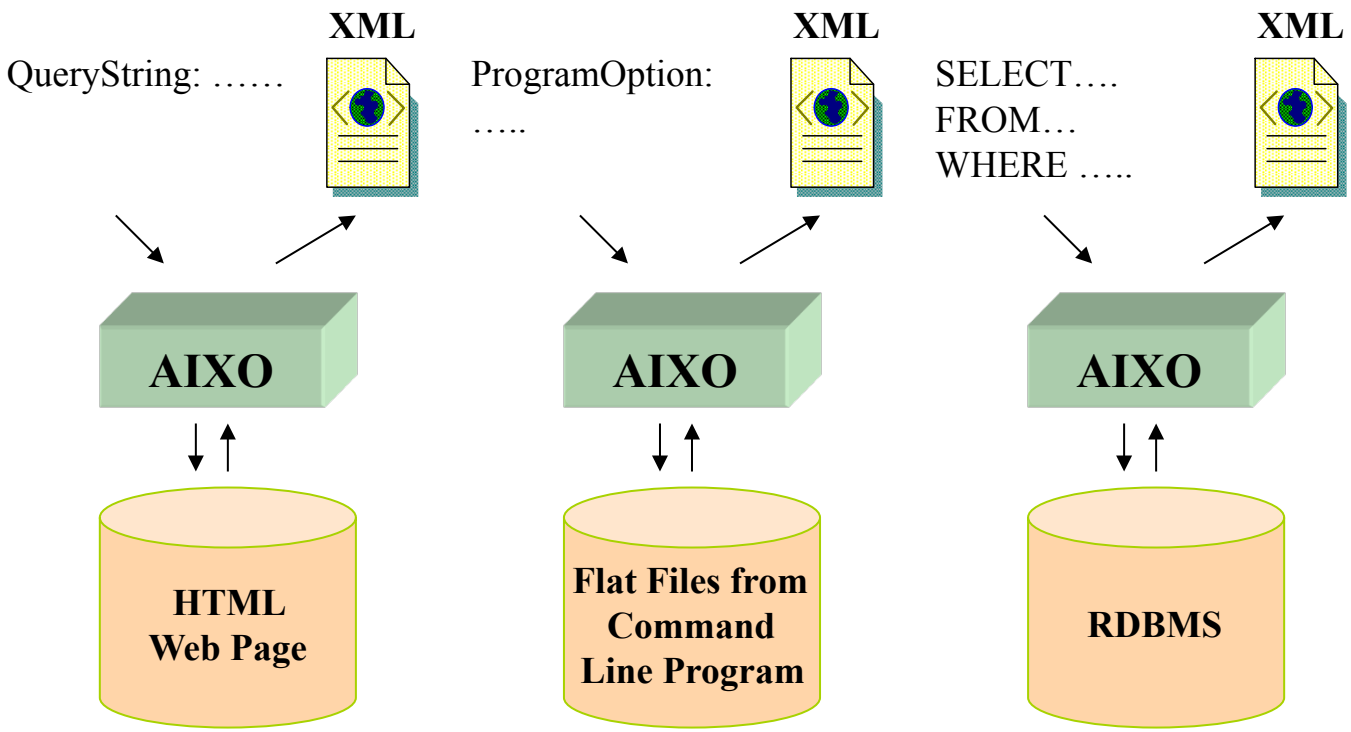
in order to automatically execute an experiment

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

Workflow-based User BioApplication



Wrapper-based System: general scenario



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:

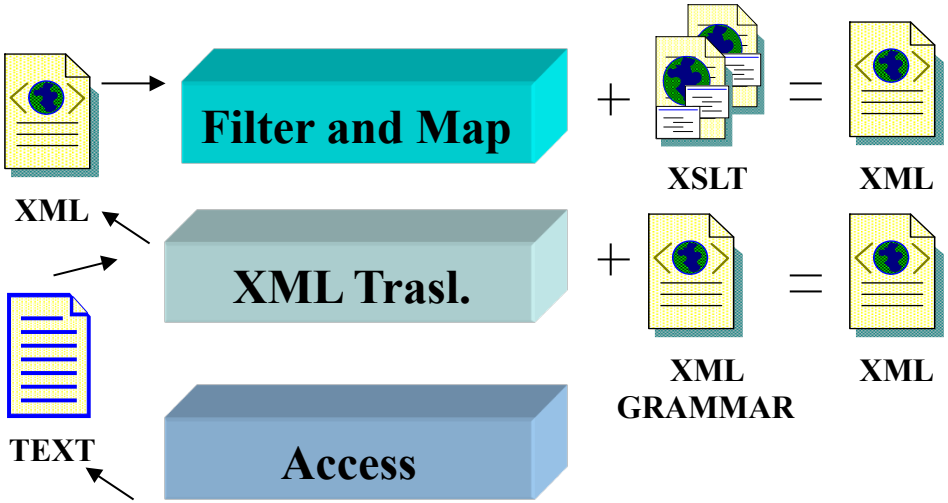
Environment: Protein DataBank web site

Data: PDB(DB): proprietary format

Tool: FASTA (Algorithm):

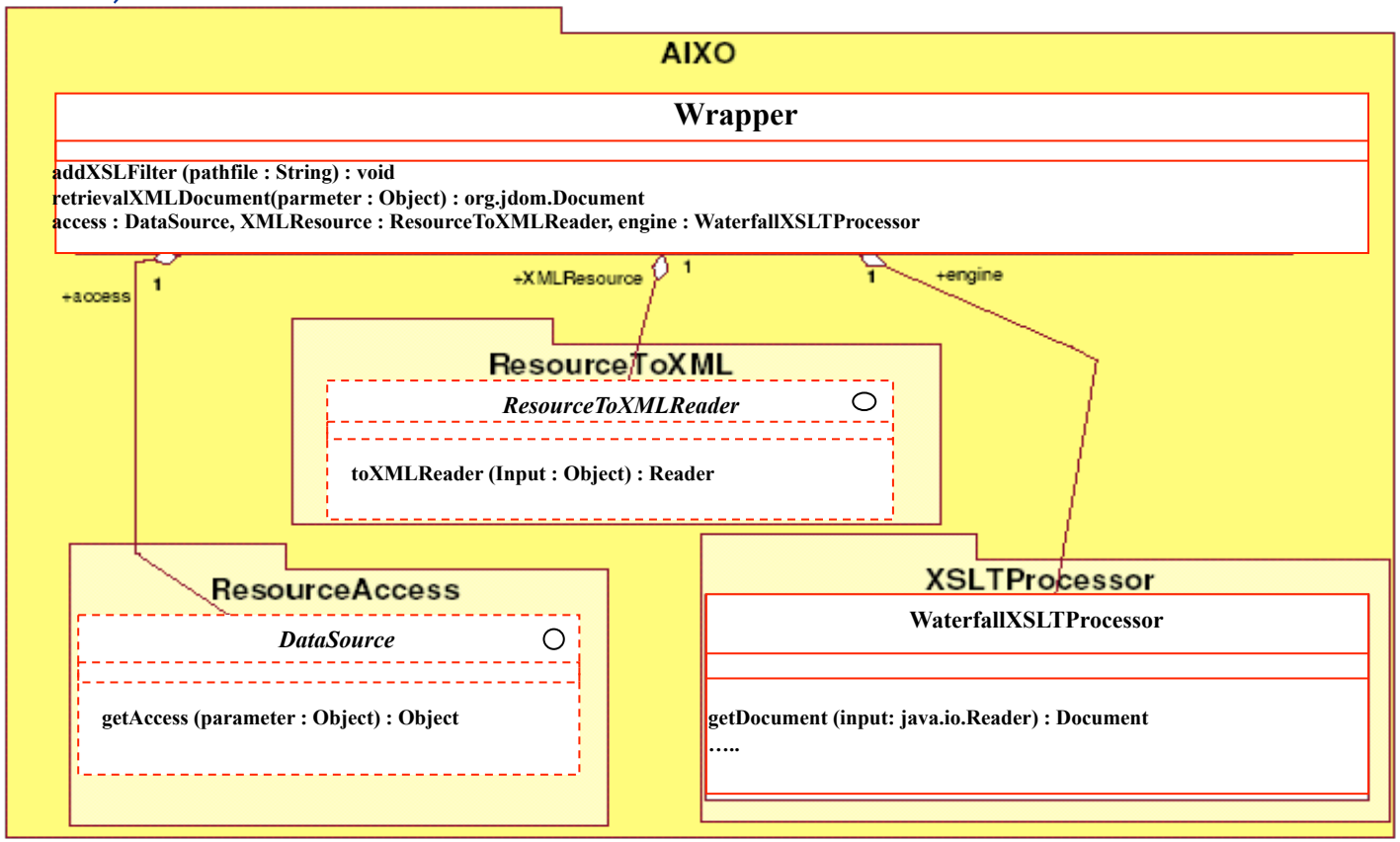
Output: FASTA Format

Wrapper-based System: Retrieval MedLine articles about P53 proteine




```
<…>
 <entry>
        <ID name="P53_HUMAN" type="STANDARD" molecule="PRT" lenght="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

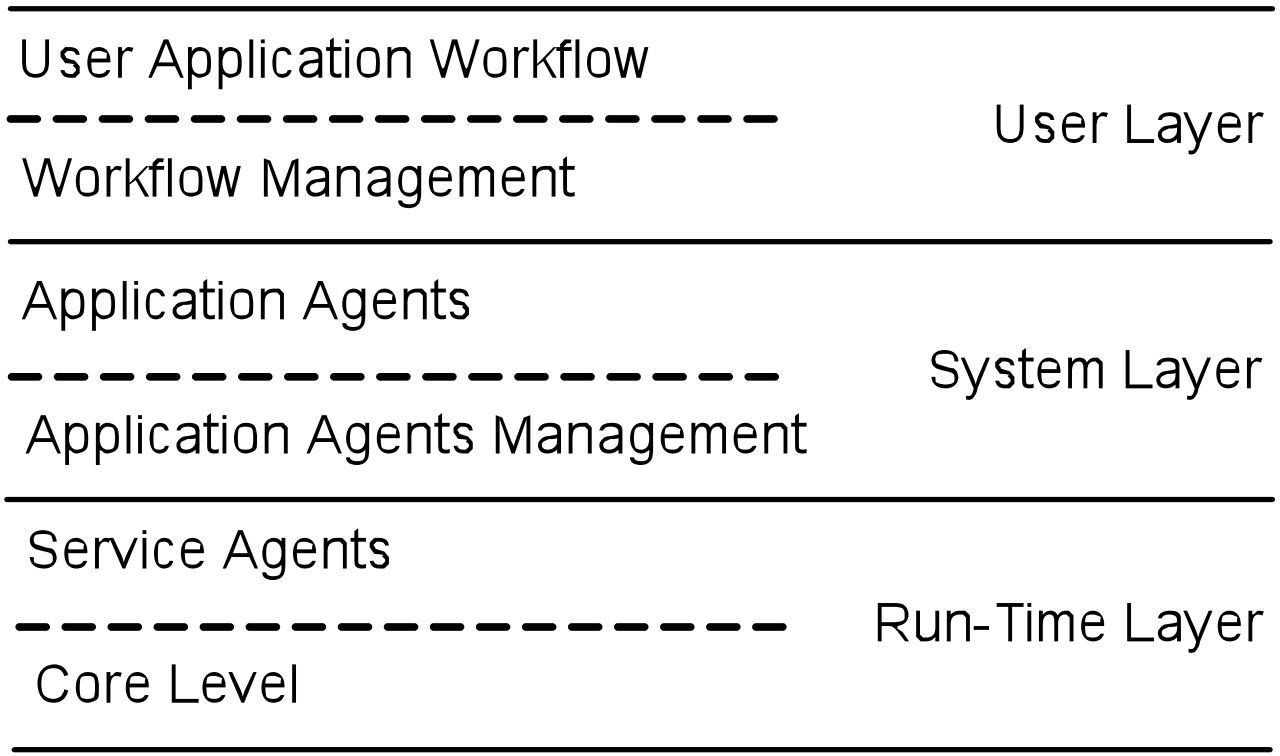
Problem: To integrate and coordinate multiple software tools for retrieving and integrating heterogeneous, distributed and frequently redundant data

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

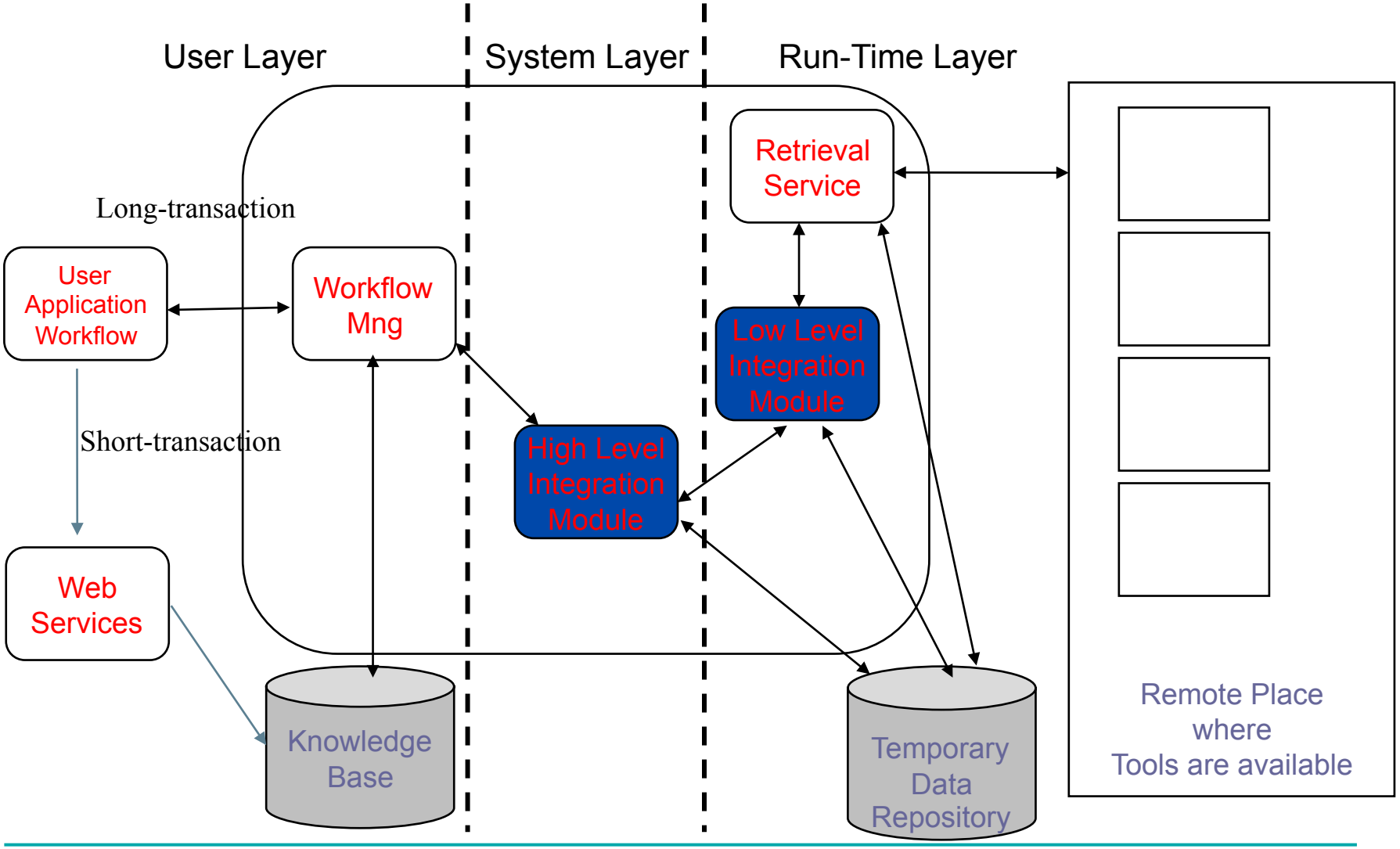
Aim: To define an *integrated environment* freeing the user from the need to know details on data repository and to coordinate the intermediate steps of an experiment (tasks)

Proposed Approach: To define an application as a workflow of tasks; to coordinate the execution of cooperative tasks by using software agent tools

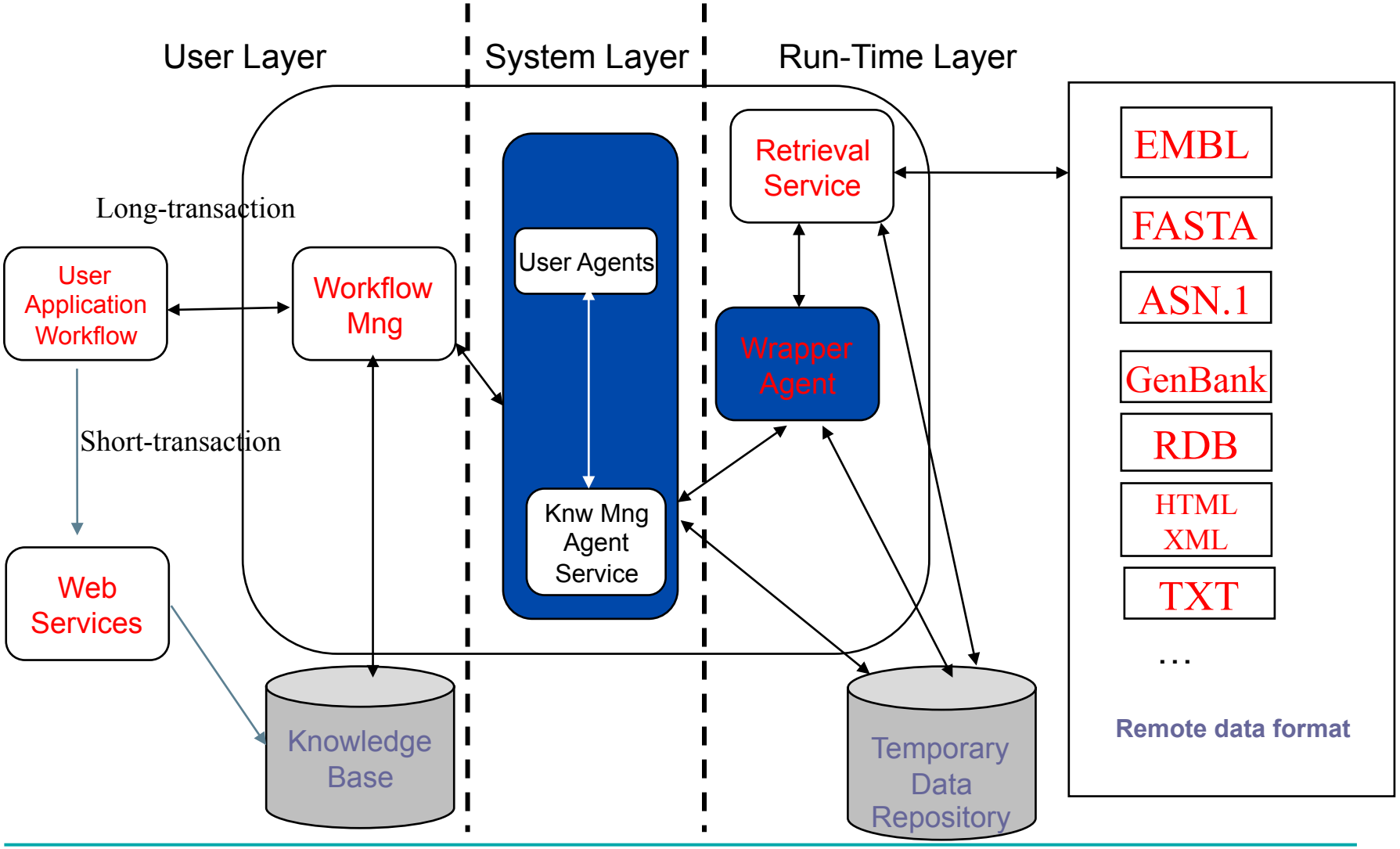
System's software architecture



A general system's architecture

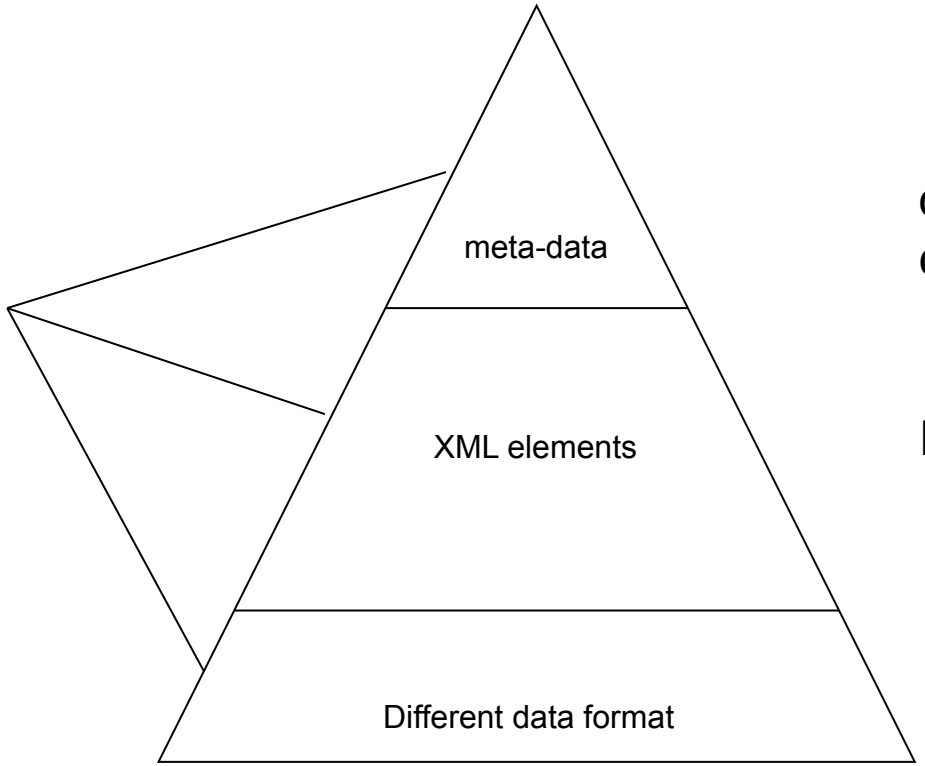


Agent-based System Architecture



From Data to Knowledge and vice versa

tool
Integration

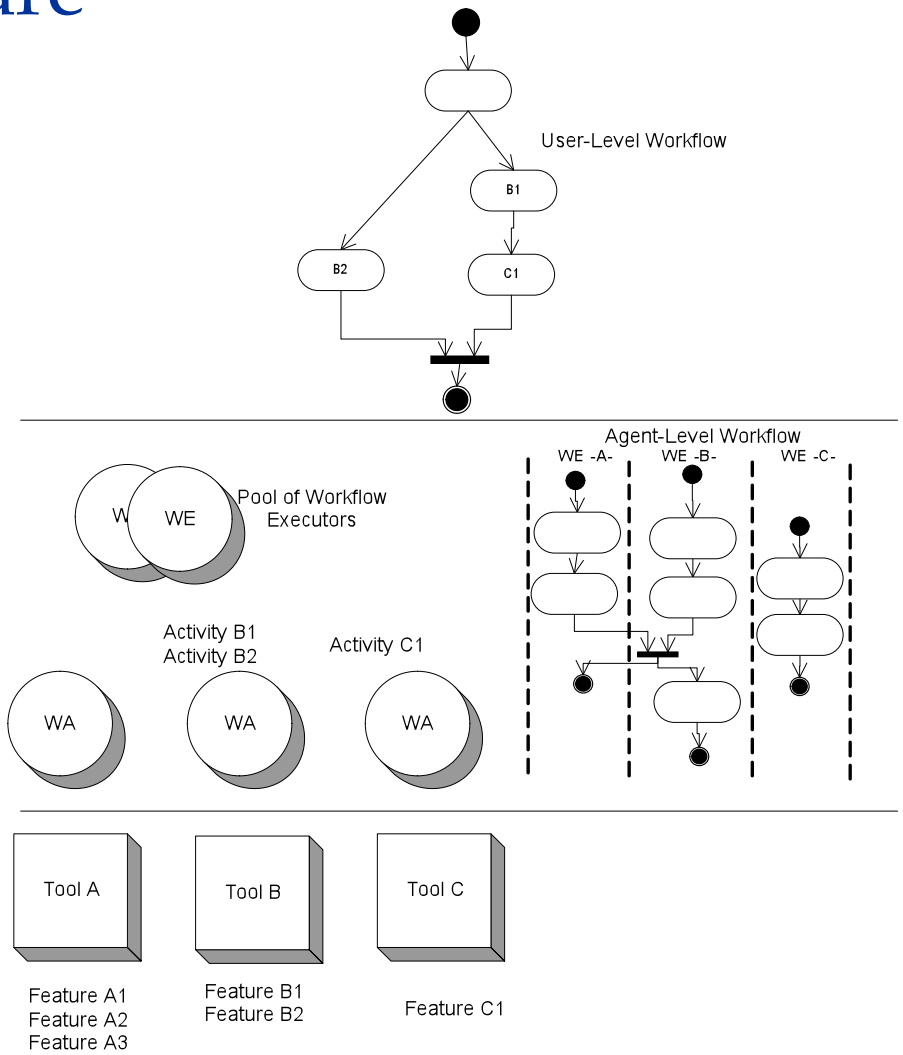


ontologies (human concepts) + workflow

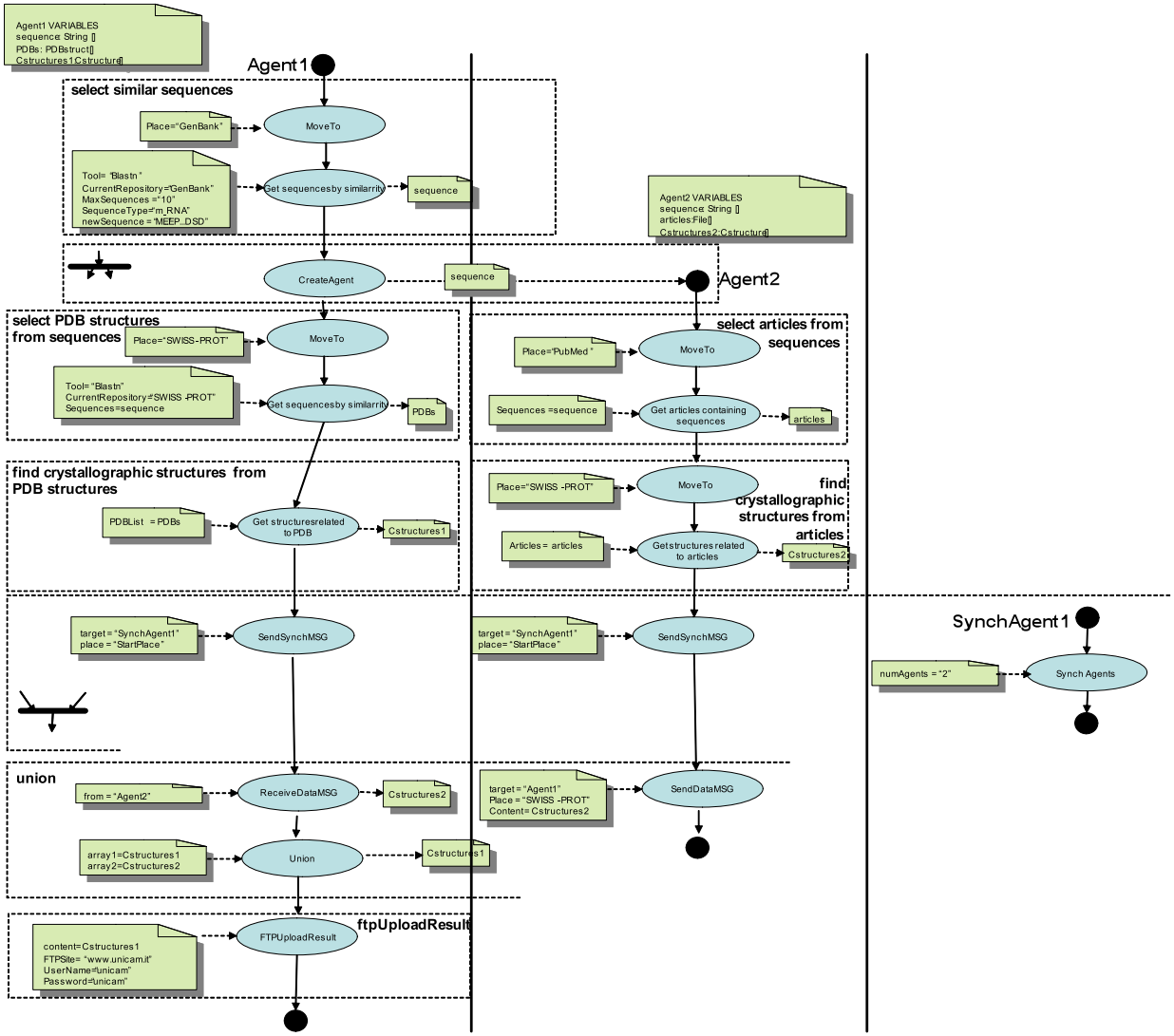
Information + coordination

data + algorithms

The Proposed Approach: an Agent-based Middleware

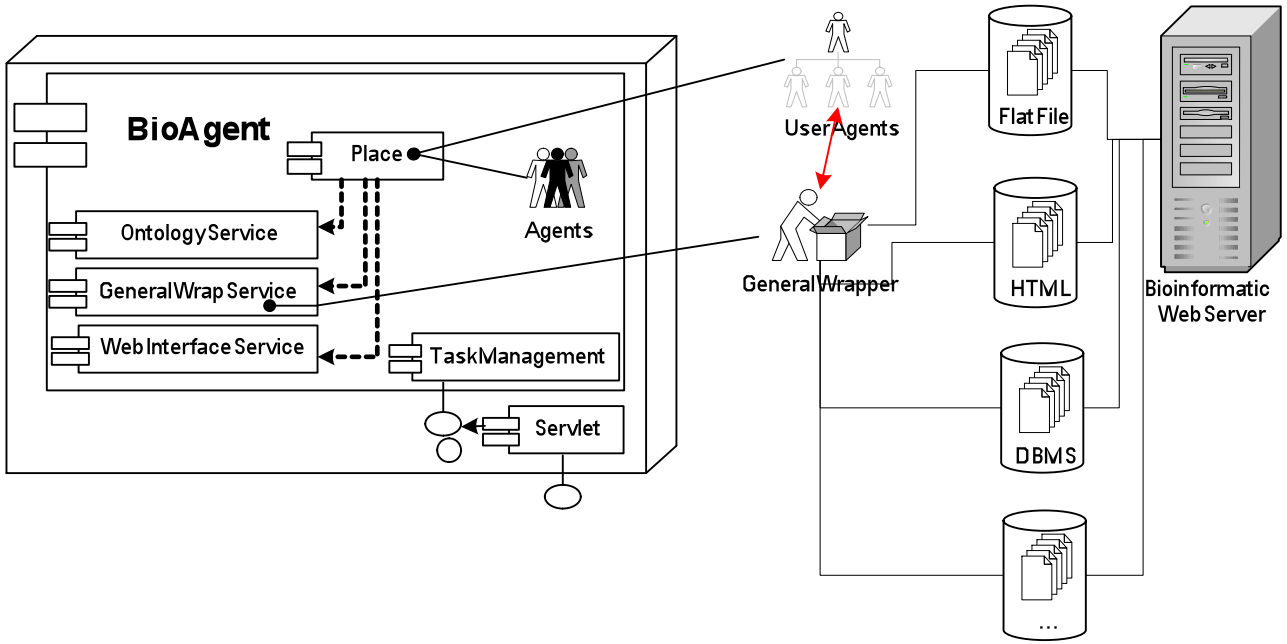


Preliminary Results: User-agent as high level *Tool Integration*



Preliminary Results: Wrapper-agent as low level Tool Integration

BioAgent


BioAgent
Place
Agents
Ontology Service
General Wrap Service
Web Interface Service
Task Management
Servlet
User Agents
General Wrapper
Flat File
HTML
DBMS
...
Bioinformatic
Web Server

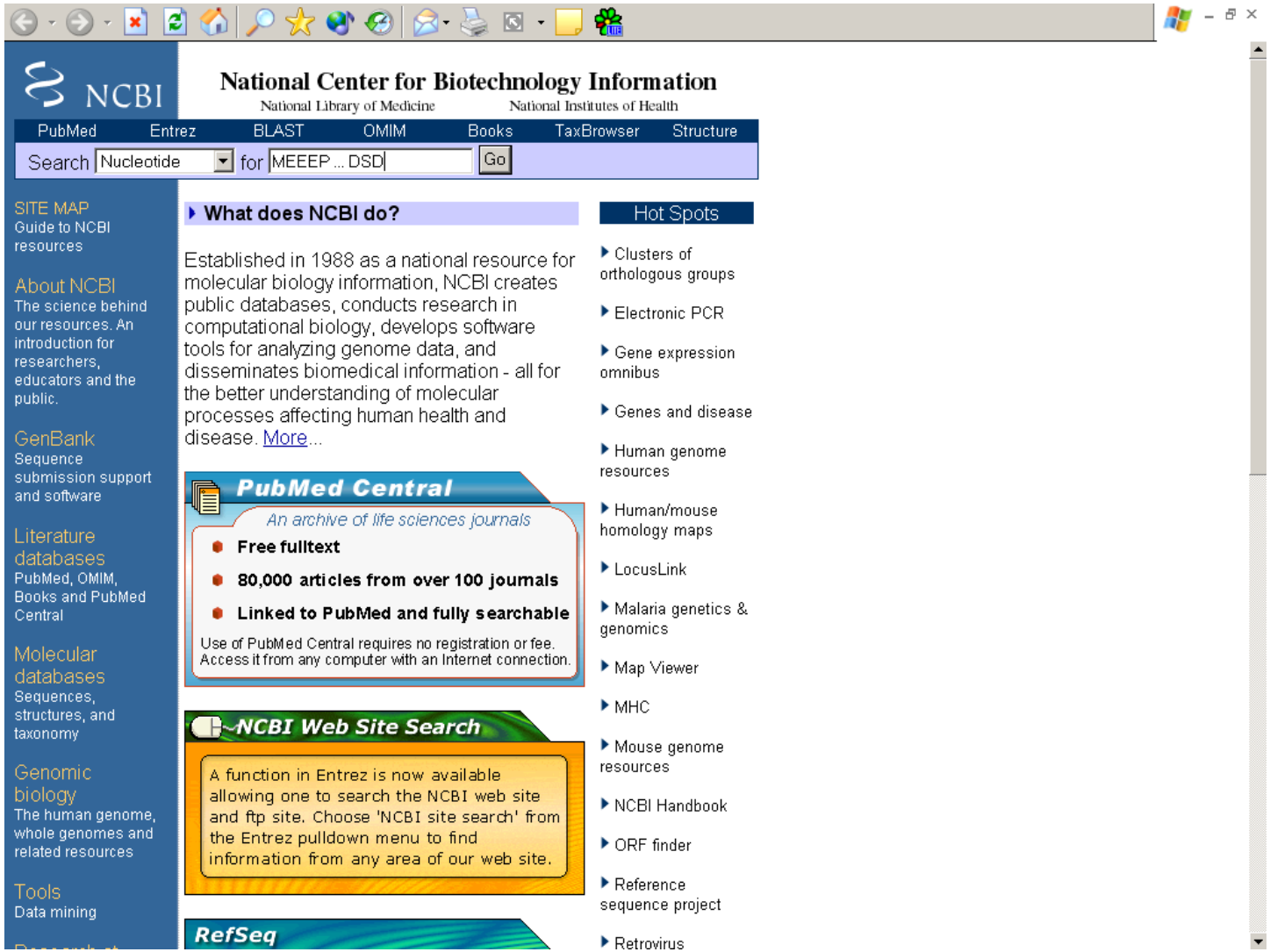
Future Activities and Conclusions

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

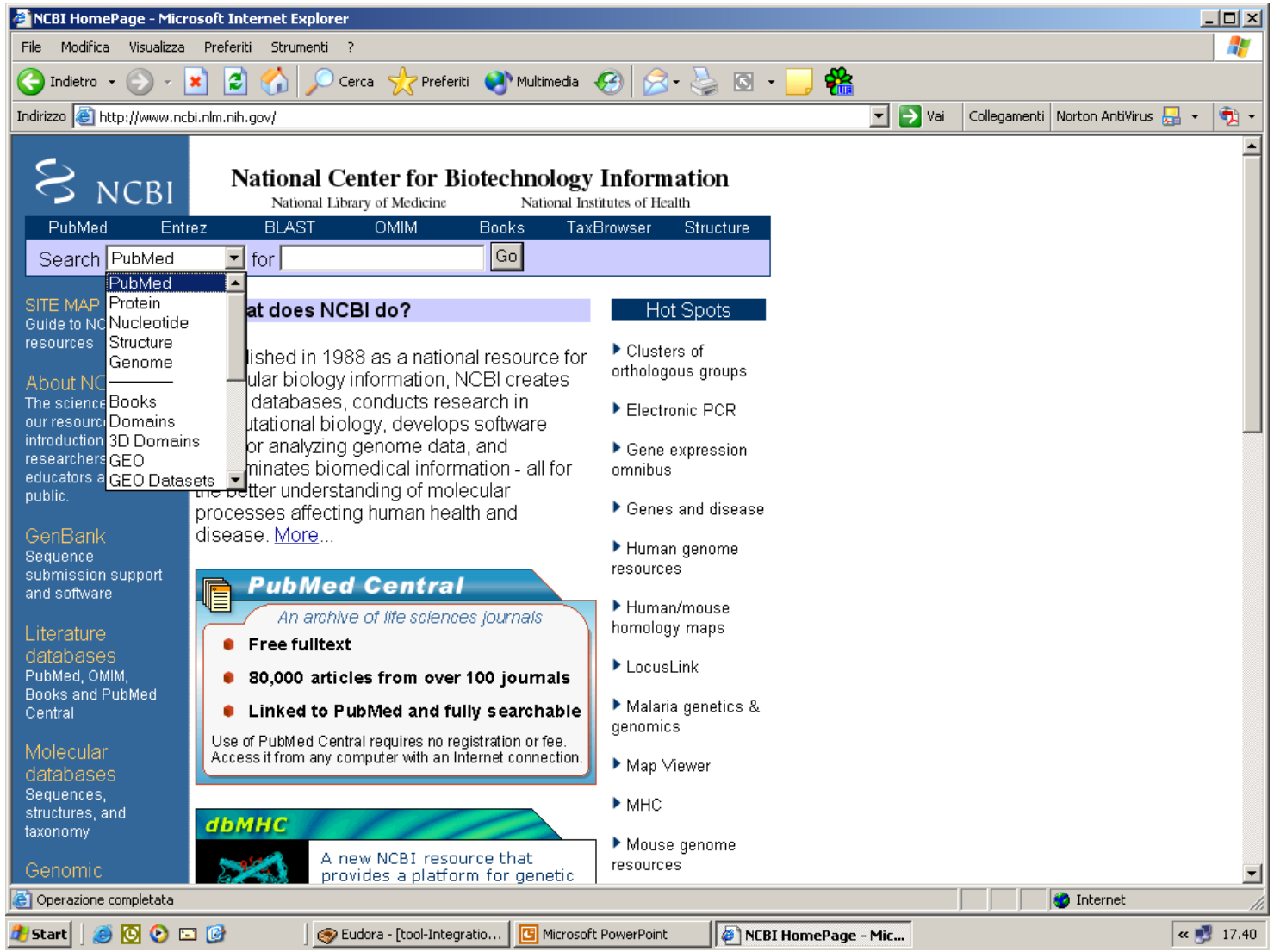
- Develop wrapper agents
- Design and develop the knowledge database to manage software tools
- Develop the compiler to allow the automatic generation of user-agents
- Evaluate the possibility to include mobility to user-agents in order to minimize the data transfer during tasks execution.

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.

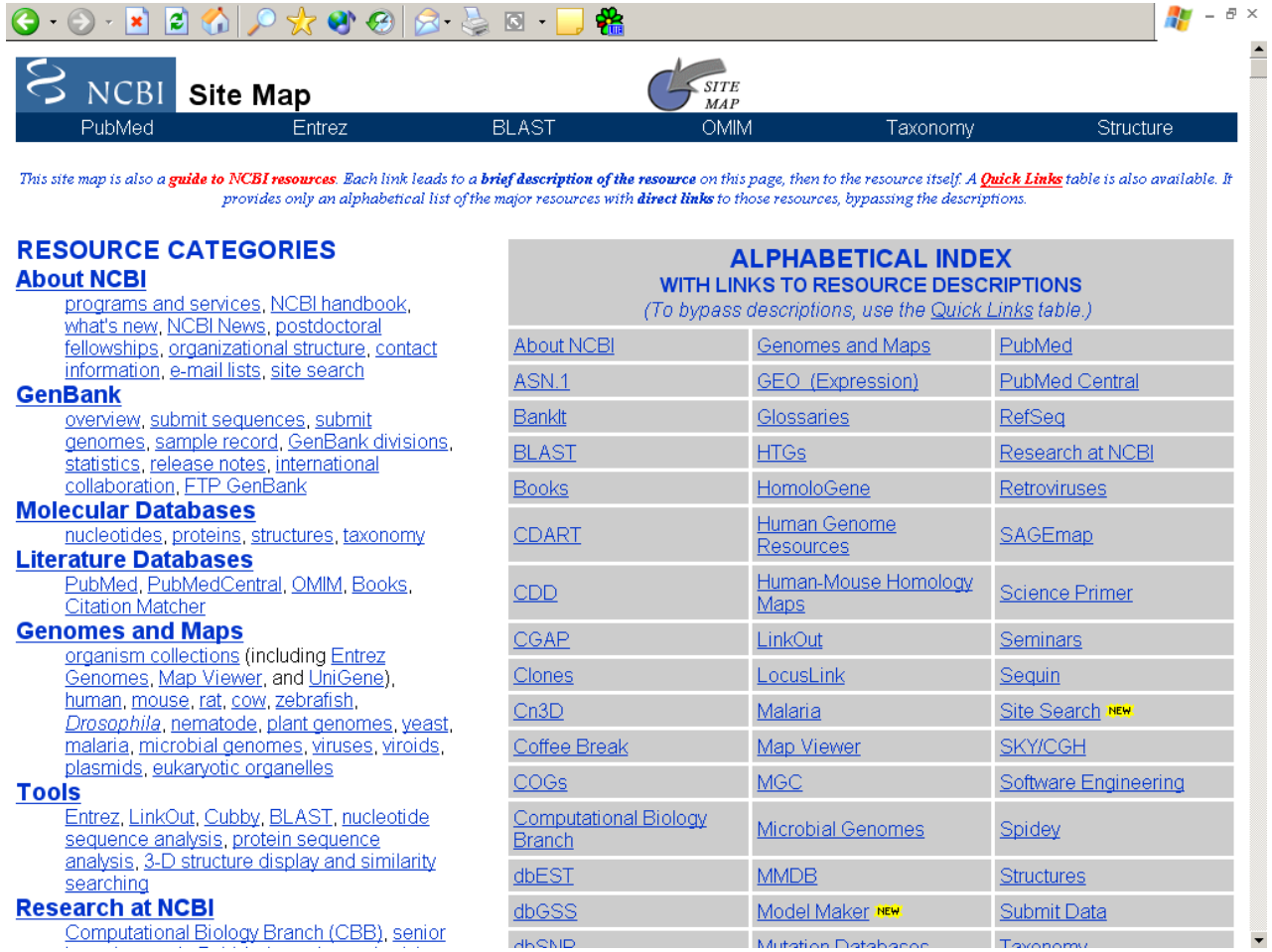
NCBI – Home page



NCBI – main databses



NCBI – site map



NCBI Site Map

PubMed | Entrez | BLAST | OMIM | Taxonomy | Structure

*This site map is also a **guide to NCBI resources**. Each link leads to a **brief description of the resource** on this page, then to the resource itself. A **Quick Links** table is also available. It provides only an alphabetical list of the major resources with **direct links** to those resources, bypassing the descriptions.*

RESOURCE CATEGORIES

About NCBI
programs and services, NCBI handbook, what's new, NCBI News, postdoctoral fellowships, organizational structure, contact information, e-mail lists, site search

GenBank
overview, submit sequences, submit genomes, sample record, GenBank divisions, statistics, release notes, international collaboration, FTP GenBank

Molecular Databases
nucleotides, proteins, structures, taxonomy

Literature Databases
PubMed, PubMedCentral, OMIM, Books, Citation Matcher

Genomes and Maps
organism collections (including Entrez Genomes, Map Viewer, and UniGene), human, mouse, rat, cow, zebrafish, *Drosophila*, nematode, plant genomes, yeast, malaria, microbial genomes, viruses, viroids, plasmids, eukaryotic organelles

Tools
Entrez, LinkOut, Cubby, BLAST, nucleotide sequence analysis, protein sequence analysis, 3-D structure display and similarity searching

Research at NCBI
Computational Biology Branch (CBB), senior

ALPHABETICAL INDEX
WITH LINKS TO RESOURCE DESCRIPTIONS
(To bypass descriptions, use the Quick Links table.)

About NCBI	Genomes and Maps	PubMed
ASN.1	GEO (Expression)	PubMed Central
BankIt	Glossaries	RefSeq
BLAST	HTGs	Research at NCBI
Books	HomoloGene	Retroviruses
CDART	Human Genome Resources	SAGEmap
CDD	Human-Mouse Homology Maps	Science Primer
CGAP	LinkOut	Seminars
Clones	LocusLink	Sequin
Cn3D	Malaria	Site Search NEW
Coffee Break	Map Viewer	SKY/CGH
COGs	MGC	Software Engineering
Computational Biology Branch	Microbial Genomes	Spidey
dbEST	MMDB	Structures
dbGSS	Model Maker NEW	Submit Data
dbSNP	Mutation Databases	Taxonomy

NCBI - Entrez

Entrez Home - Microsoft Internet Explorer

File Modifica Visualizza Preferiti Strumenti ?

Indirizzo http://www.ncbi.nlm.nih.gov/Entrez/

NCBI

Entrez
search and retrieval system

PubMed Entrez BLAST OMIM Books Taxonomy Structure

Search PubMed for Go Clear

About Entrez

SITE MAP

PubMed Help
help documentation for PubMed

Entrez Help
help documentation

Entrez Tutorial NEW
tutorial for the Entrez system

The Entrez Databases
cross-references and information

Batch Entrez
upload a file of GI or accession numbers to retrieve sequences

Making WWW Links to Entrez
linking to PubMed and GenBank

LinkOut
external links to related resources

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 Domains: domains from Entrez Structure
- GEO: Gene Expression Omnibus
- GEO Datasets: curated GEO data sets
- Journals: journals in Entrez
- MeSH: medical subject headings
- NCBI Web Site: NCBI Web site search
- OMIM: Online Mendelian Inheritance in Man
- PMC: full-text digital archive of life sciences journal literature
- PopSet: population study datasets
- SNP: single nucleotide polymorphisms
- Taxonomy: organisms in GenBank
- UniGene: gene-oriented clusters of transcript sequences
- UniSTS: markers and mapping data

Pre-computed similarity searches are available for most database records, which produce a list of related sequences, structure neighbors, as well as related articles

Operazione completata Internet

Start Eudora - [tool-Integratio... Microsoft PowerPoint Entrez Home - Micros... 17.48

PDB –Home page

NCBI - BLAST

http://www.ncbi.nlm.nih.gov/BLAST/blast_overview.html - Microsoft Internet Explorer

File Modifica Visualizza Preferiti Strumenti ?

Indietro Cerca Preferiti Multimedia

Indirizzo http://www.ncbi.nlm.nih.gov/BLAST/blast_overview.html Vai Collegamenti Norton AntiVirus

NCBI BLAST

PubMed Entrez BLAST OMIM Taxonomy Structure

Basic Overview

- What is BLAST?
- What is the best BLAST for me?
- How can I access the BLAST services at NCBI?
- What is the StandAlone WWW BLAST Server?

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 a minimal sacrifice of sensitivity to distant 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 web page. Choosing the right one depends on the type of sequence you are searching with (long, short; 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. Plus there is also addtional 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?

Operazione completata Internet

Start Eudora - [tool-Integratio... Microsoft PowerPoint http://www.ncbi.nlm.... « 17.53

NCBI – ASN.1

ASN.1 Summary - Microsoft Internet Explorer

File Modifica Visualizza Preferiti Strumenti ?

Indirizzo http://www.ncbi.nlm.nih.gov/Sitemap/Summary/asn1.html

NCBI

ASN.1 Summary

PubMed | Entrez | BLAST | OMIM | Taxonomy | Structure

NCBI Home

Site Map
brief and complete versions

About NCBI
general and contact information

GenBank
submit your sequence, general information

Molecular Databases
nucleotides, proteins, structures and taxonomy

Literature Databases
PubMed, PubRef, OMIM, Citation Matcher

Genomes and Maps
maps, the human

ASN.1, or Abstract Syntax Notation One, is an International Standards 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.

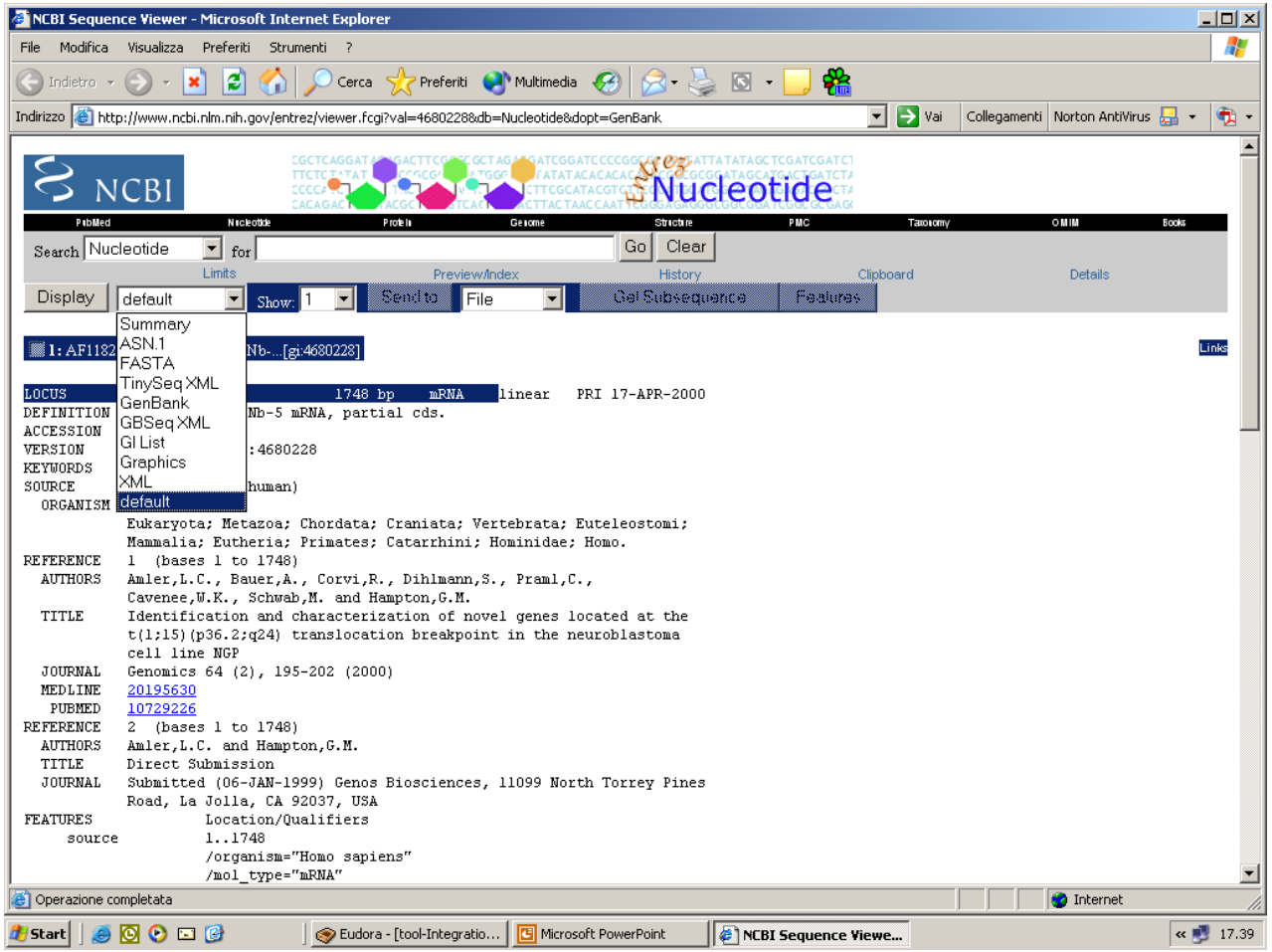
The documentation for the NCBI Software Development ToolKit contains information about the ASN.1 specification. The toolkit documentation is quite large (over 300 pages) and can be downloaded as the MS Word file sdk.doc.Z from the NCBI FTP site. Please see the readme file in that directory for more information.

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 ste contain more information about the toolbox and ASN.1.

The following **publications** might also be of interest:

Operazione completata | Internet

NCBI - fomats


NCBI Sequence Viewer - Microsoft Internet Explorer
File Modifica Visualizza Preferiti Strumenti ?
Indirizzo http://www.ncbi.nlm.nih.gov/entrez/viewer.fcgi?val=4680228&db=Nucleotide&dopt=GenBank
NCBI
Entrez Nucleotide
PubMed Nucleotide Protein Genome Structure PMC Taxonomy OMIM Books
Search Nucleotide for Go Clear
Limits Preview/Index History Clipboard Details
Display default Show: 1 Send to File Get Subsequence Features
Summary
ASN.1
FASTA
TinySeq XML
GenBank
GBSeq XML
GI List
Graphics
XML
default
Links
LOCUS 1748 bp mRNA linear PRI 17-APR-2000
DEFINITION Nb-5 mRNA, partial cds.
ACCESSION
VERSION :4680228
KEYWORDS
SOURCE human)
ORGANISM
Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
Mammalia; Eutheria; Primates; Catarrhini; Hominidae; Homo.
REFERENCE 1 (bases 1 to 1748)
AUTHORS Amler,L.C., Bauer,A., Corvi,R., Dihlmann,S., Praml,C.,
Cavenee,W.K., Schwab,M. and Hampton,G.M.
TITLE Identification and characterization of novel genes located at the
t(1;15)(p36.2;q24) translocation breakpoint in the neuroblastoma
cell line NGP
JOURNAL Genomics 64 (2), 195-202 (2000)
MEDLINE 20195630
PUBMED 10729226
REFERENCE 2 (bases 1 to 1748)
AUTHORS Amler,L.C. and Hampton,G.M.
TITLE Direct Submission
JOURNAL Submitted (06-JAN-1999) Genos Biosciences, 11099 North Torrey Pines
Road, La Jolla, CA 92037, USA
FEATURES Location/Qualifiers
source 1..1748
/organism="Homo sapiens"
/mol_type="mRNA"
Operazione completata
Internet
Start
Eudora - [tool-Integratio...
Microsoft PowerPoint
NCBI Sequence Viewe...
17.39

- PDB-ID (P53) = 1TSR
- www.rcsb.org (

DNA and nucleotide sequence

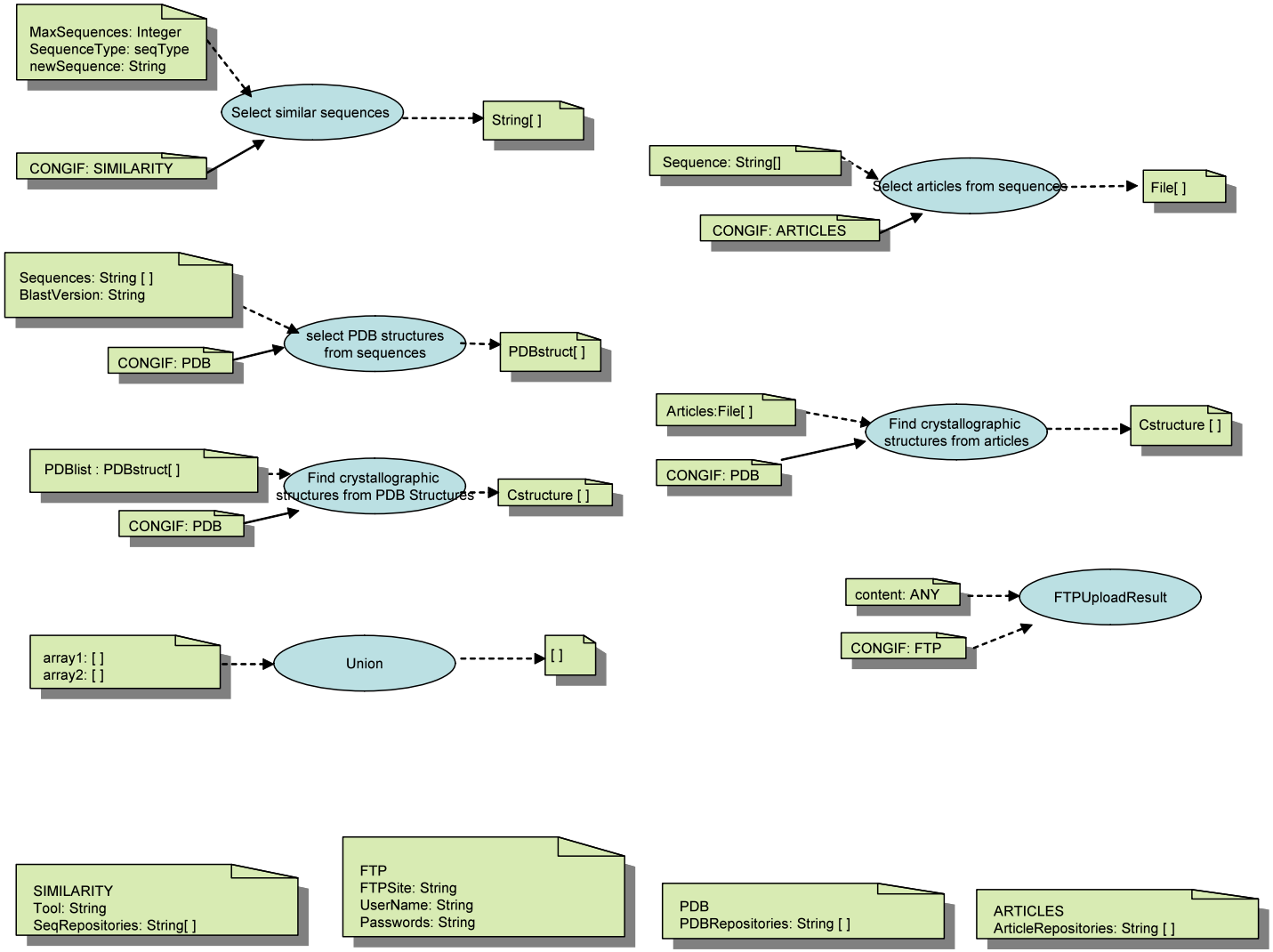
atggaggagccgcagtcagatcctagcgtcgagccccctctgagtcaggaaacattttca
M E E P Q S D P S V E P P L S Q E T F S
gacctatggaaactacttcctgaaaacaacgttctgtcccccttgccgtcccaagcaatg
D L W K L L P E N N V L S P L P S Q A M
gatgatttgatgctgtccccggacgatattgaacaatggttcactgaagacccaggtcca
D D L M L S P D D I E Q W F T E D P G P
...

atggaggagccgcagtcagatcctagcgtcgagccccctctgagtcaggaaacattttca
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

gatgaagctcccagaatgccagaggctgctccccgcgtggcccctggaccagcagctcct
D E A P R M P E A A P R V A P G P A A P
acaccggcggcccctgcaccagccccctcctggcccctgtcatcttctgtcccttcccag
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
ttgcgtgtggagtatttggatgacagaaacacttttcgacatagtgtggtggtgccctat
L R V E Y L D D R N T F R H S V V V P Y

ccagggagcactaagcgagcactgcccaacaacaccagctcctctccccagccaaagaag
P G S T K R A L P N N T S S S P Q P K K
aaaccactggatggagaatatttcacccttcagatccgtgggcgtgagcgcttcgagatg
K P L D G E Y F T L Q I R G R E R F E M
ttccgagagctgaatgaggccttggaactcaaggatgcccaggctgggaaggagccaggg
F R E L N E A L E L K D A Q A G K E P G
gggagcagggctcactccagccacctgaagtccaaaaagggtcagtctacctcccgccat
G S R A H S S H L K S K K G Q S T S R H
aaaaaactcatgttcaagacagaagggcctgactcagactga
K K L M F K T E G P D S D

ULAD


MaxSequences: Integer
SequenceType: seqType
newSequence: String
Select similar sequences
String[ ]
CONGIF: SIMILARITY
Sequence: String[]
Select articles from sequences
File[ ]
CONGIF: ARTICLES
Sequences: String [ ]
BlastVersion: String
select PDB structures from sequences
PDBstruct[ ]
CONGIF: PDB
Articles:File[ ]
Find crystallographic structures from articles
Cstructure [ ]
CONGIF: PDB
PDBlist : PDBstruct[ ]
Find crystallographic structures from PDB Structures
Cstructure [ ]
CONGIF: PDB
content: ANY
FTPUploadResult
CONGIF: FTP
array1: [ ]
array2: [ ]
Union
[ ]
SIMILARITY
Tool: String
SeqRepositories: String[ ]
FTP
FTPSite: String
UserName: String
Passwords: String
PDB
PDBRepositories: String [ ]
ARTICLES
ArticleRepositories: String [ ]

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

…

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

AIXO: Any Input XML Output, a generalized wrapper

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

PEGAA: A Programming Environment for Global Activity-based Applications