A GRID-based multilayer architecture for bioinformatics

Ezio Bartocci, Diletta Cacciagrano, Nicola Cannata,

Flavio Corradini, Emanuela Merelli,

- Mathematic and Computer Science Department, University of Camerino, Italy

Luciano Milanesi,

- Institute for Biomedical Technologies, CNR, Milan, Italy

Paolo Romano

- National Cancer Research Institute, Genova, Italy

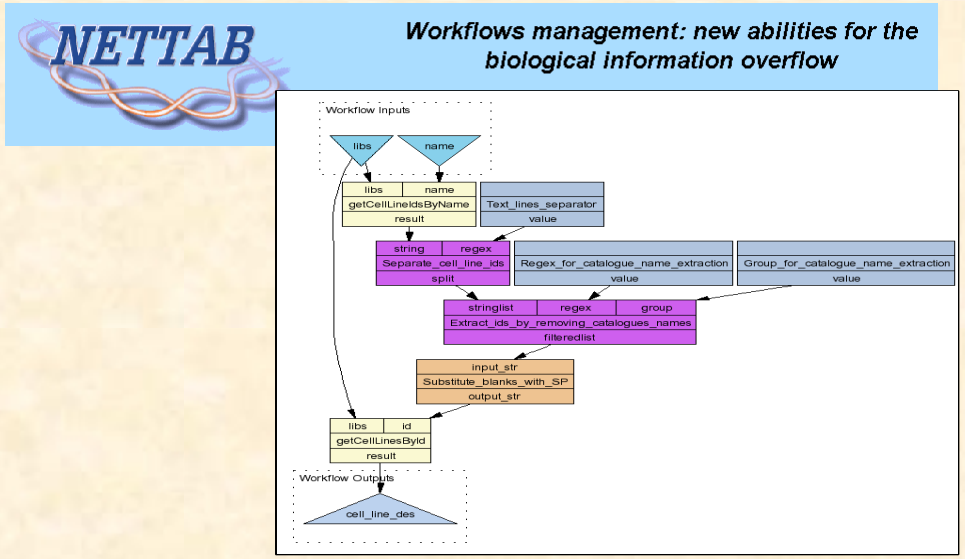
A GRID-based multilayer architecture for **bioinformatics**

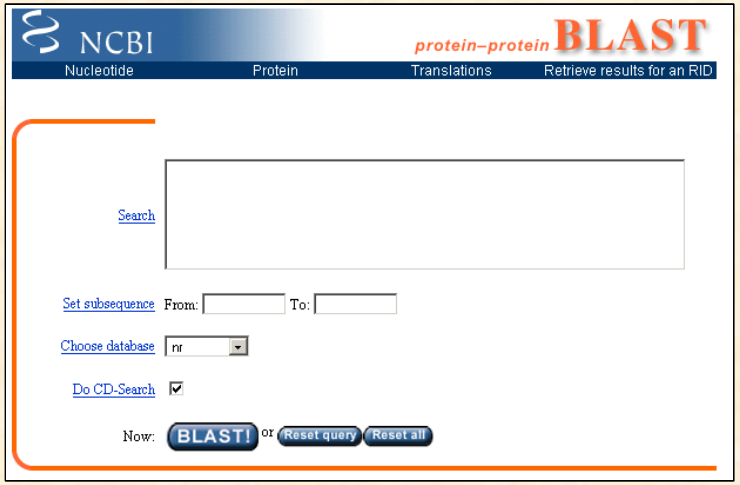
Which bioinformatics?

Bioinformatics has changed…

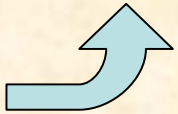
```
BLASTN 2.2.4 [Aug-26-2002]


Reference: Altschul, Stephen F., Thomas L. Madden, Alejandro A. Schaffer,
Jinghui Zhang, Zheng Zhang, Webb Miller, and David J. Lipman (1997),
"Gapped BLAST and PSI-BLAST: a new generation of protein database search
programs",  Nucleic Acids Res. 25:3389-3402.

Query= gi|8777285|gb|AB044774.1|AB044774 AB044774 Panax ginseng hairy
root Panax ginseng cDNA clone HR36, mRNA sequence
         (223 letters)

Database: /local/wwwstud/html/bioinfo3-64/EST-ginseng.fas
           21 sequences; 7481 total letters

Searching.done

                                                                 Score    E
Sequences producing significant alignments:                      (bits) Value

gi|8777285|gb|AB044774.1|AB044774 AB044774 Panax ginseng hairy r...   442    e-128
gi|8777269|gb|AB044758.1|AB044758 AB044758 Panax ginseng hairy r...    24    0.077
gi|8777271|gb|AB044760.1|AB044760 AB044760 Panax ginseng hairy r...    22    0.30

>gi|8777285|gb|AB044774.1|AB044774 AB044774 Panax ginseng hairy root
           Panax ginseng cDNA clone HR36, mRNA sequence
          Length = 223

 Score =  442 bits (223), Expect = e-128
 Identities = 223/223 (100%)
 Strand = Plus / Plus


Query: 1   ggtctccatgtttgttaaaaatatagtgctccctgctgctgaagattctgaatctatatg 60
           ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct: 1   ggtctccatgtttgttaaaaatatagtgctccctgctgctgaagattctgaatctatatg 60
```

From command line… (till 90s)



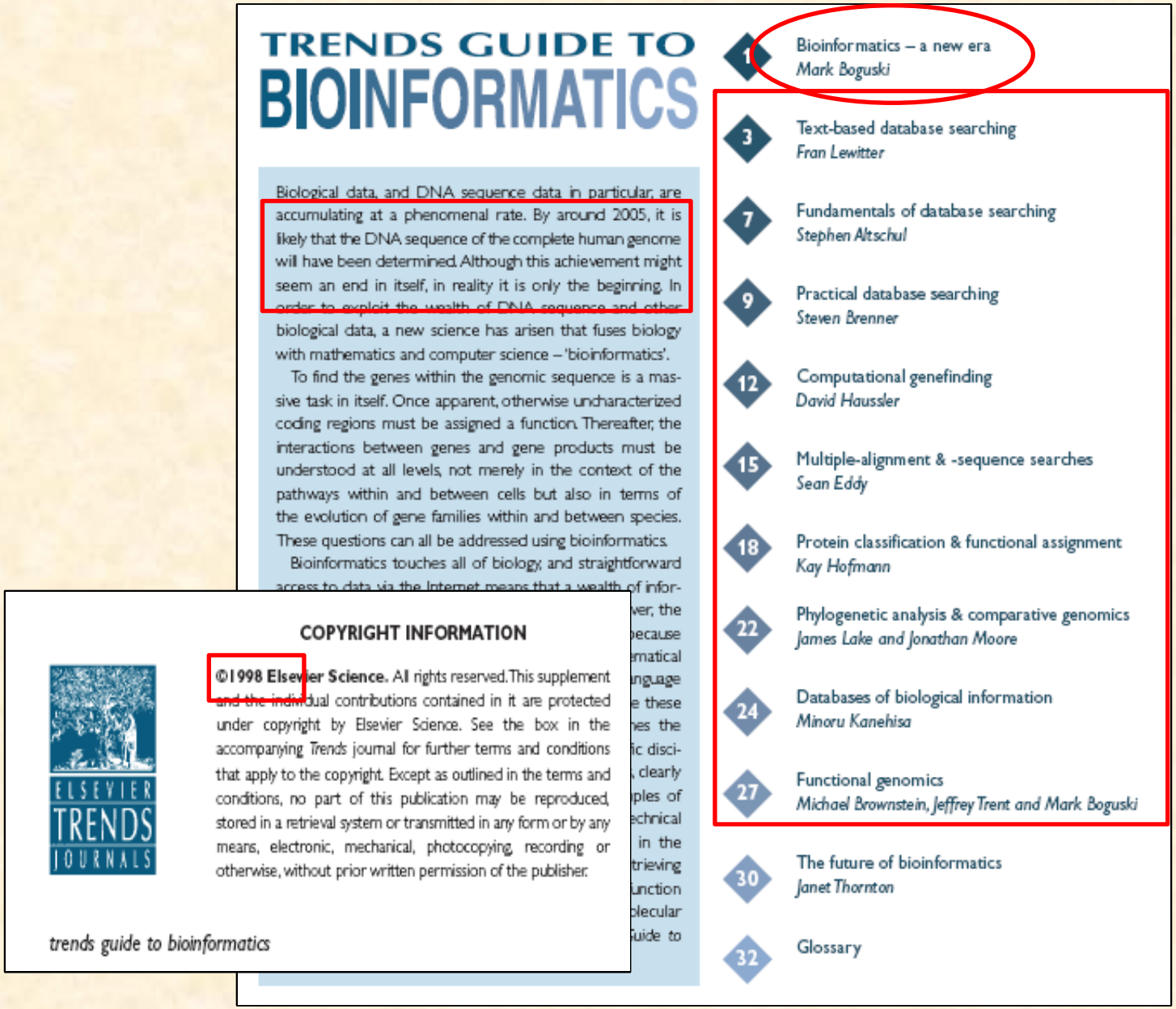


To web services and workflows (now)



to web interfaces… (with the advent of the WWW)

… very fast …



TRENDS GUIDE TO
BIOINFORMATICS

Biological data, and DNA sequence data in particular, are accumulating at a phenomenal rate. By around 2005, it is likely that the DNA sequence of the complete human genome will have been determined. Although this achievement might seem an end in itself, in reality it is only the beginning. In order to exploit the wealth of DNA sequence and other biological data, a new science has arisen that fuses biology with mathematics and computer science – 'bioinformatics'.

To find the genes within the genomic sequence is a massive task in itself. Once apparent, otherwise uncharacterized coding regions must be assigned a function. Thereafter, the interactions between genes and gene products must be understood at all levels, not merely in the context of the pathways within and between cells but also in terms of the evolution of gene families within and between species. These questions can all be addressed using bioinformatics.

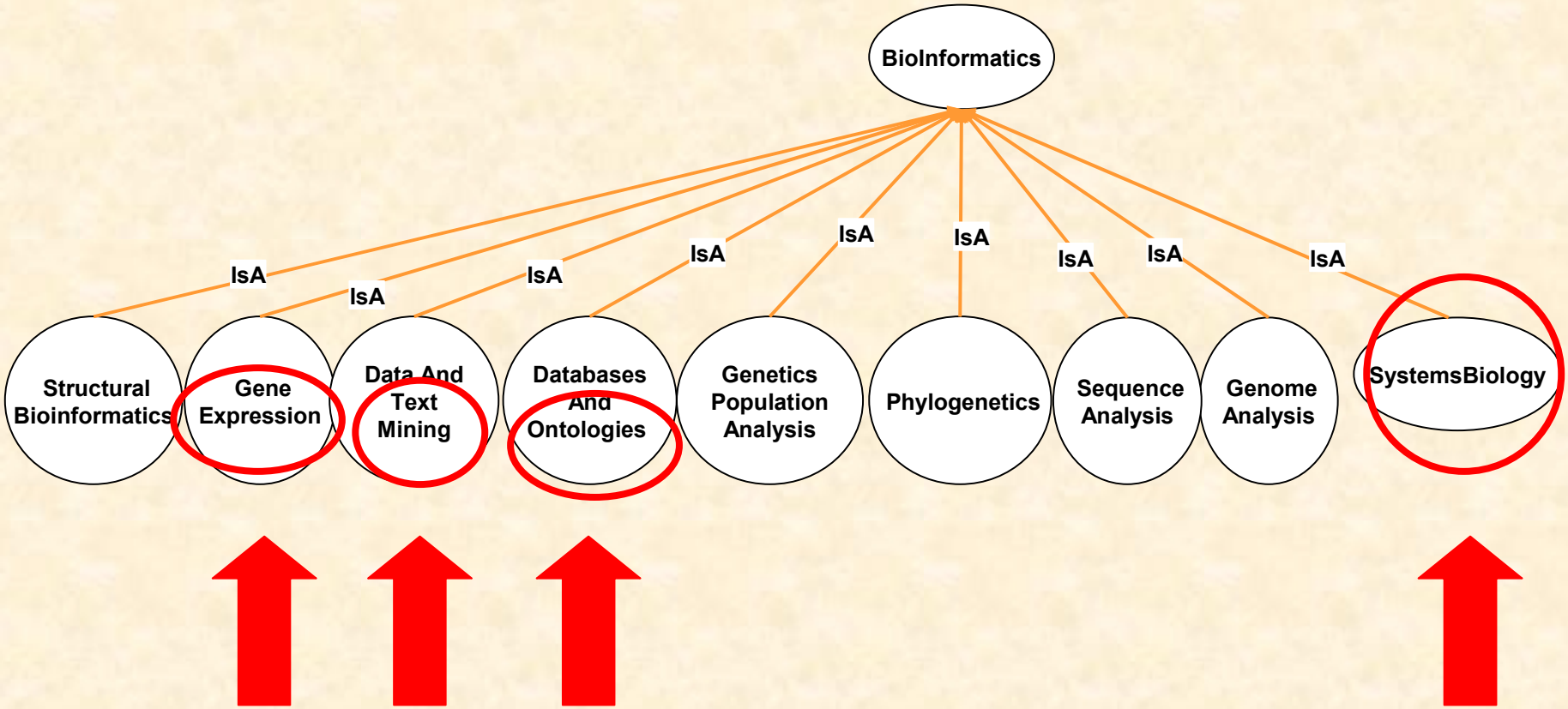
Bioinformatics touches all of biology, and straightforward access to data via the Internet means that a wealth of infor-

COPYRIGHT INFORMATION

©1998 Elsevier Science. All rights reserved. This supplement and the individual contributions contained in it are protected under copyright by Elsevier Science. See the box in the accompanying *Trends* journal for further terms and conditions that apply to the copyright. Except as outlined in the terms and conditions, no part of this publication may be reproduced, stored in a retrieval system or transmitted in any form or by any means, electronic, mechanical, photocopying, recording or otherwise, without prior written permission of the publisher.

ELSEVIER TRENDS JOURNALS

trends guide to bioinformatics



The classification introduced for articles of “Oxford’s Bioinformatics” in 2005

…and grown

(Or better burst?.....)

PSTT Vol. 2, No. 9 September 1999 | editorial

Information overflow from discovery to development

The phar
tools to d
section o

Ian Shaw
Momentum Healt
Temple Court
Cathedral Road
Cardiff
UK CF1 9HA
tel: +44 1222 78
fax: +44 1222 7
e-mail: shaw@m

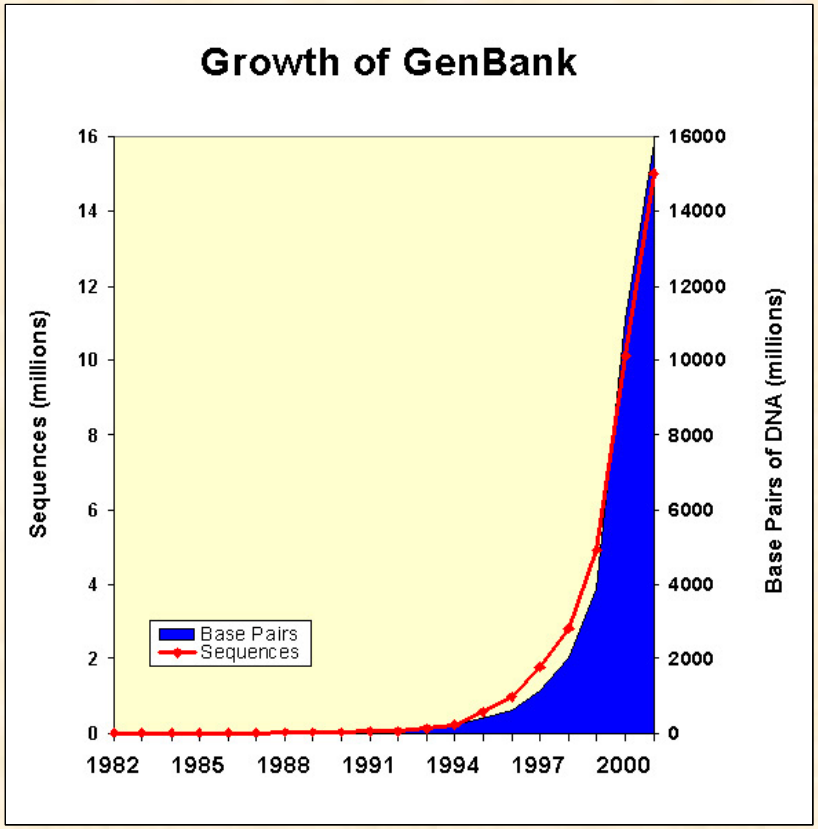
Available online at www.sciencedirect.com
SCIENCE DIRECT®
ELSEVIER
Journal of Biomedical Informatics xxx (2006) xxx–xxx
Journal of Biomedical Informatics
www.elsevier.com/locate/yjbin

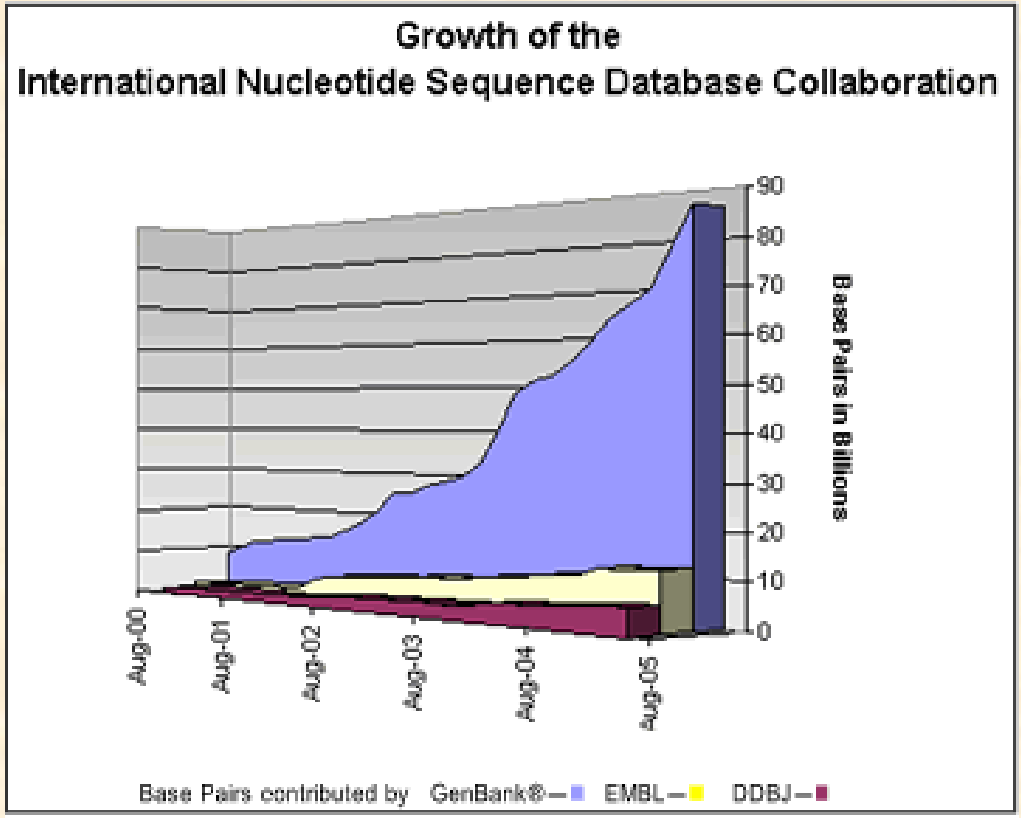
Beyond the data deluge: Data integration and bio-ontologies

Judith A. Blake *, Carol J. Bult

The Jackson Laboratory, Bar Harbor, ME, USA

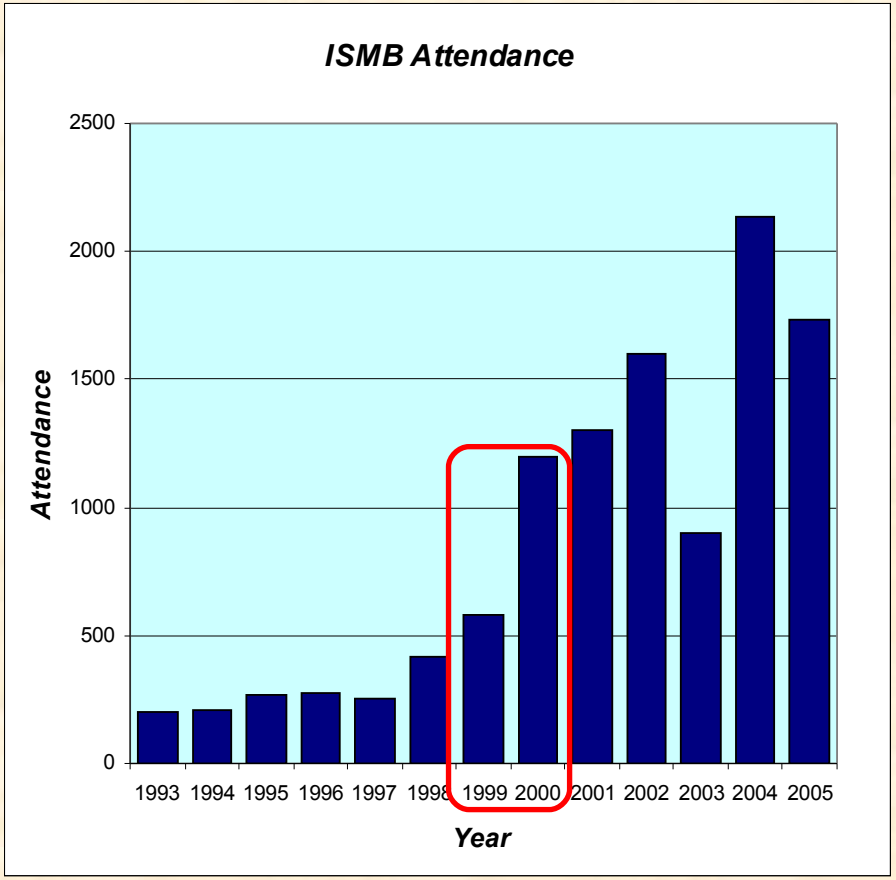
Received 31 August 2005


Growth of GenBank
Sequences (millions)
Base Pairs of DNA (millions)
Base Pairs
Sequences
0
2
4
6
8
10
12
14
16
0
2000
4000
6000
8000
10000
12000
14000
16000
1982
1985
1988
1991
1994
1997
2000


Growth of the
International Nucleotide Sequence Database Collaboration
Base Pairs in Billions
0
10
20
30
40
50
60
70
80
90
Aug-00
Aug-01
Aug-02
Aug-03
Aug-04
Aug-05
Base Pairs contributed by GenBank® EMBL DDBJ

The number of Bioinformaticians is growing…

Year	Place	Attendance
1993	Bethesda, MD, USA	201
1994	Stanford, California, USA	205
1995	Cambridge, United Kingdom	270
1996	St. Louis, MO, USA	279
1997	Halkidiki, Greece	254
1998	Montréal, Québec, Canada	413
1999	Heidelberg, Germany	580
2000	La Jolla / San Diego, CA, USA	1200
2001	Copenhagen, Denmark	1300
2002	Edmondon, Alberta, Canada	1600
2003	Brisbane, Australia	900
2004	Glasgow, Scotland, UK (With ECCB)	2136
2005	Detroit, Michigan, USA	1731



ISMB conferences attendance

…and is growing the number of their “products”!

Year	Articles	DB listed
2006	164	858
2005	137	719
2004	142	548
2003	95	386
2002	94	335
2001	73	281
2000	95	226
1999	86	201
1998	77	
1997	64	
1996	51	

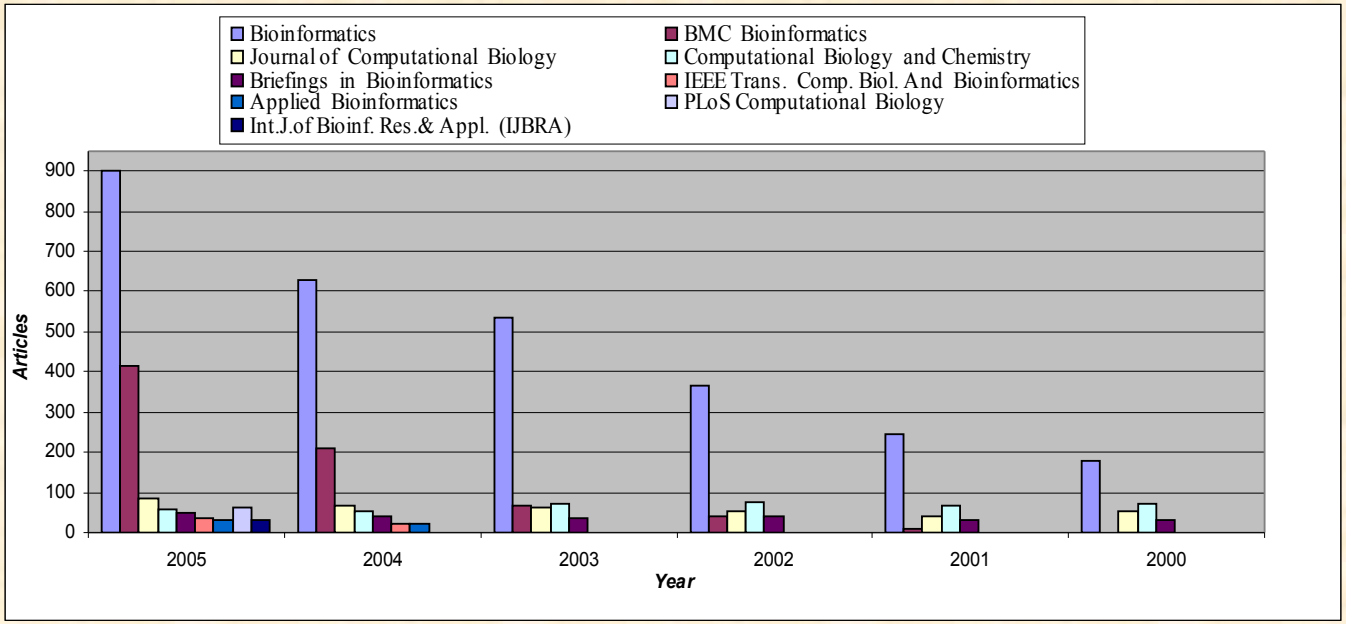
Database Issue

Year	Articles	WS presented
2005	159	166
2004	137	137
2003	106	131

Web server Issue

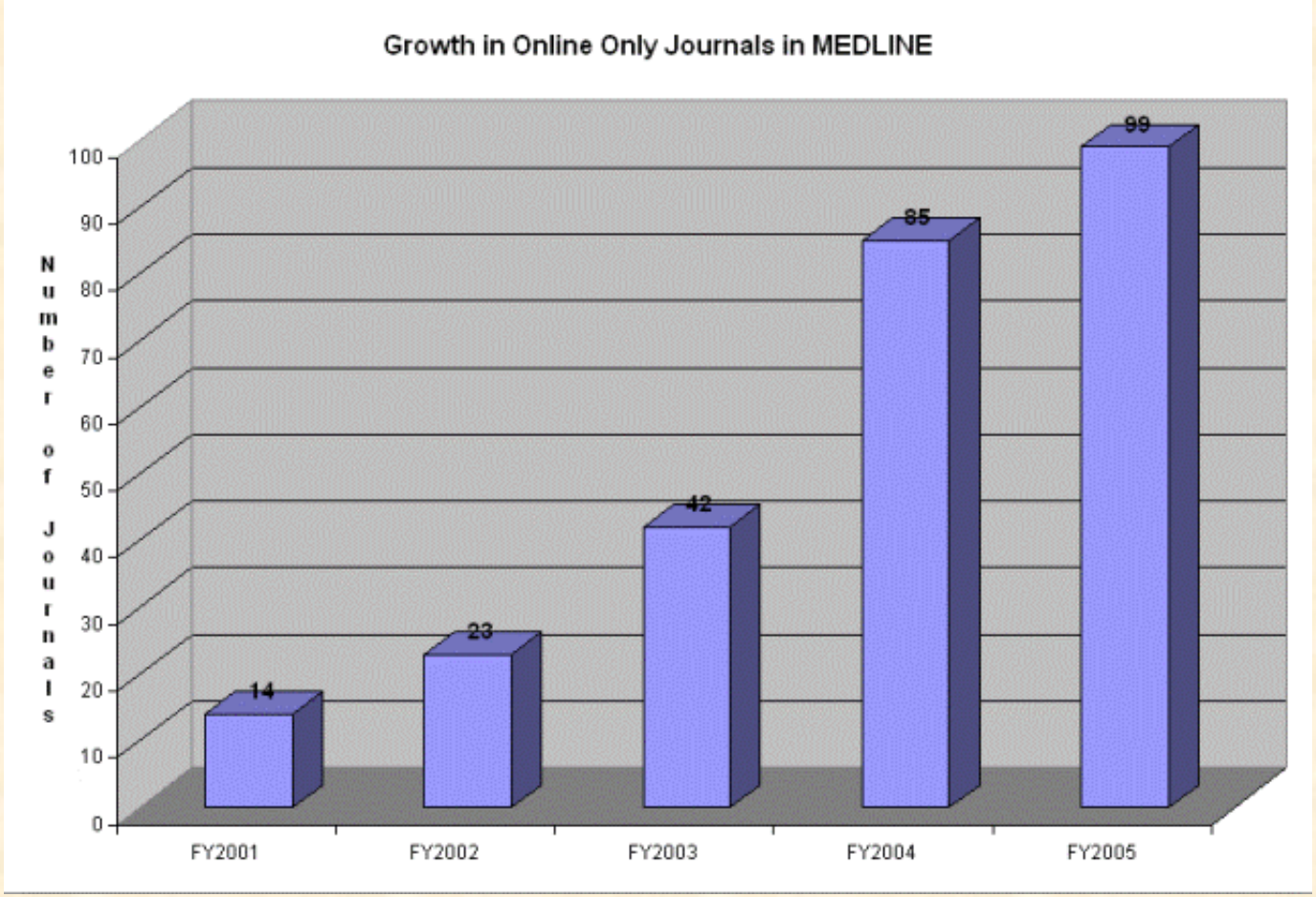
Nucleic Acid Research special issues

Year	Bioinformatics	BMC Bioinformatics	Journal of Computational Biology	PLoS Computational Biology	Computational Biology and Chemistry	Briefings in Bioinformatics	IEEE Trans. Comp. Biol. And Bioinformatics	Applied Bioinformatics	Int.J.of Bioinf. Res.& Appl. (IJBRA)
2005	900	414	83	61	53	47	34	32	32
2004	627	209	69		53	40	22	24	
2003	534	66	61		71	35			
2002	365	40	52		77	41			
2001	245	9	39		66	31			
2000	178	1	52		72	33			



Articles published "ONLY" from main bioinformatics journals

Another impressing growth



It has become really hard to find resources!

- It is difficult to track the evolution of a research area
- It is not existing any classification schema for bioinformatics resources
- For the bioinformatics domain is not existing any classification schema defined from an authority (like e.g. ACM, AMS)
- Keyword search vs Semantic search
- Take into account semantic relationships between resources
- Resources disappears (Many 404s)
- Multidisciplinarity is hard
- Easy to re-invent of the wheel

OPEN ACCESS Freely available online

PLoS COMPUTATIONAL BIOLOGY

Perspective

Time to Organize the Bioinformatics Resourceome

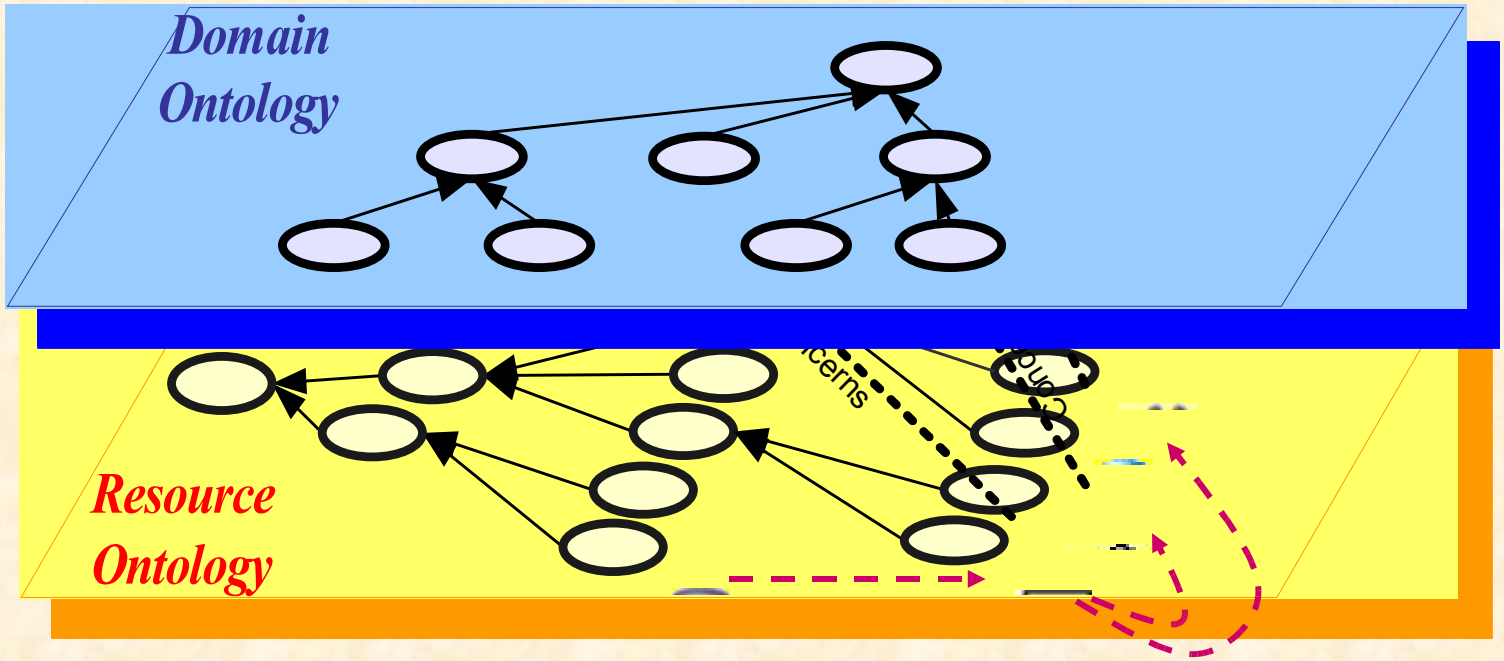
Nicola Cannata, Emanuela Merelli, Russ B. Altman*

We will be witnessing the birth of the artificial, or in-silico, scientist. — J. D. Wren [1]

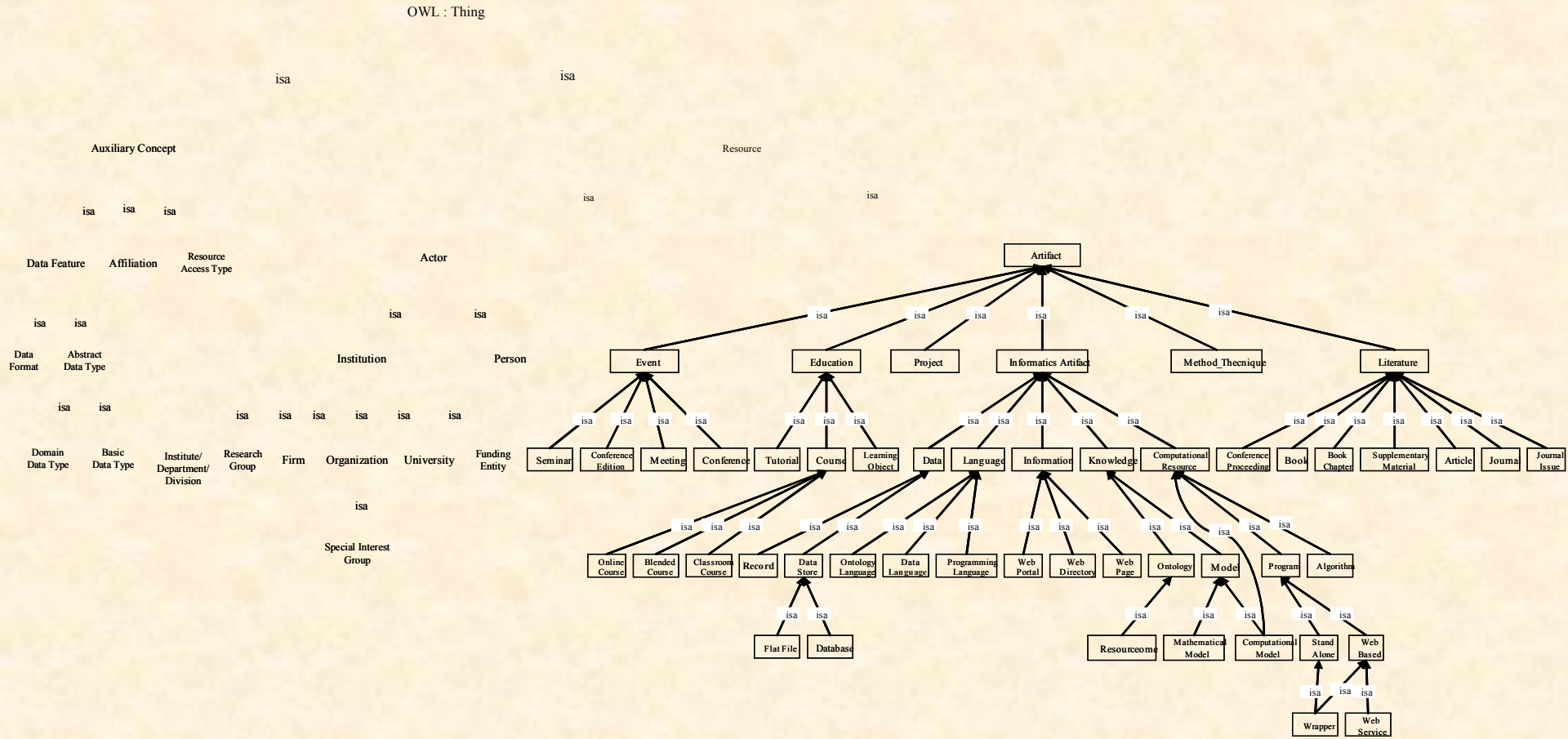
The field of bioinformatics has blossomed in the last ten years, and as a result, there is a large and increasing number of researchers generating computational tools for solving problems relevant to biology. Because the number of artifacts has increased greatly, it is impossible for many bioinformatics researchers to track tools, databases, and methods in the field—or even perhaps within their own specialty area. More critically, however, biologist users and scientists approaching the field do not have a comprehensive index of bioinformatics algorithms, databases, and literature annotated with information about their context and appropriate use. We suggest that the full set of bioinformatics resources—the "resourceome"—should be explicitly characterized and organized. A hierarchical and machine-understandable organization of the field, along with rich cross-links (an ontology!) would be a useful start. It is likely that a distributed development approach would be required so that those with focused expertise can classify resources in their area, while providing the metadata that would allow easier access to useful existing resources.

keyword searching [5]. However, the lack of standard terms makes sensitive and specific searches difficult. In addition, most search hits confound papers, Web sites, tools, departments, and people in a manner that makes extracting useful information very difficult.

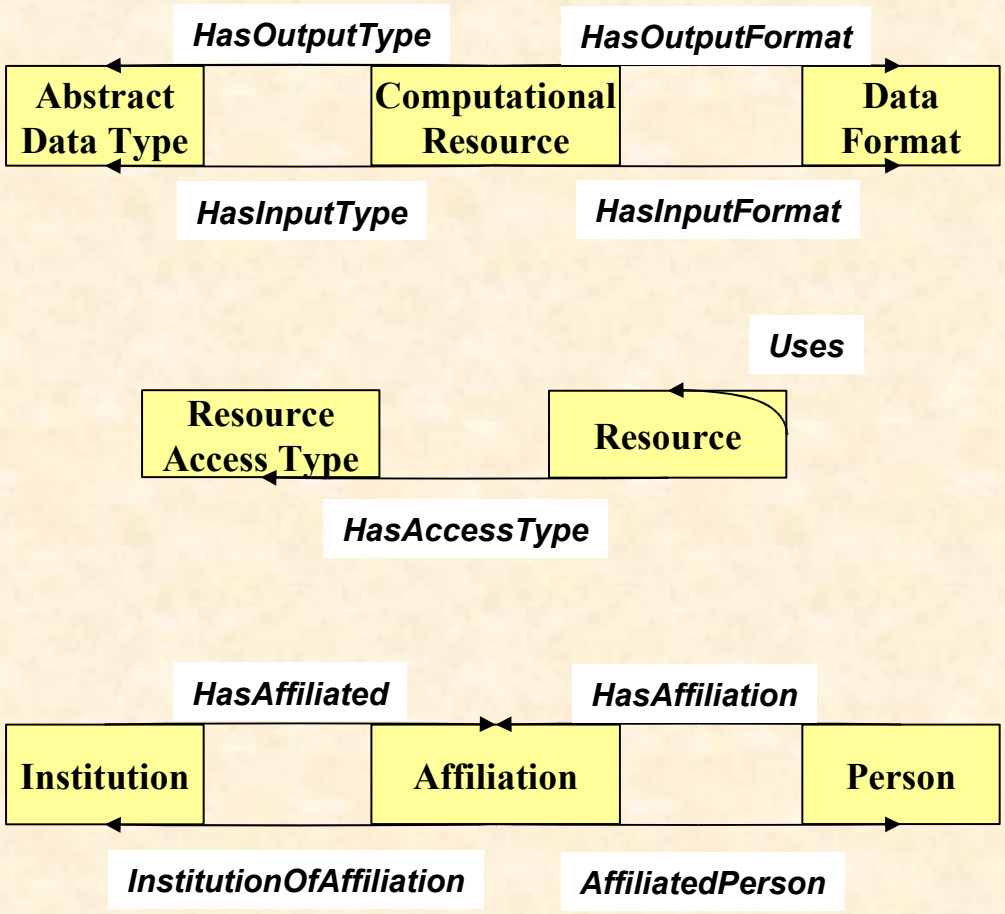
Recognizing this limitation, there have been some grassroots attempts to organize the bioinformatics resourceome. Among the most famous are the "archaeological" Pedro's List—a list of computer tools for molecular biologists (http://www.public.iastate.edu/~pedro/research_tools.html)—and the Expasy Life Sciences Directory, formerly known as the Amos's WWW links page (http://www.expasy.org/links.html). The Bioinformatics Links Directory (http://www.bioinformatics.ubc.ca/resources/links_directory/) today contains more than 700 curated links to bioinformatics resources, organized into eleven main categories, including all the databases and Web servers yearly listed in the dedicated *Nucleic Acids Research* special issues [6]. The National Center for Biotechnology Institute has tried to make access to its suite of tools transparent, with moderate success. Many Web sites can be found listing "useful sites," especially concerning special interest or limited topics (e.g., microarrays, text mining, and gene regulation). But all of these efforts are limited by the difficulty in maintaining



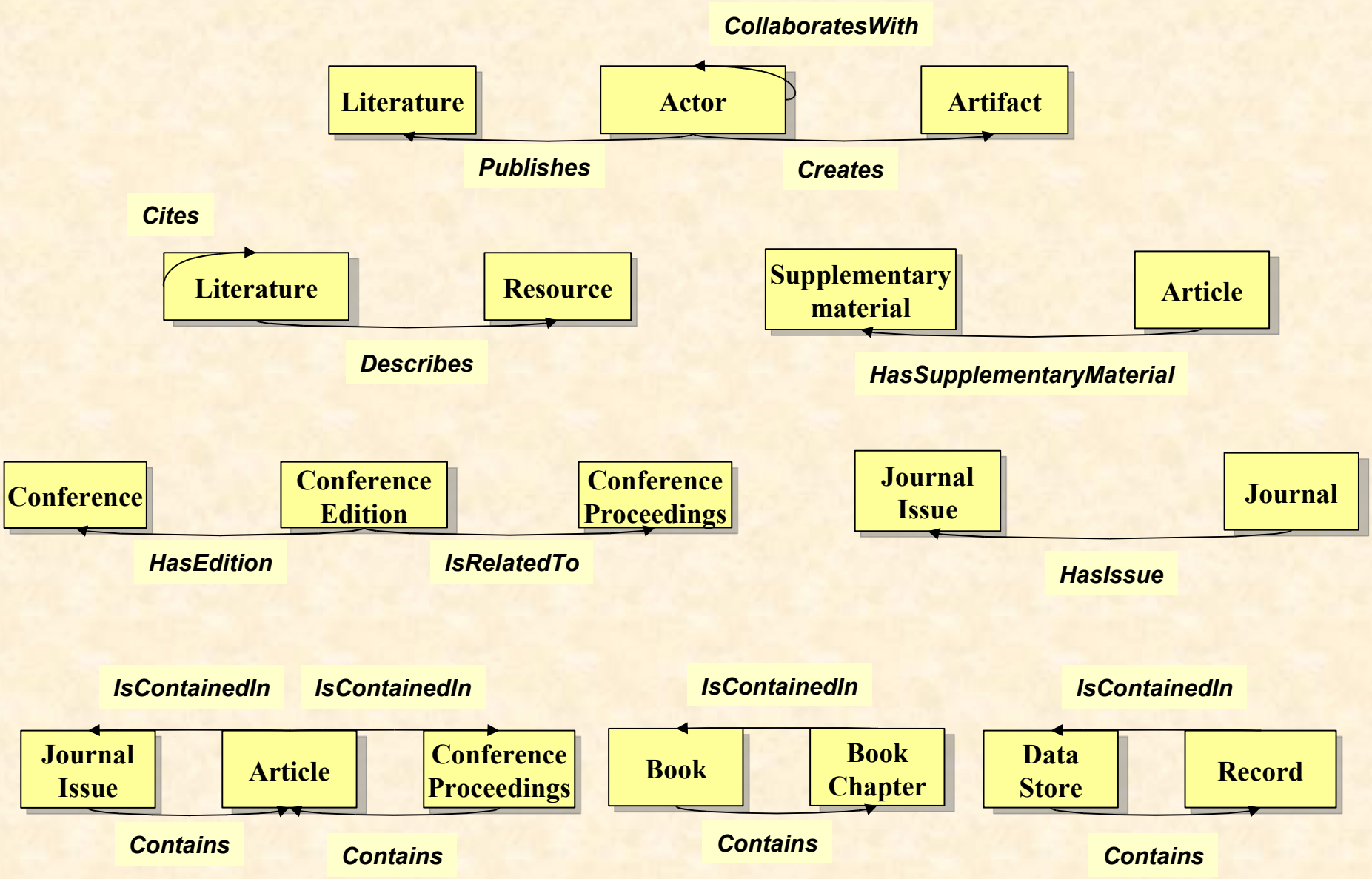
Proposal of an architecture for Resourceomes (Cannata et al. Submitted)



A proposal for a resource ontology

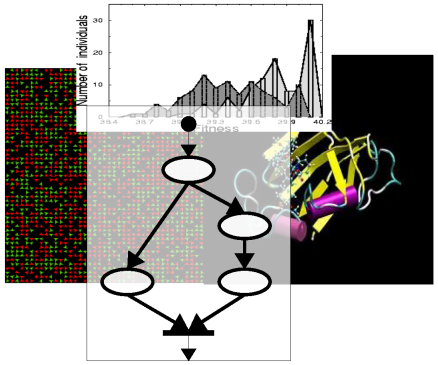


Semantic relationships between resources



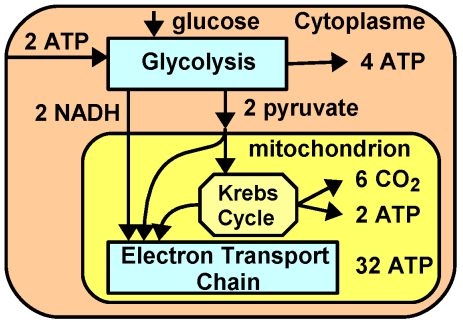
Semantic relationships between resources

A GRID-based multilayer architecture for bioinformatics



In-silico experiments

Resource Discovery



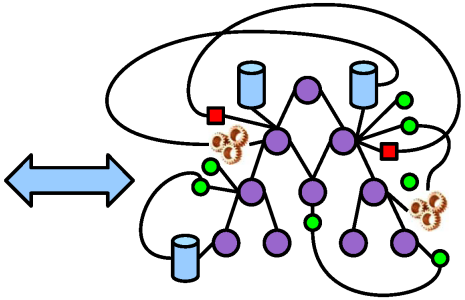
BioSimulation

WebWFlow

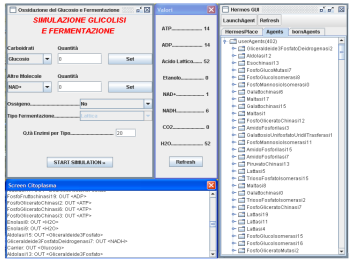
XPDL-to-MAS Compiler

Workflow Executors

BioWMS



Resourceome

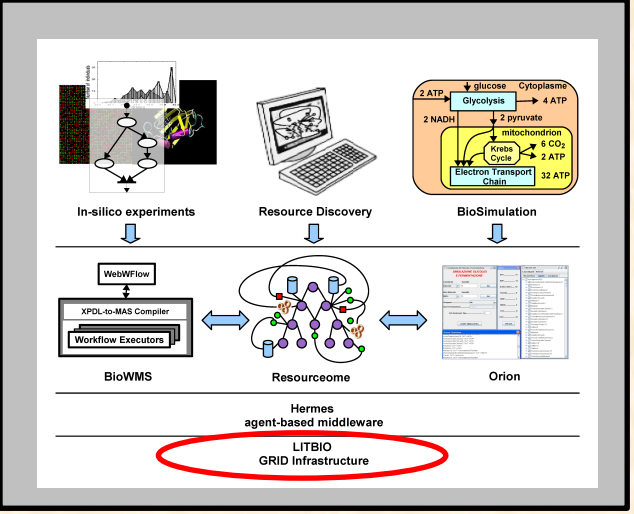


Orion

Hermes
agent-based middleware

LITBIO
GRID Infrastructure

A GRID-based multilayer architecture for bioinformatics



The Grid

- The concept of Grid is evolving (or better specializing)
- In our case it will be a “classical” Grid infrastructure provided in the LITBIO project (nodes with HPCs)

BIOINFORMATICS

myGrid: personalised bioinfo
information grid

Robert D. Stevens [1], Alan J. Robinson

The Semantic Grid: A Future e-Science Infrastructure

David De Roure, Nicholas R. Jennings and Nigel R. Shadbolt[1]

Dept of Electronics and Computer Science,
University of Southampton,
Southampton SO17 1BJ, UK

{dder,nrj,nrs}@ecs.soton.ac.uk

The Anatomy of the Grid

Enabling Scalable Virtual Organizations *

Ian Foster *¶ Carl Kesselman § Steven Tuecke *

{foster, tuecke}@mcs.anl.gov, carl@isi.edu

Abstract

“Grid” computing has emerged as an important new field, distinguished from conventional distributed computing by its focus on large-scale resource sharing, innovative applications, and, in some cases, high-performance orientation. In this article, we define this new field. First, we review the “Grid problem,” which we define as flexible, secure, coordinated resource sharing among dynamic collections of individuals, institutions, and resources—what we refer to as *virtual organizations*. In such settings, we encounter unique authentication, authorization, resource access, resource discovery, and other challenges. It is this class of problem that is addressed by Grid technologies. Next, we present an extensible and open *Grid architecture*, in which

Available at
www.ComputerScienceWeb.com
POWERED BY SCIENCE DIRECT®

FGCS
FUTURE GENERATION COMPUTER SYSTEMS

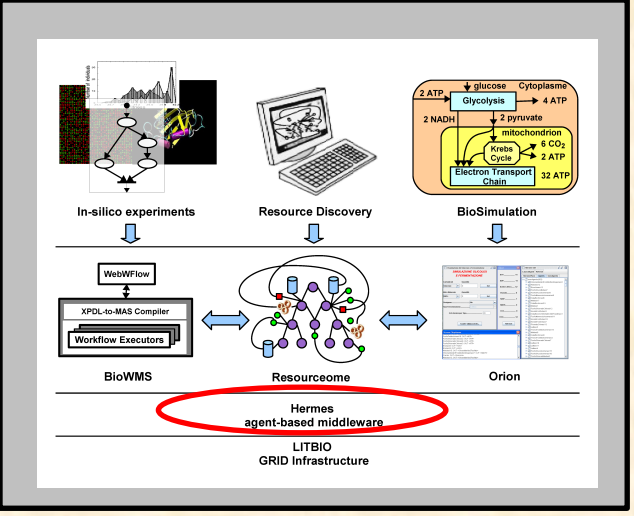
ture Generation Computer Systems 20 (2004) 101–111

www.elsevier.com/locate/future

rative assessment approach for Knowledge Grid

Hai Zhuge*, Jie Liu

ch Group, Key Lab of Intelligent Information Processing, Institute of Computing Technology, hinese Academy of Sciences, P.O. Box 2704-28, 100080 Beijing, China



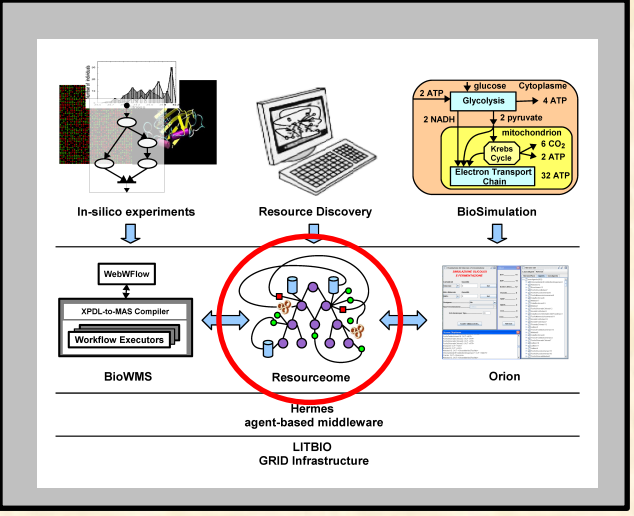
Hermes

- An agent-based middleware developed at Camerino
- Provide a common runtime support to agent based applications

What is an agent?

A computer system capable of flexible, autonomous (problem-solving) actions, situated in dynamic, unpredictable and typically multi-agent environment.

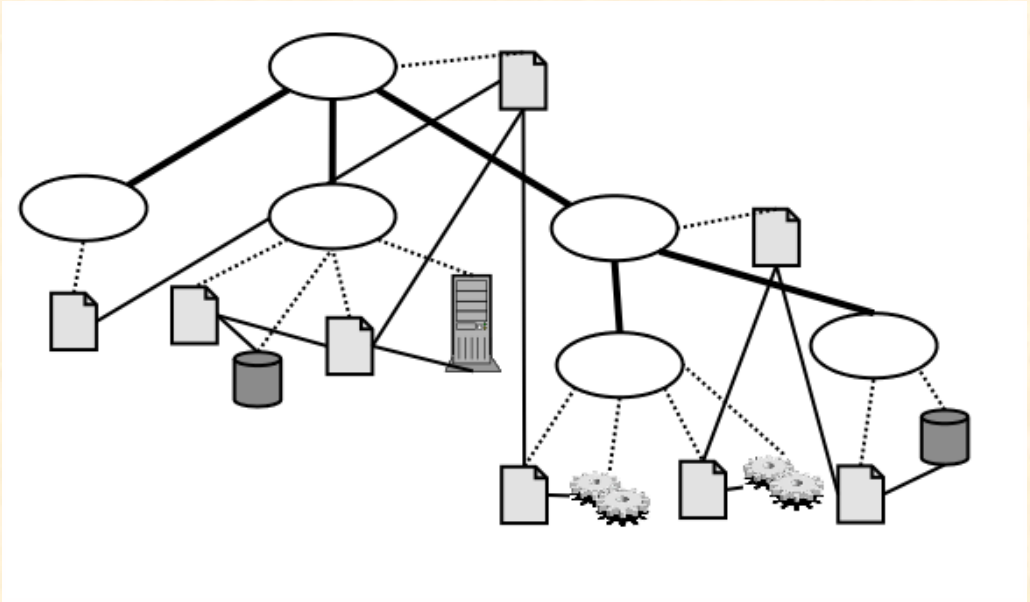
Active and reactive: responds in timely fashion to environmental change
proactive: acts in anticipation of future goals
cooperative: communicates with other agents to reach its goal
mobile: moves across distributed environments (execution platforms)
intelligent: reasons over its knowledge base and by managing ontologies

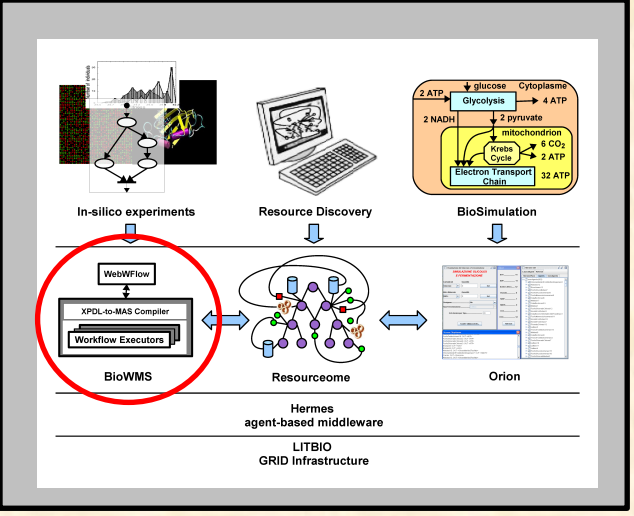


Resourceome

- A conceptual map of a domain with related resources

- It is the Pivot of the architecture
- Will permit to semantically organize workflows, activities, services
- Will permit to semantically organize systems biology resources and knowledge
- Will permit to “reason” over resources
- Agents will build, maintain and keep it “alive”

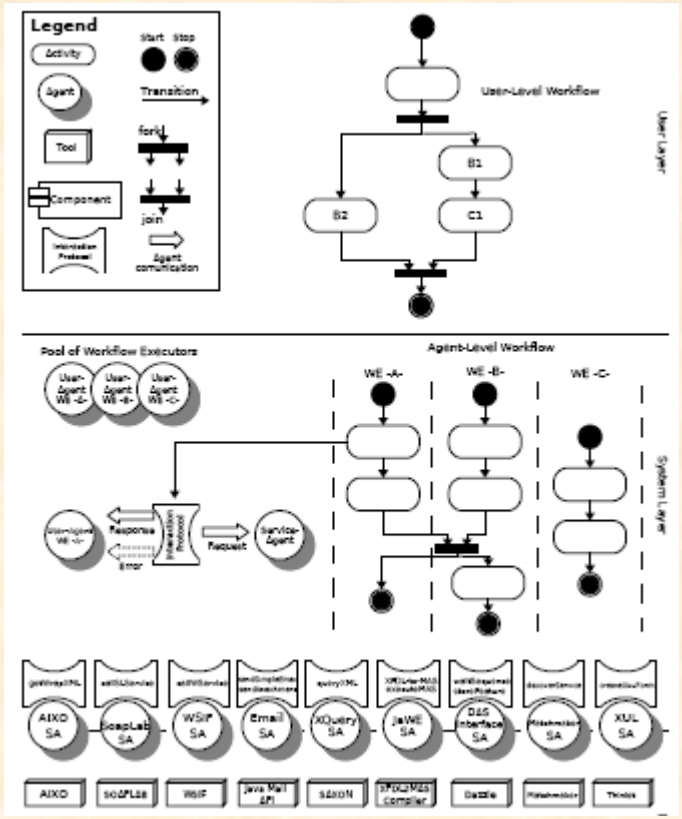


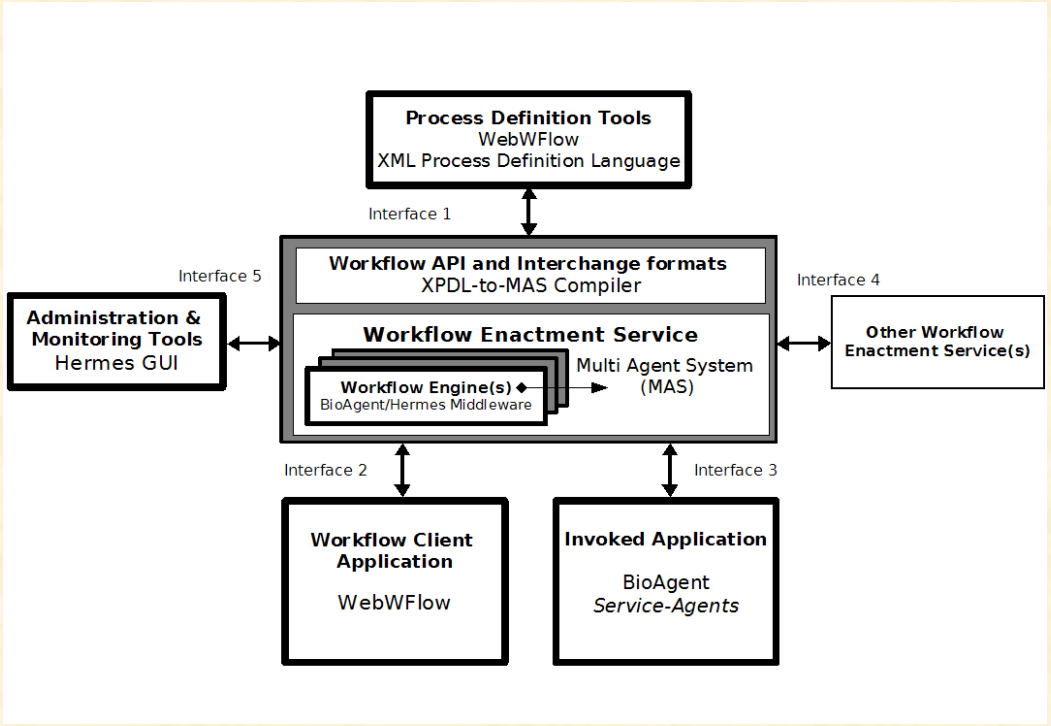


BioWMS

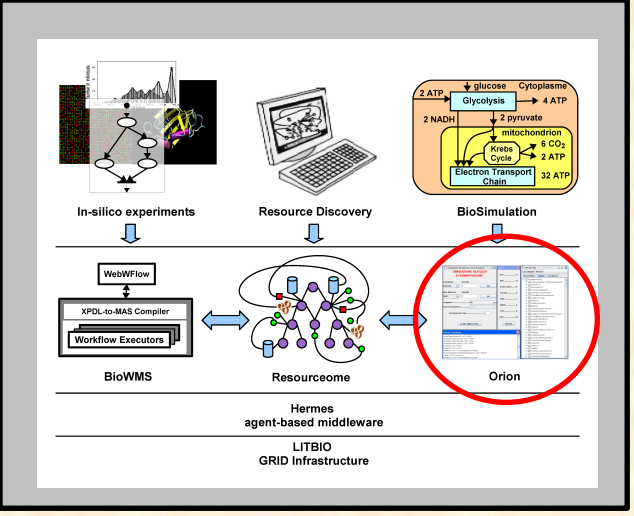
- An agent-based Workflow Management System for bioinformatics

- Workflows are defined in XPDL (standard for WfMC)
- Compiles a user workflow of activities into a MAS
- Bioinformatics services and applications are wrapped by service agents

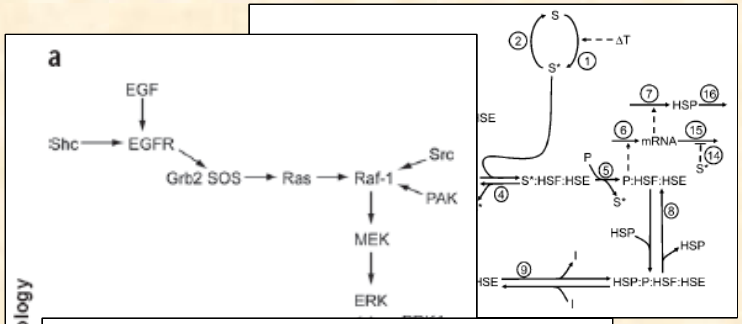


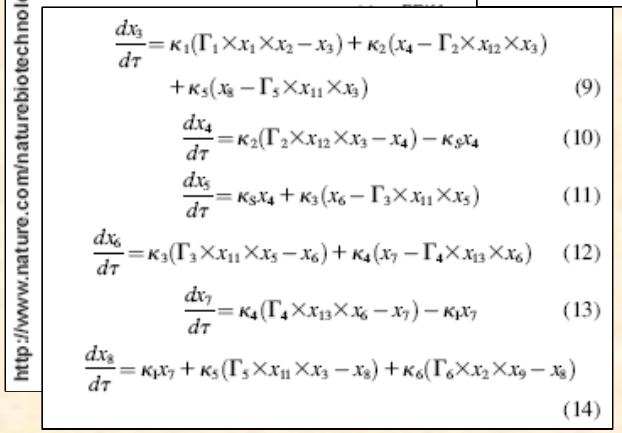


Orion



- An agent-based framework for systems biology
- Behavioural modeling of molecular entities
- We are starting with metabolic reactions…





$$\frac{dx_3}{d\tau} = \kappa_1(\Gamma_1 \times x_1 \times x_2 - x_3) + \kappa_2(x_4 - \Gamma_2 \times x_{12} \times x_3) + \kappa_5(x_8 - \Gamma_5 \times x_{11} \times x_3) \quad (9)$$

$$\frac{dx_4}{d\tau} = \kappa_2(\Gamma_2 \times x_{12} \times x_3 - x_4) - \kappa_8 x_4 \quad (10)$$

$$\frac{dx_5}{d\tau} = \kappa_8 x_4 + \kappa_3(x_6 - \Gamma_3 \times x_{11} \times x_5) \quad (11)$$

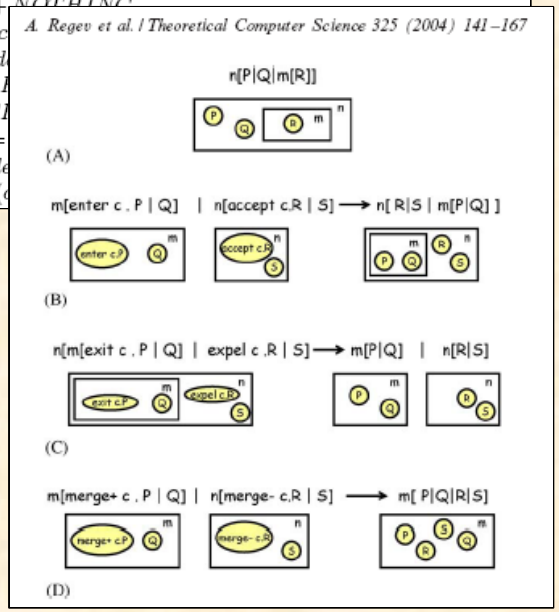
$$\frac{dx_6}{d\tau} = \kappa_3(\Gamma_3 \times x_{11} \times x_5 - x_6) + \kappa_4(x_7 - \Gamma_4 \times x_{13} \times x_6) \quad (12)$$

$$\frac{dx_7}{d\tau} = \kappa_4(\Gamma_4 \times x_{13} \times x_6 - x_7) - \kappa_1 x_7 \quad (13)$$

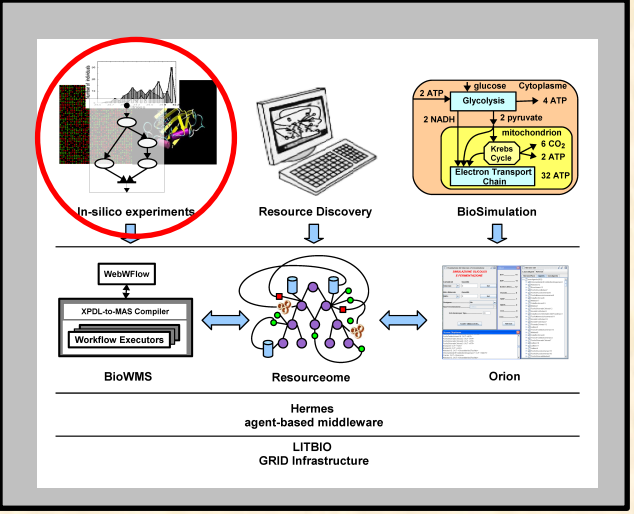
$$\frac{dx_8}{d\tau} = \kappa_1 x_7 + \kappa_5(\Gamma_5 \times x_{11} \times x_3 - x_8) + \kappa_6(\Gamma_6 \times x_2 \times x_9 - x_8) \quad (14)$$

http://www.nature.com/naturebiotechnology

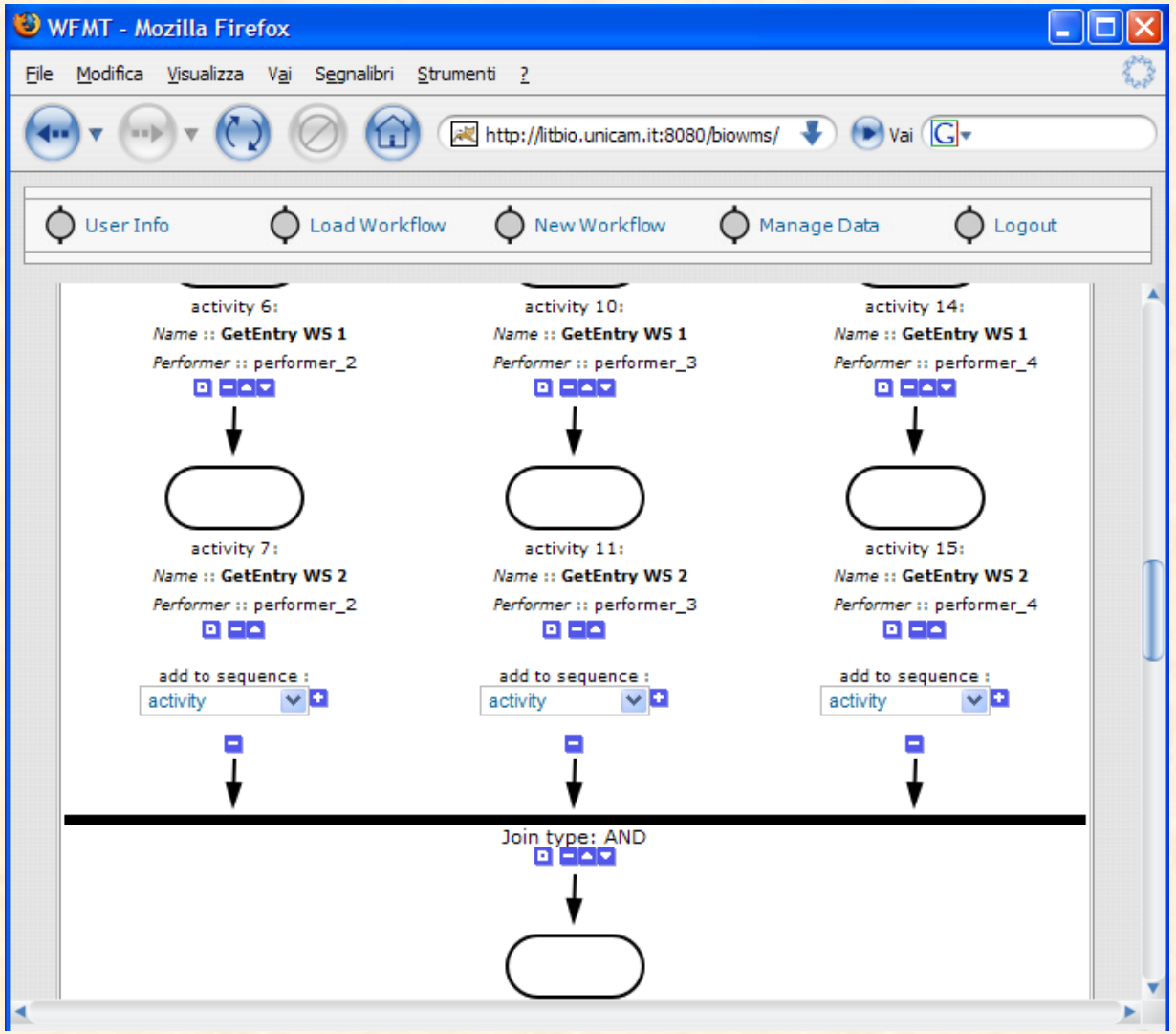
$$SYSTEM ::= CYCLIN|CDK|CDH1|CDC14|CKI|CLOCK$$
$$CYCLIN ::= (\nu\, bb)BINDING\_SITE$$
$$BINDING\_SITE ::= (\overline{lb}\langle bb\rangle, R_4).CYCLIN\_BOUND$$
$$CYCLIN\_BOUND ::= DEGCYC + DEGCKI + CYC\_CDK\_CKI$$
$$DEGCYC ::= (\overline{degp}, R_1).\overline{degc}.0$$
$$DEGCKI ::= (\overline{degd}, R_3).CYCLIN\_BOUND$$
$$CYC\_CDK\_CKI ::= (\overline{bind}\langle bb\rangle, R_{11}).\overline{bb}.TRIM$$
$$TRIM ::= DIM + NOTHING$$
$$DIM ::= (removec$$
$$NOTHING ::= (d$$
$$CDK ::= (lb(cbb), I$$
$$CDK\_CATALYTI$$
$$INACTCDH1 ::=$$
$$NEWCDK ::= (de$$
$$INACTCAT ::= (c$$



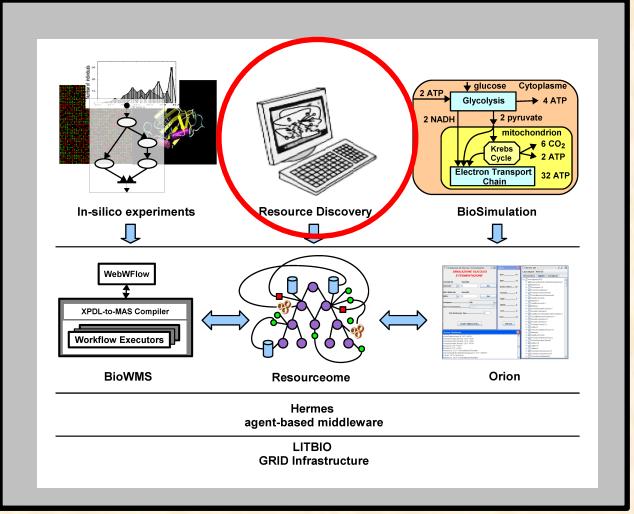
Will agents permit to “bridge” the “two” systems biologies?

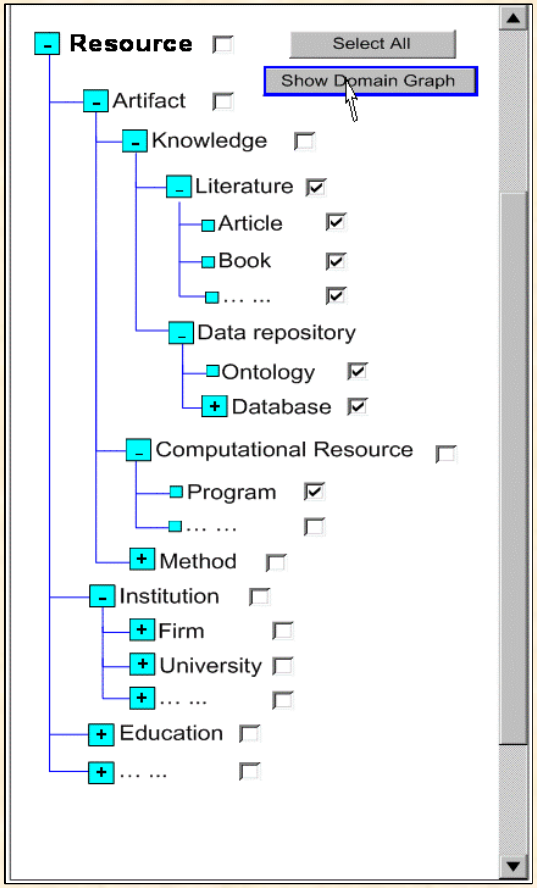


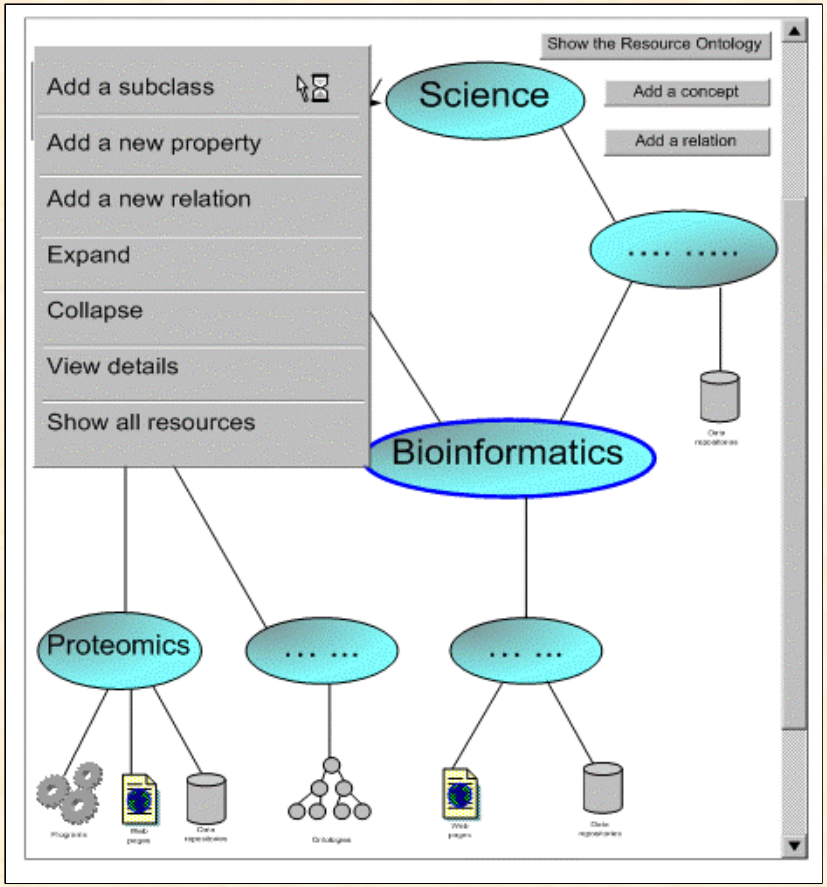
In-silico experiment

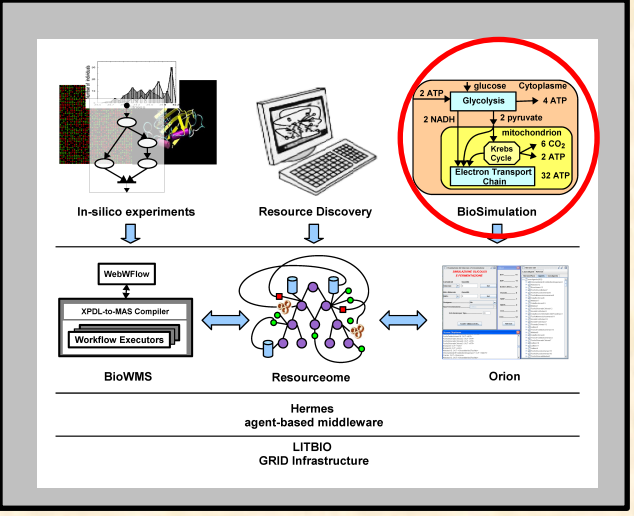


Resource discovery

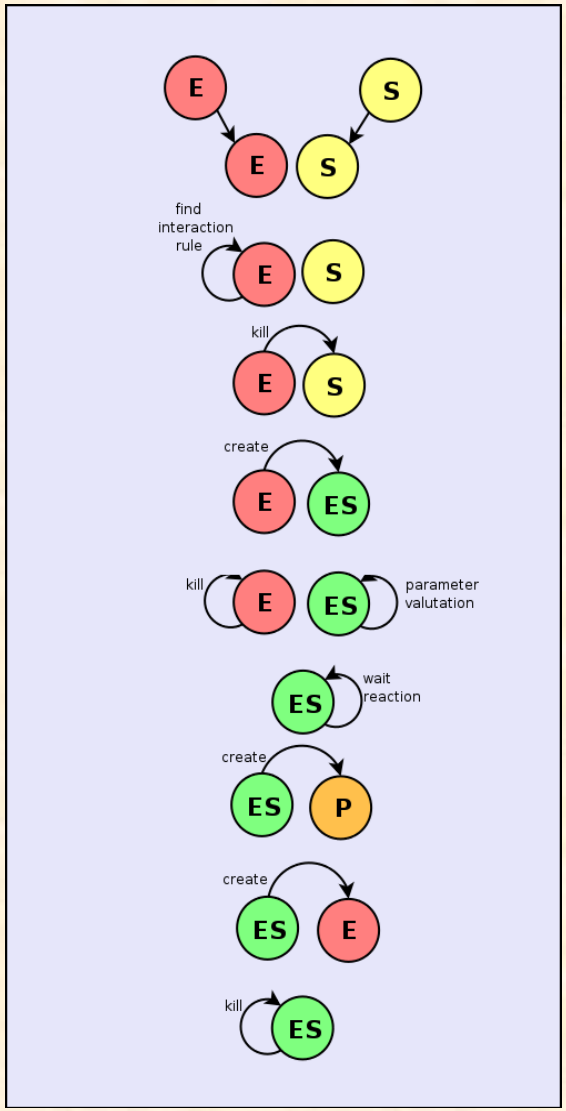






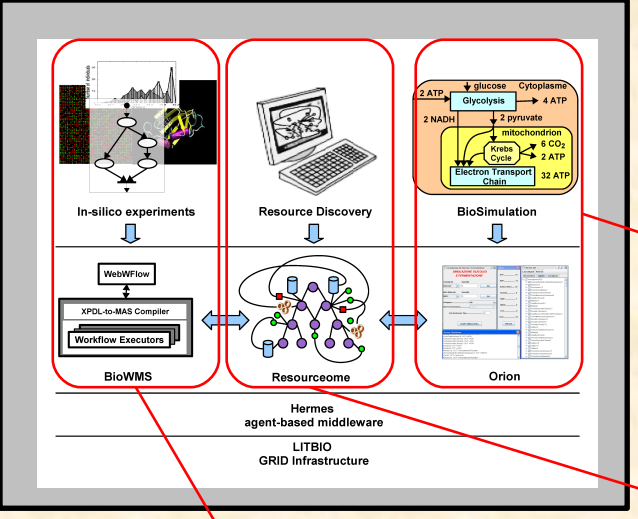


Simulation of biological process



Conclusions

- LITBIO aims to offer a Laboratory for Interdisciplinary Technologies in Bioinformatics
- Bioinformatics is fast changing and growing
- Resourceome can make interdisciplinarity easier
- We propose an agent-based multi-layer architecture which will lay on the LITBIO Grid
- User-friendly and user-assisting
- We hope it will make easier the life of molecular biologists, bioinformaticians and systems biologists of today and tomorrow



Acknowledgment

ORION

Arianna Baldoncini
Claudio Forcato
Michele Mattioni

Mauro Angeletti
Riccardo Piergallini

RESOURCEOME

Sergio Gabrielli
Luana Leoni
Francesca Piersigilli
Leonardo Vito

Rosario Culmone

BIOWMS

Lorenzo Scortichini