



BioinfoGRID
Bioinformatics Grid Application for life science



BioinfoGRID: Bioinformatics GRID Based Applications Overview



Milanesi, Luciano

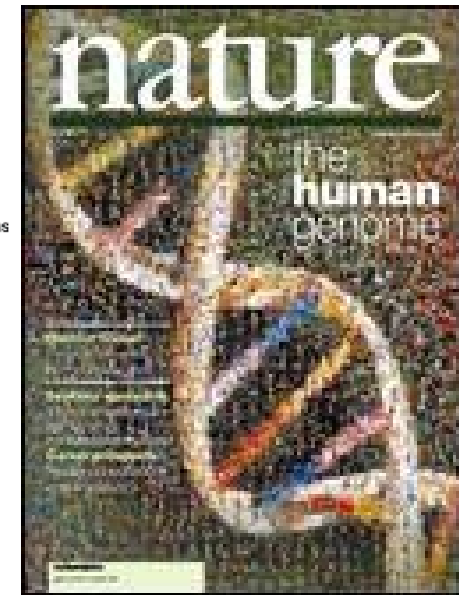
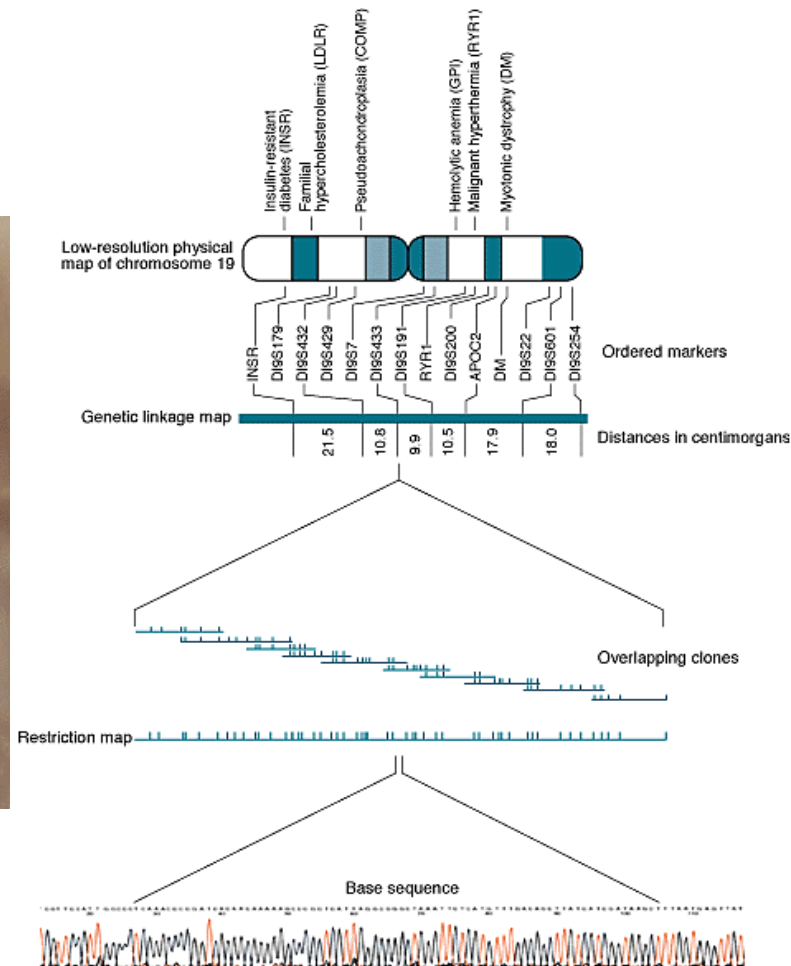
National Research Council

Institute of Biomedical Technologies, Milan, Italy

luciano.milanesi@itb.cnr.it



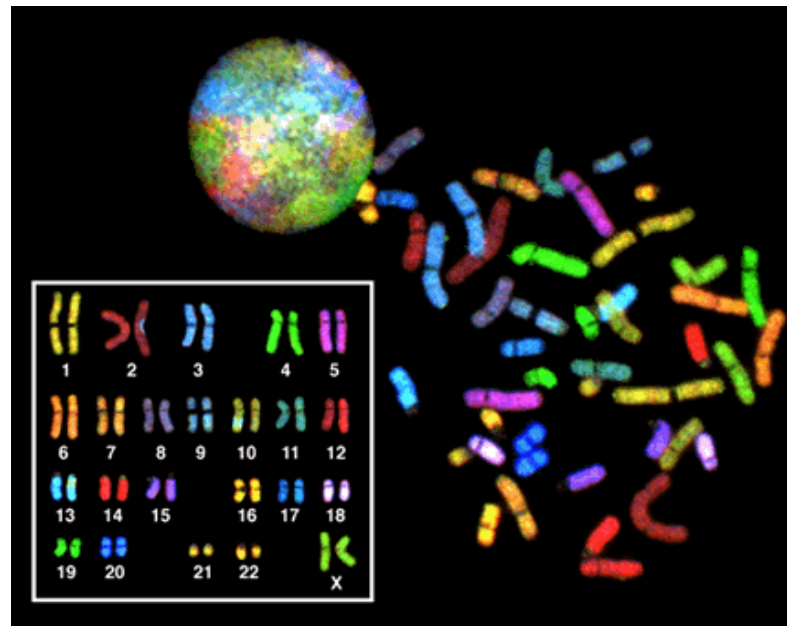
Human Genome





Post-genomic

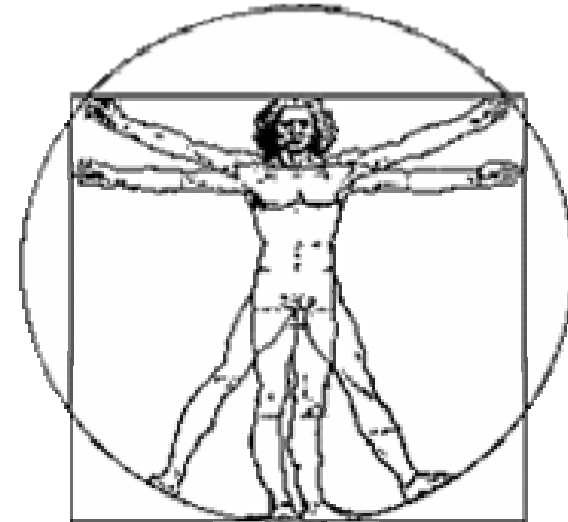
- “Post-genomic” focuses on the new tools and new methodologies emerging from the knowledge of genome sequences.
- Production and use of DNA micro arrays, analysis of transcriptome, proteome, metabolome are the different topics developed in this class.





The human organism:

- ~ 3 billion nucleotides
- ~ 30,000 genes coding for
- ~ 100,000-300,000 transcripts
- ~ 1-2 million proteins
- ~ 60 trillion cells of
- ~ 300 cell types in
- ~14,000 distinguishable morphological structures





Genome-wide analysis

- Current interest in the genome-wide analysis of cells at the level of transcription ('**transcriptome**') and translation ('proteome'), the third level of analysis is the 'metabolome'.
- The term '**metabolome**' refers to the entire complement of all the small molecular weight metabolites inside a cell suspension of interest.
- A new level of experiments are required to obtain an overall picture of **when, where, and how gene are expressed**.
- The **functional genomics** includes:
 - The analysis of gene expression profiles at the mRNA and protein levels
 - The analysis of polymorphism or mutation patterns in the genome



Human Genetic Diversity

- Any two individuals differ in about **3×10^6 bases (0.1%)**.
- The population is now about **5×10^9** .
- A catalog of all sequence differences would require **15×10^{15}** entries.
- This catalog may be needed to find the rarest or most complex disease genes.
- Less than **5%** of the **3×10^6 bases genome** encodes genes.
- A conservative estimate of the number of genes gives a value of about **25.000 genes**.
- The structural diversity within the proteins encoded by these genes is considerably greater than this small number of true genes.



- A typical gene lab can produce **100 terabytes of information a year**, the equivalent of **1 million encyclopedias**.
- Few biologists have the computational skills needed to fully explore such an **astonishing amount of data**; nor do they have the skills to explore the exploding amount of data being generated from clinical trials.
- The immense amount of data that are available, and the knowledge is the **tip of the data iceberg**.

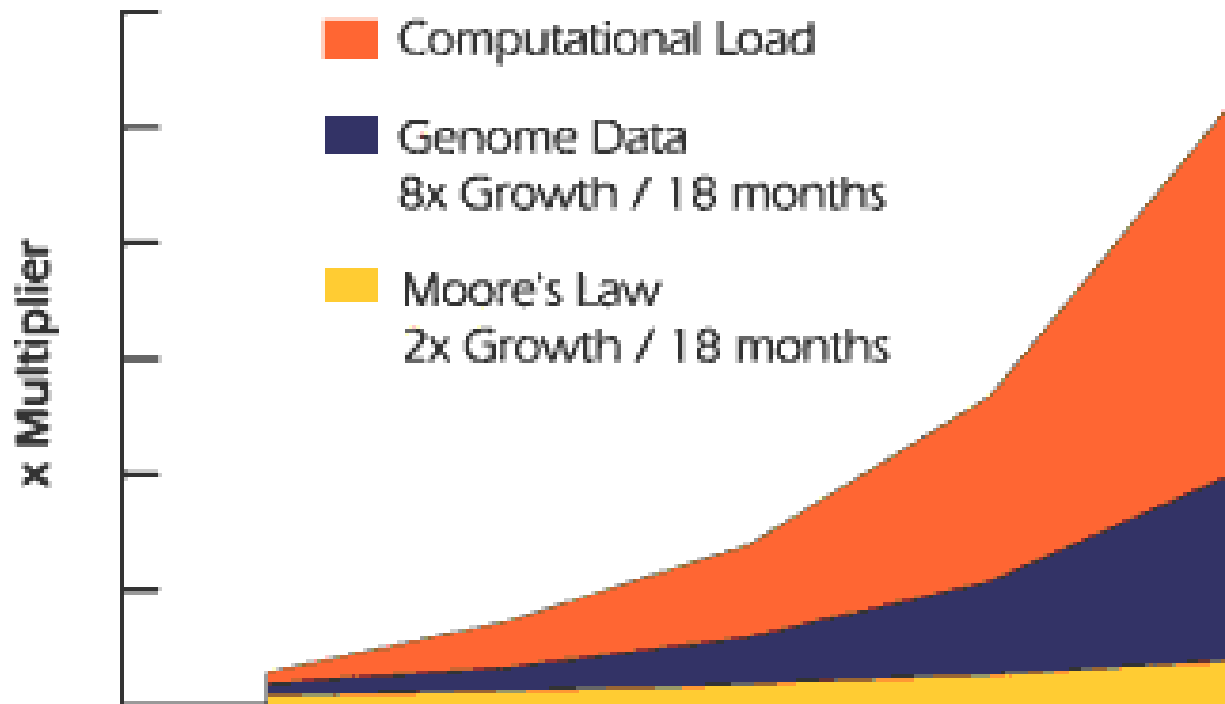
Bioinformatics: Emerging Opportunities and Emerging Gaps₁

Paula E. Stephan and Grant Black



Database explosion

- In the very beginning of the genome sequencing era, Walter Gilbert and colleagues warned of database explosion, stemming from the exponentially increasing amount of incoming DNA sequence





Complex disease Mapping

- The human genome will revolutionize medical practice and biological research into the 21st century.
- **All human genes will be studied, and accurate diagnostics will be developed for most inherited diseases.**
- Researchers have already identified single genes associated with a number of diseases, such as:
- **Cystic fibrosis,**
- **Duchenne muscular dystrophy,**
- **Myotonic dystrophy,**
- **Neurofibromatosis, and Retinoblastoma, ecc.**

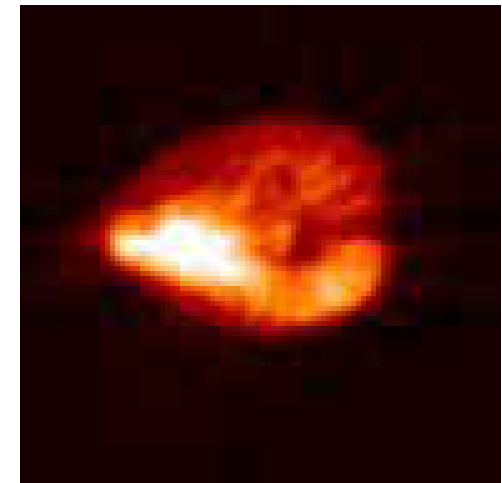
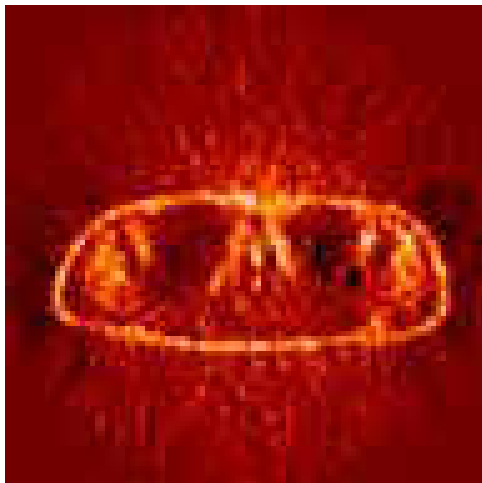


Human Genome and Medicine

- As research progresses, investigators will also uncover the mechanisms for diseases caused by **several genes** or by a **gene interacting with environmental factors**.
- The identification of these genes and their proteins will be useful in finding more-effective therapies and preventive measures.
- Investigators determining the underlying **biology of genome organization and gene regulation** will also begin to understand how humans develop from single cells to adults.

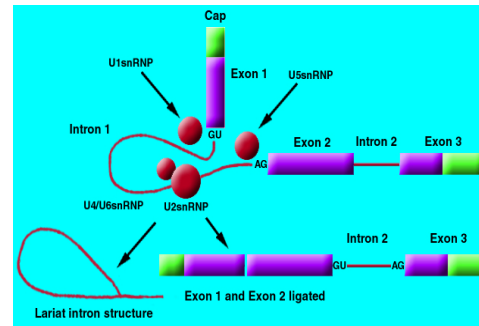


Genomics and Nuclear Medicine





Bioinformatics



The Bioinformatics analysis will produce

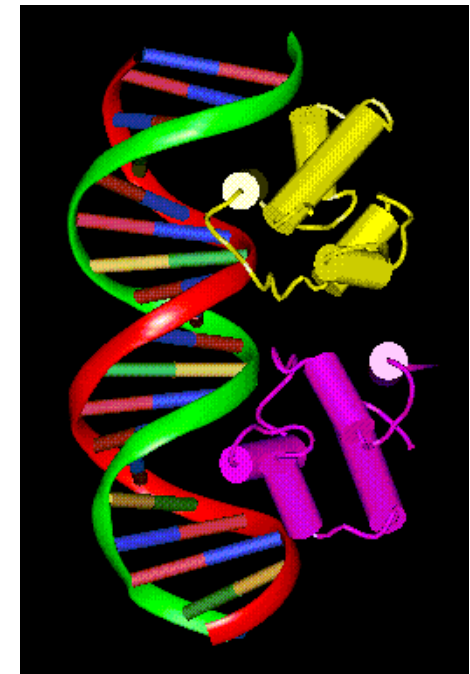
Annotated
Sequences from
Different genomes

Structural description
of interaction
interfaces



Structural Genomics

- The structural information will be used to obtain:
- Family assignment of new sequences
- Recognition of structural interaction motifs
- Prediction of interacting proteins
- The identified structural motif can be used to recognize patterns of interaction between a sequence of interest and its interactor
- The predicted complex can be modelled by molecular docking approach





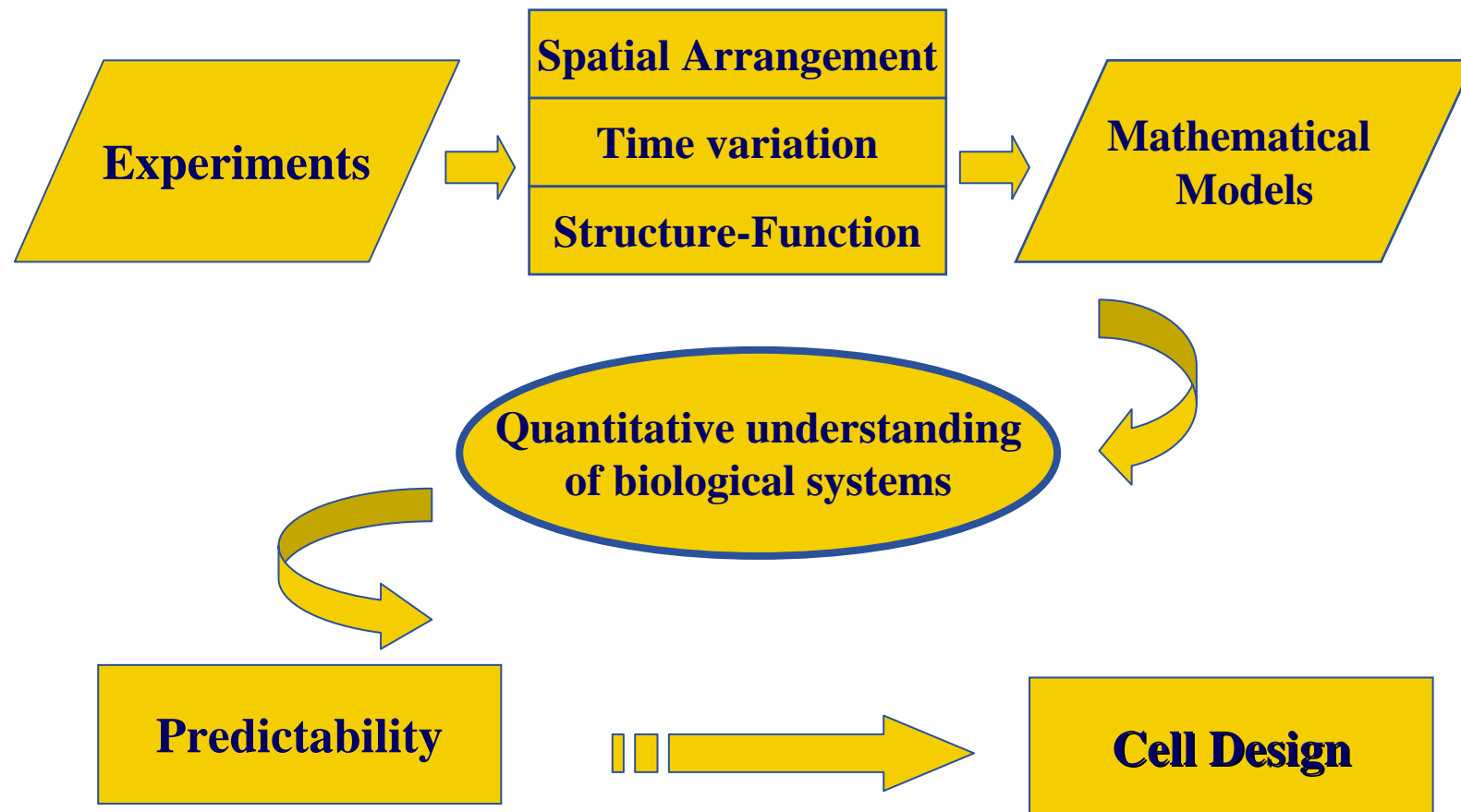
Structural Genomics

- A key development in the computational world has been the arrival of ***de novo* design algorithms** that use all available spatial information to be found within the target to **design novel drugs**.
- Coupling these algorithms to the rapidly growing body of information from structural genomics together with the new **ICT technology (eg. HPC, GRID, ecc.)**
- provides a powerful new possibility for exploring design to a broad spectrum of genomics targets, including more challenging techniques such as:
- **protein–protein interactions, docking, molecular dynamics, system biology, gene network ecc.**



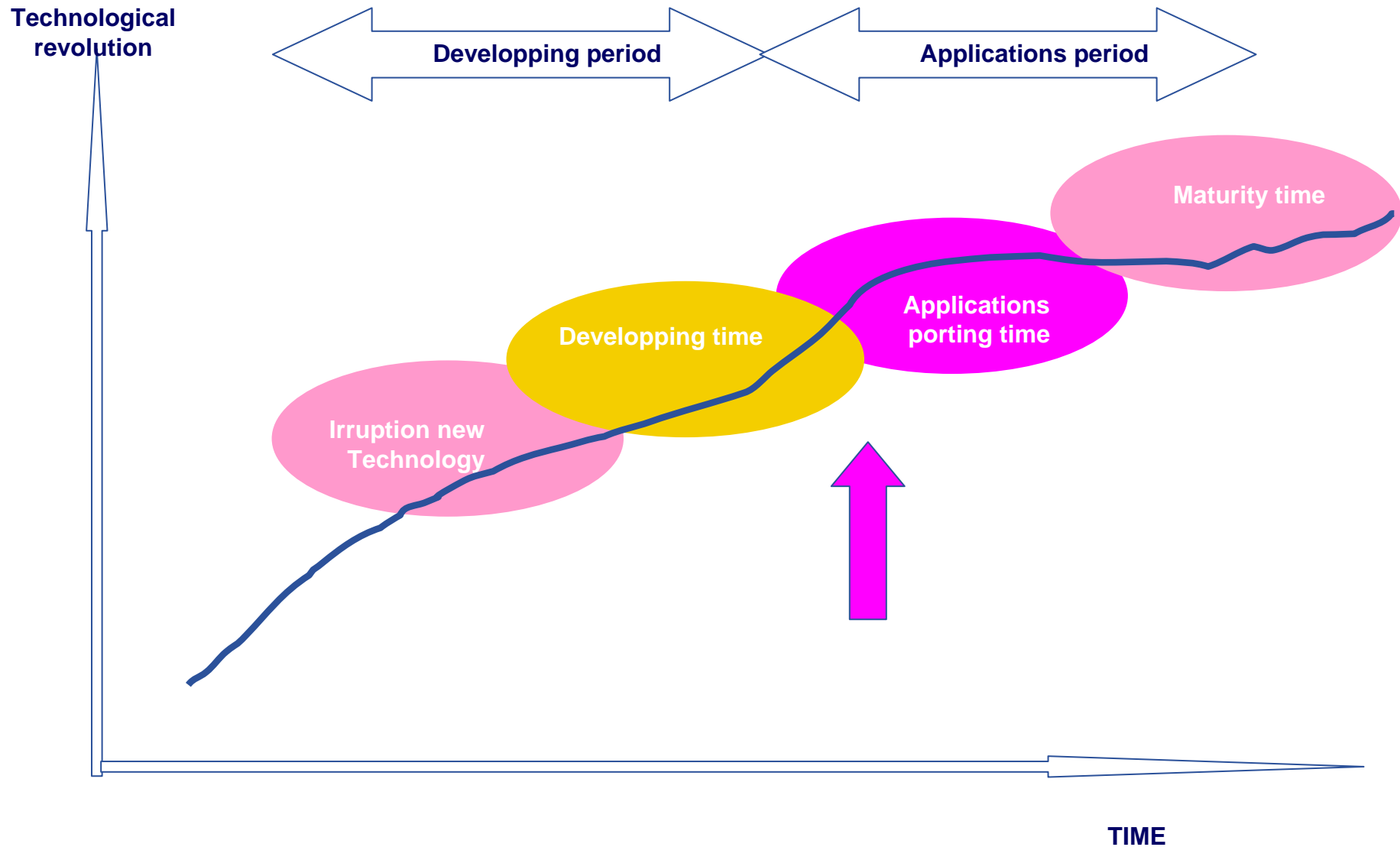
Quantitative Biology

In silico comparative analysis of genes can be used in cell cycle for structural motifs determination in the homologous protein families and for system biology





GRID Technological Revolution

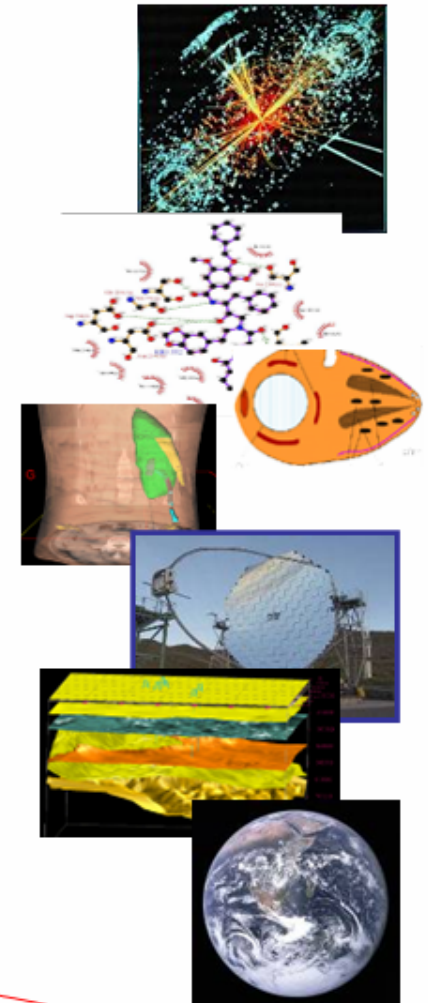


- **Manage and operate production Grid infrastructure for the European Research Area**
- **Interoperate with e-Infrastructure projects around the globe**
- **Contribute to Grid standardisation efforts**

- **Support applications deployed from diverse scientific communities**
 - High Energy Physics
 - Biomedicine
 - Earth Sciences
 - Astrophysics
 - Computational Chemistry
 - **Fusion**
 - Geophysics (supporting the Industrial application, EGEODE)
 - Finance, Multimedia
 -

- **Reinforce links with the full spectrum of interested industrial partners**
- **Disseminate knowledge about the Grid through training**

- **Prepare for a permanent/sustainable European Grid Infrastructure (in a GEANT2-like manner)**





EGEE/LCG-2 Grid Sites : September



EGEE/LCG-2 grid:

160 sites, 36 countries

>15,000 processors,

~5 PB storage

Other national & regional grids:

~60 sites,

>6,000 processors

country	sites	country	sites	country	sites
Austria	2	India	1	Russia	10
Belgium	1	Israel	2	Singapore	1
Bulgaria	4	Italy	25	Slovakia	3
Canada	6	Japan	1	Slovenia	1
China	1	Korea	1	Spain	13
Croatia	1	Netherlands	2	Sweden	2
Cyprus	1	Macedonia	1	Switzerland	2
Czech Republic	2	Pakistan	2	Taiwan	4
France	8	Poland	4	Turkey	1
Germany	8	Portugal	1	UK & Ireland	35
Greece	6	Puerto Rico	1	USA	3
Hungary	1	Romania	1	Yugoslavia	1



BioinfoGRID Project descriptions



- The **BIOINFOGRID projects** proposes to combine the Bioinformatics services and applications for molecular biology users with the Grid Infrastructure by EGEE and EGEEII projects.
- In the BIOINFOGRID initiative we plan **to evaluate genomics, transcriptomics, proteomics and molecular dynamics applications studies based on GRID technology.**
- The project start date: 1st January 2006
- The project finish date: 31 December 2007



The grid application aspects.

- The massive potential of Grid technology will be indispensable when dealing with both the complexity of models and the enormous quantity of data, for example, in searching the human genome or when carry out simulations of molecular dynamics for the study of new drugs.
- The BIOINFOGRID projects proposes to combine the Bioinformatics services and applications for molecular biology users with the Grid Infrastructure created by EGEE



EGEE

Enabling Grids for E-science



Genomics applications in GRID

Aim : use of computational GRID to analyse molecular biological data at the genomic scale

Description

- **the GRID Portal system**: unification of larger groups of bioinformatics tools into single analytical steps and their optimization for GRID
- **GRID analysis of cDNA data**: computer- aided functional annotation of cDNAs in order to optimize sensitivity and specificity



Genomics applications in GRID

- **GRID analysis of genomic databases:** integration of precomputed data, gene identification, differentiation of pseudogenes, comparative genome analysis, etc.
- **Multiple alignments:** testing of new algorithms for computationally very demanding alignment procedures, optimization for GRID.

```
PRTC      TWFLVGLVSWG-EGCGLLHNYGVYTKVSRYLWDWIHGHIRDKEAPQKSWAP-----
FA10     TYFVTGIVSWG-EGCARKGKYGIYTKVTAFLKWDIDRSMKTRGLPKAKSHAPEVITSSPLK
FA7      TWYLTGIVSWG-QGCATVGHFGVYTRVSQYIEWLQKLMRSE-----PRPGVLLRAPFP
THROMBIN RWYQMGIVSWG-EGCDRDGKYGFYTHVFRLLKKWIKVIDQFGE-----
FA9      TSFLTGLIISWG-EECAMKGKYGIYTKVSRVYVNWIKKTKLT-----
KALLIKREIN MURLVGITSWG-EGCARREQPGVYTKVAEYMDWILEKTQS SDGKAQM QS PA-----
FA11     VVHLVGITSWG-EGCAQRE R PGVYTNVVEYVDWILEKTQAV-----
TRYB1    TWLQAGVVSWG-EGCAQPNRPGIYTRVTYYLDWIHHYVPKKP-----
TRYB2    TWLQAGVVSWG-EGCAQPNRPGIYTRVTYYLDWIHHYVPKKP-----
TRYA     TWLQAGVVSWD-EGCAQPNRPGIYTRVTYYLDWIHHYVPKKP-----
KLKE     --QLQGLVSWG M ERCA L PGY PGVYTNLCKYRSWIEETMRDK-----
CTRL    TWVLI GIVSWG-TKNCNVRA PAVYTRVSKFSTWINQVIAYN-----
```



Proteomics Applications in GRID

Aim : use of computational GRIDs to analysis molecular biological data in proteomics

Description

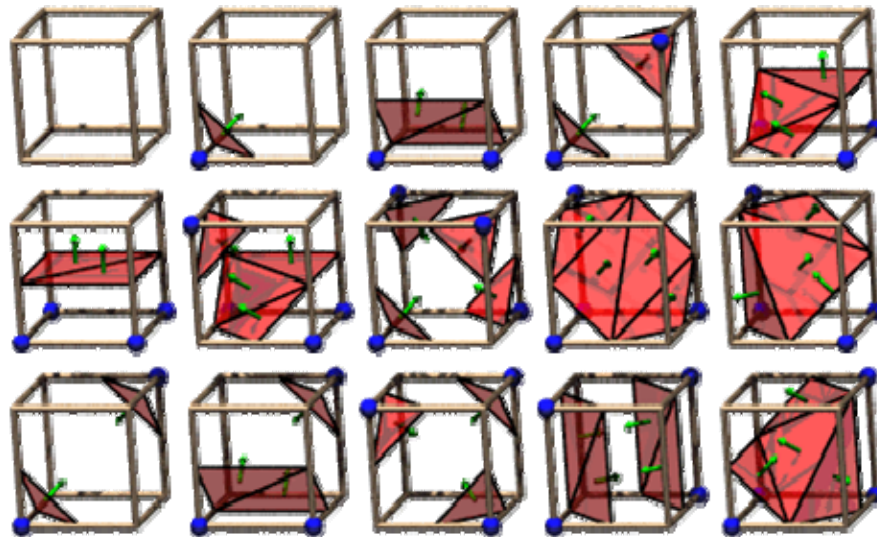
- **Perform functional protein analysis in GRID** by using the functional protein domain annotations on large protein families using GRID and related databases.



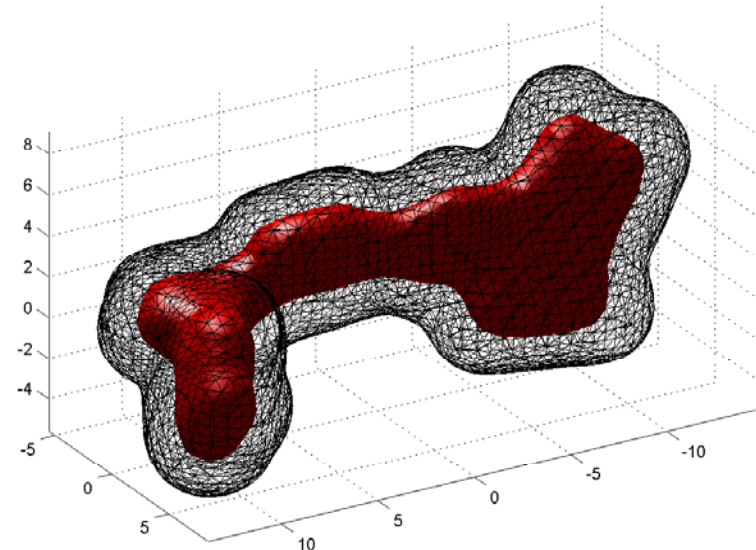


Proteomics Applications in GRID

- **Protein surface calculation in GRID.** : the grid will be used to elaborate the volumetric description of the protein obtaining a precise representation of the corresponding surface.



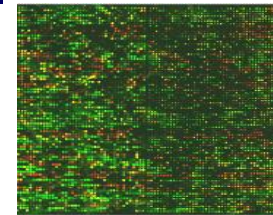
The 15 Cube Combinations





Transcriptomics applications in GRID

Aim : use of computational GRIDs to analyse transcriptomics data and to perform application of Phylogenetic methods based on estimates trees.



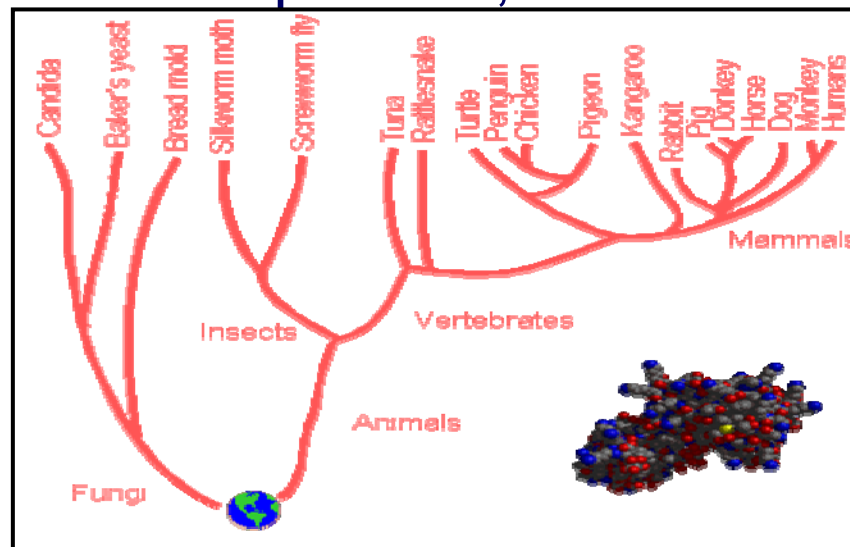
Description

- **To perform algorithmic tools for gene expression data analysis in GRID:** evaluate the computational tools for extracting biologically significant information from gene expression data.
- Algorithms will focus on clustering steady state and time series gene expression data, multiple testing and meta analysis of different microarray experiments from different groups, and identification of transcription sites.



Phylogenetic application in GRID

- **Phylogenetics** : Reconstructing the evolutionary history of a group of taxa is major research thrust in computational biology and a standard part of exploratory sequence analysis. An evolutionary history not only gives relationships among taxa, but also an important tool for inferring the universal **tree of life**, inferring structural, physiological, and biochemical properties of sequences from other similar sequences, and reconstruction of tissue evolution.





Database Applications in GRID

Aim : To manage the biological database, by using the GRID EGEE infrastructure.

Description

- **Biological database on GRID:** these databases will be complemented by others that are publicly available in Internet, by using GRID and web services where appropriate.
- **Functional Analogous Finder:** By using the GO terms and the associations to gene products it is possible to compare the total associated GO terms and their ascending parents to validate the functional analogy between two gene products

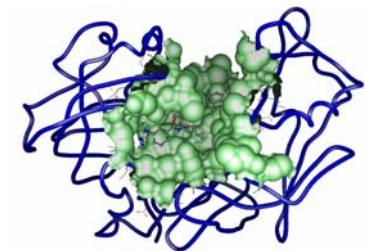


Molecular applications in GRID

Aim : The objective is to docking and Molecular Dynamics simulations, which usually take a very long time to complete the analysis.

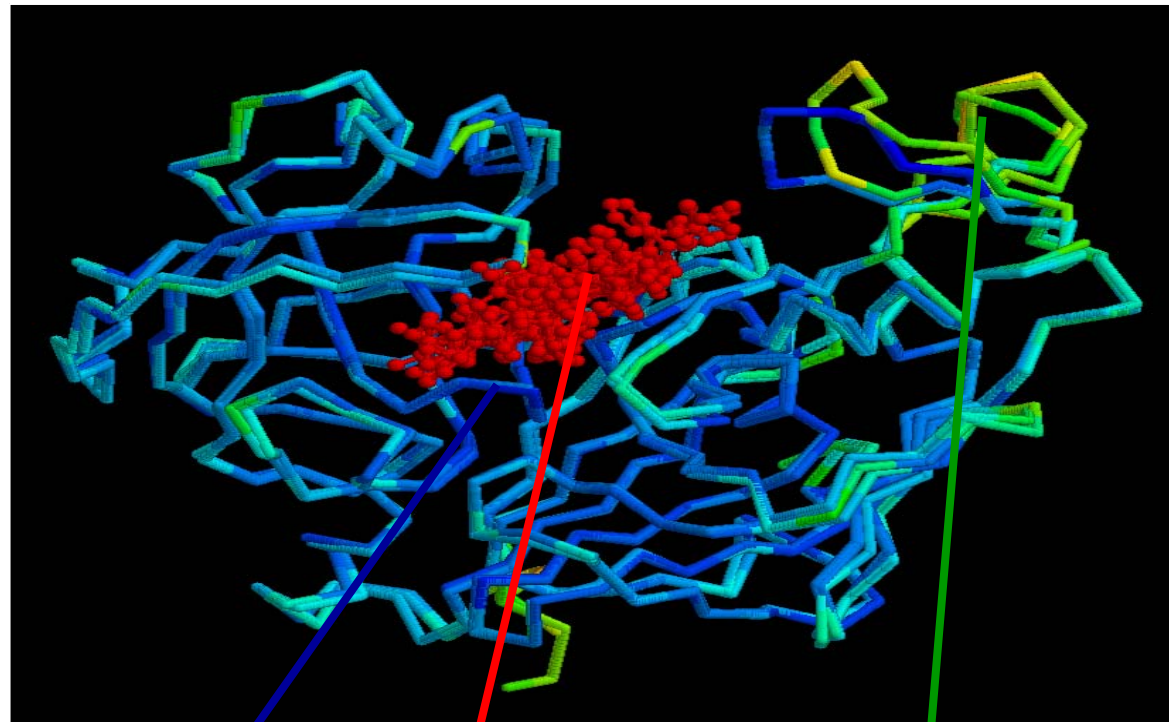
Description

- **Wide *In Silico* Docking On Malaria initiative WISDOM-II:** This project perform the docking and molecular dynamics simulation on the GRID platform for *discovery new targets for neglected diseases* . Analysis can be performed notably using the data generated by the WISDOM application on the EGEE infrastructure.





Wide *In Silico* Docking On Malaria



Active site

Ligand

Loops variation between
structures

**~40 millions complexes target-compound
were produced during the DC**

<http://wisdom.eu-egee.fr>



Docking On Influenza A Neuraminidase

- Grid-enabled High-throughput *in-silico* Screening against Influenza A Neuraminidase
- Encouraged by the success of the first EGEE biomedical data challenge against malaria (WISDOM), the second data challenge battling avian flu was kicked off in April 2006 to identify new drugs for the potential variants of the Influenza A virus. Mobilizing thousands of CPUs on the Grid, the 6-weeks high-throughput screening activity has fulfilled over 100 CPU years of computing power.
- In this project, the impact of a world-wide Grid infrastructure to efficiently deploy large scale virtual screening to speed up the drug design process has been demonstrated.

Hurung-Chun Lee



Conclusion

- New technologies have been introduced to automate the analysis, and annotation of genomic, proteomic and Systems Biology data (eg. **Web services, Workflow, Data Mining, Agent, GRID, Ontology, Semantic Web**).
- A new generation of **algorithms and data mining** needs to be developed in order to be capable of connecting the biological information of genes, proteins and metabolic pathways with the patients' disease.
- The dedicated **HPC and GRID infrastructure** will be in a position to tackle the important role of developing new strategies for production and analysis of data in the fields of biotechnology and biomedicine.
- The **massive potential of HPC and Grid technology** will be indispensable when dealing with both the complexity of models and the enormous quantity of data.



Symbiomatics

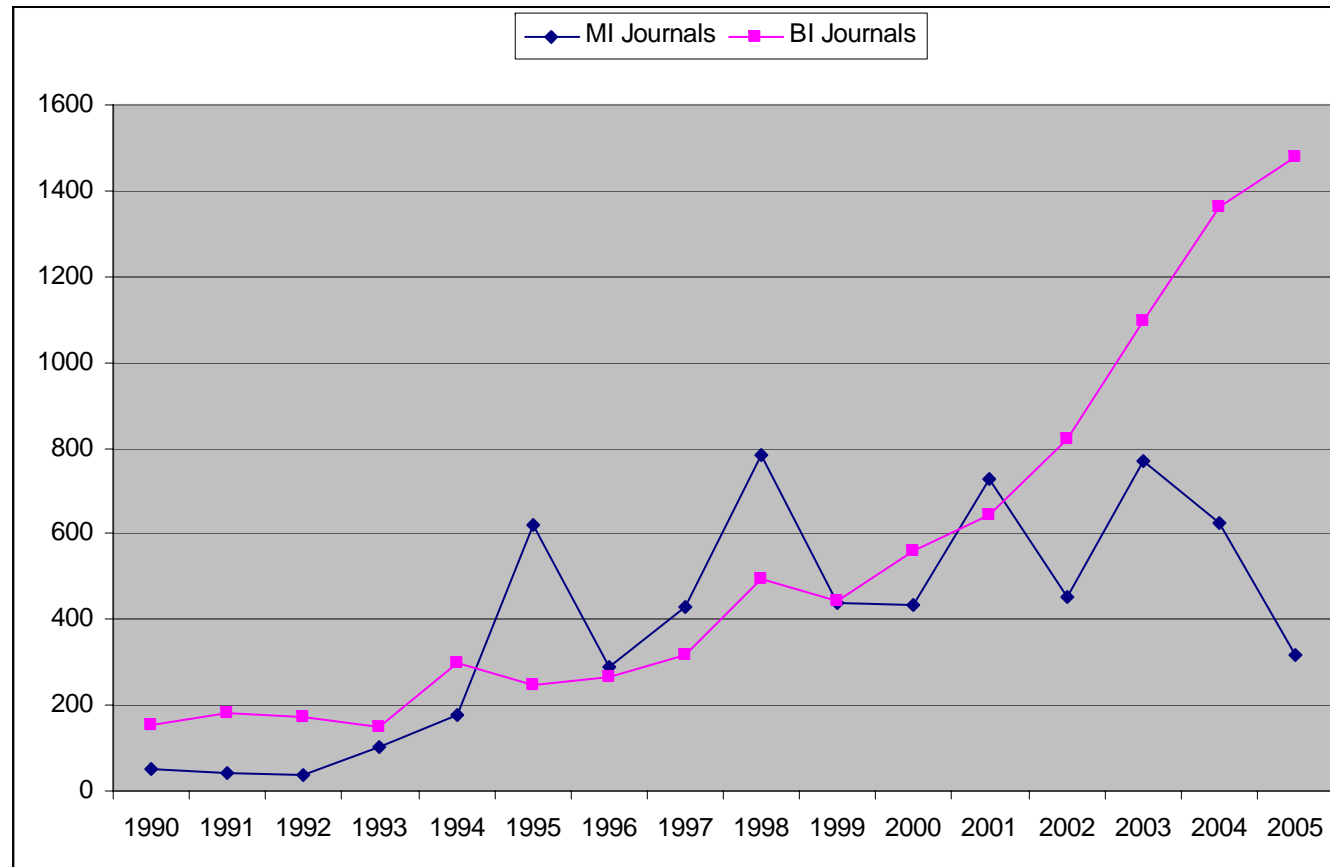
- **IST – CT – 2004 - 015862**
- **SYMBIOMatics**
- **Synergies in Medical Informatics and Bioinformatics**
- **Specific Support Action**



- **Rebholz-Schuhman D (1), Cameron G (1), Clark D (1), Beltrame F (2), Coatrieux JL (3), Del Hoyo Barbolla E (4), Martin-Sanchez F (5), Milanesi L (6), Tollis Ioannis G (7), Van der Lei J (8).**
- (1) EMBL-European Bioinformatics Institute, U.K.
- (2) Dist University of Genova, Italy
- (3) INSERM, France
- (4) Ministry of Education and Science, Spain
- (5) Institute of Health “Carlos III”, Spain
- (6) CNR-ITB – Institute of Biomedical Technologies, Italy
- (7) Foundation for Research and Technology, Greece
- (8) Erasmus Medical Center, Netherlands



Symbiotics



Distribution of the publications from the BI journal corpus and from the MI journal corpus over time



Symbiomatics

ID	Area Name	Priority
1	Medical Genetics Databases and Initiatives	HIGH
4	Gene Expression Information in Medical Diagnostics & Prognostics	HIGH
5	Modelling & Simulation of Biological Structures & Processes/Diseases	HIGH
8	Integration of data from Biosensors & Medical Devices with clinical information systems	HIGH
9	Integration of patient molecular data in Electronic Health Records	HIGH
10	Systems for Clinical Decision Making	HIGH
12	Semantic Interoperability and Ontologies in Biomedicine	HIGH
13	Technologies for Biomedical Information Integration	HIGH
16	Data Interoperability & Standards	HIGH
17	Connecting Biobanks to large scale databases to enable data mining	HIGH
21	Patient Risk Profiling and Lifestyle Management	HIGH
25	Applied Pharmaceutical Research	HIGH
28	Clinical and Ethical Issues related to biomedical data processing	HIGH



ID	Area Name	Priority
3	Therapeutic Area Focussed Initiatives	Intermediate
6	Medical annotation of biological databases	Intermediate
7	Functional and Molecular Image Processing	Intermediate
11	Molecular Information Interfaces for Physicians	Intermediate
15	Mining Biomedical Literature	Intermediate
23	Informatics to support Pharmacogenetics and Stratified Clinical Trials	Intermediate
26	Data Security and Accuracy Considerations	Intermediate
2	Proteomics Information and Analysis	
14	Multilevel Modelling and Vertical Information Integration	
18	Registries linking molecular, familial and clinical data	
19	Bio-defence information systems and networks	
20	Addressing Inf. needs from research in infectious/tropical diseases	
22	Identity Confirmation and Personal Genomics	
24	Informatics to enable Medical Device Development and Biosensors	
27	Post Marketing Surveillance of Drugs and Pharmacovigilance	
29	Health Management Inf. Systems for genomic med (inc. reimbursement)	
30	Addressing the need for training for biomedical informatics scientists	
31	Developing health information skills for researchers and carers	



Thanks

