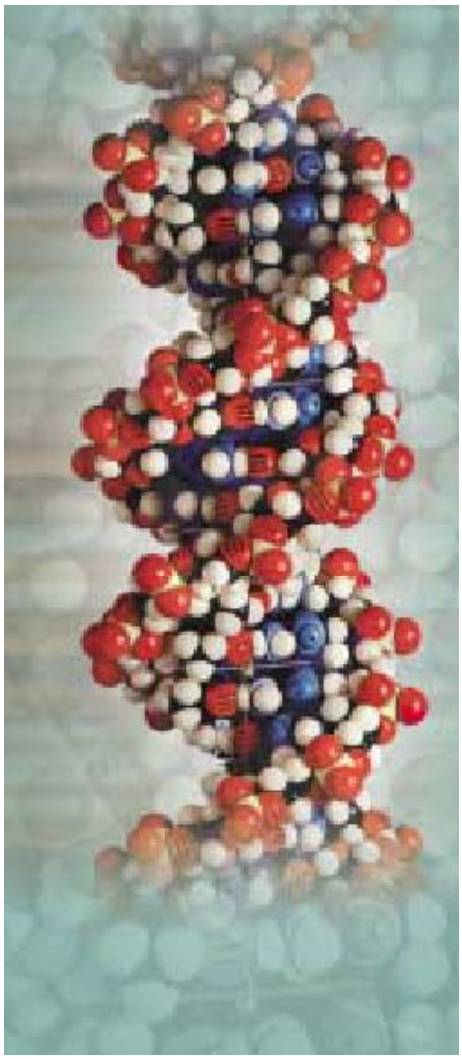
Life sciences, genomics and biotechnology for health

Bioinformatics - structures for the future



JUNE 2003

DG Research / F.4 Fundamental Genomics

Contact: Frederick Marcus or Miklos Gyorffi, Research Directorate General, European Commission
mailing address: EUROPEAN COMMISSION (SDME 8/57), B-1049 Brussels, Belgium
e-mail: Frederick.Marcus@cec.eu.int Miklos.Gyorffi@cec.eu.int
website: http://www.cordis.lu/lifescihealth/genomics/home.htm and http://www.cordis.lu/lifescihealth

Published by the EUROPEAN COMMISSION

Research Directorate-General

Workshop Report on
BIOINFORMATICS - STRUCTURES FOR THE FUTURE

for the
European Commission
Research Directorate General
Directorate F - Health Research

Based upon a Workshop
held in Brussels, Belgium on 12-13 March 2003

Diego Di Bernardo, Howard S. Bilofsky, Martin Bishop (RAPPORTEUR), Soren Brunak, Jean-Michel Claverie, Christopher Cooper, Richard Durbin, Les Grivell, Jaap Heringa (CHAIR), Marie-Paule Lefranc, Jack Leunissen, H.W.Mewes (CHAIR), Folker Meyer, Michael Nilges, Paul Schofield, Sandor Suhai, Janet Thornton (CHAIR), Anna Tramontano, Alfonso Valencia, Anne-Lise Veuthey, Martin Vingron, Gunnar von Heijne

Editors: Miklos Gyorffi, Frederick Marcus

Final Report - June 2003

TABLE OF CONTENTS

DISCLAIMER: This workshop was initiated and organised by staff of the Commission services, who participated in this workshop and who assembled and edited this report with the assistance of the rapporteur and chairpersons. They and the other invited external experts provided both written and oral contributions to this report, and all the views expressed both individually and collectively in this report are those of the external experts, and may not in any circumstances be regarded as stating an official position of the European Commission.

EXECUTIVE SUMMARY

WORKSHOP BACKGROUND

A group of top European bioinformatics experts (+ 1 USA) attended a workshop on "Bioinformatics - Structures for the Future" in order to provide a summary of the background, problem areas, current situation, guidelines and options for action by the Commission, policy makers, organisations in Member states and researchers themselves.

The scope of the workshop was to set a research agenda in the field for structuring European bioinformatics research. The workshop was implemented to explore the following areas:
a. State of the art review of existing projects that have a bioinformatics component to understand its role in the given projects:
- The extent to which these projects involve development of bioinformatics tools, application of existing tools, elementary data processing
- The degree to which experimental outcomes and interpretation rely upon the data processing facilities

b. Analysis of needs in the scientific communities that employ bioinformatics tools:
- Requirements for storage and retrieval of biological information
- Computational complexity of analyses
- Consequent need for compute power

c. Down-stream processing of genetic information:
- Implications for the discovery of medical applications (including drug development).

d. Analysis of possible scientific developments in bioinformatics:
- More interoperability of databases
- Better data access
- Better development of ontologies
- New theoretical models that permit further understanding of data

As a result of submitted papers, presentations and discussions, the workshop participants emphasised messages in 10 key areas:

1. **Areas for bioinformatics-based research in EU Framework and National Programmes**
2. **Excellence in bioinformatics and its maintenance**
3. **Free riding and funding**
4. **Increasing excellence in research projects via bioinformatics**
5. **Database development, maintenance and infrastructure funding**
6. **Standards and ontologies**
7. **Range of database information**
8. **Systems biology**
9. **Research collaborations and the pharmaceutical industry**
10. **Training**

MESSAGES IN KEY AREAS

E.1) Areas for bioinformatics-based research in EU Framework and National Programmes

A number of areas have been suggested as being suitable for funding support and encouragement both in the EU Framework Programme and as part of a coordinated programme by national and local funding agencies and charities. These topics would be suitable for inclusion in future calls for proposals.

- **Bioinformatics components of all major proposals** – It was generally felt that each major project should have an important component devoted to bioinformatics and data analysis (see E.4), including storage, retrieval and analysis of data during the life of the project, plus consideration given to long term maintenance.

- **Database and software upgrading, optimum utilisation and support** – given the increasing flood of data, important resources must be devoted to upgrading storage, analysis, access, and integrative capabilities. This needs to occur not only at a European level, but at other appropriate levels as well. This process is occurring in the context not only of moving forward from a gene centred approach (genomics) to a protein centred approach (proteomics), but also allowing for new technologies such as arrays and related problems:
- **Gene expression studies** – by a variety of methods including microarrays and RNAi [1] with the objective of understanding disease processes and classifying tumours.
- **Immunological bioinformatics** – by upgrading existing databases and analyses to make maximum use of these resources for health applications such as fighting disease including vaccine design, controlling allergies and countering bio-terrorism.
- **Haplotype analysis -** linkage and association studies to establish the molecular basis of human variation both for research purposes and for disease identification as a key to understanding the causes of disease and providing individualised health care.
- **Comparative genomics** - leading on to functional studies of identified homologues.
- **Protein identification** – from cells, tissues and body fluids e.g. plasma by MS-MS (tandem mass spectroscopy) characterisation of peptides and other methods to detect, identify and catalogue the proteins present, and to complement gene expression studies.
- **Homology models for proteins** - by an integrated approach of sequence alignment, secondary structure prediction and modelling supported by advances in modelling algorithms, to determine sequence, structure and function and relate them in order to build these characteristics into systems models.
- **Systems biology** – requires computational models able to realistically simulate the dynamic behaviour of biological processes based on molecular building blocks such as sequences, 3-D structures, interaction patterns, and physicochemical parameters (e.g. kinetics and affinity constants). The aim is to be able to understand and eventually modify the behaviour of complex regulatory networks and to test hypotheses on the functioning of biological processes.

E.2) **Excellence in bioinformatics and its maintenance**

There are several major European bioinformatics centres of excellence and world-class databases. Important reasons for their success include research excellence and a strong position in open source software and data release, access and usage, supported by adequate funding. In contrast to some national funding policies, which tend to be somewhat "protectionist," the policy of open publication and excellence attracts both internal and external funding for maintenance of the resource and continuation of the work. Fundamental principles of academic research include: credit assigned based on publication, publication allowing verification, and reuse of ideas for novel developments. Publication of data resources (primary or secondary) means allowing unconstrained reuse for unforeseen purposes.

Open data access is so necessary for science and collaboration that it is a core principle and it should be the norm for publicly funded bioinformatics. European excellence occurs at both European and local levels, and supports a hierarchy of interactions between researchers and resources that these centres of excellence serve groups that are physically nearby, national, European and world-wide. One role for the European Commission and for Member states, through groups such as ESFRI and COGENE, is to help to connect these levels with both organisational and funding support. There are also EU (Sixth Framework Programme) and various Member states' programmes aiming at creating and maintaining new institutions and organisations, for example with the EU Networks of Excellence and Integrated Project initiatives. These efforts should be identified, supported and strengthened. We have scattered activities, and the goal is to increase efficiency by combining resources, while maintaining an appropriate mix of local and European wide institutions.

[1] RNAi (RNA interference) the use of sequence-specific RNA molecules to artificially interfere with the activity of targeted endogenous genes.

E.3) **Free riding and funding**

"Free riding" (using but not contributing financially or scientifically to, American or other facilities) is short-sighted. Open international collaboration will depend upon major participants – such as the EU and its present and future member and associated states – being prepared to support an equitable share of the cost burden. The USA has the advantage of a single federal government, language and culture, and these advantages are nowhere better illustrated than in the structure of research funding through the NIH and NSF, providing vastly larger centralised funding than is available at the European level. In the world-wide context of bioinformatics, there is the temptation to try to protect local resources by privatising at the local level and relying on public resources at the higher level. This is a strategy for failure. Open publication and world class research attract further funding. Moreover, relying exclusively on external resources often means that there is not sufficient local capacity of researchers to make effective use of results, even when available without charge. Europe is leading in many fields and has led in others, but has lost the lead in some areas due to lack of funding relative to the USA.

E.4) **Increasing excellence in research projects via bioinformatics**

To be successful, each project should necessarily have a clear objective that targets a biological research problem, a theoretical and consequent computational component and an experimental validation component. Potential impacts include: optimally designed experiments and data interpretation using theoretical models and computational tools; computational tools for the discovery of novel "objects" in the genomic sequences; theoretical models to explain complex regulatory networks, their organisation and function. Bioinformatics should play a key role in experimental design. It was generally felt that a significant fraction of resources in each major project should be devoted to bioinformatics and data analysis, including storage, retrieval and analysis of data during the life of the project, plus consideration given to long term maintenance. While the percentage will depend on the nature of each experiment, experience shows that a level of 20% was typically present in large projects, and that this reached 50% or more in larger genome projects and in particular areas like microarray experiments. Smaller projects will also need significant data analysis support. It is necessary that all projects have a good percentage of bioinformatics, in order to properly store and analyse data. A key point is that all experiments in the life sciences should have a proper commitment to analyse, store and publish data.

The way in which biologists address specific problems will be deeply influenced by the availability of bioinformatics methods for the design, management and interpretation of the results. For many medical and biotechnological research areas, bioinformatics is the one of the rate limiting steps. Europe is getting much less out of the investments made already in experimental and clinical research than it otherwise would with a better coordinated bioinformatics effort. A solution to the problems with high throughput data today is to build networks of resource centres to supplement and include local, small-scale bioinformatics and the larger centres. One method of building such networks is via EU funded NoEs or IPs. Some facilities are provided by the EBI and the NCBI but there is a need for resource centres to provide bioinformatics expertise, stable environments, high-throughput and high-volume facilities with extensive training.

E.5) **Database development, maintenance and infrastructure funding**

Bioinformatics is continually evolving as biological methodology advances. Bioinformatics research to underpin biological research objectives needs to be expanded to meet the genome, proteome, metabolome, etc. challenges across a wide range of activities. Bioinformatics challenges include early involvement in experiment design involving the entire data processing pipeline, hypothesis generation, data analysis and statistics and allowing for integration with other data. Database development and management is essential for biological research and advancement. There is a need for the development of both theoretical models and algorithms to interpret and integrate the large amount of biological data now available. One of the challenges of bioinformatics for the next several years will consist in developing intelligent systems for information integration. However, human assessment should keep its place with the help of improved bioinformatics tools. The use of literature references is a key part of this endeavour with better access tools and the ability to fully analyse text based information.

There is a fundamental and longstanding problem concerning financial support for research infrastructures in Europe – particularly databases and biological research materials. Funding is available for research, for analysis of results, and for producing databases but is unavailable or inadequate for long-term support of such data infrastructure including data capture, annotation, curation, dissemination and archiving; similar problems afflict collections of biological materials. The problem has been more clearly recognised in the US where infrastructure is well financed and is made internationally available. At the European level, only very limited funding and scope for support and continued sustenance is allowed as database maintenance does not fit well with national funding models. As an example, SWISS-PROT is now funded from the NIH of the USA. Some databases require a high level of automated analysis, which requires high performance computing and complex software engineering. These are the type of infrastructures whose development and maintenance needs to be centralised in one physical location in Europe, and it is difficult to distribute it over several countries - as is usually required by EU terms and conditions when forming consortia for funding proposals. There is evidence that Europe needs a new and different funding model to maintain excellence of world class. It is recognised that for public domain databases and collections of bioresources, maintenance and related research require long-term funding, but the mechanism is not apparent in Europe. Research bioinformatics fits classical response-mode funding. Resource bioinformatics involving continuing infrastructure support is expensive and requires long term strategic planning, international co-ordination (within Europe, and globally), and funding. A possible model is found in the funding of the SNP (Single Nucleotide Polymorphisms) consortium[2] and the International HapMap initiative[3], both including industrial partners but committed to open publication of their data. An important debate on the respective roles and funding responsibilities is needed that should include organisations such as EBI, EMBnet and national bioinformatics centres in the world-wide context of major international facilities such as NCBI, and new technologies such high speed networking (GEANT) and computing distributed across organisations (GRID). A key issue is the overall level of funding needed to compete on a world-wide scale. The genomics budget of the Sanger centre of $430 million over five years is comparable with the whole EU budget for basic genomics research.

Specific funding for infrastructure, separate from that for research, is needed. Special efforts should be made, involving the EU, individual Member states, private charities and other funding agencies to find a coordinated response to this problem, with support at appropriate levels. This debate should address the full range of responsibilities for data capture, annotation, curation, dissemination and archiving and related activities such as training.

The quality and availability of infrastructure has an influence on the quality and productivity of research. It therefore needs to be recognised that database maintenance and upgrading is a continuing and necessary responsibility of funding agencies; excellence of infrastructure is an investment which will amplify the returns to the financing of research. We need better access tools and the ability to analyse text based information. Smaller funding instruments could be targeted for this role. It was suggested that a potential role for SMEs could be maintenance of databases, although there is no obvious business model for doing this. In general, discovery of ways to support SMEs should be encouraged. Europe should consider a USA SBIR-like peer review system[4].

---

2 http://snp.cshl.org Single nucleotide polymorphisms (SNPs) are common DNA sequence variations among individuals and have great significance for biomedical research. The SNP Consortium Ltd. is a non-profit foundation organized for the purpose of providing public genomic data.

3 http://www.wellcome.ac.uk/en/1/awtpubnwswnoi34ana3.html On 29 October 2002, a group of scientists gathered in Washington DC to launch the International HapMap Project - a major new initiative to create a map of human genetic variation.

4 http://www.sba.gov/sbir/indexsbir-sttr.html Small Business Innovation Research Program

E.6) **Standards and Ontologies**

Biological data are large volume and complex and therefore present new challenges. A major weakness is lack of standards. A major threat is the inability to act quickly enough to 'conquer' the data challenge. Standard and ontologies are central: 'computing standards' and 'biological ontologies' are both desperately needed. There is strong support for 'Open' standards. Science becomes gravely hampered if scientists in different locations or contexts and at different times are unable to communicate data and results reliably to one another – similar problems are facing curators of databases, public or private, hence the need to develop common languages and protocols. Ideas for tools enhancements include several needs: one European standard middleware; tools enabled to work together; centres to build interfaces for their platforms; grant proposals to be linked to standard compliance and data availability; open source to be a requirement for software development. Networking is required to foster integrative approaches and prevent the development of many solutions to the same problem. Ontologies are key, but in moving from genotype to phenotype the problem gets harder and ill-defined. Primary data may always be re-interpreted, and hence it is important never to throw it away, especially since the basic biology underlying the data is always the central element. Biological interpretation always needs the biological context, without which it is meaningless.

E.7) **Range of database information**

Identification of the full set of gene products in higher organisms is still in its infancy and many of the problems of interpretation are still unsolved. While genomics is driven by data acquisition, interpretation is still mostly hypothesis driven at the level of individual gene products. A new direction is to provide simple access to databases and analysis tools that permits building generalised models for biological systems. In the vertebrates we cannot predict all the gene products from the DNA sequence. The best approach is to use comparative genomics of a number of vertebrate species to detect conserved regions of functional significance. Gene regulation and related areas such as alternative splicing are key areas of importance and need an experimental and a computational framework that is more advanced than at present.

Key types of data include:

- *Sequence Data:* Sequence data are extremely valuable and newly determined sequences are very good value for money. The added value that the characterisation of new genomes brings to existing knowledge is very important. Unfortunately, apart from the Sanger Institute, Europe appears to have dropped out of sequencing for very large genomes, which is a strategic mistake. Currently over 170 genomes are fully or partially sequenced, and we can foresee having a thousand complete genomes sequenced in the next few years. New technology is continually reducing the cost of further sequencing. Each new genome helps researchers to detect more features - e.g. regulatory elements – in those already sequenced, and comparison across the growing number of genomes leads to fuller understanding of their evolution and structure.

- *Haplotype analysis:* Haplotype analysis for disease identification will be a key element of health care and understanding disease. Personalised treatment is a key goal. Relevant haplotype information can be gained by comparing individual genomes.
- *Homology models:* A possible solution for improving the quality of homology models for proteins may lie in an integrated approach of sequence alignment, secondary structure prediction and modelling assisted by advances in algorithms and use of appropriate physical models.
- - *Function prediction:* The aim of molecular biology was formerly to be able to predict function from structure, and structure from sequence. The reality is that we need by various methods to determine sequence, structure and function and try to relate them in order to gain understanding and to build all these elements into models of working processes. Function prediction from the structure alone remains an elusive goal. One way is to search for conservation of structural templates. We study aspects of function by looking at physical properties that can be calculated from the structure such as shape, electrostatic potential, and molecular dynamics.

E.8) **Systems biology**

Only the development of integrated models of biological processes will enable the understanding of complex biological systems. The solution of fundamental problems with these new models will require bioinformatics applications that go beyond the current work on data integration and manipulation. It is the responsibility of key practicioners to pass on to the rest of the community a clear message about the need to preserve the balance between the different areas of bioinformatics and computational biology. We are lacking a lot of basic knowledge about biological processes e.g. transcription and splicing. In systems biology, there is the potential for extensive increase in understanding. A world-wide *E. coli* project[5] is in progress and systems biology is becoming a big field. There is also the question of to what degree Europe needs or has access to the necessary large scale petaflops of computional power and terabytes of data. Given that no simple "mathematical formula" will ever be able to encompass the complex behaviour of biological systems, the ultimate bioinformatics theory is expected to consist of a computational model able to realistically simulate the dynamic behaviour of these systems, based on molecular "first principles" such as sequences, 3-D structures, interaction patterns, and physico-chemical parameters (kinetic and affinity constants). Applied science areas that can have a major impact include:

- Digital signal processing: to discover information hidden in the DNA sequence;
- System identification: to identify complex regulatory networks from experimental data;
- Network theory: to be able to understand, predict and eventually modify the behaviour of complex regulatory networks.

There are currently disequilibria within the bioinformatics area that favour informatics disciplines (i.e. database-like activities and information management systems) over data-analytical and theoretical methods, mathematical modelling and computational simulation techniques, usually referred to as "computational biology". What is already in progress is a shift in paradigm, away from the previous implicit assumption in biological research, that the scientist could directly interpret his own raw experimental data. Now mathematical models and simulation techniques are needed in order to integrate biological knowledge with the ever-increasing amount of experimental data into a formal framework (in silico model) and to test hypotheses on the functioning of biological processes. It has become painfully obvious that in the perception of many experimental biologists, research in computational biology is not considered to be essential for the future of genomics and proteomics. They are interested only in the contribution of bioinformatics as an auxiliary technique. This opinion will tend to support experimental work at the expense of theoretical work. We do need to identify the parts, but bioinformatics research and resources have to evolve from being "individual-part orientated" (genes, proteins and structures) to relatng to bigger biological pictures (such as protein-protein interactions, regulatory networks, mutation mapping to phenotype, cellular subsystems), thus aiming at generating an understanding (a *predictability of behaviour*) of biological subsystems (e.g. mitochondria, spliceosome).

E.9) **Research collaborations and the pharmaceutical industry**

There is an enormous range of collaboration in the pharmaceutical arena. A major danger observed in these collaborations but present more widely, is the tendency to store a minimal amount of data, usually in processed form. More should be permanently conserved. The more medically- and patient-orientated the research, the more confidentiality becomes a key issue. When bioinformatics reaches the level of dealing with personal data there are legal and ethical considerations that become apparent.

Even in the IPR-protection oriented world of pharmaceuticals, open source licensing is a valuable approach to encourage maximum transfer and use of data. Immunological bioinformatics is needed in the vaccine design area, in protection against bio-terrorism, and in specific disease applications e.g. asthma, autoimmunity, regulation of the immune responses. An important part of structural

[5] International E. coli Alliance, which is a consortium of different groups of scientists with the combined aim of modelling the E. coli cell in silico. The consortium includes http://www.projectcybercell.com from Canada, the http://www.iab.keio.ac.jp, in Japan, http://www.gsk.com/index.htm cell modelling group and the http://ecmc2.sdsc.edu from the USA.

bioinformatics, a key area in pharmaceutical research, in particular at a medically oriented research institute, is the study of the interaction between a protein and its ligands needed for drug design and virtual screening activities. A workshop was held in 2003 to discuss computational approaches to health related problems (http://lyon2003.healthgrid.org).

E.10) **Training**

Training is very important in national centres for each 'national research community', and also offers opportunities for private sector provision. However, training provisions are unbalanced throughout Europe. There are many courses for bioinformatics students, perhaps too many in some counties. In contrast, there is often insufficient training for biologists in the use of bioinformatics. We should be able to organise teaching of bioinformatics that is attractive to biologists. Given the lack of a general bioinformatics training in many countries at the national and also European level, most small groups need to train young scientists in bioinformatics starting from a purely biological or computational background. A framework for training in bioinformatics is needed.

WORKSHOP REPORT

WORKSHOP ORGANISATION

This workshop on "Bioinformatics - Structures for the Future" was organised by the Research Directorate-General of the European Commission, in the context of a series of workshops supporting the European Research Area (ERA).

A group of experts was invited to meet and discuss this topic, and to provide a summary of the background, problem areas, current situation, and guidelines and options for action by the Commission and policy makers and organisations in Member states, and for researchers themselves.

A workshop "Terms of Reference" and documents and references were provided before the workshop.

Attendees submitted short, highly condensed summary papers of their contributions, which are included here with the workshop summary, giving their points of view. Some were submitted as powerpoint presentations, and the text has been abstracted, and some summaries are based on notes from the presentations.

The workshop consisted of presentations by invited speakers. These presentations were followed by open discussion.

Members of the Commission services, who provided background information on relevant activities, also attended the workshop.

This workshop report was written and assembled by the Rapporteur and Chairpersons and edited by members of the Commission services, in particular the officers responsible for the workshop, based on summaries of the workshop discussions, inputs from the chairpersons and participants during and after the workshop, and the contents of the submitted papers.

The executive summary represents a large convergence of views. Where they occur, significant differences are explicitly presented as such.

TERMS OF REFERENCE

THE FOUNDATIONS -
Current and already planned activities in Bioinformatics

Unifying theme: state of the art review of running projects that have a bioinformatics component; to understand the extension and role of this component in the given projects (to what extent these projects involve bioinformatics tools development, application of existing tools, or elementary data processing, what are the centres of gravity, how do they relate the experimental outcomes to the data processing facilities, links to genomics and related research).

FUTURE INFRASTRUCTURE AND ANALYSIS REQUIREMENTS -
Hardware, Software, Archive, Data Access

Unifying theme: analysis of the needs in the scientific communities that employ bioinformatic tools. What are the actual and future needs in respect to storage and retrieval of biological information, including computational means and algorithms? What are the effects of technologies such as Array technologies and their implications for databases, standardisation and software?

FUTURE BIOINFORMATICS RESEARCH TOPICS -
Integration of knowledge, systems biology, health, pharmaceuticals, biotechnology, environment

Unifying themes:

- other application avenues of genetic information and their implications in the development of bioinformatic tools e.g. towards medical application (including drug development).
- analysis of the possible scientific developments in the field and in new research areas. What research solution might influence the appropriate development of bioinformatic tools (more interoperability, more data access, more integration at the level of ontologies, new theoretical models, other)?

SUMMARY AND CONCLUSIONS

Research Policy (EU, National, International), Research Topics, New Directions)

What are the key policy issues under discussion and what is current thinking?

- Key issues
- Problems and current solutions
- Funding agency policy options, focusing on RESEARCH policy (EU, national, regional, international)
- Research policies of the various research and innovation participants
- Statements of research policy general principles
- Specific near term policy recommendations
- Long term changes
- Future areas of discussions

WORKSHOP AGENDA, SPEAKERS AND LIST OF PAPERS

Bioinformatics - structures for the future

WORKSHOP ORGANISED BY DG RESEARCH of the European Commission, in Brussels on 12-13 March 2003 (Wed,Thu)

Centre Borschette, Rue Froissart 36, 1050 Brussels, room AB-2C

(Organisers: Miklos Gyorffi, Frederick Marcus DG Research)

contact: e-mail - Miklos.Gyorffi@cec.eu.int **;** Frederick.Marcus@cec.eu.int**)**

Wednesday 12 March

10:00 INTRODUCTION

Welcome and Introduction from Commission organisers: **Miklos Gyorffi and Frederick Marcus**
Bioinformatics and Research in Fundamental Genomics: **Manuel Hallen, Head of Unit**
Introduction of participants and **Rapporteur: Martin Bishop**

10:30 - Chair: H. Werner Mewes

THE FOUNDATIONS

Marie-Paule Lefranc - "Immunoinformatics in IMGT: a synergy between IMGT-ONTOLOGY and bioinformatics tools development for knowledge management and medical application"
Les Grivell - "eBioSci -Access and retrieval of digital information in the life sciences".
Martin Vingron - "Bioinformatics and Functional Genomics"
Folker Meyer - "Medicago"
Anne-Lise Veuthey - "Bioinformatic requirements for SWISS-PROT developments"

13:30 - Chair - Janet Thornton

FUTURE INFRASTRUCTURE AND ANALYSIS REQUIREMENTS

Michael Nilges - "Structural bioinformatics"
Martin J. Bishop : "HGMP Resource Centre Future Plans"
Christopher Cooper : "The role of IT in Bioinformatics"
Paul Schofield : "Pathbase; meeting the challenge of databasing mutant mouse pathology"
Soren Brunak : "Research-driven infrastructure"
Richard Durbin: "Future developments in primary and secondary genomic data resources"

Thu 13 March

9:00 - Chair - Jaap Heringa

FUTURE BIOINFORMATICS RESEARCH TOPICS

Jaap Heringa - COMMENTARY
H. Werner Mewes - COMMENTARY
Ilias Iakovidis - eHealth - Past and future activities of the European Commission
Gunnar Heijne - COMMENTARY
Diego Di Bernardo - "Future Bioinformatics Research Topics" AND COMMENTARY
Alfonso Valencia - "Bioinformatics - Biology by other means"
Jean-Michel Claverie - "Biology driven bioinformatics: example of the International E.coli Alliance project" AND COMMENTARY
Howard S. Bilofsky - "Complexity in the Life Sciences from the Pharma Industry's Perspective"
Anna Tramontano - "Training of bioinformaticians, present research and the needs of big pharma"
Mark Cantley - COMMENTARY

12:00 - 15:00 Session SUMMARY AND CONCLUSIONS

Chair: Miklos Gyorffi and Fred Marcus, Rapporteur: Martin BISHOP

LIST OF PARTICIPANTS

Workshop Participants (Non-Commission)

1. Diego Di Bernardo, Telethon Institute of Genetics and Medicine, Via Pietro Castellino 111, I - 80131 NAPOLI
2. Howard.S. Bilofsky, GlaxoSmithkline R&D IT, 709 Swedenland Rd, 19119 King of Prussia, PA, USA
3. Martin Bishop, HGMP Resource Centre, Hinxton, UK - Cambridge CB10 1SB
4. Soren Brunak, Centre for Biological Sequence Analysis, Biocentrum-DTU, The Technical University of Denmark, Building 208, DK-2800 LYNGBY
5. Jean-Michel Claverie, Information Génétique et Structurale, CNRS - UMR 1889, 31 Chemin Joseph Aiguier, F - 13402 MARSEILLE, Cedex 20
6. Christopher Cooper, Business development executive (EMEA), Life Sciences, IBM United Kingdom Limited, 1 New Square, Bedfont Lakes, UK - FELTHAM, Middlesex TW 14 8HB
7. Richard Durbin, The Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, UK - CAMBRIDGE CB10 1SA
8. Les Grivell, European Molecular Biology Organisation, Postfach 1022.40, D - 69012 Heidelberg
9. Jaap Heringa, Bioinformatics Unit, FEW/W&I, Vrije Universiteit, De Boelelaan 1081a, NL - 1081 HV Amsterdam
10. Marie-Paule Lefranc, IMGT, Laboratoire d'ImmunoGénétique Moléculaire, LIGM, UPR CNRS 1142, IGH, 141 rue de la Cardonille, F - 34396 Montpellier Cedex 5
11. Jack Leunissen, Dept. of Genome Informatics, Wageningen University, Dreijenlaan 3, NL - 6703 HA Wageningen
12. H.W.Mewes, GSF-Forschungszentrum fuer Umwelt und Gesundheit GmbH, Ingolstaedter Landstrasse 1, D - 85764 NEUHERBERG
13. Folker Meyer, Zentrum für Genomforschung, Universität Bielefeld, office: V6-147, D - 33594 Bielefeld
14. Michael Nilges, Bioinformatique Structurale, Institut Pasteur, 25-28 rue du docteur Roux, F-75015 Paris
15. Paul Schofield, University of Cambridge, Anatomy, Downing Street, UK - Cambridge CB2 3DY
16. Sandor Suhai, Deutsches Krebsforschungszentrum, Department of Molecular Biophysics, Im Neuenheimerfeld 280, D - 69120 HEIDELBERG
17. Janet Thornton, European Bioinformatics Institute, Wellcome Trust Genome Campus, UK - Cambridge CB10 1SD
18. Anna Tramontano, Department of Biochemical Sciences "Rossi Fanelli", University of Rome "La Sapienza", P le Aldo Moro, 5, I - 00185 Rome
19. Alfonso Valencia, Protein Design Group - Centro Nacional de Biotecnologia, Cantoblanco, E - MADRID 28049
20. Anne-Lise Veuthey, Swiss Institute of Bioinformatics, CMU 1 Michel-Sevet, CH - 1211 GENEVE 4
21. Martin Vingron, Max-Planck-Institut für Molekulare Genetik, Ihnestr. 73, D-14195 BERLIN

23. Gunnar von Heijne, Dept of Biochemistry & Biophysics Stockholm Bioinformatics Centre, Stockholm University, Sweden

European Commission Staff

Mark Cantley, DG RESEARCH, directorate E
Josefina Enfedaque, DG RESEARCH, unit F.4
Miklos Gyorffi (Workshop co-organiser), DG RESEARCH, unit F.4
Manuel Hallen, D G RESEARCH, unit F.4
Ilias Iakovidis, DG INFORMATION SOCIETY, unit C.4
Michel Millot, JOINT RESEARCH CENTRE, unit I.6 (TP 750, I-21020 ISPRA)
Sofie Norager, DG INFORMATION SOCIETY, unit C.4
Frederick Marcus (Workshop co-organiser), DG RESEARCH, unit F.4
Bernd Rainer, DG RESEARCH, unit E.4
Jacques Remacle, DG RESEARCH, unit F.4

WELCOME SPEECH

Manuel Hallen
Directorate F - Health
Directorate-General for Research, European Commission,
Head of Unit: Unit F.4 "Fundamental Genomics"

WHY THIS WORKSHOP?

TO DEVELOP A STRATEGIC VISION OF BIOINFORMATICS RESEARCH FOR RESEARCHERS, FUNDING BODIES, AND POLICY MAKERS, BY:

DETERMINING KEY ISSUES FOR THE NEXT FEW YEARS FOR

- Research Topics and New Directions
- Research Policy Makers (EU, National, Regional, International)
- Overcoming fragmentation in research structures and results

AND BY DISCUSSING AND ELABORATING

- Problems and current solutions
- Funding agency policy options, focusing on RESEARCH policy
- Research policies of the various research and innovation participants
- Specific near term recommendations
- Long term changes
- Future areas of discussions

PROBLEMS OF FRAGMENTATION

The European 6th Framework Programme is aimed at overcoming the fragmentation of European research at all levels

Bioinformatics is an essential element in life sciences because it can:

- unify and archive the results of biological research
- allow European and world wide researchers to work together to generate, store, and analyse data
- develop common research resources and tools: databases and analysis software, communication networks
- support basic research as a goal in itself
- provide a pathway from basic research to development of products, such as new medicines

FP6 Priority 1: Life sciences, genomics, biotechnology for health

One of seven major thematic priorities of FP6

The objective is to help Europe generate new knowledge by focusing on genomics and using sequence data and other results to translate it into applications that enhance human health.

Fundamental and applied research will be supported, with an emphasis on integrated, multidisciplinary, and co-ordinated efforts that

- address the present fragmentation of European research and
- increase the competitiveness of the European biotechnology industry.

Major areas of research include:

- Fundamental Genomics
- Applied Genomics and Biotechnology
- Genomic approaches to health and disease
  - Cancer
- HIV/AIDS, malaria and tuberculosis
  - Article 169 European and Developing Countries Clinical Trials Partnership" (EDCTP), concentrating on TB, AIDS and malaria.

Fundamental knowledge and basic tools for Functional Genomics in all organisms (Unit F.4 activities)

KEY RESEARCH AREAS:

TOOLS:

- Gene expression and proteomics
- Structural genomics
- Comparative genomics and population genetics
- Bioinformatics

USING ABOVE TOOLS:

- Multidisciplinary functional genomics approach to basic biological processes

EXAMPLES OF FP5 PROJECTS

- Pilot Integrated Projects
  - GENOMEUTWIN: European twins to identify genes involved in disease
  - EUMORPHIA: Human disease through mouse genomics
  - SPINE: Structural Proteomics in Europe
- COGENE: Co-ordination of Genome Research in Europe
- TEMBLOR bioinformatics projects: Integr8, Desprad, Intact, EMSD

WHY WORKSHOPS?

The European Commission instigated these workshops to allow the participants and ourselves to address issues at the European level.

Several previous workshops have been very successful in this respect:

- Mouse genomics
- Diabetes
- Rare diseases
- Cardo-vascular disease
- Structural genomics

More are planned in the future, e.g. Systems Biology, and your comments are welcome.

WHY HAVE WE INVITED YOU?

You are among the top bioinformatics experts in Europe, from a range of institutions and managing important bioinformatics facilities.

Many of you are involved in Framework Programme research.

With a group this size, we can initiate a full and open discussion.

We hope you can work together to develop a common understanding of the way forward for Europe.

AREAS FOR DISCUSSION AT THIS WORKSHOP

Current and already planned activities in bioinformatics

Future infrastructure and analysis requirements - hardware, software, archive, data access

Future bioinformatics research opics - Integration of knowledge, systems biology, health, pharmaceuticals, biotechnology, environment

---

WORKSHOP INTRODUCTION

Bioinformatics derives knowledge by computer analysis of biological molecular data. It is a rapidly growing branch of biology, highly interdisciplinary, and uses techniques and concepts from informatics, statistics, mathematics, chemistry, biochemistry, genetics, physics, linguistics and other fields. The biological data can be the information stored in the genetic code, experimental results from various sources, three-dimensional protein structures, gene expression arrays, patient statistics, scientific literature etc. An important part of research in bioinformatics is the development of methods for storage, retrieval, and analysis of these data.

A wider, general definition of bioinformatics is important. Concepts in our work, for example, do not come so much from informatics but primarily from physics and probability theory. The concept of "information in the genetic code" has its limitations - DNA can be analysed both as a text and as a molecule that makes interactions with a variety of other molecules. Interactions with and among proteins are governed by three-dimensional structures and their dynamics (and flexibility). These in turn are obviously determined by the sequence of bases, but the behaviour of DNA cannot be fully described by reducing its analysis to a one-dimensional level. The three dimensional aspect is also crucial in understanding protein sequences and protein structures.

The area of bioinformatics has gained significant importance during the EU Fifth Framework Programme FP5, where activity in this field has been mainly related to basic storage tools of the ever growing amounts of data produced by the ever more sophisticated genetic technologies in conjunction with the infrastructural needs accompanying basic genetic research. In FP6 this feature is going to be preserved, however on a higher level, having a recognised part in the work programme and being expanded in meaning to allow genesis of computational tools for the biological interpretation of the large amounts of data. In view of these developments this workshop tried to summarise the near past and present status in the field, from a European perspective, and to analyse possibilities for progress in the future in the next level of approach toward the integration and understanding of biological data.

European activity in bioinformatics is clearly of a very high level of sophistication. We tried to identify first the potential that we have here as reflected in the different activities of FP5. European research has contributed significantly to the creation of the wealth of genetic information available today and is hosting major data processing infrastructures. European excellence is present through specialised databases, resource centres and several European level institutions.

There is also a very heterogeneous scientific community that covers all aspects of today's genetic research. Europe is also engaged in several international collaborations, the main partners being USA and Japan, but other major contributors such as China and Canada should not be forgotten.

However, European research also suffers from some weak points and limitations amongst which it emerged that funding schemes are one of the most important. The wealth of data created needs permanent maintenance and present funding structures do not allow an appropriate approach. There exist very important national funding schemes, although there is little interaction between them. A commonly shared structural approach toward bioinformatics is also missing. European states and organisations have different approaches toward handling research issues in the field. Their interaction with the EU FP is limited, on one side several EU projects had as an outcome new databases (and associated datasets), on the other side the sustainability of these infrastructures viewed as being European/global is not possible within a national framework.

One of the important possibilities in the field is no doubt the EU FP6. It created a higher status for bioinformatics and has a declared scope for better integrating research in Europe. Beyond this, FP6 has also possibilities for incorporation of new informatics infrastructures (GRID). Europe is also attracting external support for some of its globally recognised resources. Bioinformatics becomes increasingly important for industry (Big-Pharma, IBM), but for other applications as well (biotechnology, health care).

Nevertheless, in the near future we are also faced with risks and problems. Generally speaking, the problem of intellectual property rights is still important. We do not yet have efficient schemes to determine whether data and software at various points in the research process should be in the public and private domains, especially concerning the usage of genetic data in health care, or other biotechnology driven applications.

Another problem is that of training - there was a shortage of well-trained bioinformaticians. Since the biotechnology bubble burst in Spring 2000, supply and demand may be mismatched. Training provision for biologists in bioinformatics remains poor.

There is also a lack of initiative in creating a pan-European approach. The existing collaborations may lack true integration, and resources are often dissipated by lack of scale or cooperation.

This report attempts to address these problems. By taking a more European and bioinformatics centred perspective, it also complements the final report of the 2001 Workshop on "New Research Tools for a Life Sciences Decade," European Communities Report EUR 20024, organised under the auspices of the EC-US Task Force on Biotechnology Research http://www.cac.es/ecusworkshop. It also takes account of the workshops (and related reports) by BTSF (See http://www.btsf.org)

Finally, the document independently, and from a European perspective, proposes many similar conclusions to those presented in "A vision for the future of genomics research: A blueprint for the genomic era," by Francis S. Collins, Eric D. Green, Alan E. Guttmacher and Mark S. Guyer on behalf of the US National Human Genome Research Institute, published in NATURE, VOL 422, 24 APRIL 2003. http://www.nature.com/nature, and also simultaneously presented by Dr. Collins at the 2003 HUGO conference. Of particular relevance are their conclusions concerning resources and computational biology. The vision for genomics research detailed there was the outcome of almost two years of intense discussions with hundreds of scientists and members of the public, in more than a dozen workshops and numerous individual consultations (http://www.genome.gov/About/Planning).

WORKSHOP CONCLUSIONS: THE MAIN THEMES

(key points in bold)

1. EUROPEAN EXCELLENCE - STATUS AND FUTURE

Introduction

There are several major European centres of excellence that are world class and their successes serve as examples of what may be achieved and their problems highlight many of the challenges. In bioinformatics, there is a hierarchy of interactions between researchers and resources that these centres of excellence serve: physically nearby groups, national groups, European and world-wide. One role for the European Commission is to help to connect these levels. We have scattered activities, and one goal is to increase efficiency by combining resources. Europe is leading in many fields and led in others, but has lost position due to lack of funding relative to the USA. Here we present a brief account of the current status of a variety of databases and software (IMGT, ENSEMBL, SWISS-PROT, PATHBASE), institutes (Sanger Institute, HGMP, EBI, EMBNET), and activities (EU-USA-Japan collaboration, Training) with some lessons to be learned and policy recommendations.

1.A. Specialised Databases - IMGT

IMGT (International ImMunoGeneTics information system®), available on the Web as http://imgt.cines.fr, is a high-quality integrated information system specialising in immunoglobulins (IG), T cell receptors (TR), major histocompatibility complex (MHC) and related proteins of the immune system of human and other vertebrate species, created in 1989. IMGT, a flagship of Europe, is unique in the world, and is recognised as the global reference in immunogenetics and immunoinformatics. While there are only a few hundred genes related to the immune systems, it must be remembered that there are 10^{12} T-cell receptor variations not encoded in the genome. A European project since 1992, it works in partnership with EBI (European Bioinformatics Institute) and other institutions. IMGT consists of sequence databases (IMGT/LIGM-DB, a comprehensive database of IG and TR from human and other vertebrates, with translation for fully annotated sequences, IMGT/MHC-DB, IMGT/PRIMER-DB), genome and structure databases (IMGT/3Dstructure-DB), Web resources (IMGT Marie-Paule page) and interactive tools. The IMGT server provides a common access to all Immunogenetics data and contains a massive 8000 pages. IMGT provides biologists with an easy to use and friendly interface.

A key point contributing to their success is the close interaction between bioinformatics and experimental biology researchers, based on a strong and coherent ontology (IMGT-ONTOLOGY). A problem is that database maintenance does not fit with national funding models. Unfortunately, the EU does not fund continuation of established infrastructures at the level needed. Its survival therefore remains continually in question.

1.B. ENSEMBL

ENSEMBL is one of the best sources of Human Genome Data in the world, available at http://www.ensembl.org. ENSEMBL is a joint project between EMBL - EBI and the Sanger Institute to develop a software system which produces and maintains automatic annotation on eukaryotic genomes. Access to all the data produced by the project, and to the software used to analyse and present it, is provided free and without constraints. ENSEMBL presents up-to-date sequence data and the best possible automatic annotation for metazoan genomes. Available now are human, mouse, rat, fugu, zebrafish, mosquito, Drosophila, C. elegans, and C. briggsae, with more to follow. **There is also a high level of automated analysis, which requires high performance computing and complex software engineering. It is the type of infrastructure whose development and maintenance needs to be centralised in one physical location in Europe. This centralisation may be complicated when distributed funding is required or implied, e.g. in EU Framework Programme research.**

New and different types of funding are required to maintain excellence on a world scale. The Sanger Institute receives extensive funding from the Wellcome Trust, but the EBI gets only 30% of the funding provided by federal funding in the USA for equivalent work. **It is recognised that public domain database maintenance and research requires long-term funding but the mechanism is not apparent in Europe.**

1.C. SWISS-PROT

The SWISS-PROT Protein knowledge base is a curated protein sequence database that provides a high level of annotation, a minimal level of redundancy and high level of integration with other databases. The Swiss Institute of Bioinformatics (SIB) and the European Bioinformatics Institute (EBI) maintain it collaboratively. The TrEMBL database contains the translations of all coding sequences (CDS) present in the EMBL Nucleotide Sequence Database, which are not yet integrated into SWISS-PROT. The databases SWISS-PROT http://www.ebi.ac.uk/swissprot/index.html and TrEMBL http://www.ebi.ac.uk/trembl/index.html are recognised as world class, so much so that they will be merged with the Protein Information Resource (PIR) at Georgetown University, based upon a U.S. National Institutes of Health (NIH) grant to establish a "unique, universal knowledge base" of protein molecules. The award, totalling $15 million over three years, will go to the establishment of a new resource http://www.embl-heidelberg.de/ExternalInfo/oipa/pr2002/pr231002.pdf called the United Protein Databases (UniProt), and European and American groups will manage it. The EU project TEMBLOR, with funding in several areas totaling nearly € 20 million over three years, is also supporting significant work related to these databases, incorporating the databases into the Integr8 http://www.ebi.ac.uk/integr8 access system for information integration. **One of the challenges of bioinformatics for the next few years will be to develop increasingly intelligent systems for further information integration. However, human assessment should keep its place with the help of improved bioinformatics tools.**

1.D. PATHBASE

The analysis of the phenotypes of mutant and transgenic rodents is key to the generation of models of human disease and to our understanding of gene function. Analysis of these systems requires considerable expertise in pathology and it is an overall aim of this project to integrate this expertise in Europe and to teach and inform scientists using transgenic and mutant rodents.
A significant problem is the availability to the scientific community of high resolution images of mutant mouse histopathology, thus the primary aim of this project is to generate a searchable database of histopathological images from transgenic and mutant mice. Production of this resource has required the development of appropriate data structures and an ontology of mouse pathology as well as innovations in database software and user interfaces. The development and implementation of ontologies underpins database interoperability and the ability to carry out accurate and sophisticated searches which are not hypothesis limiting. Over 1000 images have now been accumulated and by Summer 2003 will be on-line (http://www.pathbase.net) and annotated. In addition to the mutant image database, Pathbase is developing a reference resource for standardised and annotated images of mouse pathology and normal tissue histology. There exist few published collections of annotated colour images of systematic pathology of the mouse, and only one of non-neoplastic lesions, which are non-proprietary. Access to such selected images is important to those pathologists working on mutant mouse phenotyping, as many are either not trained as veterinary pathologists, or are new to the mouse system. This type of exercise also helps to standardise mouse pathology terminology and produce common agreed descriptions of lesions. Pathbase is now set to be integrated with other European databases (BioImage, EMMA-related databases) and with the Mouse Genome Informatics databases of the Jackson Laboratory in the USA.

The challenges which Pathbase has had to address overlap with those faced by many other bioinformatics and database projects, continuity of funding, recognition by the research community as a stable resource encouraging direct data submission, and software development, but with mutant mouse pathology a serious issue has been the lack of expertise in the European

Community and its dispersion over the whole of Europe. A major achievement of this project has been to network many pathologists working with mice across Europe, with long term effects for the training and dissemination of mouse pathology expertise. A further important factor has emerged during our development of Pathbase; that the coding of images and other data is highly labour intensive and currently cannot be done by machine. It will be important to take into account this requirement for human as well as computing resources in the future of many actively curated databases of this nature.

1.E. The Sanger Institute and its databases

Europe does succeed very well with the resources that it currently has. The Wellcome Trust Sanger Institute http://www.sanger.ac.uk is an excellent example of what can be done given adequate funding. It was founded 1992 as a genome research institute, with a primary focus on large-scale sequencing. The Sanger Institute sequenced 1/2 of the nematode worm genome, largest share of 2 yeasts, 1/3 human, and many pathogens including tuberculosis and malaria. Important reasons for their success include research excellence, that they took a strong position on open data release, and have been involved in multiple informatics resources and analyses. It is important to be able to keep raw data (e.g. a trace repository contains terabytes of data) and secondary resources are also critical. Distributed annotation is a necessary part of this process, and needs large resources. There is a large IT resource (1400 CPU, 100TB on 200 servers). **A key issue is the overall level of funding needed to compete on a world-wide scale. The genomics budget of the Sanger centre** http://www.sanger.ac.uk/Info/Intro **is of £300M core funding from the Wellcome Trust over 5 years, plus £65M IT funding and commitment to new buildings. This is comparable with the whole EU budget for basic genomics research. Good genomics resources are not cheap**! Building on its sequencing successes, the Institute renewed itself in 2000/1 with a new director (Allan Bradley). **This history shows how a policy of open publication and excellence can attract major funding.**

Attitudes to data access and IPR

- **A key to success has been to take a strong position on open data access and IPR**
- **Fundamental principles of academic research:**
  - **Credit is assigned based on publication**
  - **Publication allows verification, and, more important, reuse of ideas for novel developments**
- **Publication for data resources (primary or secondary) means allowing unconstrained reuse for unforeseen purposes**
- **Open access is good for science, and good for investigators/institutions**
- **"Fort Lauderdale" principles of tripartite responsibility[6]**
- Comments on EU Workshop Report 2001: **http://europa.eu.int/comm/research/era/pdf/ipr-bioinformatics-workshopreport.pdf** "Managing IPR in a knowledge-based economy - Bioinformatics and the influence of public policy"
  - There were concerns that the report placed too strong an emphasis on the use of IPR to protect data, databases, and software. However, the report conclusions were broader:
    - ⇒ All involved sectors, from academia to large pharmaceutical companies, strongly support a comprehensive and up-to-date publicly available and free set of biological data, with IPR control being used only where appropriate to maintain easy and universal access.
    - ⇒ Database infrastructures need clear IPR protection policies at all stages of creation, management and access. For publicly funded data resources, the policy towards IPR should be to avoid restricting access.
    - ⇒ Public funding and IPR rules should encourage collaborations, especially public/private

6 "Sharing Data from Large-Scale Biological Research Projects: A System of Tripartite Responsibility" The report on this workshop, held at Fort Lauderdale on 13-14 January 2003, is available at www.ebi.ac.uk/microarray/General/News.

ones. Open collaboration is powerful and in the public interest.

- **Conclusions from the Sanger experience**
  - **Open data access is so necessary for science and collaboration that here it is made a core principle, and it should be the norm for publicly funded bioinformatics.**
  - **Research bioinformatics is (relatively) cheap and fits classical response-mode funding; whereas resource bioinformatics involving continuing infrastructure support is expensive and requires long term strategic planning, co-ordination and funding.**

1.F. The HGMP Resource Centre

The UK Human Genome Mapping Project Resource Centre http://www.hgmp.mrc.ac.uk (HGMP-RC) provides access to leading edge tools for research in the fields of genomics, genetics and functional genomics. The HGMP-RC's mission is to: provide both biological and data resources and services to the medical research community, with a special emphasis on those relevant to the Human Genome Programme; facilitate genomic research by the provision of cost effective centralised collaborative and training facilities; encourage users to share their data, information and resources; encourage the transfer of technology from academic contexts to commercial/industrial applications. In bioinformatics, HGMP provides a wide range of resources and services: computing; integrated software; genome web; extensive training; help for users; new technologies; wet lab 'services'. From their experience it can be stated that **the holy grail of molecular biology used to be the idea that sequence can predict structure and that structure can predict function. We need to identify all the parts (molecules) encoded in a genome. Today's reality is that we need to determine sequence, structure and function and try to relate them in order to gain understanding and build all the parts into models of working systems.**

1.G. EBI - The European Bioinformatics Institute

The European Bioinformatics Institute (EBI) http://www.ebi.ac.uk is a non-profit academic organisation that is an outstation of the European Molecular Biology Laboratory (EMBL) http://www.embl.de. The EBI is a centre for research and services in bioinformatics. The Institute manages databases of biological data including nucleic acid, protein sequences and macromolecular structures. The mission of the EBI is to ensure that the growing body of information from molecular biology and genome research is placed in the public domain and is accessible freely to all facets of the scientific community in ways that promote scientific progress. The EBI serves researchers in molecular biology, genetics, medicine and agriculture from academia, and the agricultural, biotechnology, chemical and pharmaceutical industries. The EBI does this by building, maintaining and making available databases and information services relevant to molecular biology, as well as carrying out research in bioinformatics and computational molecular biology.

The EBI manages a huge range of resources, and is also in the process of unifying them and making them more effective. At the beginning of 2002, the Commission of the European Communities provided the largest ever single injection of funds into bioinformatics infrastructure and research in Europe, to the EBI, with the TEMBLOR http://www.ebi.ac.uk/Information/funding/temblor.html contract under the Quality of Life programme, which provides almost €20 million over three years to the EBI and a group of 25 collaborators in 11 other countries. The project concentrates on research and development to build the European Bioinformatics resources for the genomic era and beyond. These resources will be embedded in an integrated layer, known as Integr8, allowing biomedical researchers to fully exploit genomic and proteomic data. Integr8 will draw on databases that are maintained at major bioinformatics centres in Europe, and also on important new resources. Other major projects within TEMBLOR are: IntAct to create a database for protein-protein interaction data, EMSD to enhance the European Macromolecular Structure Database and DESPRAD to establish a microarray data repository. However, even this contract is only for three years, and is not renewable as such. **The EBI needs an NCBI-like (USA National Center for Biotechnology Information**

http://www.ncbi.nlm.nih.gov**) funding structure to ensure that databases are made available for scientific community.**

1.H. EMBNET

EMBnet http://www.embnet.org was established (1986), in the days before the widespread use of the Internet in Europe, as a service-based group of collaborating nodes throughout Europe distributing the EMBL Data Library to end-users by means of DECnet. The combined expertise of the nodes allowed EMBnet to provide services to the European molecular biology community, which encompassed more than could be provided by a single node. Following this success, a variety of nodes world-wide joined EMBnet that now offers a panel of experts available to give specialised courses. National nodes provide local training and support programmes in local languages. EMBnet supports a training programme using telematics. Staff from several EMBnet nodes collaborate in developing new biocomputing tools. EMBnet nodes provide their national scientific community with access to high performance computing resources, specialised databanks and up-to-date software. Many nodes act as redistribution centres to national research institutes. Collaborative technical expertise within EMBnet provides support for sustaining the biocomputing facilities of the member nodes. Each national node provides help to end users in their local language about all aspects of biocomputing.

The future role of EMBnet nodes was discussed. Due to the distributed and local nature of the collaborative participation, different nodes provide different quality, from dead to very lively! Different views were expressed on its future, including very critical ones, since it is considered by some to be an important resource but its best use in the world of new technologies is unclear. **The debate raised the question of how best to use distributed resources around Europe, in addition to the centralised resources.**

1.I. EU-USA-Japan collaboration and competition

Comprehensive bioinformatics databases, publicly funded and available for consultation without charge, are essential to research and innovation activities in Europe. **« Free riding » on American or other facilities is a short-sighted solution; open international collaboration will depend upon major participants – such as the EU – being prepared to support an equitable share of the cost burden. The USA has the advantage of a single federal government, language and culture, and these advantages are nowhere better illustrated than in the structure of research funding through the NIH** http://www.nih.gov**. In the context of infrastructure, the projects supported through the National Center for Research Resources are of particular relevance.**

1.J. Training

Training is very important in national centres for each ‘national research community’ and also provided to industry at a cost**.** Industry also provides its own training supplied in-house, by for-profit training centres or by academics. However, training provisions are unbalanced throughout Europe. There are many courses for bioinformatics student training, perhaps too many (the worrying factor being their future employment prospects); see http://wwwpa.in.tum.de/bioinf/PrO.doc and http://www.ebi.ac.uk/MarieCurie/index.html. But **there is often insufficient training for biologists in the use of bioinformatics. Teaching of bioinformatics should be organised and emphasised as a discipline in itself. Given the lack of a general bioinformatics training in many countries at the national and also European level, most small groups need to train young scientists in bioinformatics starting from a purely biological or computational background. A framework for training in bioinformatics is needed.**

1.K. Recommendations and Options

Organisations at national (e.g. Member State) and European (e.g. EU and other European) levels need to work together to ensure that appropriate access to bioinformatics facilities (data, software and training) is available. Failure to meet this responsibility will reduce the value of other efforts, public and private, in biological training, research, and applications. In the world-wide context of bioinformatics, there is the temptation to try to protect local resources by

privatising at the local level and relying on public resources at the higher level. This is a strategy for failure. Institutions such as the Sanger Institute have shown that open publication and world-class research attract further funding for the quality of the work.

There already exist important European institutions and databases operating at world-class levels. There are also EU (Sixth Framework Programme) and various Member State programmes aiming at creating and maintaining new institutions and organisations, for example with the EU Networks of Excellence initiatives. These efforts should be identified, supported and strengthened.

An extended debate is needed on funding responsibilities and on the relative roles of organisations such as EBI, EMBnet, national bioinformatics centres, etc., especially in the context of major international facilities, e.g. in the USA such as NCBI, of emerging technologies and projects such as GEANT and the GRID, and of new research and archiving needs. This debate must embrace the public authorities (not least as funding sources, but also for views on national needs and strategies), the public research institutes and academic centres, and the private sector industries involved both in the provision of information (e.g. publishers, database hosts, internet service providers) and users.

2. STRUCTURAL NEEDS FOR THE FUTURE

2.A. Hardware and Software Needs -

- IT infrastructure
  - Need for hundreds of terabytes for storing e.g. images
  - GRID type technology for efficient broad access to resources
- A Computer Industry Perspective

Biological data is large volume and complex – therefore there are new challenges. A major weakness is lack of standards. A major threat is inability to act quickly enough to 'conquer' data challenges. Standards are central: 'computing standards' and 'biological ontologies' – are both needed with a strong support for 'Open' standards.

A key question is whether Europe has access to sufficiently large scale computing power, with petaflops of computing power, terabytes of memory and tens of terabytes of storage, which is needed for research and analysis in bioinformatics and computational and systems biology, e.g.
http://www.cray.com/news/0211/x1announce.html
http://www.research.ibm.com/journal/sj/402/allen.html
http://h18002.www1.hp.com/alphaserver/news/sandia_celera_0101.html

2.B. New Types of Data Resources needed and/or being developed

- **Databases**
  - Comparative sequence analysis
  - Comparative gene finding ~ 10 vertebrates soon and multiple genomes will be available for many phyla. Comparative genomics already demonstrates the illuminating commonalities across phyla far wider than the chordates – e.g. the Hox genes
  - Identifying functional non-coding sequences such as regulatory sequences
  - Images: Now available e.g. FlyBase, Mouse Atlas, PATHBASE
  - Integrate sequence based functional data from microarrays, systematic RNA interference (RNAi)
  - Human variation data (e.g. haplotype map, sequencing)
  - Sequence-based human genetics (SNPs)

Several databases have been established or supported by EU funded projects, but there is no sustained funding scheme to support them later when successful. National funding for data resources is difficult to obtain since they are viewed as European/global resource. The example of SWISS-PROT is highly instructive, where major funding was eventually obtained from the NIH of the USA. **A major priority must be to resolve the structural problem of how to fund and maintain**

databases in the long term, both at the national and European level.

2.C. Access to Literature References

The use of literature references is a key part of what we do. **We need better access tools and the ability to do text based management and analysis.** One project using this approach is **E-BioSci** http://edam.prov.ingenta.com:8081/ebiosci/search/query**,** which will:

- foster optimal pooling and use of European biological archives and data collections
- stimulate the development of common protocols and methodologies for efficient searching and retrieval of information contained in bibliographic and sequence, or sequence-related databases
- provide a framework for further research into more effective strategies for linking of bibliographic with molecular, genomic and 3D-image databases.

2.D. SME (Small and Medium Enterprises) Support

The EU Commission should take note of roles of bioinformatics companies. Professional services are needed and we should consider how to support this and how to integrate them into funding. It is a difficult time for both large and small companies. For small companies money is needed for survival and they tend to be absorbed by larger companies or die. For all companies a market is needed with positive cash flow, networking with other companies and better links to scientists in academia. It may be very expensive for small companies to access to databases, software and training and they often do not have an adequate budget. Collaboration is very important for industry.-. The GRID is an important opportunity that may enable organisations to share resources. **A role for SMEs could be maintenance of databases, although there is no obvious business model for maintaining it.** One model is supporting researchers but it is difficult to be profitable based on back end servers and other services. **It is surprising that Europe does not have a USA SBIR-like (http://www.sba.gov/sbir/indexsbir-sttr.html) peer review system.**

2.E. Ontologies – encapsulating knowledge

At least one speaker gave a definition of this topical concept: "An ontology is a system of coding knowledge in such a way that it is computer readable". **Ontologies are key, and biological interpretation always needs a biological context to remain meaningful.** Function does not have a universal metric and cannot be described except in the process in which a molecule is involved. According to the context, a single molecule can have many or even innumerable functions (e.g. water).

As we move towards a knowledge-based economy, software entrepreneurs are selling systems for storing, retrieving and "mining" knowledge, including natural language databases capable of handling scientific data and natural language. "GO" is a much-cited piece of software, attributed to Michael Ashburner of Cambridge and others. The acronym stands for "Gene Ontology". More information is available at http://www.geneontology.org/, run by the Gene Ontology Consortium. The members cover 15 different species genome projects, with a goal of producing a dynamic controlled vocabulary that can be applied to all organisms even as knowledge of gene and protein roles in cells is accumulating and changing. The three organising principles of GO are molecular function, biological process and cellular component. Several speakers discussed the problems of incompatibility between different "ontologies".

2. F. General Support for Research Infrastructures

The question of support for infrastructures in bioinformatics occurs in Europe at both national and European levels. Some support is available at European level via the European Commission, both directly in research http://www.cordis.lu/lifescihealth/genomics/home.htm and in infrastructures http://www.cordis.lu/fp6/infrastructures.htm. However, intergovernmental support is also possible via the ESFRI. The mission of the new "European Strategy Forum on Research Infrastructures"

http://www.cordis.lu/rtd2002/era-developments/infrastructures_forum.htm set up by the Member States of the Union on the initiative of the Commission is to provide a multidisciplinary platform open to all EU countries to monitor the needs expressed by the scientific community. The infrastructures include major installations as well as supercomputers for biocomputing, GRID-type[7] architectures for scientific cooperation, databases for social sciences, virtual libraries or networks of ecological reserves for biodiversity. The Forum will be an incubator for European or multinational projects with the aim of developing new research infrastructures in Europe. **Attempts should be made to obtain the maximum coordination at all levels to support research with a European dimension.**

3. FUTURE DIRECTIONS IN BIOINFORMATICS RESEARCH

3.A. Genomics

The ability to predict gene products from genome sequence is rudimentary. We need to identify gene products (RNA and protein) that can actually be detected in cells. In vertebrates particularly, we need to find all the encoded molecules in vivo and in vitro, as we cannot predict them in silico. Many of the interpretative problems are still unsolved. Genomics was initially driven by data collection and much interpretation is hypothesis driven. A new direction is to provide simple tools so as to be able to access databases and perform analyses to build generalised models for biological processes. The best approach is comparative genomics leading on to functional studies with improvement of sequence analysis tools and text mining tools and new solutions for knowledge representation. Gene regulation is a key area of importance and needs well-defined experiments and a computational framework.

Sequence data is extremely valuable and new sequences are very good value for money. The added value of new genomes is still important. Unfortunately, apart from the Sanger Institute, Europe appears to have dropped out of sequencing for very large genomes. We can foresee having a thousand genomes sequenced. New technology has reduced greatly the cost of each new genome sequence. A sequence is a needed basis for many modern experiments and the extensive use of SWISS-PROT shows what a valuable resource this is to catalogue the proteins. Newly sequenced genomes do give additional information beyond that of the organism itself by increasing the value of previous data e.g. regulatory elements and comparison can lead to fuller understanding.

Haplotype analysis for disease identification will be a key element of individualised health care and understanding disease, with personalised treatment a key goal. Haplotype information can be gained by repeated genomic sequencing and will need sophisticated analysis tools to find associations with disease.

3.B. Experimental design

Bioinformatics should play a key role in the integrated design of experiments. Genomics projects must recognise that significant components should be devoted to bioinformatics. Every biological research project needs to devote substantial resources to bioinformatics, as an integral element of the project, for planning experimental procedures and the analysis and interpretation of results, as well as for their ultimate storage in accessible and comprehensible form. A key point is that all experiments in the life sciences should have a proper commitment to analyse store and publish data. It was generally felt that a significant fraction of resources in each major project should be devoted to bioinformatics and data analysis, including storage, retrieval and analysis of data during the life of the project, plus consideration given to long term maintenance. While the percentage will depend on the nature of each experiment, experience shows that a level of 20% was typically present in large projects, and that this reached 50% or more in larger genome projects and in particular areas like microarray experiments. Smaller projects will also need significant data analysis support. For small-scale projects - let's say a

[7] GRID: The internet of tomorrow with broadband networks and supercomputers with a high data storage capacity, the whole managed on a decentralised basis by scientists.

medium-sized group doing some array experiments, or a group involved in small-scale proteomics (1-2 persons running 2D gels and MS), experience is that they often need a part-time bioinformatics expert to help them with mostly rather trivial chores (writing small applications, building an in-house database to keep track, etc.). The ideal situation is when this person also does algorithm development on a more advanced level in a similar area of work on molecular biology and biomedical research. Laboratories are already hiring their own local bioinformatics experts to keep pace with the growing set of available techniques, while still deeply immersed in specific biological questions. This detailed work is only possible thanks to the availability of effective methods on the World Wide Web, reliable service providers, and the proximity of other bioinformatics experts. In time, these local experts attached to an experimental biology research group will be integrated in the quest for a single scientific goal of that group, and bioinformatics will become an essential part of the research in molecular biology and biomedicine. Indeed, **the way in which biologists address specific problems will be deeply influenced by the availability of bioinformatics methods for the design, management and interpretation of the results.**

3.C. High Throughput technologies

The new instrumentation – high throughput, automated, particularly nucleotide sequencing and arrays - is enabling biologists (and their supporting technical staff) to generate vast quantities of information. The global databases are collaborating with one another, sharing the workload of receipt, curation, annotation and storage. Admittedly, there remain unresolved subtleties in detecting genes and regulatory elements within the total genome (the software does this automatically, though not with reliability, particularly regarding gene structure and subtler regulatory relationships). The rapid development and diffusion of arrays is producing new challenges.

Arrays are available from several suppliers (e.g. Affymetrix, Amersham), and there are good review articles and company websites. The number of spots per array is increasing rapidly. At Perlegen http://www.perlegen.com/servlets/templator.Server?PAGE=about, they are speaking of millions. The biologist is in danger of drowning in this flood of data, but projects are in progress to deal with this, e.g. via the project DESPRAD http://www.ebi.ac.uk/microarray/Projects/desprad/index.html. Experiments are not readily replicable, especially not when conducted in different places, by different experimenters, with different arrays, etc. This may pose problems for referees and journal editors, as to the criteria to be applied in judging the acceptability of papers and the validity of conclusions. **Science becomes gravely hampered if scientists in different locations or contexts and at different times are unable to communicate data and results reliably to one another – similar problems are facing curators of databases, public or private, hence the need to develop common languages and protocols, e.g. MIAMI and MAGE** http://www.ebi.ac.uk/arrayexpress/Standards/index.html**.**

A solution to the problems with high throughput data analysis today is to build a network of resource centres instead of relying on local, small scale bioinformatics facilities and to use NoEs or IPs to provide various bioinformatics resources. **The resource centres could provide bioinformatics expertise, stable environments (including backup), high-throughput, high-volume analysis facilities and training opportunities.** Ideas for tools enhancements include a need for one European standard middleware and tools need to be enabled to work with e.g. BioMoby http://www.biomoby.org , which is an international research project involving biological data hosts, biological data service providors, and coders whose aim is to explore various methodologies for biological data representation, distribution, and discovery. Centres need to build interfaces for their platforms and future grant proposals need to be linked to standard compliance and data availability. Software should be available under open source licenses.

3.D. Algorithm design

Algorithm development needs to be more biologically oriented. Physics and engineering contributions are also key. Further development of algorithms is necessary to address the many unsolved problems today. Some of the problems have been around for decades and will not disappear in the next few years. The new "big" EC instruments may not be optimal for the funding of algorithmic developments: it seems difficult for the small laboratories to contribute much.

Networking on the other hand is necessary to foster integrative approaches and prevent the development of many solutions to the same problem.

3.E. Database building

In the field of genomics and proteomics, bioinformatics provides the key connection between all different forms of data gathered by new high-throughput techniques such as systematic sequencing, proteomics, expression arrays, yeast two-hybrid (y2h), and high throughput screening. We will soon have at our disposal hundreds of genomes, thousands of protein structures, protein interactions determined by y2h and tens of thousands of genes with their expression monitored in hundreds of experiments, and well over a million single-nucleotide polymorphisms (SNPs). Handling this massive amount of data requires powerful integrated bioinformatics systems. Issues related to database inter-operability, information representation and data description (the much abused term 'ontology') are currently being addressed. Also, in fields such as automatic extraction of information from the biological literature, activity has increased greatly since the first papers were published five years ago. Database building should have an XML screen so that other people can use tools themselves. We need to develop novel data structures and the ability to integrate information from various sources, e.g. BioMOBY[8] in Canada**.** Knowledge representation in combination with visualisation of text, formal representations of knowledge – ontologies plus their representation as research tools in their own right and dealing with this from engineering point of view is in its infancy, e.g. MyGrid[9] in the UK**. As came up often in the discussions, providing funding for the initial development of a database but not for the curating and maintenance leads to a waste of resources.**

3.F. Structural Genomics

Homology modelling is the most successful tool for understanding the significance of protein 3D structure. Part of the structural genomics effort goes into covering the protein structure space sufficiently well that "all" proteins can be modelled with sufficient accuracy. Better homology modelling will reduce the number of structures that need to be solved experimentally. The quality of homology models is primarily determined by the accuracy of the alignment between the sequences of the query protein and the template structure and the force field that is used in the modelling. For low to very low sequence similarity, homology models can only give qualitative features of the protein.

Function prediction from the structure alone remains an elusive goal and one way is to search for conservation of structural templates. We study aspects of function by looking at physical properties that can be calculated from the structure, such as shape, electrostatic potential, and molecular dynamics. An important part of structural bioinformatics, in particular at a medically oriented research institute, is the study of the interaction between a protein and ligands, building up drug design and virtual screening activities.

The myth of function. The holy grail of molecular biology was the idea that sequence can predict structure and that structure can predict function. We therefore needed to identify all the parts (molecules) to understand the whole (cell and organism). The reality is that we need to determine sequence, structure and function and try to relate them in order to gain understanding and build all the parts into models of working systems.

3.G. Development of Systems Biology

---

[8] BioMOBY is an international research project involving biological data hosts, biological data service providers, and coders whose aim is to explore various methodologies for biological data representation, distribution, and discovery. http://BioMOBY.org provides an online resource for modules, scripts, and schema for developers of MOBY-related software.

[9] MyGrid is a sematic web for molecular biology with a strong emphasis on ontologies http://www.mygrid.org.

There are currently disequilibria within the bioinformatics area that favours informatics disciplines (database activities and information management systems) against data-analytical and theoretical methods, mathematical modelling and computational simulation techniques, usually referred to as "Computational Biology".

What is already in progress is a shift in a paradigm implicit in biological research, where it was assumed that the scientist could directly interpret raw experimental data. Mathematical models and simulation techniques are needed in order to integrate biological knowledge with the ever-increasing amount of experimental data into a formal framework (in silico model) and to test hypotheses on the functioning of biological processes.

Future applications of bioinformatics extend towards the study of fundamental biological questions, such as macromolecular assembly, protein folding and the nature of functional specificity. Such issues extend beyond the current perception of bioinformatics as a support discipline and address aspects of biological complexity, including the simulation of cellular systems and molecular interaction networks. The contribution of bioinformatics to these areas is related to the development of concepts in theoretical molecular biology, but also to the management and representation of complex biological information.

The study of particular systems is the source of inspiration that guides the formation of general ideas from specific cases to general principles. The study of fundamental problems encourages the interdisciplinary nature of bioinformatics and allows the field to reinvent itself. It may be the interplay and parallel activities in these areas that defines bioinformatics as 'biology by other means'.

However, different areas have been developing at different rates. The technical and computational developments are very attractive for newcomers from fields such as computer science, engineering and mathematics. **The practical applications of bioinformatics are highly sought after by institutions and companies, and constitute the natural entry point for most molecular biologists and biochemists.** Perhaps the work related to the fundamental biological problems is less well regarded and requires more attention, given its importance for the future of biology as a quantitative science.

Training programmes in bioinformatics have an essential role in preparing the new generation of scientists, not only in the use of tools or in the development of computational techniques, but also in providing the necessary background to tackle basic biological questions with new methods and novel ideas. The issues related to the difficulties in combining the different flavours of bioinformatics have acquired new importance at the time of publishing these pages.

It has become painfully obvious that in the perception of many experimental biologists, research in Computational Biology is not considered to be essential for the future of Genomics and Proteomics — and they will be interested only in the contribution of Bioinformatics as an auxiliary technique. This opinion will tend to support applied work at the expense of research work in Computational Biology. **Only the development of integrated bioinformatics systems will enable the manipulation of complex biological information. The solution of fundamental problems with these new data will require bioinformatics applications that go beyond the current work on data integration and manipulation. It is the responsibility of computational biologists, and that of the Professional Societies such as ISCB, (http://www.iscb.org) and the emerging European branch (ECCB), to pass on to the biological community a clear message about the need to preserve a balance between the different areas of Bioinformatics and Computational Biology.**

a) Bioinformatics research and resources have to gradually evolve from being "individual part-oriented" (genes, proteins, structures) to trying to address bigger biological pictures (such as protein-protein interactions, regulatory networks, mutation linked to phenotype, cellular subsystems), thus aiming at generating an understanding (a *predictability of behaviour*) of biological subsystems (mitochondria, spliceosome, etc.).

b) Given that no simple "mathematical formula" will ever be able to encompass the complex behaviour of biological systems, the ultimate bioinformatics theory is expected to consist of a computational model able to realistically simulate the dynamic behaviours of these systems, based on molecular "first principles" such as sequences, 3-D structures, interaction patterns, and physico-chemical parameters (kinetic and affinity constants).

Key existing projects include:

- The realistic computer modelling of an entire *E. coli* cell is the prototype of high-visibility, ambitious, visionary project capable of federating all the components of today's bioinformatics activities (Bio-specific IT developments, data integration/visualisation schemes, algorithms, applied maths) as well as new contributions from physics and engineering disciplines. Such a project is pushed forward as a 10-year project by the ***International E. coli Alliance,*** which is a consortium of different groups of scientists with the combined aim of modelling the *E. coli* cell *in silico.* The consortium includes http://www.projectcybercell.com from Canada, the http://www.iab.keio.ac.jp/, in Japan, http://www.gsk.com/index.htm cell modelling group and the http://ecmc2.sdsc.edu from the USA. A UK academic consortium will also be created as well as European consortia in the future.
- The silicon cell consortium http://www.siliconcell.net
- The National Resource for Cell Analysis and Modeling (NRCAM), developer of the Virtual Cell, is a national resource centre supported by the National Center for Research Resources (NCRR), at the National Institutes of Health (NIH). NRCAM is developing methods for modeling cell physiological processes in the context of the actual three dimensional structure of individual cells. Approaches in computational cell biology are coupled with high resolution light microscopy to facilitate the interplay between experimental manipulation and computational simulation of specific cellular processes. NRCAM has developed a general computational tool, the Virtual Cell for modeling cell biological processes. http://www.nrcam.uchc.edu
- The E-Cell initiative http://ecell.sourceforge.net

Applied sciences areas that can have a major impact on computational biology:

- **Digital signal processing:** to discover information hidden in the DNA sequence.
- **System Identification:** to identify complex regulatory networks from experimental data.
- **Networks theory:** to be able to understand, predict and eventually modify the behaviour of complex regulatory networks.

Research projects should necessarily have:

- A clear objective that targets a biological research problem
- A theoretical/computational component
- An experimental validation component

Potential impacts:

- **Optimally designed experiments and data interpretation using theoretical models and computational tools**
- Computational tools for the discovery of novel “objects” in the genomic sequences
- **Theoretical models to explain the organisation and function of complex regulatory networks**.

3.H. Medical informatics

A workshop was held in 2003 http://lyon2003.healthgrid.org to discuss computational approaches to health related problems. The aim of the conference was to create a bigger awareness about the possibilities and advantages linked to the deployment of GRID technology in health. In this context "Health" does not involve only clinical practice but covers the whole range of information from molecular level (genetic and proteomic information) over cells and tissues, to the individual and finally the population level (social healthcare). Grid technology offers the opportunity to create a common working backbone for all different members of this large "Health family" and will hopefully lead to an increased awareness and interoperability among disciplines.

3.I. Pharmaceuticals

There is an enormous range of collaboration in the pharmaceutical arena. A major danger observed in these collaborations but present more widely, is the tendency to store a minimal amount of data, usually in processed form. More should be permanently conserved. The more medically- and patient-orientated the research, the more confidentiality becomes a key issue. When bioinformatics reaches the level of dealing with personal data there are legal and ethical considerations that become apparent.

Even in the IPR-protection oriented world of pharmaceuticals, open source licensing is a valuable approach to encourage maximum transfer and use of data. Immunological bioinformatics is needed in vaccine design area, protection against bioterrorism and in problem driven cases e.g. asthma. Access to and understanding of distributed, heterogeneous information resources are critical but it is a complex, time consuming process, because of thousands of relevant information sources. Rapidly changing domain concepts and terminology and analysis approaches confound the situation as do constantly evolving data structures, continuous creation of new data sources and highly heterogeneous sources and applications. Data and results are of uneven quality, depth and scope. Collaboration for understanding and consensus is essential.

3.J. Areas of bioinformatics-based research in EU Framework and National Programmes

A number of areas have been suggested as being suitable for funding support and encouragement, either in the EU Framework Programme (FP) and/or as part of a co-ordinated programme by national and local funding agencies and charities. Some of these recommendations confirm the correctness of actions already taken at the level of new FP6 instruments.

The research field is already supported by a number of existing projects under the EU FP5 http://www.cordis.lu/lifescihealth/genomics/home.htm :

- TEMBLOR to provide a common platform for access to and analysis of a wide range of databases, including development of ontologies, common formats, easy access tools, sophisticated analysis programmes
- Integrated projects, e.g. GENOMEUTWIN, with major bioinformatics component structure, to analyse population genomics
- Smaller projects such as IMGT to integrate immunogenetics data, eBioSci to look at the problem of text mining and access

In the next FP6 the research field is defined in first call with the topic of “Genome annotation to unify and complete the identification and analysis of the nature and function of genes” and in the second call by the topics of “A genomics grid to unify European bioinformatics resources” and “A software platform to start to analyse bioinformatics data from the systems biology point of view”.

Based on an analysis of both specific suggestions and the key ideas of the document, and also bearing in mind that the FP5 projects will terminate in 1-2 years, a number of key areas that might be part of future calls for proposals in FP6 include:

- **Bioinformatics components of all major proposals** – A general observation is that many results from European life science research programmes are seriously underutilised because of insufficient resources devoted to resulting data archiving, analysis and further usage, most especially on large throughput projects. **It was generally felt that a significant fraction of funding resources in each major project should be devoted to bioinformatics and data analysis, including storage, retrieval and analysis of data during the life of the project, plus consideration given to long term maintenance.**
- **Systems and Computational Biology** – This is a complicated area, in that it attempts to model systems of almost infinite complexity based on multiple biological data sources that are both vast and insufficient. Projects range from small scale time dependent modelling of metabolic pathways to complex protein folding calculations to time and space dependent modelling of entire cells to phsiological simulations for pharmaceutical testing. Collaborations are already forming to co-ordinate work in these areas. The best effort at the European level is to call for the development of a generalised computational biology simulation language and packages that unify all these levels,

with the flexibility to specify as boundary conditions the items not well known, and model those areas of concern where results are relevant and may be tested, and where data is sufficient for meaningful results. **It should involve a computational model able to realistically simulate the dynamic behaviours of these systems, based on molecular "first principle" such as sequences, 3-D structures, interaction patterns, and physico-chemical parameters (kinetic and affinity constants) to be able to understand, predict and eventually modify the behaviour of complex regulatory networks and to test hypotheses on the functioning of biological processes.**

- **Database and software upgrading, optimum utilisation and support** – Given the increasing flood of data, **important resources must be devoted to upgrading the storage, analysis, access, and integrative capability at a European level, moving forward from a gene centred approach to a protein centred approach**, with support for a combination of centralised and distributed facilities. Separate and appropriate funding should be considered for the infrastructural maintenance side and the new research side of this problem.
- **Immunological bioinformatics** – There is a need to upgrade existing databases and analysis to make maximum use of these resources for health applications such as fighting disease, e.g. vaccine design, allergies, autoimmunity, regulation of the immune responses, bioterrorism, etc.
- **Haplotype analysis for disease identification** will be a key element of individualised health care and understanding disease.

A possible solution for improving the quality of homology models for proteins may lie in an integrated approach of sequence alignment, secondary structure prediction and modelling, in combination with advances in the modelling algorithms. An important part of structural bioinformatics is the study of the interaction between a protein and ligands, needed for drug design and virtual screening activities.

WORKSHOP DETAILED CONTRIBUTIONS

CONTRIBUTED PAPERS, PERSONAL COMMENTARY and SESSION SUMMARIES

SESSION I: THE FOUNDATIONS (Chair: H. Werner Mewes)

Immunoinformatics in IMGT: a synergy between IMGT-ONTOLOGY and bioinformatics tools development for knowledge management and medical application

Marie-Paule Lefranc[a,b]

[a]*Laboratoire d'ImmunoGénétique Moléculaire LIGM, Université Montpellier II, UPR CNRS 1142, Institut de Génétique Humaine, IGH, Montpellier, France*

[b]*Institut Universitaire de France*

Address for correspondence

Marie-Paule Lefranc
IMGT, the international ImMunoGeneTics information system®
LIGM, UPR CNRS 1142, IGH
141 rue de la Cardonille
34396 Montpellier Cedex 5, France;
Tel: +33 4 99 61 99 65; Fax: +33 4 99 61 99 01
Email: lefranc@ligm.igh.cnrs.fr
IMGT, http://imgt.cines.fr

1. Introduction

The molecular synthesis and genetics of the Immunoglobulin (IG) and T cell Receptor (TR) chains is particularly complex and unique as it includes biological mechanisms such as DNA molecular rearrangements in multiple loci (three for IG and four for TR in human) located on different chromosomes (four in human), nucleotide deletions and insertions at the rearrangement junctions (or N-diversity), and somatic hypermutations in the IG loci (for review [1,2]). The number of potential protein forms of IG and TR is almost unlimited. Owing to the complexity and high number of published sequences, data control and classification and detailed annotations are a very difficult task for the generalist databanks such as EMBL, GenBank and DDBJ. These observations were the starting point of IMGT, the international ImMunoGeneTics information system® (http://imgt.cines.fr) [3] created in 1989, by the Laboratoire d'ImmunoGénétique Moléculaire (LIGM), at the Université Montpellier II, CNRS, Montpellier, France.

2. IMGT-ONTOLOGY concepts

IMGT is the global reference in immunogenetics and immunoinformatics. It is a high quality integrated information system, specializing in IG, TR, MHC and related proteins of the immune system of human and other vertebrates, which consists of three sequence databases, one genome database, one 3D structure database, Web resources ("IMGT Marie-Paule page") and interactive tools for sequence and genome analysis. All IMGT data are expertly annotated according to the IMGT Scientific chart.

The IMGT Scientific chart provides the controlled vocabulary and the annotation rules for data and knowledge management of the IG, TR, MHC and related proteins of the immune system of human and other vertebrates [3,4]. IMGT has developed a formal specification of the terms to be used in the domain of immunogenetics and bioinformatics to ensure accuracy, consistency and coherence in IMGT. This has been the basis of IMGT-ONTOLOGY [4], the first ontology in the domain, which allows the management of the immunogenetics knowledge for human and other vertebrate species. IMGT Scientific chart rules are based on the five concepts defined in IMGT-ONTOLOGY: IDENTIFICATION, DESCRIPTION, CLASSIFICATION, NUMEROTATION and OBTENTION.

IDENTIFICATION concept: standardised keywords. IMGT standardised keywords for IG, TR and MHC include general keywords, indispensable for the sequence assignments, and specific keywords, more specifically associated to particularities of the sequences or to diseases [3].
DESCRIPTION concept: standardised labels and annotations. 177 feature labels are necessary to describe all structural and functional subregions that compose IG and TR sequences, whereas only seven of them are available in EMBL, GenBank or DDBJ and none in PDB. Annotation of sequences with these labels constitutes the main part of the expertise [3]. Levels of annotation have been defined, which allow the users to query sequences in IMGT/LIGM-DB even though they are not fully annotated. Prototypes represent the organisational relationship between labels and give information on the order and expected length (in number of nucleotides) of the labels [3].
CLASSIFICATION concept: standardised gene nomenclature. The CLASSIFICATION concept has been used to set up a unique nomenclature of human IG and TR genes, which was approved by the Human Genome Organisation (HUGO) Nomenclature Committee (HGNC) in 1999 [1,2] and has become the community standard. The complete list of the human IG and TR gene names is available in Genew (UK), the Genome DataBase GDB (Canada), LocusLink at NCBI (USA) and GeneCards (Israel). IMGT reference sequences have been defined for each allele of each gene [1,2].
NUMEROTATION concept: the IMGT unique numbering. A uniform numbering system for IG and TR sequences of all species has been established to facilitate sequence comparison and cross-referencing between experiments from different laboratories whatever the antigen receptor (IG or TR), the chain type, or the species [5,6]. In the IMGT numbering, conserved amino acids from frameworks always have the same number whatever the IG or TR variable sequence, and whatever the species they come from. As examples: Cysteine 23 (in FR1), Tryptophan 41 (in FR2), Leucine 89 and Cysteine 104 (in FR3). The IMGT unique numbering represents a big step forward in the analysis of the IG and TR sequences of all vertebrate species. It has allowed (i) a standardised description of the allele polymorphisms [1,2] and of the IG somatic hypermutations, and (ii) the redefinition of the limits of the FR and CDR of the IG and TR variable domains. The FR-IMGT and CDR-IMGT lengths become in themselves crucial information which characterise variable regions belonging to a group, a subgroup and/or a gene. Moreover, it gives insight into the structural configuration of the domains and opens interesting views on the evolution of these sequences, since this numbering has been applied with success to all the sequences belonging to the V-set and C-set of the immunoglobulin superfamily [6].
OBTENTION concept. The OBTENTION concept is a set of standardised terms that specify the origins of the sequence (the 'origin concept') and the conditions in which the sequences were obtained (the 'methodology concept').

3. IMGT databases and bioinformatics tools development

3.1. IMGT databases

IMGT/LIGM-DB is a comprehensive database of IG and TR nucleotide sequences from human and other vertebrate species, with translation for fully annotated sequences, based on the IDENTIFICATION and DESCRIPTION concepts, created in 1989 by LIGM, Montpellier, France, on the Web since July 1995 [3]. In February 2003, IMGT/LIGM-DB contained 66,909 nucleotide sequences of IG and TR from 105 species. IMGT/LIGM-DB data, based on the DESCRIPTION concepts, are provided with a user-friendly interface. The Web interface allows searches according to immunogenetic specific criteria and is easy to use without any knowledge in a computing language. Selection is displayed at the top of the resulting sequences pages, so the users can check their own queries. Users have the possibility to modify their request or consult the results with a choice of nine possibilities. IMGT/LIGM-DB data are also distributed by anonymous FTP servers at CINES (ftp://ftp.cines.fr/IMGT/) and EBI (ftp://ftp.ebi.ac.uk/pub/databases/imgt/) and from many SRS (Sequence Retrieval System) sites [3]. IMGT/LIGM-DB can be searched by BLAST or FASTA on different servers (EBI, IGH, INFOBIOGEN, Institut Pasteur, etc.).
IMGT/MHC-DB, hosted at EBI, comprises a database of the human MHC (HLA) allele sequences, developed by Cancer Research and ANRI, UK, on the Web since December 1998, and a database of

MHC class II sequences from non human primates (NHP), curated by BPRC, The Netherlands, on the Web since April 2002.
IMGT/PRIMER-DB is an oligonucleotide primer database for IG and TR, developed by LIGM, Montpellier and EUROGENTEC, Belgium, on the Web since July 2002.
IMGT/GENE-DB is the IMGT genome database, based on the CLASSIFICATION concept, which allows a search per gene name, created by LIGM, on the Web since February 2003.
IMGT/3Dstructure-DB, based on the NUMEROTATION concept is a database which provides the IMGT gene and allele identification and Colliers de Perles of IG, TR, MHC and related proteins with known 3D structures, created by LIGM, on the Web since November 2001 [5]. In February 2003, IMGT/3Dstructure-DB contained 596 atomic coordinate files.

3.2. IMGT Repertoire

IMGT Repertoire is the global Web Resource in ImMunoGeneTics for the IG, TR, MHC and related proteins of the immune system of human and other vertebrates, based on the "IMGT Scientific chart" [3]. IMGT Repertoire is part, with other sections (IMGT Bloc-notes, IMGT Education, IMGT Index,…), of the IMGT Marie-Paule page which comprises 8000 HTML pages. IMGT Repertoire provides an easy-to-use interface to carefully and expertly annotated data on the genome, proteome, polymorphism and structural data of the IG, TR, MHC and related proteins. Genome data include chromosomal localizations, locus representations, locus description, gene tables, lists of genes and links between IMGT, HUGO, GDB, LocusLink and OMIM, correspondence between nomenclatures. tables of alleles, allotypes. 2D graphical representations or Colliers de Perles [3,5,6] permits rapid correlation between protein sequences represented by protein displays and alignments of alleles, and 3D data retrieved from the Protein Data Bank PDB. Colliers de Perles allow to easily compare V-LIKE and C-LIKE domains of proteins other than IG or TR.

3.3. IMGT Interactive bioinformatics tools

Bioinformatics development in IMGT includes sequence and genome analysis tools and relies on DESCRIPTION and NUMEROTATION IMGT-ONTOLOGY concepts. IMGT/V-QUEST (V-QUEry and STandardization) is integrated software for IG and TR. This tool, easy to use, analyses an input IG or TR germline or rearranged variable nucleotide sequences [3]. IMGT/V-QUEST results comprise the identification of the V, D and J genes and alleles and the nucleotide alignment by comparison with sequences from the IMGT reference directory, the delimitations of the FR-IMGT and CDR-IMGT based on the IMGT unique numbering, the protein translation of the input sequence, the identification of the JUNCTION and the V-REGION Collier de Perles.
IMGT/JunctionAnalysis is a tool, complementary to IMGT/V-QUEST, which provides a thorough analysis of the V-J and V-D-J junctions of IG and TR rearranged genes [3]. IMGT/JunctionAnalysis identifies the D-GENE and allele involved in the IGH, TRB and TRD V-D-J rearrangements by comparison with the IMGT reference directory, and delimits precisely the P, N and D regions. Results from IMGT/JunctionAnalysis are more accurate than those given by IMGT/V-QUEST regarding the D-GENE identification. Indeed, IMGT/JunctionAnalysis works on shorter sequences (JUNCTION), and with a higher constraint since the identification of the V-GENE and J-GENE and alleles is a prerequisite to perform the analysis. Several hundreds of junction sequences can be analysed simultaneously. IMGT/V-QUEST and IMGT/JunctionAnalysis results are crucial for the characterisation of the IG and TR genes and alleles expressed in normal and pathological situations.
IMGT/Allele-Align allows the comparison of two alleles highlighting the nucleotide and amino acid differences. IMGT/PhyloGene is an easy to use tool for phylogenetic analysis of IMGT standardised reference sequences particularly in developmental and comparative immunology. IMGT/GeneSearch, IMGT/GeneView and IMGT/LocusView are tools providing an interactive interface for genes and loci of human IG, TR and MHC and mouse TRA/TRD.
IMGT-ONTOLOGY is currently being written using XML (Extensible Markup Language) approach in IMGT-ML. By making data portable, XML is useful both internally for the integration of data and externally for sharing data with other information systems. Because of this data integration ability, XML has become the underpinning for Web-related computing. IMGT-ML defines XML schemas to encode data with XML tags respecting the IMGT-ONTOLOGY concepts. IMGT-ML schemas will be used for distributive data using the Web-services technology.

4. IMGT Web access

Since July 1995, IMGT has been available on the Web at http://imgt.cines.fr. IMGT provides the biologists with an easy to use and friendly interface. Since January 2000, more than 180,000 sites accessed the IMGT WWW Server at Montpellier. IMGT has an exceptional response with more than 120,000 requests a month. Two thirds of the visitors are equally distributed between the European Union and the United States. To facilitate the integration of IMGT data into applications developed by other laboratories, we have built an Application Programming Interface (API) (see "IMGT Informatics page" at http://imgt.cines.fr). This API includes: a set of URL links to access biological knowledge data (keywords, labels, functionalities, list of gene names, etc.), a set of URL links to access all data related to one given sequence.

IMGT distributes high quality data with an important incremental value added by the IMGT expert annotations, according to the rules described in the IMGT Scientific chart. Control of coherence in IMGT combines data integrity control and biological data evaluation.

The information provided by IMGT is of much value to clinicians and biological scientists in general [3]. IMGT is designed to allow a common access to all immunogenetics data, and a particular attention is given to the establishment of cross-referencing links to other databases pertinent to the users of IMGT.

IMGT interactive tools are particularly useful for the analysis of the IG and TR repertoires in physiological and pathological situations. By its easy data distribution, IMGT has important implications in medical research (repertoire analysis in autoimmune diseases, AIDS, leukemias, lymphomas, myelomas), biotechnology related to antibody engineering (phage displays, combinatorial libraries) and therapeutic approaches (grafts, immunotherapy). IMGT is freely available at http://imgt.cines.fr.

5. Citing IMGT

Users of IMGT are encouraged to cite [3] and to quote the IMGT home page URL, http://imgt.cines.fr when referring to IMGT in a publication.

6. Acknowledgements

We thank Gérard Lefranc for helpful discussion and Valérie Thouvenin-Contet for help in the preparation of the manuscript. We are deeply grateful to the IMGT team for its expertise and constant motivation and specially to our curators for their hard work and enthusiasm. IMGT is a registered mark of Centre National de la Recherche Scientifique (CNRS). IMGT is funded by the European Commission 5th PCRDT programme (QLG2-2000-01287), the CNRS, the Ministère de l'Education Nationale et de la Recherche.

7. References

[1] Lefranc M-P, and Lefranc G. *The Immunoglobulin FactsBook.* Academic Press, London, UK, ISBN:012441351X, 458 pages, 2001.

[2] Lefranc,M.-P. and Lefranc,G. *The T cell receptor FactsBook.* Academic Press, London, UK, ISBN:0124413528, 398 pages, 2001.

[3] Lefranc M-P. IMGT, the international ImMunoGeneTics database. *Nucl Acids Res* 2003: 31: 307-310.

[4] Giudicelli V, and Lefranc M-P. Ontology for Immunogenetics: IMGT-ONTOLOGY. *Bioinformatics* 1999: 12: 1047-1054.

[5] Ruiz M, and Lefranc M-P. IMGT gene identification and Colliers de Perles of human immunoglobulins with known 3D structures. Immunogenetics DOI 10.1007/s00251-001-0408-6. *Immunogenetics* 2002: 53: 857-883.

[6] Lefranc M-P, Pommié C, Ruiz M, Giudicelli V, Foulquier E, Truong L, Thouvenin-Contet V, and Lefranc G. IMGT unique numbering for immunoglobulin and T cell receptor variable domains and Ig superfamily V-like domains. *Dev Comp Immunol* 2002: 27: 55-77.

Les Grivell

E-Biosci – Access and retrieval of full text and factual information in the life sciences

Electronic Information Programme, EMBO, Postfach 1022.40, 69117 Heidelberg, Germany
Tel. +49 6221 8891503, Fax +49 6221 8891210,
Email: les.grivell@embo.org, Web-site: http://www.e-biosci.org

Driven both by the new discipline of genomics and increasing use of multi-dimensional imaging technologies, the amount of digital information in the life sciences is growing exponentially. Biologists are faced with the challenge of organising and integrating this torrent of biological data, held in a plethora of genomic sequence, sequence-related and other types of databases and scattered across many thousands of articles in the published literature.

As part of the response to this challenge, EMBO, the European Molecular Biology Organisation, took the lead to create the E-BioSci network, a next generation scientific information platform that will interlink genomic and other factual data with the life sciences research literature. The platform will offer scientists and other researchers new forms of navigation through an increasingly intricate and often confusing information landscape. Work on E-BioSci is currently carried out with financial support from the European Commission (Contract no. QLRI-CT-2001-30266). Together with partners drawn from different research institutions across Europe, EMBO aims to develop the platform into a freely available service for the research community.

Current partners in the EC-funded E-BioSci project include organisations providing bibliographic or sequence-related information at both national and international levels (see Table 1). The scope and nature of resources shared will expand as the platform develops. During the EC-funded project phase, efforts will be taken to develop and implement network communication protocols that allow new services to join and to exchange information with a minimum of changes in their existing technology.

Table 1: Partners in the EC-funded E-BioSci project (Contract no QLRI-CT-2001-30266)

Organisation	Role(s)
EMBO	Coordination and management; monitoring of E-BioSci platform functionality; helpline; hosting of EMBO-E-BioSci node, document repository and archive; maintenance of interactive E-BioSci web-pages; scientific quality control
Centre Informatique National de l'Enseignement Superieure (CINES)	French node; database and repository servers; access to indexes, databases and full text document archives maintained by partner organisations; testing of cross-database search engines
Consejo Superior de Investigaciones Científicas (CSIC)	Spanish node; database and repository servers; implementation of (Spanish) e-journals; E-BioSci link to Latin-American countries
Deutsches Institut für Medizinische Dokumentation und Information (DIMDI)	German node; database and repository servers; design and development of document location protocols
Edinburgh University Computing Services (EDINA)	UK node; development and test of cross-database search facilities; assessment of resource discovery models (in collaboration with BIOME, UK)
European Bioinformatics Institute (EMBL-EBI)	Development and hosting of E-BioSci root server; development of document location protocols and open client libraries; factual database management
Ingenta UK Ltd	Construction of E-BioSci presentation and access control layers
Institut National de l'Information Scientifique et Technique (INIST)	French node; database and repository servers; database design and construction taking into account use of different entry languages with homogeneity of datastructure and vocabulary

One of E-BioSci's main distinguishing features is that it links together a *distributed* set of resources. These consist of different types of scientific information, including journal full text and molecular, genomic and multi-dimensional image databases. Interlinkage of related information will be mainly achieved by semantic matching of conceptual fingerprints. This system was developed by Collexis b.v., who contributed to the development of E-BioSci's first prototype as sub-contractors to the EC-funded project.

This first prototype is in a developmental stage, demonstrating proof of principle, rather than providing a robust service on a large amount of content. Features so far implemented include Collexis full-text fingerprint-based search technology, gene symbol recognition in full text and database linkage (currently SWISS-PROT, GDB, HUGO, Unigene, Locuslink, OMIM) and cross-language querying (initially French in collaboration with INIST). Features to be incorporated in future releases will include cross-linkage to the EPO patent database, English - German cross-language queries (via DIMDI) and fingerprint-based image retrieval.
In contrast to other, more conventional bibliographic services (e.g. PubMed, Biosis), E-Biosci should be viewed primarily as a discovery tool that allows researchers to explore the semantic connections between the literature, different types of molecular datasets and image repositories. Besides allowing easy navigation through different information types, new visualisation tools will facilitate the analysis and integration of information.

Bioinformatics and Functional Genomics

Martin Vingron
Max-Planck-Institut für molekulare Genetik, Berlin

Functional genomics

Molbio experiments parallelized; Usually in conjunction with miniaturisation; High-throughput

Functional genomics: Examples

Study of mRNA levels:; DNA micro-arrays (red/green, Affy); SAGE; Protein detection & levels:; Protein chips; 2D gels & mass spec; Protein-DNA interaction: CHIP (on a chip); Protein-protein interaction:; Y2H; Complex isolation & mass spec; Metabolomics, by mass spec

Resulting data types

Abundance vectors, absent/present calls; Similarity matrix on genes - clusters of genes; (Interaction) graphs of proteins; Matrix of protein-DNA binding (affinities)

Computational support & Analysis

Functional annotation: GO, MIPS classification; Image analysis, processing of raw data; Statistical analysis ; Hypothesis generation

Basic scheme of analysis

Clustering of results ; Correlation of clusters to other categories; Example: DNA, microarray time series data; - cluster time courses; - map genes in cluster onto functional classification

Problems & pitfalls

Usually data are extremely noisy: Many false positives, many correct ones missing; No comparability across experiments on raw-data level: E.g. Fold change cannot be compared; No concept of biological controls: Lack of determination of source of variability; Multiple hypothesis testing problem: For any given set of genes we can invent a story; Result generation without (intellectual) consequences: Data disappear in repositories

Bioinformatics challenges

Early involvement in experiment design; Accompanying entire processing pipeline; Data analysis and statistics; Determination of processing level that allows for integration with other data; Hypothesis generation; The need for platforms and resource centres

Dr. Folker Meyer
Bioinformatics Resource Facility
Bielefeld University
Germany
11th March 2003

Bielefeld University, Germany
• 23.000 students; • special focus on **interdisciplinary** research; • Biology Department and Computer Science co-operate; 11th March 2003 c CeBiTec; Bielefeld University Folker Meyer.High volume data from genomics and post-genomics 3;
High volume data from genomics and post-genomics
• one genome; == several tens of thousands of SCF files; == for bacteria, one contig (hopefully); == yields thousands of genes; correlates to; • 1 microrarray with m hybridizations (each up to 300MB raw data); with thousands of spots over hundreds of arrays; • relates to Phenotypes, Bioinformatics, Databases, Proteomics,; Metabolomics,etc;

The Centre for Biotechnology @ Bielefeld University
Institute for Bioinformatics; Institute for Genome Research; Math; Chemistry 4 chairs; 2 junior groups; 2 chairs; 2 junior groups; Graduate School Bioinformatics and Genome Research; Bioinformatics Resource Facility; Biology; Science; Computer Biochemistry; Physics; Biophysics; CeBiTec; Nanotechnology; for; Institute; Institute; for; Structure; Determination
Bioinformatics and Genome Research at Bielefeld University
• undergraduate course in Bioinformatics since 1990; • in 1997 founded Centre for Genome Research; • in 1999 awarded German (DfG) grant to extend Bioinformatics; • founded International Graduate School founded in 1999; • 1999 awarded German (BMBF) "Competence Network"; • in 2002 integrated efforts into CeBiTec (CS and Biology); • 2003 Sun Centre of Excellence (in preparation); • close collaborations with: Chemistry, Physics, Mathematics

A common platform for Bioinformatics and Genome Research
Tools BioMoby Databases; Biology; Systems; Genome Bioinformatics; Algorithm; Development; Metabolome Transcriptome Proteome; Platform
An example for data integration using GenDB and GOPArc
Project genome data onto; a circular plot; GenDB, Meyer *et a*l, Nucl. Acid. Res. 2003; or KEGG pathways and functional; categories; GOPArc, Goesmann, *et a*l,in preparation
CeBiTec is an established service provider
• provides (bioinformatics) services in multiple projects; **–** in house (_ 12 groups); **–** FP5 Medicago (_ 15 groups); **–** FP5 ValPan (8 groups); **–** DFG MolMyk (_ 15 groups); **–** BMBF GenoMik (_ 25 groups); **–** EU FP6 NoE "Marine Genomics (400 scientists) in preparation; **–** EU FP6 IP "Grain Legumes" in preparation

A typical project and its problems
Metabolic pathways; Metabolome; Proteome; Genome; Biology; Design hybridize Microarray; experimental data; MS-Excel; table
Problems with high throughput data today
• data storage relies on unsuitable MS tools (Excel, Word, etc); • data lifetime is far to short; • data availability (emerging standards are not supported); • data is not stored safely; **–** no tape backup; **–** often reliance on single PC; **–** maintainers on part time non permanent basis; ! The result can be that data is lost for further research; Solution: **build network of resource centres**

Network(s) of resource centres
•...instead of local, small scale bioinformatics; • and instead of few large centres; • NoEs or IPs need various bioinformatics resources; • some are provided by EBI, NCBI etc. • need for resource centres to

provide; – bioinformatics expertise; – stable environments (e.g. backup); – high-throughput, high-volume facilities; – training opportunities; ! projects need custom bioinformatics solutions

Need to build public domain resources
includes; • data formats; • open standards; •; • databases; • open source platforms; •... ! identify and relate to commercial interests
Ideas for Tools enhancements
• need for one European standard **middleware;** • tools need to be enabled to work with e.g. BioMoby; • Centres need to build interfaces for their platforms; • Grant proposals need to be linked to standard compliance and; data availability; • Open Source, a potential requirement for software developed?

Infrastructure @ CeBiTec
• System Administrators, full time, permanent; • 120 CPU Cluster, overall _ 180 machines; • multi Terabyte SAN, incl. Backup

BIOINFORMATICS REQUIREMENTS FOR SWISS-PROT ANNOTATION

Anne-Lise Veuthey, Swiss Institute of Bioinformatics, Group SWISS-PROT, Geneva

SWISS-PROT (1) is a protein sequence and knowledge database that is valued for its high quality annotation, the usage of standardised nomenclature, direct links to specialised databases and minimal redundancy. Complementarily, TrEMBL strives to comprise all protein sequences that are not yet represented in SWISS-PROT, by incorporating a perpetually increasing level of mostly automated annotation.

Protein sequence annotation is the main occupation of the SWISS-PROT group at SIB. The predominant task of curators in the annotation workflow is the synthesis of information from multiple sources. Annotating an entry typically implies reading several papers, extracting data from several external databases, finding homologs by similarity searches, and running as many as ten or more sequence analysis programs. Making use of the data obtained by these parallel approaches to produce consistent annotation is far from trivial, and it is the raison d'être of the SWISS-PROT curation staff. It requires: (1) making the data consistent by using biological knowledge to evaluate the validity of each data item and assign priorities to conflicting items, and (2) combining the data to produce a complete annotation set regarding the protein. In the finalised entry, all relevant aspects of the protein are documented in a structured format.

Beside general annotation, dealing with protein sequences of any organism, main annotation projects with considerable progress have been initiated in order to keep track of overwhelming number of data provided by the numerous complete genome-sequencing initiatives. They concern human, microbial and plant protein annotation.

The Human Proteomics Initiative (HPI)

In 2001, the combined efforts of a number of sequencing centres and companies produced a first draft of the human genome sequence, which should be completed by the end of 2003. Such an endeavour is only a preliminary step in the understanding of human biological processes. The goal of the HPI project is to annotate all known human protein sequences and their mammalian orthologs, mainly mouse and rat. The HPI project places a special emphasis on actors playing a role in generating high levels of protein diversity, such as post-translational modifications (PTMs), alternative splicing, polymorphisms and disease-linked variants.

There are currently about 9200 annotated human sequences in SWISS-PROT. These entries are associated with about 23600 literature references; 23000 experimental or predicted PTM's, 2900 splice variants and 15400 polymorphisms (the majority of which are linked with disease states). On a chromosome basis, we have completed the annotation of all proteins encoded on chromosome 21, we still have some work to do to achieve the proteome annotation of chromosomes 20 and 22. A full description of the HPI project is provided at: http://www.expasy.org/sprot/hpi/.

High-quality Automated and Manual Annotation of microbial Proteomes (HAMAP)

In 1995, the first complete bacterial genome, *Haemophilus influenzae*, was published. Today, more than 100 archaeal and bacterial genomes have been sequenced and many more are under way. This represents more than 210,000 proteins, and such an influx makes manual annotation unfeasible. In order to be able to handle this huge number of proteins we have set up the HAMAP project (http://www.expasy.org/sprot/hamap/). Its aim is to automatically annotate a significant percentage of proteins coming from complete bacterial and archaeal proteomes while maintaining the same level of quality that we obtain through manual annotation. The targets of automated annotation are proteins with no similarity to other proteins (ORFans) and proteins that are members of protein (sub)families.

Annotation of ORFans. Various prediction tools are applied to proteins that show no similarity to known protein families. Possible transmembrane regions, signal sequence, coiled coils, ATP/GTP binding-sites, LPXTG motifs and some defined repeats are automatically annotated using rules of consistency and dependency, and without any further manual verification.

Annotation of members of well-characterised (sub)families. Proteins belonging to well-characterised

protein (sub)families can be annotated automatically using a rule system that describes the extent and nature of annotations that can be assigned by similarity with a prototype manually-annotated entry. Such a rule system also includes a carefully edited multiple alignment of the (sub)family, which is used both to propagate feature annotation from a model entry and to generate identification profiles. Species-specific rules and rules specific to the biochemical pathways are used to develop a system able to spot inconsistencies at the level of the entire proteome. Currently we have developed 753 (sub)family rules, each with at least one multiple sequence alignment and corresponding profile for identification of further family members. More than 36000 proteins from complete proteomes have been annotated (manually or semi-automatically) and integrated into SWISS-PROT so far. Almost 17500 SWISS-PROT entries (from complete or incomplete proteomes) belong to a HAMAP family.

Until now, five complete proteomes have been fully annotated in SWISS-PROT: *Escherichia coli*, *Buchnera aphidicola subsp. Acyrtosiphon pisum*, *Haemophilus influenzae*, *Mycoplasma genitalium* and *Mycoplasma pneumoniae*. The next species that will be completed are the model organisms Bacillus subtilis and Methanococcus jannaschii.

Since many proteins are common to both prokaryotes and plastids (chloroplast and cyanelle) genomes, we have included complete plastid proteomes in the frame of HAMAP. For the moment, 24 such genomes have been completely sequenced and they consist of about 2’500 proteins.

Plant Proteome Annotation Project (PPAP)

We have initiated the Plant Proteome Annotation Project (PPAP). Emphasis is currently given to the annotation of plant-specific protein families from Arabidopsis thaliana. The main problem for plant sequences encountered to date is the unreliability of the gene prediction programs. About one third of the predicted genes need correction when compared with newly released full-length cDNAs. We will broaden our scope to other species when additional plant genomes become available. SWISS-PROT contains currently 9581 sequence entries from plants, of which 1952 are from A. thaliana. More information on PPAP can be found at http://www.expasy.org/sprot/ppap.

SWISS-PROT database technical and format developments

In term of development activities a major effort has been concentrated on important issues sustaining database improvement, namely:

Development of the relational scheme, which will be used to store and maintain SWISS-PROT in the future. Such work not only addresses current needs but also takes into account the future development of SWISS-PROT both in terms of contents and of format. We are using the Unified Modeling Language (UML) for depicting the conceptual data model that describes the structure and constraints present in the data. Consequently, an XML public version of SWISS-PROT will be soon available.

Ongoing process to implement controlled vocabularies for the different type of lines. We are continuing a major overhaul of various comment line topics to make them computer-parsable, particularly, we are standardizing the format of the topics ‘ALTERNATIVE PRODUCTS’, ’COFACTOR’, ‘INDUCTION’, ‘PATHWAY’, ‘SIMILARITY’ and ‘SUBCELLULAR LOCATION’.

Implementation of cross-references to GO terms in SWISS-PROT entries and major update concerning structural information by completing PDB cross-references

Ongoing conversion of SWISS-PROT entries from all 'UPPER CASE' to 'MiXeD CaSe'.

Bioinformatics needs for SWISS-PROT annotation

SWISS-PROT is often considered as a reference for proteomics data. As a knowledgebase, its aim is to provide the researcher with far more than just a simple collection of protein sequences; indeed, it offers a critical view of what is known or postulated about each of these proteins. This includes information on the function, maturation and localisation of proteins, the molecular basis of diseases as well as many other important biological characteristics. In order to follow the exponential increase of incoming protein sequences and to maintain an optimal balance between the quality of annotation and the quantity of new sequences entering SWISS-PROT, the annotation process relies more and more on high quality computing tools ranging from nucleotide/amino acid sequence analysis tools to sophisticated data and text mining methods.

SWISS-PROT annotation takes great and constant care in correctness, at the level of the protein sequence, family assignment and function attribution, as well as in completeness, at the level of predicted and experimental sequence features.

We are currently evaluating DNA sequence analysis workbenches. The necessity of such software in the context of protein annotation may seem a paradox, but the lack of reliable gene prediction methods in eukaryotic genomes compel annotators to re-analyse genomic sequences so as to “recover” the correct protein sequences and to reconcile data provided by cDNA and genomic sequencing projects. This includes the identification of alternative splicing, polymorphism and sequencing errors in correlation with the genomic information.

The next step, i.e. family assignment, uses classical similarity search further supported by family/domain database resources (InterPro, PROSITE, Pfam, Prints, Smart, Blocks,…), which themselves rely on many efficient bioinformatics methods (pattern, profiles, Hidden Markov Models). The problem becomes more complex for subfamily assignment in multigenic families, like ABC transporters, helicases and bacterial short-chain dehydrogenases. The current tools have some difficulties in distinguishing between too similar families and this is essential as function assignment relies mostly on subfamily assignment. We are currently working on defining rules that allow the precise identification of a protein family, according to its multi-domain topology as well as its sequence and structural features. Around these rules, we plan to develop an expert system that will assist the annotation procedure by providing a plausible annotation scenario. For this, we rely on tools for PTM and subcellular location prediction that are developed by many European bioinformatic groups.

Moreover, the function assignment of a specific protein should be in agreement with the basic biological knowledge already known and regarding the organism. In the framework of the HAMAP project for microbial annotation, we are developing a rule-based expert system that controls the coherence of annotation at genome level, by using a representation of metabolic pathways described by directed acyclic graphs. The system should be able to warn the curators about missing proteins, ortholog and paralog problems, proteins belonging to a pathway which supposedly does not exist in an organism, etc...

The main source of information for a SWISS-PROT entry is provided by the biomedical literature. Querying and reading papers is the most time-consuming part of the annotation process and the development of text mining tools could be helpful to speed up this step and even enhance the coverage of information. We are currently working on a tool that retrieves the most informative abstracts from PubMed for a specific protein and then extracts informative sentences directly from these abstracts in accordance with the main annotation topics. We are also working on the database updating process, by providing an automatic procedure to keep track of new data in the literature with respect to a given protein. This text mining project will be achieved in the framework of BioMinT, an FP5 European project.

Conclusion: SWISS-PROT is widely used in functional genomics to infer information, automatically, on gene function. Developers of sequence analysis programs also use it as a source of data to train and evaluate their predictors. Thus, the correctness and completeness of SWISS-PROT annotation is an essential issue not only for the biomedical community but also for bioinformatics developments. We are trying to enhance both the quality and coverage of the database via complete genome oriented projects that intend to integrate the information provided by sequence analysis at the level of nucleotide and amino acids with the experimental results described in the literature. The biological expertise of curators still plays an essential role. Even if automated annotation methods are developed, the final human judgement will guarantee the correctness of the information and will be used to improve the expert systems continually.

SUMMARY OF PRESENTATIONS BY SESSION CHAIRPERSON

Chair: H. Werner Mewes

THE FOUNDATIONS - Current and already planned activities in Bioinformatics

Unifying theme: state of the art review of running EC funded projects that have a bioinformatics component to understand the extension and role of this component in the given projects (to what extent these projects involve bioinformatics tools development, application of existing tools, or elementary data processing, what are the centres of gravity, how do they relate the experimental outcomes to the data processing facilities, links to genomics and related research.

1. The topics are in fact highly overlapping and divisions are artificial
2. The immunodatabase is a good example of a specialised world-class database. A problem is that maintenance does not fit with funding models. The funding does not fit on a national level. However, EU does not fund infrastructure in a useful way.
3. The use of literature references is a key part of what we do. We need better access tools.
4. A key point is the interaction between bioinformatics and biology researchers. Biology produces key challenges for us. Biology is well funded because individuals can make progress.
5. In bioinformatics, we have a hierarchy of interactions, starting with nearby researchers
6. The next level is developing common solutions to common questions. Most successful tools are often those taken for granted.
7. At the national level, funding agencies try to bring groups together. A role for the Commission is to connect nodes
8. At the top of the hierarchy, we need a solution for the top resources such as the European Bioinformatics Institute (EBI).
9. We have scattered activities, and goal is to increase efficiency by combining resources. The problem is we lack fuel, so we don’t have enough to combine.
10. Concerning contributions, Europe does succeed very well with the resources it has. The Sanger Institute is an excellent example of what can be done.
11. Europe had been leading in many fields, but we lost lead due to lack of money.

FUTURE INFRASTRUCTURE AND ANALYSIS REQUIREMENTS
(13:30 - Chair - Janet Thornton)

Michael Nilges, Unité de Bio-Informatique Structurale, Institut Pasteur, Paris, France

1. An attempt to define the term bioinformatics

Bioinformatics derives knowledge by computer analysis of biological data. It is a rapidly growing branch of biology, highly interdisciplinary, and uses techniques and concepts from informatics, statistics, mathematics, chemistry, biochemistry, physics, linguistics and other fields. The biological data can be the information stored in the genetic code, experimental results from various sources, three-dimensional structures, expression arrays, patient statistics, scientific literature etc. An important part of research in bioinformatics is methods development for storage, retrieval, and analysis of these data.

I attach some importance to a wider, general definition of bioinformatics. Concepts in our work, for example, do not come so much from informatics but primarily from physics and probability theory. The concept of "information in the genetic code" has its limitations - DNA can be regarded as a text to be analysed, or as a molecule that makes interactions with a variety of other molecules. Interactions with proteins are governed by the three-dimensional structure and the dynamics (and flexibility) of the DNA. These in turn are obviously determined by the sequence of bases, but it is unnecessarily limiting to reduce the analysis to a one-dimensional level. This applies to a much larger extent to protein sequences and protein structure.

There are some key areas where "physical insight" is useful:

- prediction of genes and coding regions, using predictions of local temperature stability of double-helical DNA
- prediction of RNA secondary structure
- one of the oldest unsolved problems in computational biology, prediction of protein 3D structure. Some progress has been made in recent years; a very small protein was folded ab initio with a molecular dynamics program.

2. Bioinformatics at the Pasteur Institute

Bioinformatics has a long tradition at the Institute. The institute curates several organism-specific databases (Colibri, SubtiList, TubercuList), has developed a common web interface to many commonly used sequence analysis programs, is a partner in the development of sequence analysis software,...

The current needs and activities in bioinformatics at the institutes include

- Genome annotation (micro-organisms, anopheles)
- Maintainance of organism-specific databases (Subtilist, Colibri, Tuberculist, Leproma, PyloriGene, MypuList, ListiList, CandidaDB, SagaList, CyanoList; Anopheles)
- Other data bases (ABCISSE, CABRI)
- Web services (http://bioweb.pasteur.fr)
- Micro-array analysis
- Image analysis (microscopy)
- Functional genomics/ proteomics
- Structural genomics: target selection; structure analysis
- Structure prediction, homology modelling
- Molecular dynamics
- Small molecule docking, virtual screening

3. Structural bioinformatics at the Pasteur institute

Our projects are centred on structural bioinformatics and include algorithmic developments for NMR structure refinement, structure prediction (homology modelling), function analysis (long time molecular dynamics calculations and prediction of "other" biologically relevant structures of a molecule, protein-protein interactions and protein-ligand interactions).

The developments in algorithms for NMR structure calculations are partially funded by two EU-funded projects in FP5: SPINE, and NMRQUAL. The projects and our contributions address some of

the principal problems of NMR structure determination:
- automation to significantly speed up NMR structure determination; this has clear implications for NMR involvement in structural genomics;
- standardisation to reduce the heterogeneity of procedures used to date and to increase the reproducibility of structure determinations;
- development of a measure of quality for NMR structures - a figure of merit; this is a prerequisite for a meaningful use of NMR structures e.g. in modelling;
- data harvesting and data bases; the software has to automatically collect all relevant data for, and prepare a submission of a structure to a data base.

We develop a software package (ARIA) to address several of these issues and pursue in parallel a new development where the data analysis and structure refinement are treated together using Bayesian probability theory. Already the principal idea behind ARIA is the integration of several separate steps in NMR structure calculation: selection of data, analysis (assignment) of data, and structure calculation.

Homology modelling is the most successful protein 3D structure prediction technique. Part of the structural genomics effort goes into covering the protein structure space sufficiently well that "all" proteins can be modelled with sufficient accuracy. Better homology modelling will reduce the number of structures that need to be solved experimentally. The quality of homology models is primarily determined by the accuracy of the alignment between the sequences of the query protein and the template structure, and the force field that is used in the modelling. For low to very low sequence similarity, homology models can only give qualitative features of the protein. A possible solution for improving the quality of homology models may lie in an integrated approach of sequence alignment, secondary structure prediction, and the modelling itself, in combination with advances in the modelling algorithms and force fields.

Function prediction from the structure alone remains an elusive goal. One way is to search for conservation of structural templates. We study aspects of function by looking at physical properties that can be calculated from the structure, such as shape, electrostatic potential, and (long term) molecular dynamics.

An important part of structural bioinformatics, in particular at a medically oriented research institute, is the study of the interaction between a protein and ligands. We are building up drug design and virtual screening activities.

Edouard Yeramian, who joined the group recently, has developed a gene prediction algorithm that relies on the prediction of DNA structure: he has observed a correlation between local temperature stability of the double helix and coding regions. In the plasmodium genome, this algorithm has detected genes missed by all other algorithms.

4 Funding bioinformatics in the future

4.1 Algorithmic developments, with the emphasis of integration of several methods/ tasks.

Further development of algorithms is necessary to address the many unsolved problems today. Some of the problems have been around for decades and will not disappear in the next two years. The new "big" instruments may not be optimal for the funding of algorithmic developments: it seems difficult for the small laboratories contributing much (or as it was said the most) of the algorithms to enter big networks. Networking on the other hand is necessary to foster integrative approaches and prevent the development of many solutions to the same problem.

4.2 Development and maintenance of data bases

They are on the one hand the basis of much of our work and the work of the researchers in the laboratories. As it came up often in the discussions, providing funding for the initial development of a data base but not for the curating and maintenance leads to a waste of resources.

4.3 Systems biology

One or very few projects in systems biology should be funded well.

Molecular biology needs bioinformatics

Martin John Bishop
UK HGMP Resource Centre, Hinxton, Cambridge CB10 1 SB, UK
mbishop@hgmp.mrc.ac.uk http://www.hgmp.mrc.ac.uk

Biological data - molecules; Sequences; Structures; Gene expression; Proteomes; Pathways - signalling; Evolution; Computer analysis – methods; Comparison; Modelling; Co-regulation; Mass spectrometry; Knowledge bases; Phylogenetics

Molecular biology is about information

Central dogma; DNA; <-> RNA; -> protein; -> phenotype; <- DNA; Molecules; Processes; Central paradigm; Genome repository; <-> RNA world; -> Protein sequence; -> Protein structure; -> Protein function; -> Phenotype; <- Fed back to genome; Information processing

Comparative method

Life is more uniform at the molecular level than we might have imagined; Genes from bacteria can be informative in human biology; Model organisms are extremely informative; yeast, nematode, fruit fly; rice, maize, thale cress (a brassica, *Arabidopsis thaliana*); puffer fish, zebra fish, chicken, mouse, rat

Genes and proteins

The number of human genes is about 35,000; Alternative splicing; Protein cleavage; Post-translational modification (glycosylation, phosphorylation); Estimate there may be up to 350,000 human proteins

HGMP Resource Centre

Research Division; Functional Genomics Group; Comparative Genomics Group; Gene Expression Group; Post-transcriptional Regulation Group; Neurological Disorders Group (Integrated transcriptomics and proteomics); Bioinformatics Division; MRC geneservice (Babraham)

HGMP Resource Centre

Research Division; Bioinformatics Division; MRC geneservice (Babraham); DNA Services (Genotyping,SNPs); Reagents (BACs,PACs,ESTs,CDNAs,RNAi); Media Services; RNA Expression; Cloning Services

HGMP Resource Centre

Research Division; Bioinformatics Division; Applications Group; Protein Group; Development Group; Systems Group; MRC geneservice (Babraham)

Bioinformatics research

Bioinformatics research to underpin biological research objectives needs to be expanded to meet the post genome challenge across a wide range of activities; sequencing, mutation detection and SNPs, functional genomics, comparative genomics, gene expression and microarrays, proteomics including 2-D gels, mass spectrometry, protein interactions, in situ localisation

Phenotype and genotype

Genetic linkage mapping requires at least two variants of a characteristic; The only way to map morphological characters; Genetic Linkage User Environment (GLUE) is provided by HGMP

Man and mouse

Human and mouse dysmorphologies can be related; Mapping is possible in either species followed by gene identification for defect; Large mutation programs exist for mouse; DHMHD - dysmorphic human mouse homology database catalogues results

Genotyping and SNPs

Linkage studies have been successful using microsatellite genome screening; Automated SNP methodology evaluated; Complex disease studies aided by pyrosequencing for pooled material; Bioinformatics aspects have been evaluated

Functional genomics and proteomics

A very wide range of experimental methodologies are available in house; New techniques are being developed; Bioinformatics underpins all these activities

Microarray collaborations

We are collaborating on gene expression studies; We offer the necessary arrays for experiments; We are offering related training and analysis services

EMAGE database

The Edinburgh Mouse Atlas Gene Expression (EMAGE) is a community database - Richard Baldock

and Duncan Davidson; Mouse embryonic anatomy; Patterns of gene expression mapped to anatomy; GRID demonstration project

Protein structure and function

A variety of experimental techniques are in use; Library of discriminating elements developed from SCOP; Work on prediction of ligand-binding in progress

Protein interaction networks

Developing new high throughput methodology (Y2H); Needs development of new bioinformatics tools ; Needs expertise in pathways databases

Training

LINKAGE and SNPs; EMAGE course ; MICROARRAY course being developed; GENE PREDICTION; PROTEIN STRUCTURE; PROTEIN FUNCTION AND INTERACTIONS;

GENOME WEB

Up to date; Relevant; Fully searchable; Fully verified; Extensive

INTEGRATED ANALYSIS

BLAST; NIX; PIX; GLUE; PIE; MAGI; PINT

APPLICATIONS GROUP

Gary Williams; Lisa Mullen; Frank Dudbridge; Tim Carver; Linkage Analysis; Radiation Hybrid Mapping; Sequence Ready Clone Maps; Genome Databases; Polymorphisms; Sequence Analysis; Gene Prediction; Expression Profiling; Phylogenetic Analysis; Integrated Tools - GLUE, RHYME, NIX, PIE;

PROTEIN GROUP

Alan Bleasby; Jon Ison; Claude Beezley; Hugh Morgan; Damien Counsell; Protein Sequence Analysis; Protein Structure Analysis; Protein Structural Modelling; Proteome Databases; Tools for Peptide Sequence Determination; Protein Cellular Localisation; Protein Functional Studies; Pathways and Protein Interactions; Integrated tools and databases - PIX

DEVELOPMENT GROUP

Phil Gardner; Naran Hirani; Yagnesh Umrania; Tony Brookes; Jayne Vallance; Lee Cave-Berry; Microarray Project; Annotation - AnnDB; Limas; Fugu genome finishing; RDBMS skills

EMBOSS and JEMBOSS development

A comprehensive analysis package is being developed; There is much work remaining to be done; JEMBOSS is the graphical user interface that also needs further development

NETWORK / JANET SERVICE

HINXTON CAMPUS; LONDON; 34 Mbps link for backup; ; Geoff Gibbs; Peter Tribble; Terry Stewart; Steve Gamble; ; CAMBRIDGE ; Gigabit Ethernet

SERVERS

More than 100 servers; 1, 4 and 8 cpu SMP; Sparc and Intel; Solaris and Linux; Databases doubling every 14 months

HELPING THE USER

Information discovery – completeness; Communication – multiple sites; Ontology – uniformity?; Software integration – ease of use; Reasoning about results; Monitoring – repeat queries

NEW TECHNOLOGIES

Web services; GRID; Object-orientated computing; Multi-agent systems

The myth of function

The holy grail of molecular biology was the idea the sequence can predict structure and that structure can predict function. We need to identify all the parts (molecules). The reality is that we need to determine sequence, structure and function and try to relate them in order to gain understanding. We need to build all the parts into models of working systems.

Christopher Cooper – IBM

The Role of Information Technologies (IT) in Bioinformatics

Introduction

There are some fundamental challenges facing the IT industry in the role they are required to play in providing infrastructure solutions and services in Bioinformatics. This discussion aims to outline some of the foci taken by the IBM Life Sciences (LS) organisation to help address the IT needs of the industry.

These play an important part in dealing with the rapid advances in biotechnology that are changing the face of drug discovery and development. The fast data access and accumulation demands of the emerging biotechnology industries require high performance computing data integration, and storage systems that can provide advanced scalable solutions from single query data access to collaborative Web research, encompassing discovery methods and pattern matching.

How and Why an IT organisation would invest in the Life Sciences

One of the main features of the industry is the drive towards open standards and platforms. Taking some of these key initiatives, IBM has invested in excess of $1B in the Linux operating system, and in particular in Bioinformatics, the $100M investments in the Life Sciences organisation to further develop the particular needs of IT in this industry. The IBM commitment to the life sciences industry is defined by the establishment of the specialized IBM Life Sciences Solutions Group dedicated to rapidly bringing leading-edge technology out of the laboratory and into the marketplace for customers and Business Partners in the fields of pharmaceutical research, biotechnology, genomics, proteomics, health and other life sciences. In addition, a dedicated IBM Global Life Sciences Consulting and Solutions practice has been established to focus IBM service capabilities and expertise, as well as intellectual capital on helping customers and Business Partners to migrate their R&D units into even more efficient and competitive operations. IBM Research continues to pursue strategic exploration of technologies applicable to life sciences research and product development, including Pattern discovery and matching, functional and structural genomics, proteomics, Visualization, technical knowledge management and nanotechnology. The IBM Computational Biology Center and the IBM Deep Computing Institute are two key research centres housing teams of scientists working on projects involving computational biology, chemistry and material science. The long-term projects at the IBM Computational Biology Center foster IBM collaboration with life sciences companies to bring scientific expertise directly into the development of life sciences solutions.

Detailed studies are in progress aimed at providing new clues for medical diagnostics, the synthesis and design of novel materials and the analysis of genes and their relationships.

The IBM Deep Computing Institute provides business decision-making capabilities to analyze and develop solutions for complex and difficult problems. Combining these capabilities with advances in algorithms, analytic methods, modelling and simulation, data management and software infrastructures enables valuable scientific, engineering and business solutions. Joining life sciences research with information technology, IBM is uniquely positioned to help the industry solve the challenges of R&D information management. IBM is actively pursuing strategic business partnerships with life sciences companies whose complementary skills, knowledge and resources can help build value-rich solutions for the industry. There are extraordinary challenges and opportunities ahead for the companies engaged in the life sciences industry. The scientific challenges in this emerging industry are matched equally by the challenges associated with managing data integration and developing the computing technology and tools needed to provide solutions for the laboratory.

Combined with IBM's core strength in providing robust technology and global services, the IBM Life Sciences business unit delivers innovative, scalable, infrastructure solutions that are unique within the industry. IBM Life Science Solutions provides the IT infrastructure that researchers in biotechnology, pharmaceutical research, genomics, proteomics, and healthcare need to turn data into scientific discovery and new treatments for disease. The solutions segments are: Data and knowledge management, high performance information infrastructure, clinical development, devices and diagnostics, and medical imaging solutions.

The Way Ahead - Facing the challenges of IT in Bioinformatics

The primary challenges of the Bioinformatics industry faced by IT are nothing new. They include:

- Open standards and platforms (as previously outlined)
- Security – enabling industry players to take part in accessing public databases from their own sites without the threat of risking unwanted access to their own intellectual property
- Federation of Data – providing middleware tools/methodologies which enable access and data mining of multiple heterogeneous data sets stored in a variety of locations
- Integrity of Data – ensuring captured and shared data is of an acceptable standard
- Connectivity – as the volume of data increases exponentially and more and more groups/individuals require access to the data, technologies like grid may be suitable methods to be deployed, but network bandwidths and latency may become a bottleneck

There are numerous publications referencing many of these challenges; the following is an example:
C.A. Goble, R. Stevens, G. Ng, S. Bechhofer, N.W. Paton, P.G. Baker, M. Peim, and A. Brass. **Transparent Access to Multiple Bioinformatics Information Sources.** IBM Systems Journal, 40(2):532 - 552, 2001.

Taking a `SWOT' analysis approach to these challenges, the following could be defined:

- Strength – the volume of data generated
- Weakness – the current lack of common standards for data access/federation
- Opportunity – the technologies are much more robust than they have been
- Threat – key influencers/players do not act quick enough to define standards of commonality to harness and manage the data, determine the information in the data collected and generate the knowledge from the information

In Summary: so what have we learned so far about Life Sciences Data Management and Integration (particularly in Bioinformatics)?

1. Performance (high bandwidth and low latency) and scalability (large data volumes and large numbers of objects) are top priority
2. To meet stringent performance and scalability requirements, users will be continuingly using and developing ad hoc data management strategies
3. Users want the power of relational databases (search and retrieval) along with the high performance of file-systems (to move data at high bandwidth and low latency)
4. As users begin to perform work in 'virtual' organizations, they are finding it important to share data across organizational boundaries, federating independent instances of their data repositories
5. Data volumes are growing fast, and there is an increasing need for long-term data retention. Users value performance and cost-effectiveness in addressing these problems.
6. Data integration is the key to successful deciphering meaning

Pathbase: A Database of Mutant Mouse Pathology

Paul N.Schofield and the Patbase Consortium
Department *of Anatomy, University of Cambridge, Downing Street, Cambridge, CB2 3DY, UK.*
ps@mole.bio.cam.ac.uk

Introduction

Extant web pathology resources focus on specific domains such as tumour pathology (http://www.ita.fhg.de/reni/index.htm or http://tumor.informatics.jax.org (Bult et al., 2001; Naf et al., 2002); see http://www.ncifcrf.gov/vetpath for a collection of useful links) but, whilst these are often invaluable, there is currently no general web-based reference resource that either covers the field of mouse pathology as a whole, or provides a means for standardising the nomenclature used to categorise the pathological lesions of mutant mice. Furthermore, existing databases of mutant mice such as TBase (http://tbase.jax.org/) use text rather than pictures to describe abnormal phenotypes.

This paper describes Pathbase, a new web-accessible database that attempts to fill these gaps. It includes a reference site with an ontology (see below) of mouse pathology (MPATH). This ontology has been produced by pathologists and is designed both to act as a reference facility and to catalogue histopathologic images of mutant mice. The database also includes a repository for such images and users are invited to add images for the benefit of the community. Here, we describe the facilities available as part of the Pathbase project together with the informatics infrastructure of the database.

The Database

Both Pathbase and EMpathy are held in Sybase, a standard relational database, and incorporate a common pathology nomenclature. This includes ontologies of mouse developmental and adult anatomy (EMAGE and MA) and pathology (MPATH) together with a set of short controlled vocabularies (CVs), that are used for both archiving data and querying the database (Figure 1). The CVs are essentially lists covering limited domains of knowledge and can be considered as hierarchies one level deep. As such, they are the simplest example of an ontology which is most simply described as a complete and formalised description of a domain of knowledge that can be interpreted by a computer (Stevens et al., 2000). The anatomical attributes of the image are coded by using either he time dependent embryo anatomy developed by the EMAGE consortium (http://genex.hgu.mrc.ac.uk) (Bard and Winter, 2001; Bard et al., 1998) for images of the embryonic menifestation of developmental lesions, orthe mouse adult anatomy (MA) developed by the gene expression database at the Jackson Laboratory. The standard anatomical nomenclature (International Committee on Veterinary Gross Anatomical Nomenclature***;*** 1983) whilst highly detailed, is too large to implement for database searching.

MPATH is a new ontology that currently contains 457 classes of mouse pathology arranged as a Directed Acyclic Graph (DAG) that extends to a depth of 6 levels. Each class can be viewed as a leaf attached to a higher-level node by being “an instance of” that higher level. Each item in the hierarchy has an MPATH ID that can be used for database interoperability and the ontology itself is accessible both from Pathbase and from the Gene Ontology Consortium’s GOBO site (Ashburner et al., 2000; The Gene Ontology Consortium, 2001) where bio-ontologies are archived. The MPATH ontology covers all currently known categories of mouse pathology, and, as it is actively curated, can be expanded in the light of future knowledge. It has recently been retstructured to take account of the NIH Mouse models of human cancer consortium recommendations on haematopoietic neoplasms and will be maintained in accordance with reviews form this body as they are published.

Informatics and software

Pathbase is currently running on an Apache web server (1.3.x) (www.apache.org) with a Sybase database (www.sybase.com) as a backend. PHP (www.php.net) creates the web pages dynamically and serves the data from the database. Such a system, which uses Gentoo Linux as the underlying operating system, is fast, stable and secure. To ensure that the system response time is fast, submitted images are automatically converted to JPGs and thumbnails, with the former being stored outside the database. As no searching is done on the actual images, Sybase is kept relatively simple and this allows the response time of Pathbase to be rapid under heavy user load. The ontologies themselves are

stored as flat files and converted to the required formats and hierarchies on the fly through PHP scripts. The same ontologies and CVs are used for searching and for archiving data. Searching the database can most simply be done through the ontologies and controlled vocabularies. However, as there is free text associated with the annotations of the images, there is also a string search facility that allows a user to search the database through user-produced keywords.

Databases

Pathbase and EMpathy share a range of common features: both use the same controlled vocabularies and ontologies and are directly linked to PubMed, the literature database, and to TBase, the database of transgenic mice. They are however separate entities.

Pathbase

The core part of this database is a set of images that can be searched via the controlled vocabularies). These images are mainly from mutant mice including transgenics, knockouts, chemical- and radiation-induced and spontaneous mutations. In many cases these records will carry a Tbase number, linking the two resources. We aim in the future to include representative data from strain specific background lesions. Data acquisition proceeds in two modes. Curators will request images as papers appear in the literature and will begin a retrospective data acquisition operation within the next year. Alternatively users may send or upload their own images via user-friendly interfaces, ftp transfer, email or portable media such as CDs. Images can be uploaded in all common formats with an optimal resolution of at least 300 pixels per inch and a maximum file size of 8MB. The Pathbase consortium is also able to accept transparencies and slides which will be scanned and returned. Images in the database may be downloaded as JPG files, but TIFFs will only be available on request due to the large uncompressed file size.

Sending images to the databases requires that the submitter also send key metadata and the entry interface page for this has been made as simple as possible through the use of the ontologies and controlled vocabularies that merely require the user to click on an appropriate term. Searching for images uses the same ontologies and controlled vocabularies and is again designed to be simple.
Images are linked to other core web resources. Those carrying a Tbase number are directly linked to the data file for this mutation in Tbase; images are linked via their tissue names to GXD, the mouse gene expression database (Ringwald et al., 2001), while the literature associated with the record can be accessed via PubMed.

Pathbase also allows users to annotate image records directly. Although moderated by the curators, such additions to the information on the records will build up a community expertise resource not so far attempted in a database of this kind.

EM*pathy*

The purpose of this database is to provide a reference and teaching resource for mouse pathology. The key feature of the reference component is a formal ontology of known pathological lesions that is linked to a set of reference images (as yet incomplete since some of the disorders are very rare). The ontology itself can be downloaded from the Pathbase or the GOBO site (see above) and the pathology IDs used for cross-computer queries (interoperability). EMpathy also holds other useful resources such as an illustrated guide to conducting a mouse necropsy (which can also be downloaded as a pdf file).
EMpathy is organised by organ and tissue, and shows users the abnormalities associated with each site. Each pathology record consists of a set of reference images and notes on the aetiology, diagnostic criteria and, where appropriate, reference to the background incidence of the disorder, and examples of the occurrence in genetically modified mice. There are currently about 200 such pathological lesions listed and we intend to double the number over the next year.

Discussion

The first stage in the production of biology databases covered the archiving of sequence data in its many forms. The next stage was the production of databases that unified particular groups by providing them with a community resource (e.g. Flybase). The current need is for databases that carry image data that will provide visual as well as textual data and that is ideally linked to formal knowledge systems; Pathbase is one of a group of such databases that is now available. In terms of informatics, its one notable feature is its use of formal ontologies for searching, and considerable effort has gone into making the pathology ontology (MPATH) as comprehensive and authoritative as possible. It is important to stress that the Pathbase ontology is not intended as an alternative diagnostic framework to the existing nomenclatures such as SNOVET (Palotay, 1983), but as a tool for data retrieval providing a set of inclusive terms into which all pathological lesions can be fitted. SNOVET now contains over 100,000 terms and is not appropriately structured to describe the pathology of transgenic mice as many of its terms are dependent on aetiology and anatomical location. Whilst the 'lumping' of pathological terms in the MPATH ontology does remove some diagnostic precision, this has the advantage that small differences in opinion as to the precise diagnosis, or usage of terms, which vary between different traditions of pathology, do not affect the accuracy of a search. Such precise diagnostic terms can still be entered into the free text field and searched independently.

This ontology underpins EMpathy, the reference and teaching site constructed by the curation team, and we hope that users with a limited background in histopathology will find its many facilities useful, even though the archive of images associated with the pathological descriptions is still incomplete. We also hope that users who feel that the site requires other facilities or that it could be more helpful in particular ways will provide feedback to the curator.

Pathbase itself is still only a facility. While it holds many images and their associated information, data acquisition is an open-ended process (unlike an ontology which is intended to encapsulate a well-defined domain of knowledge). Its success will depend on users being prepared to use the facility to store their own data and thus on a certain degree of altruism. To help here, the curators have tried to make it as easy as possible for users to upload or send images and hope that users who are publishing interesting pictures of mutation associated abnormalities in both embryonic and adult rodents will also submit them to the database.

Our intention is that Pathbase will remain actively curated. Pictures that are submitted by users will be rapidly made available to the field. New images and additions to the literature found by the curators will be entered into the database and, if appropriate, added to the reference site. We hope, in particular, to be able to complete the set of pathology reference images and invite anyone with high quality definitive images of specific pathologies to share them with the community.

Pathbase is intended to have two uses. The first is to help biologists with a limited knowledge of mouse pathology to assess and compare their data, and to provide access to more expert advice. The second is to allow the mouse community to share their knowledge of image-based pathology. The first of these aims should be met by the curation team, the latter requires an involvement by the user community. We hope that they will take advantage of this opportunity.

Issues and lessons arising from the development of Pathbase

Pathbase is the first open access resource of its kind in the world so while many of the issues we have faced have not been previously addressed, most will be met with by other biological databases.

- Databases must maintain a global profile and user constituency to be useful and legitimate. Europe may lead, but we have to be inclusive in our interactions with other databases and projects world-wide to fulfil our objectives.
- There are issues of nomenclature, definition and standardisation particular to pathology, which have not previously been addressed.
- There are problems with getting scientists to submit unsolicited data; individual authors need to be assured that the database is stable for at least the medium term, expertly curated, and will be regarded as an authoritative resource by other scientists. The issue of establishing trust and legitimacy is much more important than we initially imagined.

- It is important to establish multiple, diverse data pipelines.
- Requirements for interoperability with other databases – query systems, data migration, common data structures and ontologies. These need consensus of the community and careful development.
- There is a requirement for intensive expert curation – the number of curators is a potential bottleneck in populating the database.
- The establishment and operation of an expert review panel is vital for this kind of database.
- Biological expertise in mouse pathology is scant and dispersed across Europe and the rest of the world leading to a lack of experts and frequently published mis-diagnoses in the more molecular literature, eg. Cell.
- Realisation of the seriousness of the requirement for pathology training and reference resources for mouse pathology.
- We have been surprised at the reluctance of pathologists to use the full features of databases or informatics tools, requiring that we pay special attention to the design of any GUI.
- Importance of developing bioinformatics tools based on ontologies and other description logics to use the pathology data to generate new hypotheses about gene function in disease.
- The importance of developing a disease 'ontology' or description logic based system for mice in the near future.
- Realisation of the need for large amounts of storage given the size of the data objects (images) stored.

References

Ashburner, M., Ball, C.A., Blake, J.A., Botstein, D., Butler, H., Cherry, J.M., Davis, A.P., Dolinski, K., Dwight, S.S., Eppig, J.T., Harris, M.A., Hill, D.P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J.C., Richardson, J.E., Ringwald, M., Rubin, G.M. and Sherlock, G. (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet*, **25**, 25-29.

Bard, J. and Winter, R. (2001) Ontologies of developmental anatomy: their current and future roles. *Brief Bioinform*, **2**, 289-299.

Bard, J.L., Kaufman, M.H., Dubreuil, C., Brune, R.M., Burger, A., Baldock, R.A. and Davidson, D.R. (1998) An internet-accessible database of mouse developmental anatomy based on a systematic nomenclature. *Mech Dev*, **74**, 111-120.

Bult, C.J., Krupke, D.M., Naf, D., Sundberg, J.P. and Eppig, J.T. (2001) Web-based access to mouse models of human cancers: the Mouse Tumor Biology (MTB) Database. *Nucleic Acids Res*, **29**, 95-97.

Consortium, T.G.O. (2001) Creating the gene ontology resource: design and implementation. *Genome Res*, **11**, 1425-1433.

International Committee on Veterinary Gross Anatomical Nomenclature (1983) *Nomina anatomica veterinaria* Ithaca, N.Y.

Naf, D., Krupke, D.M., Sundberg, J.P., Eppig, J.T. and Bult, C.J. (2002) The Mouse Tumor Biology Database: a public resource for cancer genetics and pathology of the mouse. *Cancer Res*, **62**, 1235-1240.

Palotay, J.L. (1983) SNOMED-SNOVET: an information system for comparative medicine. *Med Inform (Lond)*, **8**, 17-21.

Ringwald, M., Eppig, J.T., Begley, D.A., Corradi, J.P., McCright, I.J., Hayamizu, T.F., Hill, D.P., Kadin, J.A. and Richardson, J.E. (2001) The Mouse Gene Expression Database (GXD). *Nucleic Acids Res*, **29**, 98-101.

Stevens, R., Goble, C.A. and Bechhofer, S. (2000) Ontology-based knowledge representation for bioinformatics. *Brief Bioinform*, **1**, 398-414.

Søren Brunak
Center for Biological Sequence Analysis, Biocentrum-DTU, Technical University of Denmark,
Building 208, DK-2800 LYNGBY

Research Driven Infrastructure

In all areas of biological and medical research, the role of the computer has been dramatically enhanced in the last ten year period. In the next ten years this development is likely to continue – leading to an even more dramatic change in the way biological research is carried out. While the first wave of computational analysis focused on sequence analysis, where many highly important unsolved problems still remain, the current and future needs will in particular concern sophisticated, large scale integration of extremely diverse sets of data: gene sequences and their control regions, information on genetic variation, gene expression profiles, protein-protein interaction networks, temporal knowledge on protein complex formation, and signaling cascades, to mention but a few.

In terms of infrastructure Europe is not well prepared for this situation. In terms of the more traditional bioinformatics Europe has also not established an infrastructure that allows for optimal exploitation of the substantial resources invested in biomedical and biotechnological research. The bioinformatics infrastructure in Europe is characterized by duplication of many of the older approaches, a highly non-uniform quality of the options in relation to novel techniques in the different countries, despite the fact that most countries – as a consequence of the diversity of the academic and industrial research – need largely the same opportunities for researchers in the public and private sectors.

The need for a coordinated, bioinformatics infrastructure in Europe is significant. The number of experimentalists who contact skilled bioinformaticians is not decreasing, and it is clear that with new, powerful methods appearing at great pace, experimentalists are less and less able to carry out analyses of their data on their own using state-of-the art methods. Part of the reason is of course also that novel types of data are produced by high-throughput techniques all the time, but this is also likely to be the case for the next couple of decades. Experimentalists at universities and hospitals know how to obtain funds for an Affymetrix scanner, or Ciphergen protein chip equipment – but not, on their own, how to analyse the data, and in particular how to integrate these data with other data sources. In the past 30 years a hit in the database has largely meant detecting sequence or structural similarity, in the future hits will involve extremely complex combinations of data structures, and dynamical pattern matching across massive amounts of data.

Europe suffers in general from fragmentation of public research effort, and from a low level of interregional co-operation, among companies and public research institutions from different regions of several countries. For many medical and biotechnological research areas bioinformatics is the one of the rate limiting steps, the fragmentation within bioinformatics has therefore far reaching consequences. Europe is getting much less out of the investments made already in experimental and clinical research than it otherwise would with a better coordinated bioinformatics effort.

As an example of this situation one can take a look at the EMBnet nodes, which were established for data distribution before the advent of the internet, and in particular before it came into common use within life science research. In many European countries this network of nodes is by national funding agencies and research ministries viewed as a key component in the infrastructure for bioinformatics and computational biology. The quality of the services varies:

1. Some nodes are non-existent or down e.g. Greece, Italy (temp. ?)
2. Some nodes are just dead e.g. Sweden, Norway (good bad examples)
3. Some nodes are active, but do rather old-fashioned stuff e.g. Denmark, …

4. Some nodes are good e.g. UK, Belgium, Israel, Holland, …

The non-existent or dead nodes do presumably little harm as researchers in these countries will not be misled, and will search elsewhere for methods for data analysis. The biggest problem remains with the large group of intermediate quality nodes, which offer many non state-of-the-art methods to their users. Experimentalists often have quite unrealistic expectations of what bioinformatics can do and not do for their particular problem. Even if the nodes were primarily offering high quality service, the EMBnet node organization is a prototypical example of European duplication of effort, which normally is something the European Union strives to avoid.

Although there are several good examples of the opposite, most of the EMBnet nodes are not associated with strong bioinformatics research environments. Most of the high-quality bioinformatics infrastructure and service components in Europe are in fact associated with strong basic research efforts within the area. The primary example is the European Bioinformatics Institute, where research and service co-exist and drives each other in highly synergistic ways. Many other European bioinformatics research groups also contribute to the infrastructure in significant ways. For example, my own group in Denmark, the Center for Biological Sequence Analysis (www.cbs.dtu.dk), offers around 35 different methods and databases, where the WWW site receives 500,000-700,000 page views per month. The situation is similar in many other countries, where key infrastructure components are offered by groups primarily funded via research budgets. This kind of research driven infrastructure is likely to become even more important in the future, where new, sophisticated tools for data integration and systems biology will be very hard to construct without intimate links to basic research. Today, obsolete infra-structure in some countries prevents more research-driven, up-to-date funnels and portals for the experimentalists. It is important to acknowledge that although infrastructure to a high degree depends on large-scale database efforts supported by stable funding, it also is based on basic research in bioinformatics. The European Union should fund the area in ways where stable funding for database infrastructure is secured, at the same time acknowledging the infrastructure based on research in bioinformatics and systems biology.

Future developments in primary and secondary genomic data resources

Richard Durbin
Wellcome Trust Sanger Institute
rd@sanger.ac.uk

Outline

Sanger Institute genomics and bioinformatics data resource programmes; Future perspective for genomics data resources; Browsers such as ENSEMBL integrate many data resources; Comparative genomic sequence data; Public access primary data resources: trace repository; Human variation data/human genetics; ?In situ image data will become important in coming 5 years; Data access and IPR; “Fort Lauderdale” principles of tripartite responsibility for community data resources

Wellcome Trust Sanger Institute

Founded 1992 as a genome research institute, with a primary focus on large scale sequencing; Sequenced 1/2 worm, largest share of yeasts, 1/3 human, many pathogens, e.g. TB, malaria (with others); Took a strong position on open data release, and have been involved in multiple informatics resources and analyses; Building on sequencing success, the institute renewed itself in 2000/1 with a new director (Allan Bradley), £300M core funding from the Wellcome Trust over 5 years, plus £65M IT funding and commitment to new buildings. ; ~650 staff: >20% informatics, <10% in informatics division

Current and Future Sanger Projects

Sequencing: Zebrafish and pathogen genomes. Variation/resequencing (e.g. exons); Genotyping; Haplotype Map project (HapMap): phase II of the Human Genome Project is to characterise common variation; Genetics; Cancer Genome Project - identify somatic mutations directly; Mouse, worm and human genetics: large scale e.g. exon trap, RNAi and smaller; Image data sets; Systematic monoclonal antibodies on tissue microarrays; Systematic *in situ* expression data in mouse embryos 3/4-dimensional; Informatics; ENSEMBL, Pfam, and further resource projects; Support for experimental programmes and related research

Sanger Public Informatics Data Resources

Primary resources; Trace Repository - exchange with NCBI; (primary data -> EMBL, dbSNP, ArrayExpress); Secondary resources; ENSEMBL: animal genomes with automated annotation; Joint project with EBI, compute and web site at Sanger; Pfam: 5200 curated protein families hit 72% proteins; home at Sanger - collaboration with WashU, Karolinska; WormBase: model organism database for *C. elegans*; Now joint with CalTech, Cold Spring Harbor, WashU; Rfam, GeneDB, Human variation resource, …

www.ensembl.org

Distributed Annotation Server (Stein et al.)

ENSEMBL: easy access to the data via …

A web-based genome browser (extensively customizable); Entry points via text search, sequence search, or location; Integrate other data with links out - increasing rapidly for public and priviate use (UCSC site used similarly); DAS support to allow flexible dynamic addition of external data; A web-based system for data export and data mining; EnsMART: flexible generation of data tables; 'Dumps' of sequence and other data sets for you to download; Direct access to the databases; Online or recreate your own copy and customise ; Open software: a Perl-based object layer (Java coming)

Trace Repository: Reads and traces are the primary sequence resource

Unfinished sequence is in EMBL/Genbank; The true raw data are sequence traces; Public mouse sequencing used Whole Genome Shotgun; 6 months to a year delay to first assembly; Openness and sharing led to new data source:; http://trace.ensembl.org/; http://www.ncbi.nlm.nih.gov/

Trace Repository: Current state and uses

~150 million traces; 56M mouse, 33M rat, 12M zebrafish, 17M human, 5M chimp, 4.7M/1.9M Ciona, 2.9M/1.9M pufferfish, chicken, xenopus,…; ~5TB data increasing at an accelerating rate; Access to more sequence data; E.g. Chimpanzee, a pufferfish, *D. pseudobscura* are only available this way Single read data e.g. ESTs, STSs have much more data available in traces; Analysing variation; SNPs = single base mismatches can be verified; Human variation, mutation detection,…; Download reads

and quality values; In bulk based on search or genome/centre; Individual trace files available via API

The (more) global future

Comparative sequence analysis; Despite a slower start than expected, comparative gene finding will improve in the next couple of years, particularly using multiple species data. ~ 10 vertebrates soon, multiples in many phyla. Progress will also be made in identifying functional non-coding sequences such as regulatory sequences; Data resources and browsers; Will integrate sequence based functional data such as from microarrays, systematic RNA interference etc. (happening already); Human variation data (e.g. haplotype map, resequencing); Will form the basis for a new era of sequence-based human genetics; ?Images; Just starting, e.g. FlyBase, Mouse Atlas, …; Microarray, MassSpec are perhaps tools like gels, not systematic

IT infrastructure at Sanger

Current resources; 1400 CPU compute farm; ~100 TB disk, mostly SAN: keep all data and tracking info; ~200 CPU in server clusters for core data and compute; Complex web site: web hits approaching 1,000,000 per day (mainly ENSEMBL, but others growing); Plans; Increase farm progressively, and 100s TB for e.g. images; GRID type connectivity and technology for efficient broad access to resources ; New 1000 m2 Data Centre as part of new building

Data access and IPR

A key to success of many Sanger informatics endeavours has been to take a strong position on open data access and IPR; Fundamental principle of academic research; Credit is assigned based on publication; Publication allows verification, and, more important, reuse of ideas for novel developments; Publication for data resources (primary or secondary) means allowing unconstrained reuse for unforeseen purposes; Good for science, and good for investigators/institutions

“Fort Lauderdale” tripartite principles

Wellcome Trust/NIH meeting 2/03 on data release, with international participants building on genome sequence core; Reaffirm Bermuda principles for rapid release of large scale sequence data; Propose similar unrestricted pre-publication data release for large scale **publicly funded** community resource projects; Requires involvement of three parties; Data generators: release quality-controlled data rapidly; Funding agencies: support generator’s analysis and central DBs; Users: respect data generator’s rights and contributions, and assist equity through peer review process etc. cf OECD Follow-up Group on Issues of Access to Publicly Funded Research Data

EC Workshop Report 2001: Managing IPR in a knowledge-based economy - Bioinformatics and the influence of public policy

Almost non-overlapping participant list; “All involved sectors, from academia to … large pharma companies, strongly support a comprehensive and up-to-date publicly-available and free set of bioinformatics data, with IPR control being used only where appropriate to maintain easy and universal access.”; “Database infrastructures need clear IPR protection policies at all stages of creation, management and access.”; **For publicly funded data resources these should be to void IPR;** “Public funding and IPR rules should encourage collaborations, especially public/private ones”

Collaboration in the open is powerful and in the public interest

Conclusions

Sanger Institutes forward interests; Genomic and protein sequence sets and analyses (continuing); Human variation (inherited and somatic) and genetics; Open data access is so strong for science and collaboration that we make it the a core principle, and support that it should be the norm for publicly funded bioinformatics; Research bioinformatics is (relatively) cheap and fits classical response-mode funding; resource bioinformatics is expensive and requires long term strategic planning, co-ordination and funding

Outline

Finding the genes; Why this is hard; Overview of current methods; Comparative gene finding; Presenting the genes and other information in ENSEMBL; Openly available to industry and academia; More than just human: 9 animal genomes now covered; EnsMart: data mining resources from ENSEMBL; The Trace Repository; 140 million reads create new opportunities; Prospects for the future

ENSEMBL

Joint Sanger Institute/EMBL-EBI project
(European Bioinformatics Institute); Analysis of public human and other animal genomes, available without restriction
http://www.ensembl.org/; Designed to keep up with updated sequence versions and data sets as they appear (moving to monthly release cycle); Automated analysis requires high performance computing and complex software engineering

ENSEMBL provides …

Easy access to sequence data; A gene set for each genome; For known genes, predicted structure and location in the genome sequence; Prediction of novel genes, all with supporting evidence; Further information and links for genes and gene products; Annotation of other features of the genome; Targeted connections to other genome resources world-wide

GeneView gives further info and links Comparative information to mouse ENSEMBL gene prediction pipeline Current human gene stats in ENSEMBL

Release 10.30.1 February 3 2003; 22,980 ENSEMBL genes based on evidence; Approximately 15,000 Refseq (i.e. already known); A few thousand from other species homology or paralogy; A few thousand from extended cDNA (e.g. MGC); ESTs give more but typically fragmentary and are noisy; 204,094 exons and 27,628 transcripts (1.2 per gene); 73,128 Genscan predictions; What is missing?; Comparative gene prediction with mouse (Nature, December 2003) suggests a few thousand extra genes; More exons and transcripts…

Comparison to previous releases

February 2003 release on NCBI 30 sequence; 2.815 Gb, N50 1.7 Mb, 22,980 genes, 204,094 exons; March 2002 release on NCBI 28 sequence; 2.795 Gb, N50 460 kb, 24,179 genes, 199,740 exons; January 2002 release on NCBI 26 sequence; 2.795 Gb, N50 300 kb, 29,181 genes, 191,656 exons; Gene number decreasing, exon number (and total protein sequence) increasing; Genome sequence is improving in quality (nearly finished!); Gene identification system is improving

Results: HTML or tab-delimited file

The Trace Repository

Early access to the primary sequence data

What is the Trace Repository good for?

Access to more sequence data; Chimpanzee, a pufferfish, *Drosophila pseudobscura* are only available this way (always some genomes in this state); Single read data e.g. ESTs, STSs have much more data available in traces; Analysing variation; SNPs = single base mismatches can be verified; Human variation, mutation detection,…; Download reads and quality values; In bulk based on search or genome/centre; Individual trace files available via API

Accessing traces by search

BLAST searches ~$3x10^{10}$ cells/s; Trace search of 1000bp = 10^{11} x 10^{3} cells; SSAHA (Jim Mullikin, Zemin Ning) does this in <1 second; Tuned near-exact matches; Trades off memory against CPU; There is a protein to DNA version; There is a new generation of algorithms for this type of problem; BLAT, PatternHunter; Different variants used for many purposes.

The EBI's activities and plans

Janet Thornton
prepared by Cath Brooksbank, Scientific Outreach Officer
EMBL - European Bioinformatics Institute,
Wellcome Trust Genome Campus, UK - Cambridge CB10 1SD

Introduction

The European Bioinformatics Institute (EBI) is a non-profit academic organisation that forms part of the European Molecular Biology Laboratory (EMBL) and is based in Hinxton, just outside Cambridge, UK.

The mission of the EBI is to ensure that the growing body of information from molecular biology and genome research is accessible freely to all facets of the scientific community in ways that promote scientific progress.

It achieves this mission through four broad activities — (1) supplying information to biologists across the globe through its databases and other resources; (2) state-of-the-art bioinformatics research; (3) training and (4) supporting industry.

(1) Services

Databases

The explosion in genomic sequence and functional genomics information far exceeds the capacity of individual scientists to keep abreast of the data. If we are to fulfil the promise of genomics, we need to collect, store and curate it in ways that allow its rapid retrieval, and we need to build tools that allow us to mine and analyse the data.

The EBI is one of the few places in the world that has the resources and expertise to perform these vital functions. Building on more than 20 years' experience in bioinformatics, the EBI maintains five core molecular databases **(Table 1)**. We accept data submissions from scientists across the globe; perform quality control and annotation of the data; and make it publicly available. Annotation adds enormous value to our databases, providing researchers with detailed information on any gene or protein that they need information on. Much of our annotation is done manually by curators, although some of our databases are annotated automatically, and automatic annotation is an active area of our research.

Table 1: The EBI's core databases

Database	Purpose
EMBL-Bank	One of only three co-ordinated repositories world-wide for storing and retrieving DNA and RNA sequences.
SWISS-PROT (with the Swiss Institute for Bioinformatics) and TrEMBL	SWISS-PROT is a unique database of protein sequences; it is highly annotated (manually) and integrated with other databases.
	TrEMBL is a database of automatically 'translated' protein sequences from EMBL-Bank. It contains more sequences than SWISS-PROT but lacks SWISS-PROT's high degree of annotation.
MSD	One of only three repositories world-wide for collecting, storing and retrieving information on the three-dimensional structure of proteins.
Ensembl* (with the Wellcome Trust Sanger Institute)	A system for annotating the genomes of higher organisms, including the human genome.
ArrayExpress*	A repository for information on gene expression, generated by the use of high-throughput DNA microarrays.

*Established in the past 3 years.

In addition to these five core databases, the EBI hosts approximately 160 other biological data resources. One of these is the Gene Ontology project, which is creating a defined vocabulary to

describe the functions, processes and cellular components that are associated with gene products. This vocabulary is vital for annotating other databases, because it facilitates uniform queries across them. InterPro is a powerful tool for the classification of proteins based on protein families, domains and functional sites. We are also developing new databases to serve the rapidly expanding field of proteomics. We are already developing IntAct — a database for cataloguing protein complexes and their interactions. Other collaborative proteomics projects, moving towards providing analytical tools for 'systems biology', are in the early planning stages.

Securing funding for the maintenance and improvement of established databases is a difficult challenge. For example, the rate of submissions to EMBL-Bank continues to grow exponentially and yet EMBL-Bank the number of curators has remained constant for several years and our core funding from EMBL is spread increasingly thinly as we develop more resources. However, we have been fortunate enough to obtain several large grants to develop new data resources. In 2001, the EU provided the EBI with the largest ever single injection of funds into bioinformatics infrastructure in Europe. The TEMBLOR contract has provided funds to develop the ArrayExpress database of microarray-based gene expression data, create new tools for the Macromolecular Structure Database, create the IntAct database of protein–protein interactions, and develop an integrative layer to facilitate searching across all our resources (Integr8). Other contracts with the EU include BioBabel, which has allowed us to develop tools to standardize database content, through the production of controlled vocabularies; and SPINE, which has funded the development of structural genomics across Europe. In 2002 we also received funding from the NIH to develop UniProt, a unified protein database that will incorporate data from SWISS-PROT, TrEMBL and the US-based database PIR. The Wellcome Trust is another major funder of the EBI's activities, including the Ensembl project (animal genomes) and a significant proportion of the MSD group's work.

Nevertheless, it is far from clear these resources will be maintained in the future. The development of a funding model that will safeguard data is strategically vital for Europe.

The toolbox

As well as providing access to its databases, the EBI provides an extensive range of bioinformatics services — the toolbox — to the research community. The toolbox allows researchers to perform tasks such as comparing DNA sequences or protein structures generated in their labs with those in the public databases.

Some of our tools facilitate the submission of new types of data; for example, EM-dep allows researchers to submit their 3D electron microscopy data to the Macromolecular Structure Database; and MIAMExpress allows researchers to submit microarray data, in a form that complies with community standards, to ArrayExpress.

Demand for these services continues to grow as the amount of biological information in the databases increases. The services provided by our toolbox are used across the globe, by academic researchers, educators and the commercial sector.

(2) Research

As well as providing services to the research community, the EBI has a growing research base. We currently have four teams dedicated to research (**Table 2**), and another new group leader is due to start in October 2003. Our research groups are improving the understanding of genomic and proteomic data by developing new computational approaches, algorithms and data resources. The close association between the EBI's research and services divisions is mutually beneficial: the research groups help to develop our services, and the high concentration of bioinformatics expertise at the EBI expedites our research. In 2001 the UK government allowed EBI staff to apply for research council funding; the research councils and the Wellcome Trust now provide a significant proportion of funding to our research teams, and we hope that our research base will continue to grow over the coming years.

Table 2: The EBI's research teams

Group leader	Summary of research
Current research groups	
Janet Thornton	Using structural biology and molecular modelling to understand and predict how proteins perform their biological function, interact with other biological molecules, and have changed during evolution.
Christos Ouzounis	Using computational methods to automatically annotate genome sequences, classify protein function, understand transcription and the evolutionary relationships between organisms, and map metabolic pathways.
Nick Goldman	Computational methods for studying evolution through the analysis of DNA and amino acid sequences.
Dietrich Schuhmann	Text mining and information extraction. Joined the EBI from LION Bioscience AG, Heidelberg, in May 2003.
Future research group	
Nicholas Le Novère	Systems biology, particularly as it relates to neurons in health and disease. Currently at the Institut Pasteur, Paris, France. Joining EBI in October 2003.

It is important to emphasise that our services groups also perform a significant amount of research and development; a great deal of effort goes into improving established resources and developing new ones. A key part of this work is the development of global standards for bioinformatics.

(3) Training and outreach

Training is a core part of the EBI's activities and falls into two categories.

Training future bioinformaticians

Bioinformatics is a rapidly expanding discipline that touches on all areas of biology — pure and applied. There is a global shortage of trained bioinformaticians, and the EBI fulfils an important function in working towards reducing this shortage. Our bioinformatics training falls into the following areas: PhD studentships; postdoctoral fellowships; short-term research placements through our visitors' programme; short-term placements for PhD students from other institutions in Europe through EU-funded Marie Curie fellowships; training days for masters' and PhD students from other institutions; and hosting international conferences.

It is EMBL policy to provide training for young researchers across Europe and then encourage them to return to their home country to strengthen its national research base. Expertise generated at the EBI is, therefore, disseminated world-wide. This is of great benefit to the bioinformatics community but does pose a challenge to the stable maintenance of our data resources.

Training our users

As our databases and tools grow increasingly sophisticated, so does the need to train users how to get the most out of them. A major part of our current end-user training falls under the auspices of our industry programme (see section 4, below); we have held more than 50 1–2-day Industry Programme workshops since the programme began in 1998.

Funding permitting, we hope to be able to offer more training to academic end-users and also to small-to-medium enterprises in the near future. The EU-funded TEMBLOR contract has already enabled our Macromolecular Structure Database (MSD) group to present a road show throughout Europe; the purpose of this is to train structural biologists how to use the wide range of new structural biology services that have been developed around MSD. In the future, we hope to be able to offer a more extensive range of workshops for biologists, spanning the full spectrum of our services.

We also hope to secure funding to train those who teach bioinformatics at the undergraduate and graduate levels, and we regularly run special workshops (generally 1/2–1 day) to introduce research students to our resources.

Finally, we provide comprehensive online help pages and e-mail support to our users, and are developing an educational website for them.

Outreach

Outreach — not only to our users but also to the general public — is an increasingly important part of our mission. Activities include publication and distribution of press releases and news briefs; production of literature that explains our activities to non-bioinformaticians; exhibiting our resources at international conferences; hosting international conferences and workshops; and educating school teachers and students about the importance of bioinformatics to modern biology.

(4) Supporting industry

Advances in bioinformatics are having a major impact on industry, particularly the biotechnological, chemical, agricultural and pharmaceutical sectors. The EBI's Industry Programme is designed to help industry to adapt quickly to, and maximize the benefits from, developments in this fast-growing field. A total of 18 companies subscribe to the Programme. Membership provides direct access to the EBI's expertise, through research funded by the programme, quarterly meetings, and regular workshops. Members of the Industry Programme have significant input into the content of the workshops and quarterly meetings. In turn, funding from the programme has pump-primed several exciting projects at the EBI, including the development of our microarray resources.

The popularity of the EBI's industry programme with large companies has prompted us to consider how we can best serve small-to-medium enterprises (SMEs), and we are now planning the development of a programme aimed specifically at these companies. Like our Industry programme, our SME programme will need to be self supporting.

Achieving excellence through collaboration

The EBI openly welcomes collaborations with other institutes and almost 70% of the papers that we have published are the result of collaborations. We are currently participating in 14 EU-funded collaborative projects and are involved in a further 17 applications for Framework 6 grants. Many of these applications are collaborations with experimentalists, providing vital contact with the community that we serve.

Formal collaborations between databases are especially important, allowing researchers world-wide to access accurate and up-to-date information from a number of host databases. All of our core databases participate in such collaborations. For example, EMBL-Bank has an international collaboration with GenBank and DDBJ to exchange data every day, and MSD similarly exchanges information with the Protein Data Bank. The development of data standards is often an important part of such collaborations, and EBI staff members are heavily involved in groups that are working towards such standards, including the Microarray Gene Expression Data Society, the Proteomics Standards Initiative and the Gene Ontology Consortium.

SUMMARY OF PRESENTATIONS BY SESSION CHAIRPERSON

Chair - Janet Thornton

FUTURE INFRASTRUCTURE AND ANALYSIS REQUIREMENTS - Hardware, Software, Archive, Data Access

Unifying theme: analysis of the needs in the scientific communities that employ bioinformatic tools: what are the actual and future needs in respect to storage and retrieval of biological information, including computational means and algorithms? What are the effects of technologies such as Array technologies and their implications for databases, standardisation and software?

Session

- Martin Bishop – National Resource
- Chris Cooper – Computer Industry Perspective
- John Schofield – Pathbase: Image Database
- Soren Brunak – Research bioinformatics
- Richard Durbin – Large Genome Centre

I National Provision of Bioinformatics Infrastructure

- In Bioinformatics services HGMP provide:

– Compute resources; integrated software; genome web; extensive training; help for users; new technologies
– Charge for training
– Centre also does research and provides wet lab 'services'.
– Cost ~1.5 million euros per annum for bioinformatics

Discussion Points on National Bioinformatics Infrastructure Provision

- **Alternative models for national provision** from fully funded to full cost recovery from users. In the Netherlands, they have recently moved from cost-recovery to funded model. In Germany DKFZ run cost-recovery service but this has been difficult.
- **Training very important** in national centres for 'national research community' and also provided for industry (but charged)
- **Role of EMBnet nodes** discussed – Different nodes provide different quality – from dead to very lively! Different views expressed on future – either disband or re-invigorate
- **Training Provision** unbalanced. There are many courses for bioinformatics student training (perhaps too many) – but insufficient training for biologists in use of bioinformatics

II 'Computer Industry Perspective'

- Life Sciences now very important to computer industry (e.g. IBM increased Life Sciences personnel from 25 in 2000 to >2000 now).
- Biological data is large volume and complex – therefore new challenges
- Major weakness is lack of standards
- Major threat is inability to act quickly enough to 'conquer' data challenge

Discussion Points

- Standards are central: 'computing standards' and biological 'ontologies' – both needed desperately
- Strong support for 'Open' standards.

III New Type of Data resource: Pathbase

- New types of data: Images
- Major Problem and bottleneck is lack of 'phenotype' ontologies (Pathology & anatomy)
- Major effort has been in development of data structure
- Currently 600 images, mainly from 'high throughput labs; still in development
- Links to other data resources important – but yet to be made
- Data resource supported now (by EU grant)– but no future funding yet secured

Discussion

- EU starts databases but then does not continue funding, even when success!
- National Funding for data resources difficult to obtain since viewed as European/global resource

IV Research

- Funding for data resources has 'overwhelmed' funding for research in bioinformatics
- Research into methods is critical for future of biology
- European research is fragmented/'duplicated'
- Recommendations:
  - Split funding for infrastructure and research
  - Create network of research bioinformaticians in Europe

Discussion

- Consensus that funding models in bioinformatics are not meeting current needs for database resources either large (centralised) or specialised (distributed) – these are expensive and need longer term strategic planning, co-ordination and funding
- Fall between national, European and global
- US funding has been stable and considerable
- Research bioinformatics critical – also needs support, but EU funding models are potentially OK – if not hijacked by need to fund resources
- Better integration and exploitation of EU expertise is needed, including experimentalists

V Genome Centre

- Importance of keeping raw data (trace repository – 5TB data)
- Secondary resources are also critical
- Distributed annotation – DAS
- Large IT resource (1400 CPU; 100TB; 200 CPU servers)
- Open data, software and collaborations

Discussion

- Comparison of Sanger funding with whole of EU budget!
  - Good Bioinformatics is not cheap!
- Discussion of Open Resources: Advantages & disadvantages – US has a much stronger record than EU

FUTURE BIOINFORMATICS RESEARCH TOPICS
(Thu 13 March 9:00 - Chair - Jaap Heringa)

==========

H. Werner Mewes - Comments
Future of bioinformatics

Bioinformatics is not an extension of its past.

Future : what is in a genome?; - genome analysis is the hunt for sequence/structure/function using homology; - there is no straightforward method for sequence to function.

Association of attributes; Information inferred by prediction; We must collect, structure, analyse.

The best resource is SWISS-PROT.

Methods in genome annotation are difficult, and functional classification is hard.

Oneomics: Hypothesis driven vs systematic approaches are required for databases, algorithms, knowledge

Many databases are from 1970s. Large institutions are trapped in their functions; We need better data structures, and we must make data computable. Often data structures are not suitable for complex data.

Comparative genomics involves information transfer from one object to another; - i.e. annotation ; - We need to find out what are rules, and the level of confidence required: algorithms, statistics, careful evaluation

There is a context problem, which istime dependent - Database maintenance is dynamic, whereas most information is static. No centralised system will cover all requirements. We require resources for content maintenance and development of new databases.

Experimental data: We need stable and curated data resources – like yeast aradopsis and e.g. biobus. We need general layers for analysis.

Bioinformatics as an integrative science

Jaap Heringa
Integrative Bioinformatics Institute VU (IBIVU), Faculty of Sciences & Faculty of Earth and Life Sciences, Free University, Amsterdam, The Netherlands
heringa@cs.vu.nl, www.cs.vu.nl/~ibivu, Tel. +31-20-4447649

What is it about?

Bioinformatics has been defined early on as "studying informational processes in biological systems" (Hogeweg, Utrecht University; early 1970s). This is possibly one of the widest definitions around but it correctly puts the emphasis on biology. In contrast, an incorrect idea about bioinformatics, predominantly in the US but now also in Europe, is that the field is about applying computational algorithms and mathematical formalisms in biology. This definition, and the fallacy that bioinformatics is a new science, has the undesirable effect that the field might become inundated and even dominated by computational experts who lack the necessary bioinformatics or even biological background. Within some other disciplines, where bioinformatics is now conceived as a helpful science, it is even thought that the task of bioinformatics is "taking care of the computational infrastructure and data management", which reduces the scientific challenge to basic IT.

The history of science might be divided in four eras with the following main focuses:

- Anatomy, architecture -- a fitting illustration is Rembrandt's paiting "The anatomical lesson of Professor Tulp" (1632)
- Dynamics, mechanics – putatively started with "Newton's apple" (1724)
- Informatics -- Cybernetics (Wiener, 1948) has been defined as the science of control in machines and animals, and hence it applies to technological, animal and environmental systems
- Genomics, bioinformatics -- the full blown information challenge of living systems

Many researchers argue that the human mind has been adapted very well to recognise patterns in 2-D or 3-D and that we have well-developed skills for estimating and dealing with dynamic processes. However, model systems prove time and again that our intuition for information processes is quite rudimentary, leading to an enormous intellectual challenge for getting on top of the information interplay involved in cellular complexity.

Integrating bioinformatics data and methods

Understanding the fundamental information flows resulting from molecular processes at the cellular and super-cellular level will require the integration of all cellular genomic levels: sequence, transcriptome, proteome, metabolome, and physiome. This is sometimes referred to as *vertical genomics*. To scale up further, these integrated levels will need to be incorporated in a model system at the organismal level.

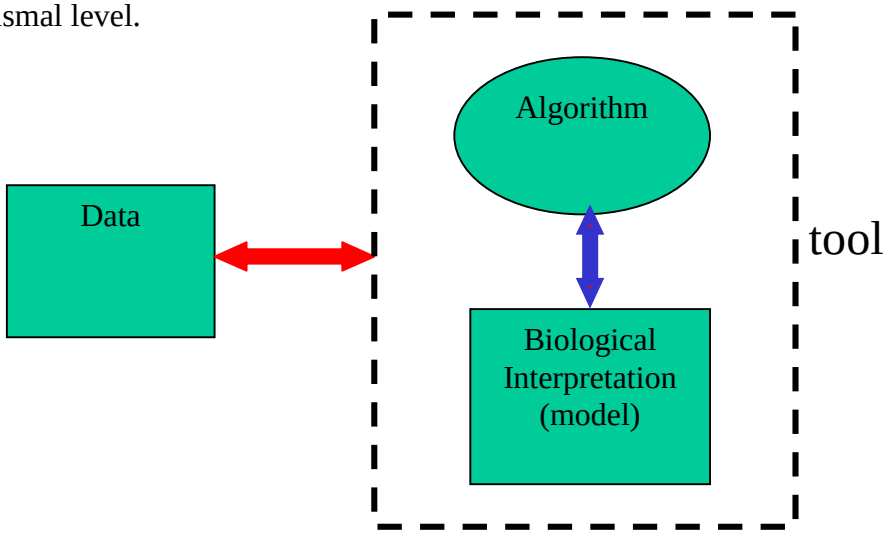


Fig. 1. *Schematic outline of a bioinformatics tool. Any such tool has a biological component, as it manipulates biological data.*

Figure 1 shows a general outline of a bioinformatics method. Many bioinformatics algorithms have sophisticated data interpretation capabilities or contain an elaborate model with is used to analyse the data. However, also the simplest algorithm not conceived for any biological implementation will by definition contain a biological model, whenever it is applied to biological data. An example is string matching: even the simplest matching scheme, for example only based on identities, represents a biological model. The problem with many standard computer science algorithms is that, when applied to biological data, the biological component of such methods can be plainly incorrect.

Data integration is often conceived as implementing interfaces between databases (Fig. 2), such that annotation for an entry in one database can be taken directly from a second database. For example, one clicks on the name of a protein sequence in one database and gets a picture of a 3-D structure in another database; clicking on the structure could then reveal the cellular location and function of the protein as deposited in a third database, etc. While these developments are significant and extremely useful, they do not constitute integration of the inference methods, such that the biological interpretation has to be performed by the individual researcher.

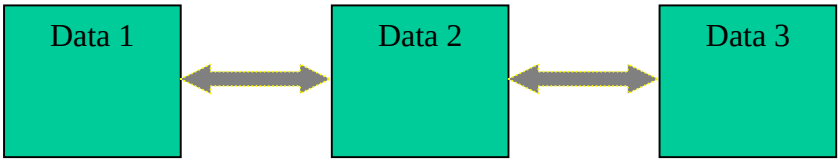


Fig. 2. Data integration. Often taken as producing (clickable) links between databases.

In order to build up integrative knowledge of systems using genomic data sizes, it is crucial that the bioinformatics inference engines are integrated, such that the data integration takes place *through* the bioinformatics analysis methods (Fig. 3). This will require the development of sophisticated and flexible interface structures, and constitutes a grand challenge for bioinformatics research, particularly given the rapidly growing and heterogeneous genomic datasets.

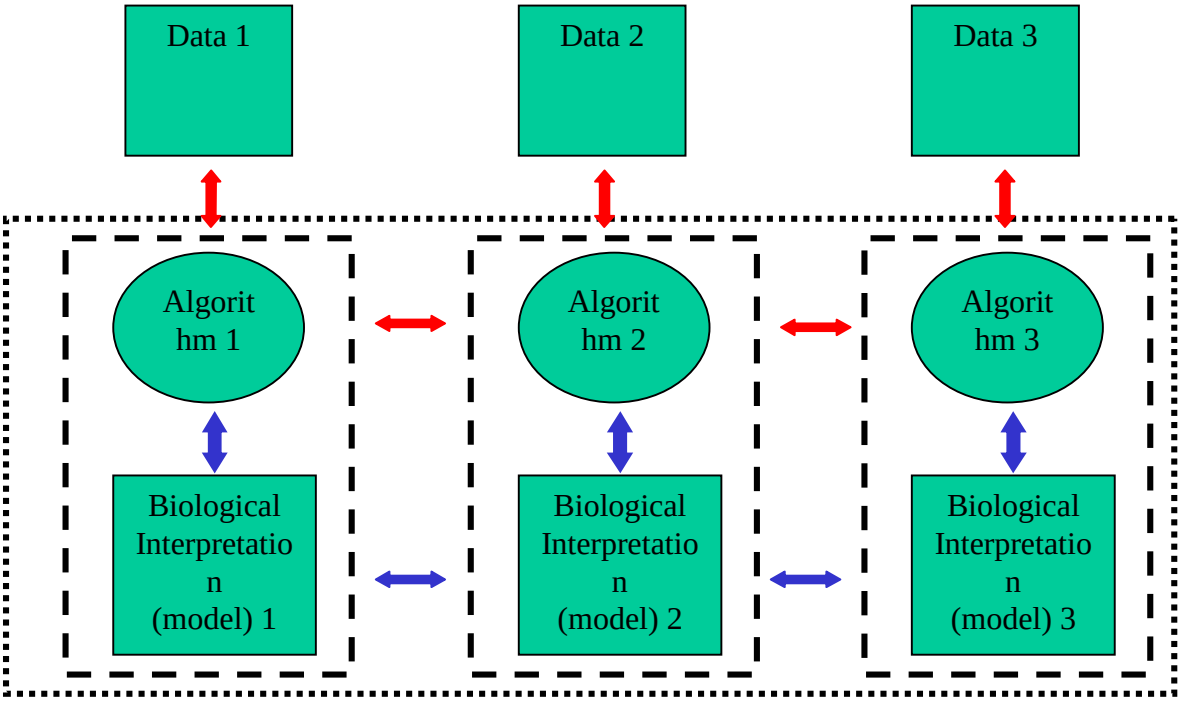


Fig. 3. Bioinformatics method integration.

A few research projects conducted in our laboratory, such as the integration of multiple alignment methods and those for secondary structure prediction, have underlined the complexities. Shaping and

tuning of any newly developed bioinformatics method is normally a complicated and iterative process, but dealing with the increased complexities of the biological signals here makes this development process even more challenging.

The low-key examples of method integration in our lab deal with fairly homogeneous data (e.g. multiple alignment and secondary structures can both be represented as strings of letters). But this already turned out to be difficult to do, as the resulting integrated methods were not easily tunable. The big question then becomes: How can we scale up to knowledge-integrating and inference engines?

- We have some formalisms (ontologies, distributed databases) but we need to develop many completely new formalisms and also new technology.
- There is a clear role for the EC to increase funding for such bioinformatics initiatives

Integrative Genomics/bioinformatics approaches in The Netherlands

Many universities in the Netherlands over the last 3 years have started to create bioinformatics chairs, and at present, these positions have mostly been filled such that about 15 academic groups are currently active. Some universities are establishing more bioinformatics positions to enlarge the critical mass.

Apart from the Centre for Medical and Biomolecular Research (CMBI) at Nijmegen University, which over the last few years has been the largest national centre, the other Dutch universities with one or more bioinformativs research teams include Maastricht, Leiden, Nijmegen, Amsterdam (VUA and UVA), Utrecht, Groningen, Delft and Rotterdam. The same picture across Dutch universities holds for genomics research, and many collaborative initiatives between genomics and bioinformatics research are being organized at the university level.

The Free University Amsterdam (VU) recently founded the Integrative Bioinformatics Institute VU (IBIVU), and in its mission statement, integration at various levels is prominent:

> "*The new Centre has a focus on 'Integrative Bioinformatics'. Integrative Bioinformatics aims to integrate information from all the levels of genomics that are involved in important biological, medical, ecological and/or societal functions. Depending on the function under study, these levels may include genome sequence, transcriptome, proteome, metabolome, metagenome, ecosystem character, population studies, and behavioral patterns. The Centre will bring together the activities of excellent VU scientists from all scientific disciplines that can contribute to integrative bioinformatics. This leads to yet another profile of integration, i.e. that of maximum combination of mathematical, computer-science, biophysical, biochemical, genetic, medical and biological expertise. In keeping with the multi-disciplinary nature of the proposed Centre and to ensure maximum national and international visibility, four Faculties of the Vrije Universiteit (Sciences (FEW), Earth and Life Sciences, the VU Medical Centre (VUMC), Psychology and Pedagogy (FPP)), one company (i.e. Biological Detection Systems) and three National Genomics initiatives will ultimately be involved. The new Centre will organize, streamline and integrate Bioinformatics research at the Vrije Universiteit (VU), including fundamental bioinformatics research.*"

The IBIVU functions in loose collaboration with a number of other inter-faculty institutions at the Free University such as the Centre for Neurobiology and Cognitive Research (CNCR), the Centre for Research on BioComplex Systems (CRBCS) (Systems Biology), and the VU Medical Centre (Microarray, CGH data).

The IBIVU takes part in a few virtual Centres of Excellence (Dutch incarnations of the current FP6 funding organisation) funded by the Netherlands Scientific Research council (NWO), which are about bioinformatics research or have a clear bioinformatics component in their research mission. These are

- *NBIC – Netherland BioInformatics Centre* – a pan-Dutch bioinformatics centre with a large bioinformatics service component.
- *Centre for Medical Systems Biology* (Leiden, A'dam, R'dam) – A Systems Biology approach to disease.
- *Ecogenomics* (A'dam, Wageningen, National Institute For Health and Environment (RIVM) – Analysing soil and its resistance against pollution through simultaneous expression profiling of a population of soil microorganisms.

The NBIC also initiated activities to organise and streamline bioinformatics teaching across the Netherlands, called BioASP.

The current Dutch Centres of Excellence funded by the NWO where bioinformatics plays an integral role are:

- *Cancer Genomics Consortium* [DCGP]
- *Centre for Biosystem Genomics* [CBSG], focuses on plant genomics (potato, tomato)
- *Kluyver Centre for Genomics of Industrial Fermentation* [Kluyver]
- *Centre for Medical Systems Biology* [CMSB], focuses on multifactorial diseases
- *Netherlands Proteomics Centre* for proteomics as an emerging horizontal genomics discipline

Some Dutch academic/industrial initiatives with a significant bioinformatics component include consortia with the following research focus:

- *Nutrigenomics* -- exploration into the prevention and care of nutrional inroads in vascular disease, diabetes, hypertension and obesity
- Interaction between the immune system and food; a functional genomics approach to *celiac disease*
- Mechanisms of life-threatening *virus disease* and new leads for treatment and vaccines
- Genomics of *host – respiratory virus interactions*: towards novel intervention strategies
- Ecogenomics: Functioning of *ecosystems* targeted at sustainable environmentally friendly and healthy products (ecology, toxicology and sustainable innovation)

It is clear that these national initiatives and those at the Free University Amsterdam are well in line with the keywords of the Thursday morning session:

- Integration of knowledge
- Systems Biology
- Pharmaceuticals
- Biotechnology
- Environment

Ilias Iakovidis

eHealth - Past and future activities of the European Commission

Directorate General Information Society
Components and subsystems. Applications.
eHealth Unit

Topography : A functional anatomy for human and veterinary medicine. *Meninges: T-A1110*
Morphology : Terms used to name and describe structural changes in disease and abnormal development. *Inflammation, NOS: M-40000*
Function : Terms used to describe the physiology and pathophysiology of disease processes. *Fever: F-03003*
Living Organisms : Living organisms of etiological significance in human and animal disease. *Streptococcus, NOS: L-25100*
Physical Agents, Activities, and Forces; A compilation of physical activities, physical hazards and the forces of nature. *Work-related activity, NOS: A-70100*
Chemicals, Drugs, and Biological Products; Including pharmaceutical manufacturers. *Penicillin, NOS: C-54000;*
Procedures : A classification of healthcare procedures. *Prescription of drug, NOS: P2-08050*
Occupations : Developed by, and used with permission from, the International Labour Office in Geneva, Switzerland. *Forestry worker: J-63230*
Social Context : Social conditions and relationships of importance to medicine. *Non-smokers: S-32080*
Diseases/Diagnoses : A classification of the recognised clinical conditions encountered in human and veterinary medicine. *Meningitis: DA-10010*
General Linkage/Modifiers : Linkage, descriptors, and qualifiers to link or modify terms from each module. *Clinical stage I: G-E100; Neck:T-D1600; with: G-C008; muscle stiffness: F-11320;*

Bioinfomed Study; In November of 2001 a study was launched to continue the findings of the conference of December 14,2001:
Synergy between Research in ; Medical Informatics, Bio-Informatics and Neuro-Informatics
30 experts worked 1 year to present the potential of the synergy between Medical Informatics and Bioinformatics and proposed a roadmap for collaboration called "**Synergy between Medical Informatics and Bio-Informatics: Facilitating Genomic Medicine for future healthcare**

HealthGRID & FP6

Application of the existing GRID and GRID-like technology in the Health sector for timely and secure access to (distributed) patient data
Electronic Health records, Regional Health Information Networks
interoperability of databases of heterogeneous content (biology and medicine) for research purposes enabling new knowledge discovery (research, drug design), better guidance and information (healthcare professionals)
computing intensive applications and knowledge discovery imaging, simulation and modelling
Workshop January 16-17, 2003 http://lyon2003.healthgrid.org HealthGRID applications in the *e*Health unit
*e*MOLECULE; Molecular biology databases - knowledge discovery; Molecular Medicine (e-Pharmacology)
*e*CELL; Pathway simulations, virtual cell - computing power
*e*INDIVIDUAL' Medical imaging Combination of genetic and clinical data
*e*POPULATION Environmental Influences

Gunnar von Heijne

Dept of Biochemistry & Biophysics Stockholm Bioinformatics Centre, Stockholm University, Sweden

Commentary

The Commission and the Parliament should be able to figure out how to provide stable European funding for critical databases like SWISS-PROT.

Problem with specialised databases: if there is no mechanism for continuation grants given that a project is sccessful initially, providing start-up funding may be money down the drain.

Groups working on algorithms: the current EU funding-schemes are OK. If a new algorithm turns out to be useful to the community it will survive even if the grant expires.

On research area, we all discuss systems biology and high throughput. There is a danger of going too far in this direction and ignoring biology - only a minority of the wet lab people do high throughput. Most biology is still small scale and hypothesis-driven. Europe is still competitive in algorithm-development; we must maintain this edge.

Concerning the fraction of general life sciences projects that should be devoted to bioinformatics and data analysis - for the large-scale projects, people seem to converge to around 20% or so. For small-scale projects - let's say a medium-sized group doing some array experiments, or a group involved in small-scale proteomics (1-2 persons running 2D gels & MS), my experience is that they often need a part-time bioinformatics expert to help them with mostly rather trivial chores (writing small applications, building an in-house database to keep track, etc.). The ideal situation is when this person also does algorithm development on a more advanced level in a similar area of research - my own experience is that the bioinformatics students often find it quite rewarding to get involved in a real experimental project to see how far their skills and methods hold when confronted with all the nitty gritty of day-to-day labwork.

This direct contact can only be established if there is a core of bioinformatics people at the institute or university that has strong links to some of the experimental groups. huge centres are so preoccupied with their own projects that they rarely have the time or interest to do this. Also, they're so large that size in itself becomes an impediment to the outsider.

Future Bioinformatics Research Topics: from data management to data interpretation

Diego di Bernardo
Telethon Institute of Genetics and Medicine, Naples, Italy

TIGEM: a short introduction

Research is focussed on the identification and function of "disease genes". Funded by the non-profit Telethon Foundation; Staff: More than 100 people including: researchers, technicians and students. Director: Andrea Ballabio

TIGEM & Bioinformatics

Bioinformatics team:; Researchers: Diego di Bernardo, Elia Stupka; 3 technicians, 6 PhD students; Projects involving Bioinformatics:; Identification and Analysis of interspecies conserved sequences of 1022 disease genes (ongoing); Reverse-engineering of complex regulatory networks (ongoing); Human curated annotation database of disease genes (future project); Computational identification of regulatory sequences; Prediction of gene tissue-specific expression using EST (completed)

Bioinformatics in Life Sciences

Advances in biotechnology: low-cost, low noise and repeatable measurements ; The quantity and quality of experimental data has grown enormously ; ; Informatics disciplines (i.e. database-like activities and information management systems)

Disequilibrium

Informatics disciplines: database-like activities and information management systems ; ; Computational Biology: data-analytical and theoretical methods, mathematical modelling and computational simulation techniques

PubMed Search Results

A paradigm shift

In biological research it was assumed that raw experimental data could be directly interpreted by the scientist. ; Theoretical models and computational tools to explain the experimental data are needed in order to help the discovery process

Solutions

The problem was partly addressed in FP6; "*...to enable researchers to better decipher the functions of genes and gene products as well as to define the complex regulatory networks that control fundamental biological processes…"* from Thematic priority area 1: Advanced genomics and its applications for health-Fundamental knowledge and basic tools for functional genomics in all organisms. Increase collaboration between physics and engineering disciplines with life sciences.

Applied sciences areas that can have a major impact:

Digital signal processing: to discover information hidden in the DNA sequence. ; ; **System Identification:** to identify complex regulatory networks from experimental data. ; **Networks theory:** to be able to understand, predict and eventually modify the behaviour of complex regulatory networks.

Research project features:

To be successful each project should necesseraly have:; A clear objective that targets a biological research problem; A theoretical/computational component; An experimental validation component

Risks

Algorithm/Models not based on biological knowledge; Lack of experimental validation; ; Useless for biological/medical research; Waste of resources

Potential impacts

Optimally designed experiments and data interpretation using theoretical models and computational tools

Computational tools for the discovery of novel "objects" in the genomic sequences; Theoretical models to explain complex regulatory networks organisation and function.

Commentary - Diego Di Bernardo
Telethon Institute of Genetics and Medicine, Via Pietro Castellino 111, I - 80131 NAPOLI

The draft agenda of the workshop “Bioinformatics - structures for the future” focuses on the main challenges facing Bioinformatics research in the near future. We would like to contribute to the discussion in the session “Future Bioinformaticsd Research Topics”.

The topic of the contribution we would like to present has its centre of gravity in the FP6 thematic priority area 1: **Advanced genomics and its applications for health - Fundamental knowledge and basic tools for functional genomics in all organisms**: *Gene expression and proteomics to enable researchers to better decipher the functions of genes and gene products as well as to define the complex regulatory networks that control fundamental biological processes.*

In the last decade biology and genomics have evolved from empirical sciences to quantitative sciences, thanks to advances in biotechnology that allowed low-cost, low noise and repeatable measurements. As a consequence of the availability of this novel technology, together with an increase in governmental and private spending in the biotechnology and life sciences sector in general, the quantity and quality of experimental data has grown enormously. There is currently disequilibrium within the bioinformatics area which favours informatics disciplines (i.e. database-like activities and information management systems) against data-analytical and theoretical methods, mathematical modelling and computational simulation techniques, usually referred to as “Computational Biology”. This problem has been partly addressed in the FP6.

What is already in progress is a shift in a paradigm implicit in biological research, where it was assumed that the scientist could directly interpret raw experimental data. Mathematical models and simulation techniques are needed in order to integrate biological knowledge with the ever-increasing amount of experimental data into a formal framework (in silico model) and to test hypotheses on the functioning of biological processes. This has been recognised in the FP6, however computational biology can be extremely helpful not only for integration and analysis of biological knowledge, but also in the discovery process. This can be achieved by increasing collaboration between physics and engineering disciplines with life sciences. Engineering disciplines that can have a major impact in this field are:

(1) **digital signal processing** to discover information hidden in the DNA sequence.

(2) **system identification** to identify complex regulatory networks from experimental data.

(3) **network theory** to be able to understand, predict and eventually modify the behaviour of complex regulatory networks.

In conclusion, database development and management is essential for biological research and advancement. In addition to this, there is a need for the development of both theoretical models and algorithms to interpret and integrate the large amount of biological data now available.

Editorial

Alfonso Valencia
Protein Design Group, National Centre for Biotechnology CNB-CSIC, Cantoblanco, Madrid E-28049 Spain, Tel: +34 91 585 45 70, Fax: +34 91 585 45 06, E-mail: valencia@cnb.uam.es,

The presentation was based on the editorial by Alfonso Valencia entitled:

"**BIOINFORMATICS: BIOLOGY BY OTHER MEANS**",

BIOINFORMATICS, *Vol. 18 no. 12 2002, Pages 1551–1552.*

This article may be accessed at:
http://bioinformatics.oupjournals.org/cgi/reprint/19/7/795?ijkey=54OK0qWunkYEI&keytype=ref

His key premise is that the success of bioinformatics in its application to genomics and proteomics has complicated the relationship of computation with experimental biology. There is a need to attend to our pressing needs of bioinformatics applications without forgetting other, perhaps less evident but equally important, aspects of computation in biology." The article considers the areas:

- **MAINSTREAM BIOINFORMATICS**
- **BIOINFORMATICS IN THE STUDY OF SPECIFIC BIOLOGICAL PROBLEMS**
- **BIOINFORMATICS IN THE STUDY OF GENERAL BIOLOGICAL PROBLEMS**

He concludes that "It is our responsibility as a community and the Professional Societies such as ISCB, (http://www.iscb.org) and the emerging European branch (ECCB), to pass on to the rest of the community a clear message about the need to preserve the balance between the different areas of Bioinformatics and Computational Biology."

From Bioinformatics to Computational Biology

Jean-Michel Claverie

Structural and Genomic Information, CNRS-UPR 2589, Marseille cedex 20, France

It is quite ironic that the uncertainty about the number of human genes (28,000–120,000) appears to increase as the determination of the human genome sequence is nearing completion (Claverie 2001). This paradox reveals deep epistemological problems, and suggests that "bioinformatics"—a term coined in 1990 to define the use of computers in sequence analysis—is no longer developing in directions most relevant to biology.

After the pioneers who established the basic concepts of molecular sequence analysis (Fitch and Margoliash 1967; Needleman and Wunsch 1970; Chou and Fasman 1974), most computational biologists of my generation (the second one) embarked on their journey into the emerging discipline with the ambition to turn it into the ***bona fide*** theoretical branch of molecular biology. Having a physicist's background, I suspect that many of us had the vision of establishing bioinformatics in a leadership role over experimental biology, similar to the supremacy that theoretical physics enjoys over experimental physics. Somewhere along the line, it seems that bioinformatics lost this ambition and became sidetracked onto what physicists would call a "phenomenological" pathway. Let us follow the example of particle physics for a little longer. There, theoretical research has two phases (which, in fact, run in parallel). In the first phase (so-called phenomenological), a large number of physical events are recorded in huge raw databases, classified into separate groups based on statistical regularities, and then utilized to identify the most recurrent objects. Optimal database design, fast classification/ clustering algorithms, and data mining software are the main area of development here. The level of knowledge gained from this phase is, for instance, that objects A and B often appear together except when C is around, or when parameter X is lower than a certain threshold; it is mostly statistical in nature. The parallel with the current state of bioinformatics is clear.

However, theoretical physics also has a subsequent, totally different phase, aiming at discovering the basic (few) rules (e.g., E = mc 2) underlying the relationships between the objects, their individual properties, and thus finally explaining the statistical distributions of the events recorded in the databases. Once known, these rules considerably simplify the description of the database content and, more important, have a predictive power: the realm of the theory may encompass objects or events that have not been ob-served previously. This part of theoretical endeavor is entirely missing in current bioinformatics. As a consequence, we are still not able to agree on the number of human genes despite having the complete sequence of the human genome at hand. Identifying precisely the 5' and 3' boundaries of genes (the transcription unit) in metazoan genomes, as well as the correct sequences of the resulting mRNA ("exon parsing") has been a major challenge of bioinformatics for years. Yet, the current program performances are still totally insufficient for a reliable automated annotation (Claverie 1997; Ashburner 2000). It is interesting to recapitulate quickly the research in this area to illustrate the essential limitation plaguing modern bioinformatics.

Encoding a protein imposes a variety of constraints on nucleotide sequences, which do not apply to non coding regions of the genome. These constraints induce statistical biases of various kinds, the most discriminant of which was soon recognised to be the distribution of six nucleotide-long "words" or hexamers (Claverie and Bougueleret 1986; Fickett and Tung 1992). Initial gene parsing methods were then simply based on word frequency computation, eventually combined with the detection of splicing consensus motifs. The next generation of software implemented the same basic principles in simulated neural network architectures (Uberbacher and Mural 1991). Finally, the last generation of software, based on hidden Markov models, added an additional refinement by computing the likelihood of the predicted gene architectures (e.g., favoring human genes with an average of seven coding exons, each 150 nucleotides long) is added (Kulp et al. 1996; Burge and Karlin, 1997)). These ab initio methods are used in conjunction with a search for sequence similarity with previously characterized genes or expressed sequence tags (EST). Sadly, it is often claimed that matching back cDNA to genomic sequences is the best gene identification protocol; hence, admitting that the best way to find genes is to look them up in a previously established catalog!

Thus, the two main principles behind state-of-the-art gene prediction software are (1) common statistical regularities and (2) plain sequence similarity. These concepts are quite primitive. For instance, the idea of analyzing the frequency of groups of letters was actually introduced by Arab scholars around A.D. 700 to break substitution ciphering. Thus, the legendary cryptanalyst al-Kindi (Singh 1999) could still grab the essence of modern bioinformatics without much difficulty. Moreover, the above concepts are intrinsically conservative and introduce a bias in favor of the detection of genes similar to those already known.

But the most fundamental limitation of the current approaches is that they bear no relationship to the actual molecular mechanisms of gene expression; when a human cell triggers the transcription of a given region of its genome, it is not because of its homologue in yeast, or because the transcripts (once translated) will lead to a meaningful (three-dimensional folding) amino acid sequence. Thus, current approaches are not on the pathway of a theoretical understanding of the genome, and have no predictive power beyond the realm of immediate analogy. This limitation is well illustrated by the difficulty we have in locating non-protein coding ("RNA") genes (such as *Xist* and *H19*), which probably have essential regulatory roles. The number of non-protein coding genes is unknown, and they might constitute a significant fraction of yet anonymous EST clusters. The same fundamental problem is also attested by the near-zero performance of current methods to locate core promoter regions, as well as all other regulatory segments (Stormo 2000).

The current approaches are fundamentally limited by the fallacious analogy that the human genome is a text to be ***deciphered***. This vision, very popular with the media, is also pervading the scientific policy in the field. One often hears that bioinformatics must become "multidisciplinary," must attract more computer scientists and mathematicians, etc., in the hope that fancier computational techniques will crack the code. However, computer crypto-analysis techniques only work at the level of symbols; the final understanding of the meaning of a message remains the privilege of its intended recipient—a human brain. For example, a simple frequency analysis will recognize a simple Cesar shift ciphering in the following message: *ZfmmpxEvdlxjmmnffuUbsabobu2241qnbuNjbnjCfbd*i, leading to its decoding to: *YellowDuc*kwillmee*tTarza*nat1130pma*tMiamiBeac*h. However, this is not truly useful if we do not know the meaning of the predefined code words: *YellowDuc*k, *Tarzan*, and *MiamiBeach* (e.g., a given boat, a certain admiral, and a precise geographical location). Similarly, for whatever DNA sequence we decipher, we have to determine its meaning from the cell's point of view.

Thus, I believe that the days of abstract DNA "numerology" are over, and theoretical biologists with a strong interest in the intricacies of the cell machinery should now reinvest the field. We now need to make educated guesses on the meaning of the code words by concentrating on the way they might interact with each other. The statistical features recognised at the level of DNA sequence have now to be related to chromatin structures, kinetic properties, and physicochemical principles of macromolecular interactions.

The huge amount of information acquired recently on the spatial organisation of large molecular edifices such as transcriptional complexes and the spliceosome have now to be incorporated into a radically new type of bioinformatic approach. For example, one could try to develop *in silico* promoter detection algorithms mimicking the formation of the pre-transcriptional complex in response to the proper sequential recognition of individual sequence motifs. Standard approaches used until now (such as finite state machines, neural networks, or Markov models) cannot implement all the properties of such a contextual recognition process. For those wishing to remain at a more abstract level, designing new algorithms at least logically consistent with known biochemical and cellular processes is also a worthwhile direction of research.

This suggests that qualitative progress should now be given priority over the incremental improvement of current methods. Trying to achieve a reasonably accurate detection of human genes without reference to coding potential, sequence similarity, or any property of the gene product, is certainly a good benchmark problem—both of tremendous practical and fundamental interest—on which to focus the development of new approaches.

Similarly, the realistic computer modeling of an entire *E. coli* cell, pushed forward as a 10-year project by the ***International E. coli Alliance (IECA)*** is a good prototype of a high-visibility, ambitious, visionary project capable of federating all the components of today's bioinformatics

activities (Bio-specific IT developments, data integration/visualization schemes, algorithmic, applied maths) as well as new contribution from physics and engineering disciplines. Such a project**,** already in search of preliminary funding in the US, Japan and Canada, should have a European branch.

Instead of computer scientists and mathematicians who look at the DNA as if it was the tape of a Turing machine, we now need a generation of computational biologists with a solid background in such fields as transcription, development, enzymology, microbiology, structural biology, etc. This will help bioinformatics to become a truly successful branch of biology, in pursuit of a satisfactory understanding of the function and evolution of genomic sequences in their cellular context.

REFERENCES

Ashburner, M. 2000. *Genome Res.* **10:** 391–393.
Burge, C. and Karlin, S. 1997. *J. Mol. Biol.* **268:** 78–94.
Claverie, J.-M. 1997. *Human Mol. Genet.* **6:** 1735–1744.
Claverie, J.-M. 2001 *Science* **291:** 1255-1257
Claverie, J.M. and Bougueleret, L. 1986. *Nucleic Acids Res.* **14:** 179–196.
Chou, P.Y. and Fasman, G.D. 1974. *Biochemistry* **13:** 222–245.
Fickett, J.W. and Tung, C.S. 1992. *Nucleic Acids Res.* **20:** 6441–6450.
Fitch, W.M. and Margoliash, E. 1967. *Science* **155:** 279–284.
Kulp, D., Haussler, D., Reese, M.G., and Eeckman, F.H. 1996. *Ismb* **4:** 134–142.
Needleman, S.B. and Wunsch, C.D. 1970. *J. Mol. Biol.* **48:** 443–53.
Singh, S. 1999. *The Code Book.* Doubleday, New York, NY.
Stormo, G.D. 2000. *Genome Res.* **10:** 394–397.
Uberbacher, E.C. and Mural, R.J. 1991. *Proc. Natl. Acad. Sci.* **88:** 11261–11265.

Commentary

Jean-Michel Claverie

Structural and Genomic Information, CNRS-UPR 2589, Marseille cedex 20, France

I found this meeting very productive, probably because of its exceptionally well-chosen attendees in representing the high quality and variety of bioinformatic activities currently going on in Europe.

The workshop was quite unanimous on the followings aspects:

1) Bioinformatic Resources Centres (data centre/databases) and Bioinformatic Research activities should not have to compete against each others for funding through EC instruments. There are both equally valuable activities, complementary to each other. For the research side, it was pointed out that much valuable bioinformatics progress come from small laboratories, not only from a few big centres. These relatively isolated laboratories should be encouraged to establish more links and thus gain access to funding via the EC instruments: Networks of Excellence or Integrated Projects.

2) It was reiterated that to be recognised as "Bioinformatics", research projects (in applied maths, statistics, data mining, algorithmics, or any other IT developments) have to be CLEARLY linked to well identified Biological/Biomedical problems: e.g. identifying genes, predicting their function or their regulation, designing a vaccine. This was the unanimous concept of "Biology-driven bioinformatics".

3) Finally, "System Biology" was the other unanimous keyword. What this meant, I think, is twofold:

a) Bioinformatic research and resources have to gradually evolve from being "individual part-oriented" (genes, proteins, structures) to trying to address bigger biological pictures (such as protein-protein interaction, regulatory network, mutation to phenotype, cellular subsystems), thus aiming at generating an understanding (a *predictability of behaviour*) of biological subsystems (mitochondria, spliceosome, etc.),

b) Given that no simple "mathematical formula" will ever be able to encompass the complex behaviour of biological systems, the ultimate bioinformatic theory is expected to consist of a computational model able to realistically simulate the dynamic behaviours of these systems, based on molecular "first principle" such as sequences, 3-D structures, interaction patterns, and physico-chemistry parameters (kinetic and affinity constants).

In my opinion, the realistic computer modelling of an entire E. coli cell is the prototype of high-visibility, ambitious, visionary project capable of federating all the components of today's bioinformatics activities (Bio-specific IT developments, data integration/visualisation schemes, algoritmics, applied maths) as well as new contribution from physics and engineering disciplines. Such a project is pushed forward as a 10-year project by the ***International E. coli Alliance (IECA),*** already in search of preliminary funding in the US, Japan and Canada. This ambitious project will also foster progresses in techniques of biotechnological/biomedical interest such as high throughput parallel measurement of protein abundance and activity, as well as mathematical simulation techniques, with the long-term goal of replacing the need for animal/cellular testing. IECA's activities already elicit considerable interest from the pharmaceutical industry (e.g. GSK). I am pushing for the funding of a European branch of this ambitious and truly international project that appears well adapted to the current EEC instruments (Integrated project).

Howard Bilofsky, PhD
GlaxoSmithkline R&D IT, 709 Swedenland Rd, 19119 King of Prussia, PA, USA
(5 Dec 2002 PRISM Forum)

Pharmaceutical R&D IS Managers Forum; Scope is the use of Information Technology to impact R&D Processes; The mission of the PRISM Forum is to; share pre-competitive information and best practices ; define requirements for standards to support information exchange across the R&D process. The Forum is open to individuals able to represent their companies with respect to the above ; Meets twice a year, normally once in Europe and once in the USA. Last few meetings had reps from: **Biovitrum, Lilly, AZ, BMS, GSK, Novartis, Schering-Plough, Wyeth, Roche, J&J, Pfizer, Amgen, Lundbeck;** 2003 Meetings - Princeton (Spring) and Madrid (Fall)

Objectives

Recognise that Pharma R&D represents practical compelling challenges for Informatics?; These are drivers (Scientific and Business) for new:; BIx (OmIx, CIx, MIx, …) techniques ; *Domain Knowledge-based Informatics*; Knowledge Engineering, Management & Collaboration; Modeling and Simulation

Challenges to Well-Founded R&D Decision Making

Deep Understanding; Consensus; Comprehensive, Timely and Relevant Information; Complex Diverse Knowledge Domains; Verifiable Models

Challenges

Access to and understanding of distributed, heterogeneous information resources is critical but; Complex, time consuming process, because: 1000’s of relevant information sources; Rapidly changing domain concepts and terminology and analysis approaches; Constantly evolving data structures ; Continuous creation of new data sources; Highly heterogeneous sources and applications ; Data and results of uneven quality, depth, scope; But still growing; Collaboration for understanding and consensus is essential - internal, external...

Collaboration - Most Broadly

Within the Organisation; across the org functionally and geographically (world-wide); along the pipeline and up the hierarchy; Externally With Others:; Pharmas ; Biotechs ; CROs; Clinical Investigators ; Academics; Advisors; Regulatory Agencies

e-Collaborations

Opportunities and Constraints

Data (and Computationally)-Intensive Sharing of Complex (Content and/or Context); Complex Knowledge Domain and Business Processes Models; Secure, Confidential, Standards-based, Able to Audit and Validate; Efficient, Economic, Robust, Scalable, Replicable, Widely Available

Argues for Life Science focussed...

Deep Extensive Ontologies, Models and Standards (not ends in themselves); Open Source; Intelligent Integration based on Mediator Architectures; Data GRID

Acknowledgements

Helpful discussions and innovative suggestions from numerous colleagues in GSK especially David Searls, Richard Fritzson and W. David Benton and PRISM especially Chris Jones, CERN.
Feedback to: Howard.S.Bilofsky@gsk.com

Anna Tramontano -

COMMENTARY: Present bioinformatics research in Europe and training needs; Recent advances in protein structure modeling

Department of Biochemical Sciences "Rossi Fanelli", University of Rome "La Sapienza",
P le Aldo Moro, 5, I - 00185 Rome

Comments on present bioinformatics research in Europe and training needs

The organisation of bioinformatics research in Europe is based on both large, service providing, groups and small research groups. The latter, in some cases located within the larger service providing Institutions, have been extremely active in Bioinformatics European research. The strengths of these small groups, that should be reinforced and exploited, include their proven ability to develop many useful tools and their role as key connection points between computational and experimental biologists. Thanks to this close connection, these groups have very often shown the ability a very creative use of existing tools in novel areas and they have been instrumental in identifying novel routes of interest for experimental life sciences.

In Europe the level of co-operation and networking between these groups is quite advanced, sometimes fostered by joint funded research, more often driven by the needs of the biological “clients” that each of the Bioinformatics groups needs to serve. This is a quite unique aspect of bioinformatics research in Europe that should be supported and expanded.

There are several problems that Bioinformatics research groups face:

1. Under-funding. Local government support is quite variable within Europe so that effective collaboration between different national groups can be hampered by a very diverse level of support.
2. Scope of activities. The requests posed by the surrounding experimental community often are extremely heavy for local groups. Bioinformatics has many diverse aspects. On one side they require a very diverse set of expertise, on the other their concurrent use is often the key to success in a post-genomic project. Incidentally, we believe that the need to meet the diverse requests of the experimental collaborators is the underlying reason why these groups have developed a high level of networking and collaboration within Europe.
3. European level funding. While these groups heavily rely on the existence of service provider Institution, so far they often had to compete with them for resources.
4. Training. Given the lack of a general bioinformatics training at the national and European level, most small groups need to train young scientists in bioinformatics starting from a purely biological or computational background.
5. The relationship between small bioinformatics groups and the local experimental community has included collaboration with industrial, especially pharmaceutical, partners. Genomic research will provide pharmaceutical companies with an increased number of potential targets for therapeutical intervention. The bottleneck for effective use of this information will move from target identification to target validation. Bioinformatics research groups will need to devise appropriate computational experiments to help prioritise experiments on the potential targets. This will require a close relationship between Bioinformatics and other branches of science, such as pathology, physiology, cell biology, etc.
6. There are clear signs of a fast development of closer relationships with the medical community. While this is a welcome development, it will represent yet another area in which local groups will have to become proficient and effective.

The above analysis clearly points to actions that could be taken to ensure that the competitiveness of European research is not compromised:

1. Schemes for funding basic research for small networks of national groups, taking into account the different level of funding provided by different countries.
2. Platforms for exchanging expertise and collaborative projects, for example by fostering the creation of a road map of Bioinformatics in Europe
3. Separate funding schemes for service and basic research
4. Framework for training in bioinformatics.
5. A funding scheme for joint projects between Bioinformatics research groups and pharmaceutical company.
6. Strategies for taking the opportunity of medical informatics as a platform for collaboration rather than competition between bioinformatics and the medical community.

Comments on recent advances in protein structure modeling

There are very interesting novel results in protein structure modelling that allow us to start questioning some of our previous basic assumptions, for example the assumption that constructing a three-dimensional model is a step-wise process and that each of the modelling steps can be optimised separately.

As of today, the main development that seems to be the key to build better models is based on the idea of constructing several models for each target protein and of selecting the most likely one only at the end of the complete model building procedure. In other words, rather than optimising independently each of the steps of the procedure, the most successful methods have adopted the strategy of funnelling into each subsequent step not only the optimal but also the sub-optimal intermediate results and to evaluate the final models at the atomic level. This represent a first degree approximation to a full multi-parameter optimisation procedure, which is clearly a very complex computational problem and we should start developing strategies to tackle it effectively in the near future.

SUMMARY OF PRESENTATIONS BY SESSION CHAIRPERSON

Chair - Jaap Heringa

FUTURE BIOINFORMATICS RESEARCH TOPICS - Integration of knowledge, systems biology, health, pharmaceuticals, biotechnology, environment

Unifying themes:

-possible other application avenues of processed genetic information and their implications in the development of bioinformatic tools: is it towards medical application (including drug development), etc.

-analysis of the possible scientific developments in the field and in new research areas: what research solution might influence the appropriate development of bioinformatic tools (more interoperability, more data access, more integration at the level of ontologies, new theoretical models, else ?)

A number of keywords dominate the discussion.

Integration of knowledge:

Ontologies are key, but moving from genotype to phenotype gets harder

Never throw away primary data.

Remember that biology is always a central element

Biology interpretation always needs the biological context.

DATABASE MAINTENANCE NEEDS TO BE FUNDED.

Genomics

Identification of genetics elements is still poor. We should collect all the data we can. Many of the interpretative problems are still unsolved.

Genomics was initially driven by data collection, and interpretation is still mostly hypothesis driven. A key question is always the interaction with biologists.

For genomics, we need many more databases and more biological knowledge.

There is a problem with static format of databases. We may need to restructure or create more secondary databases.

Systems Biology

A key question is what is it?

The other question is should it be funded at EU level. It is certainly an important area.

Medical informatics: Problems of ontologies is difficult.

It could provide examples for bioinformatics, where people have worked on common problems. However, bioinformatics has achieved much wider usage. When bioinformatics reaches the level of dealing with people oriented data, the same problems may appear.

A danger in training : Do not separate the teaching from the basic discipline

Experimental design

Bioinformatics can play a key role in experimental design.

Algorithm design

They need to be more biologically oriented.

Physics and engineering contributions are also key.

Level of knowledge

We are missing a lot of basic knowledge, e.g. transcription

In systems biology, there is a lot of progress. A world-wide E.coli project is being studied.

Systems biology is becoming a big field :

Pharmaceuticals

There is an enormous range of collaborations in pharma. A major danger is the tendency to store a minimal amount of data, usually in processed form. We should store a wide range of data.

The more medical becomes the research, the more confidentiality becomes a key issue.

Open source is a valuable approach.

Environment

This is a key area of research, e.g. Netherlands.

Example of the wide range of uses.

We need to streamline methods of data analysis.

Five to ten years, what will be new and fundamental

Web services will be increasing

A new direction is to provide simple access to tools so as to be able to access all databases and tolls quickly.

GENERAL DISCUSSION AT END OF MEETING

Chairs - Martin Bishop, Miklos Gyorffi, Frederick Marcus

Subjects for Calls for proposals

1. Knowledge representation – includes ontologies, modelling
2. Algorithmic development – It is good to organise data ; but we must interpret it
3. Something that links bioinformatics to health, like clinical genomics
4. Text managing and collecting and analysing text based information, ability to integrate information from various sources, e.g. Biomoby in USA, Knowledge representation in combination with visualisation of text
5. Building generalised model for biological systems
6. Comparative genomics leading on to functional studies
7. Data integration is key: a–data ; b-function ; then we need web services to develop this, connecting to all fields ; e.g. clinical
8. Integrative and analysis tools and algorithm development
9. Improvement of sequence analysis tools and text mining tools and new solutions for knowledge representation
10. Intelligent image retrieval
11. Mathematical and statistics methods for bioinformatics analysis
12. Computational proteomics - there is a need to maintain SWISS-PROT lead
13. Immunologogical bioinformatics – need in vaccine design area and also protection against bioterrorism
14. Formal representations of knowledge – ontologies plus and use of representation as research tools in their own right and dealing with this from engineering point of view is untapped
15. Application driven immunological bioinformatics e.g. asthma, autoimmunity, regulation of immune responses.
16. European branch of international E.coli project
17. Systems biology in general
18. Gene regulation combining experiments and a computational framework.

Establish the molecular basis of human variation

Best working methods

i. Methods that determine biological function
ii. Database building should have an XML screen so other people can use tools themselves, to help everyone
iii. It was generally felt that a significant fraction of funding resources in each major project should be devoted to bioinformatics and data analysis, including storage, retrieval and analysis of data during the life of the project, plus consideration given to long term maintenance. While the percentage will depend on the nature of each experiment, experience shows that a level of 20% was typically present in large projects, and that this reached 50% or more in larger genome projects and in particular areas like microarray experiments. Smaller projects will also need significant data analysis support.
iv. Integrative algorithms in best sense. An example is functional assignment.
v. Whenever a data selection is started, we are bound to flat file philosophy. We should shape databases to biological basis. We need to develop novel data structures

General Commentary

a) EU Commission should take note of roles of bioinformatics companies. Professional services are needed. We should consider how to support this and how to integrate them into funding : For small companies money is needed for survival
b) For companies, a market is needed, and networking is needed. Better links to scientists is needed. Life is also very expensive for small companies for access to databases.

c) Collaboration is very important for industry. Even collaboration within the Commission is important. The DATAGRID is an important opportunity. Money is important, but it is not the major motivation. Companies have broader perspectives.
d) We need NCBI-like supported structure where databases are available for scientific community.
e) It is surprising that Europe does not have a USA SBIR-like (http://www.sba.gov/sbir/indexsbir-sttr.html) peer review system. The role of companies is acknowledged. CRAFT did exist in last FP, and had problems.
f) New genomes gives you information on the old ones. It improves the value of old data. New technology has reduced greatly the cost of each new sequence. Each new genome helps to understand regulatory elements. Only by comparison can you understand fully.
g) The haplotype analysis for disease identification will be a key element of individualised health care and understanding disease, and we will need sophisticated analysis tools. Personalised treatment is a key goal.
h) A lot of haplotype information can be gained by comparative genomics.
i) It is necessary that all projects have a good percentage of bioinformatics, in order to properly store and analyse data. Very few genomes are properly annotated, and this represents a major opportunity.
j) A better analysis of future needs may well depend on current funding rounds.
k) A role for SMEs could be maintenance of databases, although there is no obvious business model for maintaining it. One model is supporting researchers, but it is a bad business. One company model is based on back end servers and other services. It is a rough time for small companies. Software development can be considered.
l) Sequence data is extremely valuable. New sequences are very good value for money. Marginal value of new genomes is still important. Unfortunately apart from Sanger, Europe dropped out of sequence. We can foresee having a thousand genomes sequenced. A sequence is a needed basis for many modern experiments. SWISS-PROT searches show what a valuable resource this is.