What is bioinformatics?

• The science of storing, retrieving and analysing large amounts of biological information

• An interdisciplinary science involving biologists, biochemists, computer scientists and mathematicians

• At the heart of modern biology
1993
The term “Bioinformatics” starts to appear in the Literature

- 1987
  - Nucleic Acids Res. 15: 7581–7592
  - Friedrich Miescher Institut, Bioinformatic group

- 1989
The molecules of life: the central dogma
Understanding Life: from molecules to systems

• Parts Dictionary

• Thesaurus (interactions)

• Complete Networks (Circuit Diagrams)

• Atlas – what happens where

• How it works & when (simulations)

• How it goes wrong - & how to put it right
From Molecules to Organisms

Genome → Protein/DNA → Cell → Embryo

Fruitfly → Mouse → Human health
The flood of biological data

- **EMBL-Bank**
- **Ensembl**
- **ArrayExpress**
- **UniProt**
- **PDBe**
- **InterPro**
Why is biological information important?
Genome-wide analysis of crop plants

- Population growth and climate change are major challenges to food security.
- Traditional routes to crop improvement are too slow to keep up with this increase in demand.
- Understanding plant genomes helps us identify which species will be most tolerant to drought, salt and pests while still providing optimum nutrition.
Matching the treatment to the cancer

- One in ten women in the EU-27 will develop breast cancer before the age of 80.
- If we can identify patterns of genes that are active in different tumours, we can diagnose and treat cancers earlier.
Tracking the source of infectious disease

• Methicillin-resistant MRSA (*Staphylococcus aureus*) infection is a global problem.

• Transmission of individual clones can be tracked using small variations in DNA sequence.

• This technology can be used to identify the source of new outbreaks across continents and within wards.
Barcoding life

- DNA barcodes are short sections of DNA that we use to identify an organism.
- The Barcode of Life Initiative is developing DNA barcoding as a global standard for identifying species.
- Applications include:
  - Protection of endangered species
  - Sustaining natural resources through pest control
  - Food labelling
Repurposing drugs for neglected diseases

- Schistosomiasis is a parasitic infection that affects 210 million people in 76 countries.
- Resistance is developing to the one available drug.
- We look at the Schistosome genome to identify the targets of existing drugs.
- Candidates can be tested for anti-schistosomal activity or used as leads for further optimisation.

Image courtesy of CDC Public Health Image Library
Personal Genomes

• The era of personal genome sequencing is upon us.
• Sequence data will not cross national boundaries.
• Every national health system will need expertise to interpret it and treat patients accordingly.
• Individuals need to be sure that their personal biological data are in safe hands.
Supporting industry

- Cosmetics
- Forestry
- Biotech
- Fisheries
- Big pharma
- Computing
- Environmental protection
- Agri-food
- Medical devices
- Biofuels
The Bioinformatics Challenges

• Technical: handling & integrating the data

• Academic: Understanding what it means

• Translational: Using knowledge to make predictions and benefit society
The technical challenge: Data Flood

- Computer speed and storage capacity is **doubling every 18 months** and this rate is steady.
- DNA sequence data is **doubling every 9 months** and this rate is increasing.

![Graph showing the doubling of new biological data and computer storage and processor power over time.](image-url)
The fundamental challenge: From Molecules to Organisms

Genome & Variation

Computational Biology

Human/Disease Data
From Molecules to Organisms

Protein/DNA

Structural Biology

Gene Expression

Macromolecular Complex

Subcellular Structure

Cell Biology

Developmental Biology

Cell

Organism

Computational Biology
The translational challenge: From molecules to medicine

Molecular components
- Genomes
- Nucleotides
- Transcripts
- Proteins
- Complexes
- Pathways
- Domains
- Structures
- Smal molecules

Integration
- Cells
- Tissues and organs
- Biobanks

Translation
- Human individuals
- Therapies
- Disease prevention
- Early Diagnosis
- Human populations
Biology is changing

- Data explosion
- New types of data
- High-throughput biology
- Emphasis on systems, not reductionism
- Growth of applied biology
  - molecular medicine
  - agriculture
  - food
  - environmental sciences…
EMBL

- Basic research in molecular biology
- Administration
- EMBO

- 1500 staff
- >60 nationalities
From Systems Biology to Information Biology

- Challenge in systems biology is shifting from data production towards developing new ways to use data to produce novel biological information.

**Information Biology** combines two concepts;

The requirement for computational methods to cope with both data quantity and biological complexity.

Understanding information flow (logic) from genotype to phenotype in biological systems.
Bridging dimensions: from molecules to cells to organisms

Challenge: Determine relationships between structure and function

- Bridging molecular and cellular resolution: from protein-protein interactions to networks in cells
- Structural Cell Biology: a structure-based, mechanistic understanding of cellular processes
- From cells to organisms: dynamic organization and imaging
Biology in four dimensions

Developing and applying new technologies for real-time live imaging at the molecular, cellular and organismal level to study dynamic processes in space and time
Unravelling Biological Complexity

Challenge: understand information flow in biological systems

- Biological functions arise from the combined action of multiple molecular or cellular components
- Living systems behave non-linearly: one input often produces multiple outputs

Strategies

- High-throughput technologies: record multiple components of a system
- Computational methods: analyse experimental data
- Detailed analyses: quantify critical aspects of system behaviour
- Modelling and simulation: quantitatively describe biological form and function at the sub-cellular, cellular and supra-cellular level and derive predictions about emerging properties that can be tested experimentally
Exploring Biological Variation

Challenge: link genetic variation to phenotypic variation

- **Evolution: inter-species variation**
  - Next generation sequencing provides new capacity for comparative study of genetic variation between species, including unexplored species

- **Genetic variation: intra-species variation**
  - Unprecedented data on intra-species variation
  - Goal: Develop a general methodological framework to analyse the contribution of genetic variation to any phenotype

- **Disease models and mechanisms**
  - Use human sequence data and model systems to understand disease
Bioinformatics and Computational Biology

Challenge: analysing, integrating and exploiting quantitative data

• Crucial enabling techniques for systems biology and information biology
• Generating new knowledge and hypotheses to guide experiments

Plans 2012-2016

• Develop new ways to analyse, integrate and represent of heterogeneous data generated in large quantities by experiment
• Build increasingly accurate models and simulations of a variety of biological processes
Research in Bioinformatics

All biologists will need to be computational biologists
Survey: time spent on computational activities

- 2009 survey shows “wet lab” researchers at EMBL HD spend ~40% of their time on computational activities.

Number of groups

Average: 33%

Average: 40%

Number of groups

% people allocated per group

0 2 4 6 8 10

% computational work per group

0 2 4 6 8 10

Percentage of people in a group with >=50% of their time devoted to (bio)computing or data analysis

Percentage of all time of all people in a group with bioinformatics/computational activities
Many computational biologists are beginning to generate their own data.
Research themes at EBI

**Genomes**
- Nick Goldman
- Ewan Birney
- Paul Flicek

**Text mining**
- Dietrich Rebholz-Schuhmann

**Proteins**
- Janet Thornton
- Rolf Apweiler
- Gerard Kleywegt

**Transcriptomes**
- Anton Enright
- John Marioni*
- Alvis Brazma

**Chemistry**
- Christoph Steinbeck
- John Overington

**Pathways and systems**
- Nicolas Le Novère
- Nick Luscombe
- Paul Bertone
- Julio Saez-Rodriguez*
Bioinformatics in multi-disciplinary Proteins

Genomes
- Ewan Birney
- Paul Flicek
- Nick Goldman

Transcriptomes
- Alvis Brazma
- Anton Enright
- John Marioni

Proteins
- Janet Thornton
- Rolf Apweiler
- Gerard Kleywegt

Pathways and systems
- Nicolas Le Novère
- Nick Luscombe
- Paul Bertone

Text mining
- Dietrich Rebholz-Schuhmann

Chemistry
- Christoph Steinbeck
- John Overington

biology/medicine
chemistry/chem eng
maths
physics
Examples of EMBL-EBI research

What is the molecular basis of ageing?

How do the neurons of someone with Parkinson’s disease signal differently from healthy neurons?

What makes a stem cell decide to become skin or muscle?

Which of these proteins will make good targets for drugs?

Which of these changes to a genome’s structure drive cancer?
New Challenges for Bioinformatics Research

- **New Sequencing Technologies**
  - Human variation
  - Intra vs Inter species variation
  - Dynamics of Genome Evolution
  - Linking population and molecular studies
  - Variation & Disease
  - Biodiversity
    - Barcode for life
    - Plant and animal breeding

- **Multi-’omics Data**
  - Data integration technologies
  - Time series
  - Atlases/Tissue and Cell type specificity

- **Biobanks**
  - Molecular and Phenotypic data
  - ‘Molecular-based phenotypic studies
  - Multi-omic data sets

- **Opening up of Literature**
  - Linking literature to databases and vice versa
  - Text Mining
  - The semantic web

- **Immense Complexity of Control networks**
  - New modelling techniques
  - Commissioning of experiments

- **Development of Cloud Computing and Grid technologies**
  - New methods for handling data

- **New Image Technologies**
  - Analysis

- **Move of Medicine towards ‘Molecular Technologies’**
  - Molecular Diagnoses
  - Pharmacogenomics
  - Target Identification
The translational challenge: From molecules to medicine

**Molecular components**
- Genomes
- Nucleotides
- Transcripts
- Proteins
- Domains
- Pathways
- Structures
- Small molecules

**Integration**
- Complexes
- Tissues and organs
- Biobanks
- Cells
- Human individuals

**Translation**
- Human populations
- Therapies
- Disease prevention
- Early Diagnosis
We need to build bridges between the biological information and the clinical data
Bridges need strong anchors at both ends!
In 2009, the House of Lords Science and Technology Committee Report on Genomic Medicine was published. Recommendations 23 and 24 dealt with the issue of bioinformatics and specifically recommended:

“... the establishment of a new **Institute of Biomedical Informatics** to address the challenges of handling the linking of medical and genetic information in order to maximise the value of these two unique sources of information. Such an institute would bridge the knowledge, culture and communications gap that currently exists between the expertise in NHS IT systems and bioinformaticians working on genome research. The Institute would guide the NHS in the creation of NHS informatics platforms that will interface with databases containing personal genetic data and with publicly available genome databases”

and

“... the Department of Health should establish a **centre for national training in biomedical informatics** (within the Institute of Biomedical Informatics) with the aim of providing training that bridges the gap between health records information technology and genome informatics, and ensuring the delivery of an expert workforce for the NHS”
The role of a National Biomedical Informatics Institute (BII)

The National Biomedical Informatics Institute should inspire and lead the application of genomics to healthcare in the UK by:

- Developing and providing biomedical data and informatics services (focussed on genetics and genomics data) that are fit for purpose for use in the NHS

- Performing investigator-led world class research in biomedical informatics to enable translational medicine

- Providing training in biomedical informatics for researchers and NHS staff

- Providing an interface to industry, including technology transfer
Position of a National Biomedical Informatics Institute

- Research
- International Data
- UK Research
- Standards
- Data
- Biomedical Informatics Institute
- Anonymised data
- Translation
- NIHR
- OSCHR
- SMEs
- TSB
- Standards
- Interpretation
- RCP Archive
- Standards
- Test
- Standards
- Patient
- Clinical Practice
- Education
- User training
EBI’s role...

A repository for public knowledge and data

- Instituto de informática biomédica (ES)
- Biomedical informatics Institute (UK)
- Institut für biomedizinische Informatik (DE)
- Instituts d’Informatique Médicale (FR)
From the perspective of a National BII

Institut für biomedizinische Informatik (DE)

GINI Cohort München

DKFZ, Cancer Cohort

16 Lander Insurance Frameworks

683,000 Beds in ~5,000 Hospitals
Whilst this is being developed

- UCL, UCH Hospital
- Sanger Institute
- Estonia BioBank
- Berlin Max Plank, Charitie
- ITFoM
- CRG, Barcelona, IMI
- Dundee, Tayside Cohorts
- Elixir, BBMRI etc
An Infrastructure for biological information

Flow of data from laboratory

Lab => Local database => National database (?) => International Database
The cost of storing the data is small compared to the cost of generating it

Annual cost of generating new protein structure data in labs around the world

Annual cost of maintaining it in a central database
EBI’s Databases: molecules to systems

- **Genomes**
  - Ensembl
  - Ensembl Genomes
  - EGA

- **Nucleotide sequence**
  - ENA

- **Protein Sequences**
  - UniProt

- **Protein activity**
  - IntAct, PRIDE

- **Chemical entities**
  - ChEBI

- **Chemogenomics**
  - ChEMBL

- **Functional genomics**
  - ArrayExpress
  - Expression Atlas

- **Literature and ontologies**
  - CiteXplore, GO

- **Protein families, motifs and domains**
  - InterPro

- **Macromolecular**
  - PDBe

- **Pathways**
  - Reactome

- **Systems**
  - BioModels
  - BioSamples
All data resources are part of a global collaborative effort
ELIXIR: Europe’s emerging infrastructure for biological information
Why ELIXIR?

• Creating a robust infrastructure for biological information is a bigger task than EMBL-EBI – or any individual organisation or nation – can take on alone.

• Biology has by far the largest research community:
  • ~3 million life science researchers in Europe
  • >4 million web hits a day at EMBL-EBI alone
  • 1 million unique users per year

• We need to involve other European partners
ELIXIR’s mission

To build a sustainable European infrastructure for biological information, supporting life science research and its translation to:

- medicine
- environment
- bioindustries
- society
A distributed pan-European infrastructure
ELIXIR Nodes

- National Coordination Nodes
- Data nodes
- Compute nodes
- Biological domain nodes
- Links to other BMS projects (See later)
- Tools infrastructure nodes
- Training centres for users of ELIXIR resources.
54 Node suggestions; 23 countries

ELIXIR Node proposals as of Apr 2011
ELIXIR Networks

Data Network

Compute Network

Tool I/S Network

BMS Network

Training Network

NGS Network
The Future

• More Data
• More Complex problems
• More Modelling
• More applications and relevance to society

More Bioinformatics, involving many more scientists around the world