EMBL-EBI: Providing Bioinformatics Research Infrastructure for the Life Sciences

Rolf Apweiler
Director, EMBL-EBI
www.ebi.ac.uk
What is EMBL-EBI?

• Europe’s home for biological data services, research and training
• A trusted data provider for the life sciences
• Part of the European Molecular Biology Laboratory, an intergovernmental research organisation
• International: >600 members of staff from >60 nations
• Home of the ELIXIR Technical hub
Our Funders

- EMBL-EBI is primarily funded by EMBL member states
- Other major funders:
  - European Commission
  - Research Councils UK
  - National Institutes of Health
  - Wellcome Trust
  - Industry
How we are organised: overview

Services

- Genes, Genomes & Variation
- RNA, Protein & Metabolite Expression
- Proteins & Protein Families
- Molecular & Cellular Structures
- Chemical Biology
- Molecular Systems
- Cross-Domain Resources
- Archival Resources

Research

- Genomics - Sequence analysis methods
- Proteins & Protein Structures - Multi-dimensional statistical analysis
- Chemical Biology - Data-driven biological discovery
- Pathways & Systems – from plant biology to mammalian development and disease

Training Programme • Industry Programme •
Systems Applications • Systems Infrastructure • Administration • Web Production •
Web Development • External Relations

Rolf Apweiler and Ewan Birney
Directors
Our mission

Deliver excellent research

Deliver scientific services

Train the next generation of scientists

Engage with European industry

Coordinate bioinformatics in Europe
EMBL-EBI Service Mission

To enable life science research and its translation to medicine, agriculture, the bioindustries and society by providing biological data, information and knowledge
What services do we provide?

- Labs around the world send us their data and we…
- Archive it
- Classify it
- Analyse, add value and integrate it
- ...provide tools to help researchers use it
- A collaborative enterprise
- Share it with other data providers
Data growth in 2016

- Nucleotide sequence data
- Genomes (all species)
- Gene expression data
- Protein sequence data
- Protein families, motifs, and domains
- Macromolecular structures

Growth by platform

- Sequence
- Mass Spectrometry
- Array

Date

2003 2005 2007 2009 2011 2013 2015 2017
Data resources at EMBL-EBI

Genes, genomes & variation
- Ensembl
- Ensembl Genomes
- GWAS Catalog
- Metagenomics portal

Gene, protein & metabolite expression
- Expression Atlas
- Metabolights
- PRIDE
- RNA Central

Protein sequences, families & motifs
- InterPro
- Pfam
- UniProt

Molecular structures
- Protein Data Bank in Europe
- Electron Microscopy Data Bank

Chemical biology
- ChEBI
- ChEMBL
- SureChEMBL

Systems
- BioModels
- BioSamples
- Enzyme Portal
- IntAct
- Reactome

Molecular Archives
- European Nucleotide Archive
- European Variation Archive
- European Genome-phenome Archive
- ArrayExpress

Literature & ontologies
- Experimental Factor Ontology
- Gene Ontology
- BioStudies
- Europe PMC
Major Database Collaborations

- Our data resource teams collaborate with organisations throughout the world
- Internal collaboration essential
- Integration of public molecular data for the global scientific community
- Network makes open data possible and sustainable.
Big data, big demand

>37 million requests to EMBL-EBI websites every day

Scientists at over 5 million unique sites use EMBL-EBI websites

EMBL-EBI handles > 14 million jobs on average per month

> 200 petabytes of EMBL-EBI storage capacity
Training
For scientists working at all levels

www.ebi.ac.uk/training
The EMBL International PhD Programme

First-class training and education

• Internationality
• Dedicated mentoring
• Early independence in research
• Europe’s ‘best in class’ PhD programme for the life sciences
• 200 students from over 40 countries (around 30 at EMBL-EBI)
• EMBL-EBI graduates gain their degree from the University of Cambridge
Around 50 Postdocs at EMBL-EBI

Interdisciplinary fellowships

• **EI₃POD** projects bring together scientific fields that are usually separate, or transfer techniques to a novel context.

Genome Campus fellowships

• **ESPODs**, joint EMBL-EBI and Sanger Institute fellowships, combine experimental and computational approaches

Computational biomedicine fellowships

• **EBPODs** are joint projects with EMBL-EBI, NIHR Cambridge Biomedical Research Centre and the University of Cambridge School of the Biological Sciences
Advanced Bioinformatics Training

Online training in bioinformatics
• Providing free access to EMBL-EBI courses and materials.
• Empowering individuals to choose when, where and how they learn.

On-site courses
• Delivering courses to develop practical skills and knowledge.
• Led by EMBL-EBI experts and hosted in our purpose-built training suite.
Research

www.ebi.ac.uk/research
Data-driven discovery

The Genome Campus: a unique setting for research.

• Super-rich technical environment;
• One of the world’s highest concentrations of expertise in genomics and Big Data in biology;
• Research and service activities are mutually supportive.
Research groups at EMBL-EBI
New group leaders: Translational bioinformatics
Industry engagement
Support and collaboration

www.ebi.ac.uk/industry
Supporting industry

• EMBL-EBI is a champion of open data in the life sciences.
• We support industry by providing freely available data services.
• Our Industry Programme provides neutral ground where companies can meet and explore new developments and shared challenges.
• We welcome interactions with industry through:
  • Pre-competitive collaboration
  • Standards development
  • Technical development
  • Services development
Industry Programme members
Bioinformatics goes translational
The Open Targets example
Supporting the coordination of biological data provision in Europe

EMBL-EBI as an ELIXIR Node

www.elixir-europe.org
ELIXIR

Supporting life science research and its translation to medicine, agriculture, bioindustries and society by building a sustainable European infrastructure for biological information.
ELIXIR members

• ELIXIR members host Nodes, including EMBL-EBI, which represent centres of excellence in bioinformatics.

• ELIXIR aims to empower its member states to integrate data and services, and to maximise their bioinformatics expertise.
Looking ahead
Human Cell Atlas: Data Coordination

- Open initiative to chart genetic properties of all human cells, build new reference map of body
- Billions of cells to be profiled over 5 years > petabytes of data
- Collaborating to build cloud-based Data Coordination Platform for open, immediate access
- Several new positions at EMBL-EBI, hackathons
- End-to-end pipeline (beta) in place
- scRNAseq pipeline mid-2018, support for images
The Human Cell Atlas Data Coordination Platform
Bioinformatics goes translational
Genomics: from research to healthcare

Research

• English language
• Light-weight legal
• Similar systems
• Open data
• Publications
• Grant funding

Practicing Medicine

• National language
• Heavy legal framework
• Different systems
• Closed data
• Not published
• Contract funding
Nurses in hospitals collect data

Few tools for data analysis since data not available

Rarely stored in a computer-friendly systematic way

Clinicians interpret data & treat patients

Data analysis limited to a few research hospitals

Little data sharing

A not-so collaborative enterprise
Long term goal for clinical data in Europe

• Ideally each European nation state should establish an “Institute for Biomedical informatics” and EMBL-EBI should interact with them

• In bigger nations, this is likely to be a network, but with a centre of gravity

• EMBL-EBI (and ELIXIR) will handle research style data, sharing it between researchers (including clinical researchers)

• These “BMI” institutes will have responsibility for exploiting molecular reference data and ensure “open research” data flows back to enhance the core reference data at EMBL-EBI and is made available to researchers
EMBL-EBI long term goal

Skill set:

- Honest Broker for data sharing
- Long-horizon information management
- Integration across biological domains
- Large scale disk, compute, network
- Knowledge extraction/management
- Algorithms and Statistics
- International and Consortium interactions
Global standards: the GA4GH

- GA4GH is the standards-setting body for genomics in healthcare
  - Embraces federated approach
  - Setting community standards early
- Cloud: Analysis carried out where the data ‘lives’
- “You’re already using it!”: SAM/BAM/CRAM/VCF formats
- Tools: htsget – the first step away from file-based access
- Rare disease diagnoses: Matchmaker Exchange
- Federated discovery: GA4GH Beacons
Federation

Open research data

Aggregate data globally
Download, analyse locally

Healthcare data with research use

Analyse data locally (via VMs)
Collate analyses
Thank you