Global Biomolecular Information Infrastructure and Australia

Graham Cameron
Director
The EMBL Australia Bioinformatics Resource
What is bioinformatics?

- Methods, data, IT to exploit biomolecular information to understand living systems
- Biology, biochemistry, computer science, mathematics
- At the heart of modern biology
- Requires extensive information infrastructure
DNA

About 12 base pairs

There are about 3 billion in the human genome
<table>
<thead>
<tr>
<th>Genomes contain genes</th>
<th>Genes are transcribed</th>
<th>Transcripts translate to protein sequences</th>
<th>Proteins form complex three dimensional structures</th>
<th>Proteins interact with each other and with small molecules</th>
</tr>
</thead>
</table>

Central Dogma
Societal impact

• Basis of all living systems
• Obvious application - human health and wellbeing
  – Human biology
  – Biology of human pathogens – viruses, bacteria, parasites

• The entire biosphere
  – From crop plants to rain forests
  – From coral reefs to brewers yeast
  – From farmed fish to fruit trees
It’s all about molecules

• Different molecules are required:
  – in different tissues
  – at different developmental stages
  – in response to environment

• Levels of transcription and translation:
  – influence molecular interactions and pathways
  – can be a consequence/cause of pathologies

• Genetic differences cause differences in gene products:
  – can be benign
  – can cause pathologies
  – can influence propensity to pathologies
A bit of a shake up

• Genome is about 3 billion base-pairs
• <1% of the genome is genes
• Some of the rest controls how, when and where genes are switched on and off
• Is the rest “junk”?
Complexity

• We expected about 100,000 genes
• We found about 20,000
• A lot of the complexity must be elsewhere

• Actually about 20% of the non-coding DNA is known to be functional
• A further 60% is transcribed
• 90% of disease causing SNPs are outside coding regions
Understanding, diagnosis, treatment

• By analysing genomes, genes and gene products we can
• Understand the mechanisms of health and disease
• Diagnose disease
• Design drugs to treat disease
Using the information

<table>
<thead>
<tr>
<th>Healthy</th>
<th>Diseased</th>
</tr>
</thead>
<tbody>
<tr>
<td>High Yield</td>
<td>Low Yield</td>
</tr>
<tr>
<td>Disease Resistant</td>
<td>Disease prone</td>
</tr>
<tr>
<td>Salt Tolerant</td>
<td>Not Salt Tolerant</td>
</tr>
</tbody>
</table>

Suppose a genome variation seems important
Using the information

<table>
<thead>
<tr>
<th>Healthy</th>
<th>Disease prone</th>
</tr>
</thead>
<tbody>
<tr>
<td>High Yield</td>
<td>Disease prone</td>
</tr>
<tr>
<td>Disease Resistant</td>
<td>Disease prone</td>
</tr>
<tr>
<td>Salt Tolerant</td>
<td>Not Salt Tolerant</td>
</tr>
</tbody>
</table>

Can we influence the processes in which they are involved?
Using the information

<table>
<thead>
<tr>
<th>Healthy</th>
<th>Diseased</th>
</tr>
</thead>
<tbody>
<tr>
<td>High Yield</td>
<td>Low Yield</td>
</tr>
<tr>
<td>Disease Resistant</td>
<td>Disease prone</td>
</tr>
<tr>
<td>Salt Tolerant</td>
<td>Not Salt Tolerant</td>
</tr>
</tbody>
</table>

Can we influence the processes in which they are involved? (breeding, drugs, nutrients, genetic engineering)
Hypertrophic Cardiomyopathy (HCM)

- Most common cause of sudden cardiac death in young athletes
- Caused by a mutation in one of nine sarcomeric genes
- Inherited autosomal dominant (~1 in 500)
- Test and intervene
  - Beta blockers
  - Antiarythmics
  - Pacemakers
Familial Cholesterolemia

- Genetic basis detectable
- Statins prescribed
- Blood clots avoided
Similarity

• We usually work out what a molecule does by finding something similar to it which is already understood
  – Algorithms which “understand” evolution
  – Algorithms which “understand” molecular behaviour

• Computationally expensive
Info ecosystem

• Large datasets for whole database searching, eg:
  – 1000 genomes project – 200 terabytes (March 2012)
  – Sequence Read Archive – 400 terabytes (?)
  – Also many more modest data collections (dozens)

• Curated and made available scientists throughout the world

• N.B. Size does correlate directly with importance or cost
DNA

Europe ENA

USA Genbank

Japan DDBJ
Protein Sequences

Europe EBI

UniProt

Switzerland SIB

USA Georgetown
Molecular Structures

wwPDB

Europe PBDc

USA PDB

Japan PDBj
Scientific and patent literature
Scientific and patent literature

Data submission
Serving the data
General search and browse tools

Specialist interfaces

Databases

Web Site

The status quo

User
General search and browse tools

Specialist interfaces

Web Site

Global

The status quo

User
General search and browse tools

Web Site

Specialist interfaces

The status quo

User

Knowledge?
IMB/UQ context

Mark Ragan....

- EBI Mirror – (Gavin Graham + team)
- NCI SF in bioinformatics
  - Compute (Gerald Hartig + team)
- QFAB
  - Expertise (Jeremy Barker + team)
The EBI

- 23000 cores
- 75 Terabytes of memory
- 19 Petabytes of storage
- $60 million a year operation
- 550 staff (3/4 dedicated to information infrastructure)
- NCBI (USA) is of comparable size
EBI Usage

- Up to 6 million web hits per day
- Over 3 million unique user hosts in 2011

Australia’s EBI usage
- About 60 thousand EBI hits per day
- 6000 unique hosts in May 2012
- 40000 unique hosts in 2011
Complexity
General search and browse tools

Databases

Specialist interfaces

European Bioinformatics Institute (EBI)

Web Site

User
All of these data resources started as independent entities
Complexity
Diversity of databases

- A few kilobytes to a petabyte
- 50 accessions per week to 5 a second
- $5 - $100,000 spend per accession
Technical Solution

Service Oriented Architecture

Genomes
Genes
Variation

Proteins
Structures
Interaction

Chemicals
Drugs
Medicine

User
General search and browse tools

Specialist interfaces

Web Site

European Bioinformatics Institute (EBI)

User
Aside (important)

• Technology can’t solve the problem without some help from the content
  – Persistence and scope of identifiers
  – Ontologies and controlled vocabularies
  – Content analysis (e.g., text mining, sequence comparision)
Changing IT strategy

- EBI leases space in specialist computer centres (reasons not the subject of this talk)
- Implement everything in virtual machines
- Do we really need to own the computers?
Cloud computing

- Service oriented architectures
- Virtualisation
- Cloud computing
- Ensembl in the Amazon Cloud (Europe)
- 1000 genomes data (USA)
- SRA Google cloud
What we want from eResearch

Software

Storage

Network

Data

compute
What we want from eResearch

- Software
- Data
- Storage
- Compute
Australian Engagement
Sharing the task

Europe
- ENA
- EBI
- SIB
- PBDe

USA
- Genbank
- PDB

Japan
- DDBJ
- PDBj
Sharing the task

Become a member of an existing data curation collaboration?
Sharing the task

Unlikely to make sense
Sharing the task

Become involved a new data curation collaboration?
Sharing the task

Europe ENA

USA Genbank

Japan DDBJ

Europe PDBe

USA PDB

Japan PDBj

Europe EBI

Switzerland SIB

USA Georgetown

Australia

?
Optimise Australian access

- Mirror *(as necessary for performance)*
- Training *(✔)*
- User support *(✔)*
- Networks and bottlenecks *(engage with the Australian eResearch to optimise IT capability)*
Australian science in the databases

• The generic databases are already global (✔)

• Specialist content
  – Australian specific organisms or topics
    • Australian dwelling species (✔)
    • Coral reefs (✔ ?)
    • Livestock (✔ ?)
    • Crops (✔ ?)

• Data of Australian interest curated and contributed by Australia

• Australian ambition should not be constrained by this
Australian Data
Australian Data
Australian ingenuity

Ingenuity – tool building R&D (✓)
Australian engagement

- Sharing the task
- Optimise Australian access
- Australian science visible in the databases
- Australian technical ingenuity benefits the infrastructure
- Australia involved in global agenda-setting