Using Linked Data To Inform the Drug Discovery Process

James Snowden (UCB)
UCB: A patient-centric biopharma leader

- 2010 revenue: €3.2 billion
  - 2011 half-year revenue: €1 679 million
- More than 8 500 employees globally
- Operations in more than 40 countries
- R&D Spend: 22% of revenue

Stephanie, living with rheumatoid arthritis
Working In Pharma

 Pharma characterised by opportunities & challenges

- Novel therapeutic approaches providing new opportunities
  - How to translate these into actual therapeutics
- Huge number of known proteins / targets
  - Most accessible targets already targeted
- Huge amount of freely available data to utilise (Omics)
  - How to effectively utilise and integrate this
- Pharmacogenomics offers promise of more targeted medicine & improved therapeutic effect
  - Still unproven, step up in biological understanding needed
Biology has become an information science
Therapeutic Target Selection

Disease can be considered a dysregulation of this interaction network
  • Therapeutics aim to restore balance

Selection of therapeutic targets
  • Continual process to keep pipeline replenished
  • Increasing set of criteria to assess by
  • Fail fast, fail early – High attrition rate of targets

Twin demands of assessing more targets & in more detail
Information Searching

Current approach
- Inefficient: 1-2 people researching, large time requirement
- Non-comprehensive, favourite resources searched
- Data retrieved is non-accessible (Word / PowerPoint)
  - Person becomes data hub

Takes scientists away from the lab where they can research new targets via experimentation
- Less time to assess targets, fewer targets looked at...
- Bottleneck in pipeline

Lots of potential to improve this
New Targets Support

- Lots of data sources we can bring to bear in this area
  - Sequence, structure, literature, omics
  - Also commercial vendor information
- Not feasible for a single person to interact with all of these
- Informatics identified Semantic Web as approach to solve this
- Long term vision of moving towards a New Target ID platform
Traditional technologies struggling to handle data challenges in our domain

- Semantic web technology identified as solution
  - Phil Ashworth has been main advocate

Dynamic area

- Understand how it should be used
- Bring people along with us (internal & external)

Linked Data a key aspect of approach
Integrating Biological Data

- RDF allows us to handle complexity of data and capture concepts and align them

- Biological data needs to be integrated
  - Lots of silos of data (~1100) covering different aspects of biology
  - Provides better answers to questions / Ask new questions
  - Linked Data allows us to do this integration
Linked Data has allowed for the 1st time a unified query environment over key biological datasources, encompassing multiple aspects of biology (e.g. sequence, structure, disease)
Very good coverage of key biological data sources in LOD cloud
• Bio2RDF has been crucial in this area

Provides access to important concept areas for NT support
• Starting point for semantic approach

Focus on target and disease resources (initially)

Two systems designed to utilise LOD
• Target Information Portal (TIP) – *Target*
• Disease Information Portal (DIP) – *Disease*
Key Concepts

- Single starting point (UniProt ID / Diseasome entry)
  - Network aspect permits this
- Use LD principles to connect out to other resources
- Retrieve key data from each data source
- Collate information from non-LOD sources (RSS feeds etc.)
- Generate RDF data model of information extracted
Benefits

End user systems

- ‘Portal’ interface providing single point of access to scientists
  - Single / simple access popular with scientists
  - Data from best of breed systems

Data Infrastructure

- Re-model data into target and disease concepts
  - Can be accessed by any application
- Additional targets and diseases enriches network
OMIM Information

CLONING

Efforts to understand neural secretion have focused on molecules unique to synapses. Bajjalieh et al. (1998) cloned a rat CDN encoding syningetic vesicle protein-2 (SV2), which was functionally expressed in the rat nervous system. The protein was found to be homologous to a family of proton cotransporters. Proton-cotransporters raised the possibility that SV2 might mediate the SV2 cleavage for similar CDNs. In rat brain, Bajjalieh et al. (1998) characterized the SV2B protein was recognized by the anti-SV2 antibody. Expression of SV2B was limited to the highest levels in the cortex and hippocampus, whereas the highest level of expression

Structure Information (PDB)

PDB entry

Title: Solution structure of the first Ig-like domain from human SV2B protein

Sequence:

Structure:

Involvement in regulated insulin secretion. They observed that SV2A and SV2C are
Initial iteration

- SPIN script developed in TopBraid
- Advantages
  - Simple starting point, fast, multiple data types
- Disadvantages
  - Hard coded, still relying on human input to guide search
v1 Overview

[Diagram showing various EP sources connected to TopBRaid (SPIN) and RDF]

- UniProt EP
- PDB EP
- HGNC EP
- Homologene EP
- PubMed EP
- OMIM EP
- GeneID EP

Connections:
1. UniProt EP → TopBRaid
2. PDB EP → TopBRaid
3. HGNC EP → TopBRaid
4. Homologene EP → TopBRaid
5. PubMed EP → TopBRaid
6. OMIM EP → TopBRaid
7. GeneID EP → TopBRaid
8. TopBRaid → RDF

RDF connection indicates the output of the TopBRaid (SPIN) system.
Guided Concept Search

Protein ID

P = Protein
G = Gene
A = Annotation
S = Structure
D = Disease
Guided Concept Benefits

- 6 – 10 min to create dataset
- 250K Triples of a relevant targeted data set delivered
- Very easy to create
  - No Inferencing Required
  - No Linkset Knowledge Required
  - Just need to know which Concepts are required
- Does not need to be modified when new datasets or linksets added
Future Direction

- Evolution towards complete NT support platform
- More concept areas (pathways, omics)
- Enable data environment that makes all relevant data available
- Inform / guide target selection
Summary

- Application of Linked Data to aid pharmaceutical research
  - Provides mechanism to access available data
  - Enables integration of biological data
  - Allows more targets to be assessed & in more detail

- Approach identified to move towards complete target support / identification environment
  - Linked Data principle central to this