MoA Central: A Massively Orthogonal Search Engine for Mechanism of Action & Toxicity Studies

Douglas Selinger, Ph.D.
PreClinical Safety Informatics (PCSi)
ACS 2015, August 19th, 2015
The Future of Drug Discovery

1. Data-driven

“There are known knowns, there are known unknowns, and there are unknown unknowns. Then there are things in databases.”*

2. Mechanism-based

“Nothing in Biology Makes Sense Except in the Light of [insert mechanism here]”**

*I made this up, but it was inspired by Donald Rumsfeld’s “known unknowns” statement (2002).
**Adapted from Theodosius Dobzhansky. The mechanism he was referring to was evolution. (1973)
The challenge

More data should mean better medicines.
The problem

Data dissemination is *massively* inefficient.
The solution

Search
Why a search engine and not a predictive algorithm?

MoA Central is fundamentally a search engine

- **Predictive algorithms**
  - Often black-box
  - p-values are problematic
  - Not clear how to integrate multiple data types
  - The definition of a “target” is ambiguous enough to cause serious problems for predictive algorithms.

- **Search engine**
  - Actual biological data is presented. Conclusions are made by the scientist.
  - Strong prioritization means there is always a manageable amount of data to review. No matter how much data is searched, there is always only one thing at the top of the list.
  - Easily handles multiple heterogeneous data types. Highly scalable. Could be considered a “massively orthogonal” approach.
  - Highly successful algorithms are available, e.g. Google’s PageRank.

- **Find a good predictive algorithm? Include its results in the search engine!**
Can we ‘Google®’ our data?
Yes!!!

The web is a **searchable** collection of connected web pages.

We can create a **searchable** web of compound target data.

*Graph representation of the internet*

*MoA Central graph for the query “rapamycin”.*
MoA Central Vision

MoA = Mechanism of Action

A single search engine for **ALL** data (experimental or *in silico*) relevant to compound target/MoA.
What target ID/MoA data does Novartis have?

*Just a sampling...*

**HitHub**
Compound activities (DMP)
Data collected from multiple internal and external sources.

**CLE**
Cell line encyclopedia (Onc)

**PFVS**
Protein Family Virtual Screening (GDC, EMV)

**Bayes**
Bayesian target prediction (DMP)

**RGAP**
Reporter Gene Assay Profiling (GNF)

**LMF**
Transcriptional signatures (DMP)

**TIP**
Target ID Proteomics (DMP)

**HTSFP**
HTS fingerprints (DMP)

**CLiP**
Cell Line Proferation (DMP)

**ECFP**
2D structural similarity (DMP)

- **in silico**
- **Databases**
- **Omics**

**HIP-HOP**
Yeast chemical genetic profiling (DMP)
MoA Central search strategy

*What are the targets of compound X? (or similar compounds)*

1. Query compound
2. Structurally similar compounds
3. Phenotypically similar compounds
4. Consensus targets

- Rapamycin (sirolimus)
- E.g. “Rapalogs” (only Rapamycin shown)

Other mTOR & PI3K pathway modulators

- mTOR
- FKBP1A
- RPTOR

Data types:

- ~200 million data points
- >15 data sources
- ~80,000 gene & compound sets

- 2D structural similarity
- More to come...

- Transcriptional profile
- Yeast chemical genetic profile
- Mammalian cell line proliferation profile
- High throughput screening profile
- Etc.

- Biochemical activity databases
- Chemical proteomics
- Yeast chemical genetics
- Mammalian chemical genetics
- In silico target prediction methods
- Etc.
Focal graphs: structured evidence

Nodes (circles) are compounds or genes; edges (lines) are relations

Edge examples (black arrows)
- Compounds are structurally similar
- Compounds have similar transcriptional profiles
- Compounds kill similar cell lines
- Gene product is inhibited by compound
- Gene is overexpressed in sensitive cell lines
- Etc.

Focal graphs highlight where the preponderance of evidence is pointing.
Focal graphs: prioritization & segmentation

Rapamycin (sirolimus); Labels sized by indegree or PageRank; Targets/Compounds colored by walktrap community finding algorithm

Visualization and graph metrics are from Gephi (gephi.org).
Focal graph analysis of Rapamycin (sirolimus)

Which graph analysis measures best point to target/MoA?

No labels

OutDegree

InDegree

Equal

Google PageRank

Eigenvector Centrality

MoA graph v2; visualization in Gephi (gephi.org)
MoA Central architecture

Search results can be pre-calculated and stored for fast visualization

- Single search
  - Parallelize steps within search
    - If not already calculated

- Batch searches
  - Parallelize searches in the cluster/cloud
MoA Central: a compound target search engine

MoA Central fetches molecular target data for small molecules. The query is automatically expanded to include compounds with a similar structure or biological profile.

**Compound similarity analyses**
- 2D Chemical similarity (ECFP)
- Cell line proliferation (CLIP)
- Transcriptional gene signatures (LMF)
- Reporter gene assay profiling (RGAP)
- Yeast chemical genetic profiling (HIP-HOP)
- HTS Fingerprints (HTSFP)

**Compound target data**
- IC50 (HitHub: Avalon, MAGMA, GVK...)
- *In silico* target prediction (Bayes)
- Yeast chemical genetics (HIP)
- Mammalian cell line sensitivity (CLIP-CLE)
- Target ID chemical proteomics (TIP)
MoA Central

Generate custom search results for compounds or lists of compounds

- Enter compound name/ID, select from suggestions
- Multiple compounds can be entered in succession to generate a single result with possible shared targets
- Enter a search name (optional)

Users receive an e-mail when the search has started, and another e-mail with a link to the results when it’s finished.
Graph metrics are calculated in R using the igraph package.
MoA Central: a compound target search engine

Gleevec (imatinib): supporting evidence for targets (direct)
MoA Central: a compound target search engine

Gleevec (imatinib): supporting evidence for targets (indirect)
MoA Central: a compound target search engine

*Gleevec (imatinib): most similar compounds*
MoA Central: a compound target search engine

Gleevec (imatinib): evidence supporting compound similarity
Integrated set analysis tells you if these targets/compounds:

- Are linked to a toxicity or adverse event
- Cause or reverse an in vivo phenotype
- Share a mechanism
- Share a protein domain
- Are members of the same complex
- Are members of the same signaling pathway

This information itself is collated from a large number of data sources.
Set enrichment analysis
76,689 unique gene and compound sets across multiple domains

- **Genes**
  - **In vivo**
    - MGI gene to phenotype (8,083 sets)
    - CTD gene to disease (4,908 sets)
    - Metabase gene to toxicity (by species, all species) (2,117 & 1,561 sets respectively)
  - **Pathways**
    - Broad canonical pathways (1,332 sets)
    - Magma siRNA screens (372 sets)
    - Magma “Best of Pathways” siRNA screens (66 sets)
  - **Gene ontology (from Broad MSigDB)**
    - Biological function (872 sets)
    - Cellular component (269 sets)
    - Molecular function (398 sets)
  - **Biophysical**
    - Interpro gene to protein domain/family (10,185 sets)
    - Metabase protein complexes (436 sets)
  - **Mixed**
    - Broad MSigDB complete collection (10,746 sets)

- **Compounds**
  - **In vivo**
    - Integrity compound to therapeutic area (8,289 sets)
    - GVK GoStar compound to toxicity (by species, all species) (11,263 & 8,115 sets respectively)
    - GVK GoStar compound to adverse event (1,928 sets)
  - **Mechanisms**
    - Integrity compound to mechanism (8,620 sets)

Coming soon:
Compounds linked to toxicities, pathologies and biomarkers in preclinical models.
MoA Central: a compound target search engine

Gene and compound panes are opportunities to link to additional info
MoA Central: Mechanism of Action/Toxicity Analysis

Investigating skin toxicity observed preclinically with Compound X

1. Search for compound

2. MoA Central identifies links to the primary target (target x) as well as to PI3Kinase, via the Cell Line Profiling platform (CLiP).

3. A search for ‘skin’ reveals 21 possible links between Compound X and skin phenotypes/toxicities.

Analysis done in collaboration with Rie Kikkawa & Daher IbrahimAibo, PCS Pathology
Target characterization via tool compound pharmacology

Analysis done in collaboration with Jeremy Jenkins, Yuan Wang, & Ben Cornett (DMP)
Data dissemination made *massively* efficient *should mean* better medicines.
Will there be a tipping point?
Speculations

- How much data is needed before we have a high confidence target ID/MoA hypothesis for 80% of our compound archive?

- Will phenotypic screening compound hit lists be immediately interpretable in terms of their likely target/MoA?

- Will we run virtual screens where we identify the compounds most likely to:
  - Hit a given target?
  - Cause or reverse a given phenotype?

- Can we apply focal graphs to other questions, e.g. “if I modulate this gene, what is most likely to happen?” Could this then be used for *in silico* target selection?
What a compelling compound target hypothesis might look like:

- A close structural analog (ECFP) binds to the target in a chemical proteomics experiment (TIP).
- 5 compounds which hit the same HTS screening assays (HTSFP) are known to be potent inhibitors of the target (HitHub).
- Based on its structure, it’s predicted to hit this target by two different in silico target prediction methods (Bayes, PFVS).
- Overexpression of this target in mammalian cell lines leads to compound resistance (CLiP-CLE).
- Reduced dosage of the yeast ortholog of the target causes sensitivity (HIP) to 2 compounds with closely related structures (ECFP).

Furthermore, from set analysis:

- Targets linked to this compound are known to cause a phenotype relatable to the phenotypic screening assay in which the compound was originally identified.
- Targets linked to this compound share a common domain, which further strengthens the direct target hypothesis.
- Compounds linked to the query compound cause a specific toxicity, giving an early warning of things to look out for.

Tipping point?

How much data do we need before we have a compelling target hypothesis for 80% of our compound archive?

- in silico
- Databases
- Omics
Prospective New Data Types

*From anywhere and everywhere*

1. *in silico* models...
2. Publications...
3. Collaborators...
4. Seminars...
5. Conferences...
6. Databases...
7. Updates to existing databases/analyses...

Data dissemination made *MASSIVELY* efficient
The Future of Drug Discovery

MoA Central

1. Data-driven

“There are known knowns, there are known unknowns, and there are unknown unknowns. Then there are things in databases.”*

2. Mechanism-based

“Nothing in Biology Makes Sense Except in the Light of [insert mechanism here]”**

*I made this up, but it was inspired by Donald Rumsfeld’s “known unknowns” statement (2002).

**Adapted from Theodosius Dobzhansky. The mechanism he was referring to was evolution. (1973)
For more information, please visit my Research Gate profile:

https://www.researchgate.net/profile/Douglas_Selinger
Thanks!

MoA Central Core Team
- Doug Selinger (PCS)
- Varun Shivashankar (NX)
- Azita Ghodssi (Program Office)
- Igor Mendelev (NX)
- Mustapha Larbaoui (PCS)
- Mike Steeves (NX)
- Melissa Wilbert (PCS)

The PCS informatics team

Key collaborators
- Steve Litster (NX)
- Jeremy Jenkins (DMP)

Target ID technology platforms
- Target ID chemical Proteomics – TIP (DMP)
- Cell line profiling - CLiP (DMP)
- Cell line encyclopedia - CLE (Onc)
- Reporter Gene Assay Profiling - RGAP (GNF)
- HIP-HOP yeast chemical genetics (DMP)
- Transcriptional signatures – LMF (DMP)
- Bayesian target prediction (DMP)
- Protein Family Virtual Screening – PFVS (GDC, EMV)
- HTS fingerprints – HTSFP (DMP)
- HitHub (DMP)

Philippe Marc
(PCS informatics, Global Head)

Page Bouchard
(PCS, Global Head)