The Case for Use Cases

The integration of internal and external chemical information is a vital and complex activity for the pharmaceutical industry.

David Walsh, Grail Entropix Ltd

Costs of Integrating Internal and External Data

- It is expensive!

- Generating quality, comprehensive external data is expensive.
- Finding data that can be integrated is expensive

- External chemical data needs to be accurate, comprehensive and contain metadata (e.g. ownership and bioactivity (qualitative/ quantitative)

It’s difficult, so Why?

- Deriving knowledge on competitor company’s molecular approaches is valuable competitor intelligence

- Determination of the creative processes and medicinal chemistry rules that drive lead optimisation is important to assist internal drug discovery processes

Why?

- Describes what has already formed part of prior art
- Allows decisions to be made on freedom to operate
- Describes areas of chemical space that is congested, affecting issues of lead optimisation.

Why?

- Allows modelling of molecular properties in a comparable way to internal molecules

- Determination of comparable physicochemical properties between internal and external molecules

How?

- It is important to provide externally derived molecules in formats which are capable of being integrated with internally derived molecules

- It may also be important to recognise that internal and external molecules are essentially different.

Pfizer Use Case Study

- I will make some reference to Pfizer's use cases, which were derived up to 3 years ago, and may have been modified and amended over time.....

- Requirements were for a large molecule collection of exemplified molecules from patents, supplemented by data on pharmaceutically relevant exemplified molecules

- An investigation of available approaches to investigate Markush space.

Exemplified Compounds from Patents

IBM Database

Status

- Chemical structures and meta data extracted from the document text of WO, EP, and USPTO documents (1976 to present, 4.5 million unique patents)
- Automated updating of internal DB as new documents are published (2-4 week lag)
- Consortium with 9 pharmaceutical and chemical companies - reduces costs and drives technical improvements
- Image to structure conversions for all patents back to 2000

Value

- Pfizer owns the data (8 million unique structures) that can be used and stored as project teams see fit
- Data can be easily integrated with internal and external datasets and manipulated by internal tools

Downsides

- Automated process: errors from poor name to structure conversions and “junk” can creep into data stream
- Only structures that are named in the text are indexed for patents pre 2000

Exemplified Compounds from Patents

GVK Biosciences GoStar Database

Status

- Chemical structures and bioassay data hand curated from patents and journals
- Database is updated quarterly
- Pfizer licenses the database, but does not own the contents
- GVK GoStar database integrated into CCT protocols for patent analysis

Value

- High quality structures and bioassay data
- Little overlap with structures in IBM patent database
- Data can be easily integrated with internal datasets and manipulated by internal tools

Downsides

- Limited coverage of the patent literature (about 1/10th that of IBM database)

Cross-disciplinary Teams

- It is also important to develop cross disciplinary teams to develop use cases across a range of discovery chemists, informatics and patent practitioners, as their combined requirements are important to maximise the use of information and their individual usage or data handling skills and knowledge are vastly different.

- The cross disciplinary teams are capable of extending the variety of use cases, as well as extracting the meta knowledge of data sources and perspectives for using chemical information, and also allows ownership to be distributed between silos in the organisation.

Cross Disciplinary teams

- Medicinal chemists
- Information Scientists
- Patent Attorneys
- Cheminformatics

- People who have experience of Chemical and Patent databases

- People who have skills in patents, chemical nomenclature

- People who have experience of the structure and nature of documents

Cross Disciplinary Teams

- Advocates within various Drug Discovery teams, who can assist colleagues with the development and acceptance of new tools.

Use Cases 1

- Integrate third-party and in-house sources of data from chemical patents
  - Chemical structures, text, and numeric data (biological data, where available; clinical data)
  - Incorporate new and delete existing data sources (anticipate change)
  - Operate within third-party guidelines for acceptable use of data
  - Easily updatable and adaptable systems
  - Concordance between Pfizer patents and bioassay on those compounds

Use Cases 3

Patent chemistry landscape:

1) Identify structures (exemplified and Markush) in the patent literature (and the RIF) that are relevant to a TA project series. Include external biological data, if available.

2) Inform potential IP overlap of a project's compounds/series against the entire patent landscape (exemplified and Markush). Identify and visualize top ranked patents and molecular feature overlap (patent ranking and feature analysis toolset).

3) Given a list of PF numbers, identify patents which claim or disclose these Pfizer compounds.

Target chemistry landscape:

Identify biologically relevant molecules (exemplified and Markush) that have activity against a target of interest, or class (family) of targets. Support a workflow to identify key compounds for follow-up and the most similar matches within the file. Provide visualization of property and chemical space landscape including data and assignee information (with option to include internal compounds) to facilitate analysis of competitor's activity in the target space.

Patent alerts:

Identify patents by similarity (or substructure) to a TA project series, or target, indication, or research area. Implement a push deliver system to automatically alert user as relevant info is published

Patent chemistry landscape

Scope: Identify structures in the RIF and chemical patents that are similar to a TA project series. Include biological data, if available.

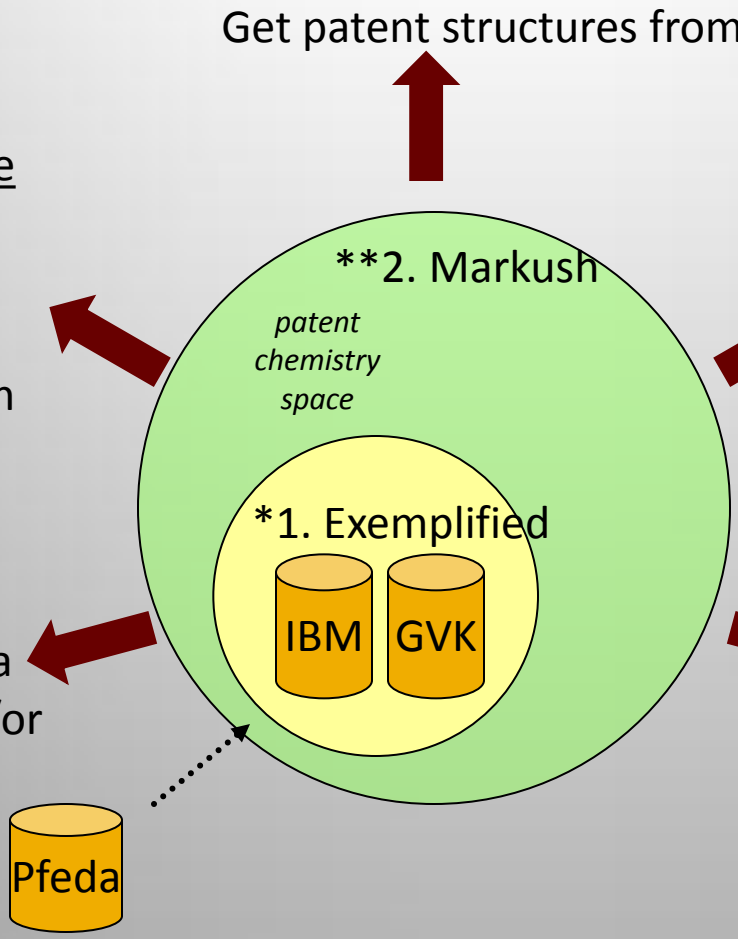
Inputs: Hit structures for a project

Outputs: Related structures in the RIF and patent literature with related biological data

Process: Similarity and/or substructure searching against internal and external structure databases. Join the structures retrieved from the internal and external databases with other non-structure data (activity data, patent number, patent assignee, filing date, data source, etc.). Need to consider tautomers, stereochemistry, differences in fingerprints applied to similarity searching in internal versus external data sources.

Presentation of the output in a format compatible with the desktop applications (e.g., PCAT, Spotfire, MoViT).

Current tools and proposed integration with Markush



Chemistry Landscape
- Substructure search
- Similarity search
- Exact-match search

Patent Ranking
- Rank order a series of patents by feature similarity to a query structure

Target Landscape
- Patents related to a specified target and/or query structure

Feature Analysis
- Detailed analysis and visualization of substructural features within patent structures compared with those for a set of query structures

*IBM database of 8 million structures from patents
*GVK database of 3 million structures from patents
**Thomson-Reuters database of 1.2 million Markush from patents

Information Overload

- Information Overload
  - More than 70,000 chemical patents are filed every year around the world

- Is there an easier way to identify those patents that have pharmaceutical relevance and eliminate the redundancy between equivalent patents?

Search Pfizer patents:

Concordance between Structure IDs and the patents where they are published

Scope: Given a list of PF numbers, identify patents which cover these Pfizer compounds.

Inputs: A list of PF compound Ids, or PF structures

Outputs: The PF compounds and patents which cover these compounds

Process: Search for the presence of the PF compounds within a database of curated patent structures (e.g., IBM patent database). This is an exact structure match. Need to account for tautomers, stereoisomers when performing the search.

Important in companies that have acquired other companies!!

Patent Alerts

Scope: Identify patents by similarity to a TA project series, or target, indication, or research area.

Inputs: A series of compounds from a TA project, or a target, indication, or research area

Outputs: A formatted summary of patents relevant to the search input.

page | discussion | view source | history

This article is a wiki. Want to improve it? Log in and change it or Report a Problem

PI3K

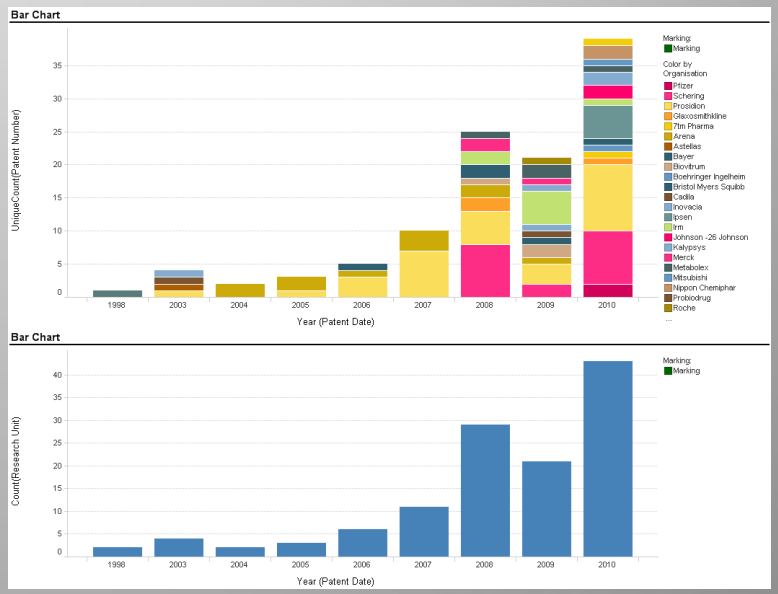
Patent Alert | Research Unit | Indication | Target | Organisation

This page lists all the patents that identify PI3K as the target.

RSS

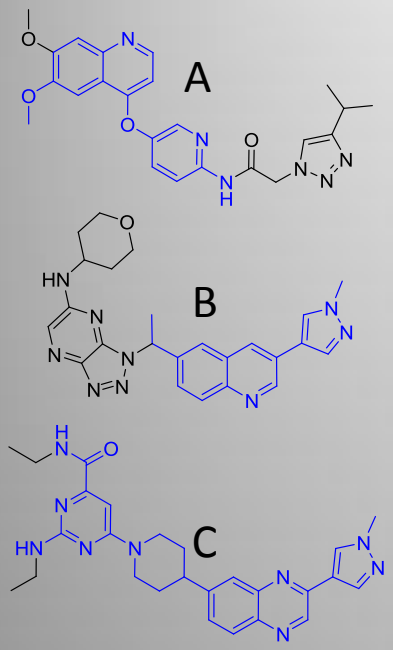
Patent Alerts:

Patent Number	Organisation	Target	Title	Date	Indication	RU	Structure Clipping
WO10080996	Curis	PI3K	Phosphoinositide 3-kinase Inhibitors With a Zinc Binding Moiety	15 July 2010		Allergy and Respiratory	
WO10079423	CSIR India	PI3K	Inhibitors of Phosphatidylinositol-3-kinase (Pi3) and Inducers of Nitric Oxide (No)	15 July 2010		Allergy and Respiratory	
WO10074586	Pathway Therapeutics	PI3K PI3K Delta	Pyrazolo[1 5-a]pyridine and Imidazo[1 2-a]pyridine Derivatives and Their Use in Cancer Therapy	1 July 2010		Allergy and Respiratory	
WO10061903	Shionogi	PI3K	Pyrimidine Derivative and Pyridine Derivative Both Having Pi3k Inhibitory Activity	3 June 2010		Allergy and Respiratory	



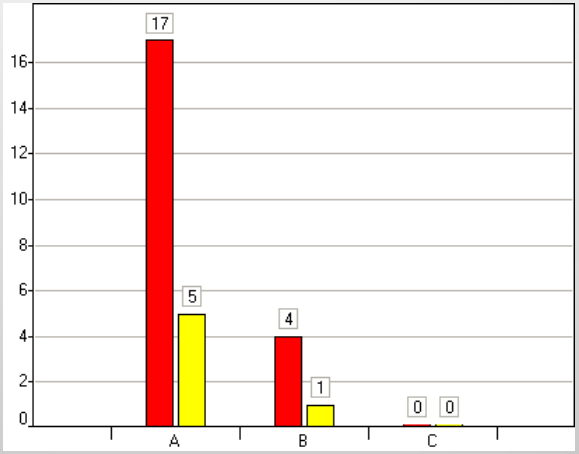
Project Applications: Chemistry Landscape

Pan-Trk project: Analysis of IP in evaluating a second series.

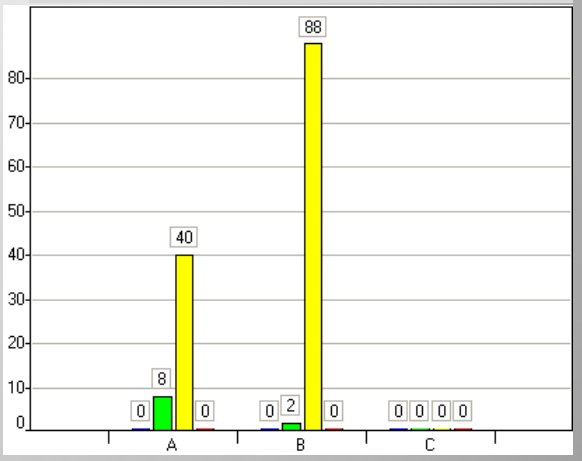


Substructure and Similarity Search

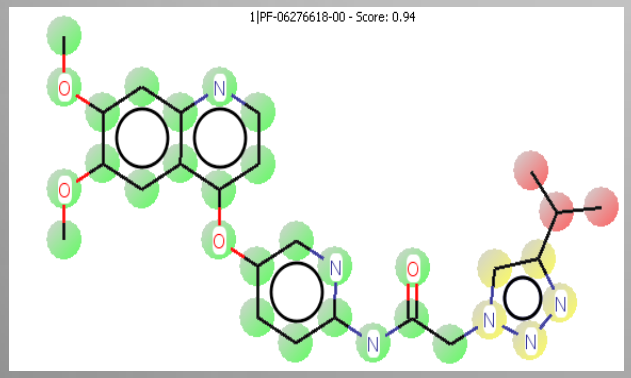
Num of Unique Patents and Assignees

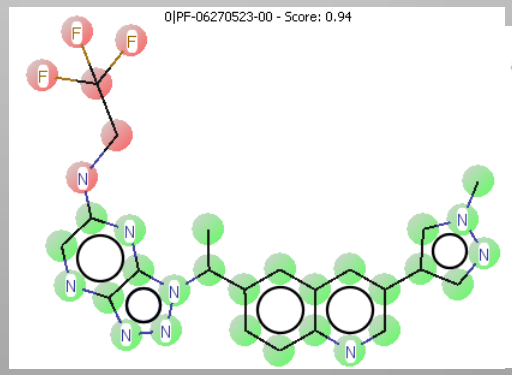


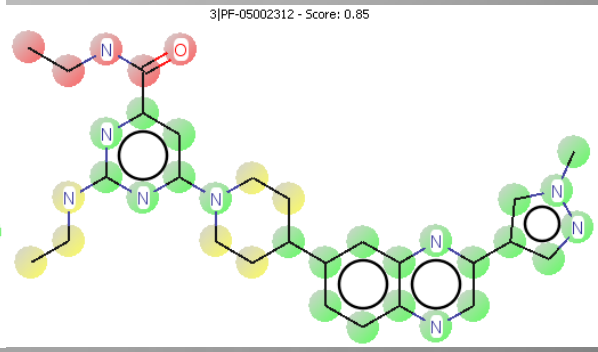
Num Similarity in Bins: 90, 80, 70, 60



Feature Analysis on Identified Patent IDs



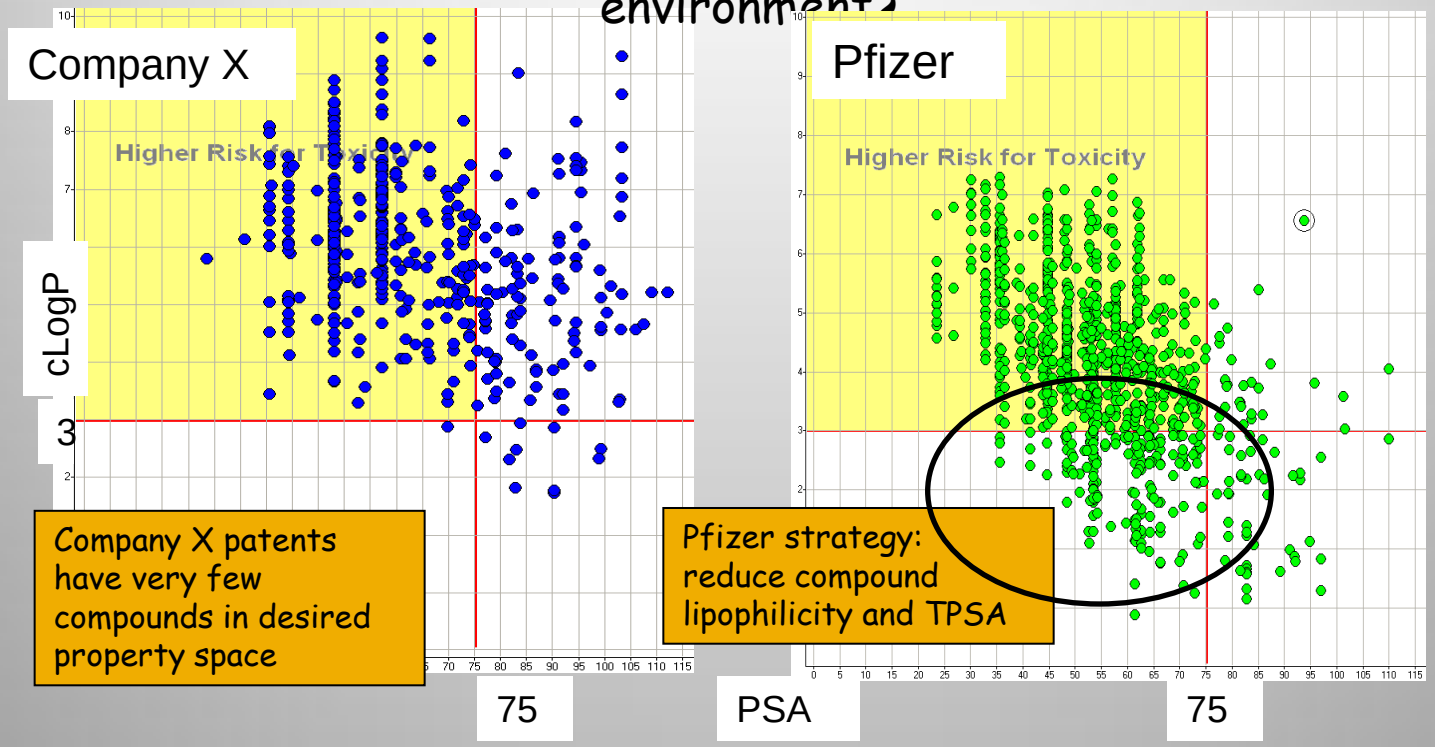


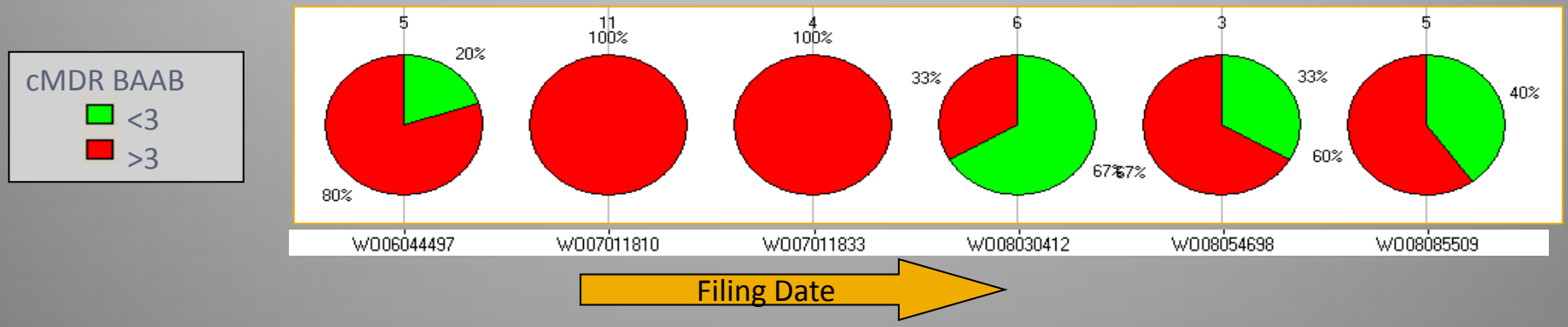


Currently investigating alternatives to substructure and similarity searching (Bayesian models) to improve identification of patents of interest.

Property Landscaping

How do the properties of our chemical matter relate to the external environment?



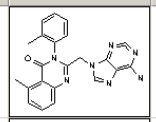
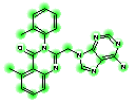
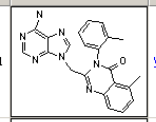
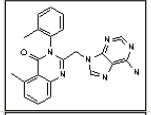
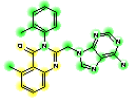
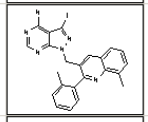
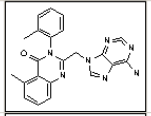
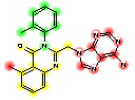
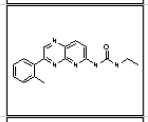
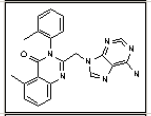
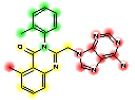
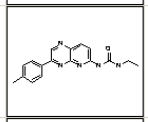
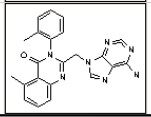
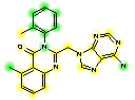
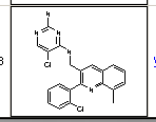


Substructure Search, Similarity and Feature Analysis

Substructure and Similarity are well known concepts.

Is there a need for other methods of “relatedness” based on connectivity?

One such is Feature Analysis – driven by the need to formalise the markush definitions within patents by regression to “superatoms” and why they cover molecules of relevance.

Id	Structure	FEATURES_ATOMPROP	BestScore	BestMatch_stx	PatentNumber
PF-04475495 Instance 1			1		WO05112935
PF-04475495 Instance 2			**0.89**		**WO200800118454**
PF-04475495 Instance 3			0.77		WO200800138878
PF-04475495 Instance 4			0.77		WO200700079999
PF-04475495 Instance 5			0.73		WO200800118455