Open for Innovation ®

KNIME

Guided Analytics in Action: Patent analysis from your browser[1]

Greg Landrum

KNIME.com AG

[1] and some other stuff

Agenda

- A few more words on “What’s new” and “What’s cooking”
  - KNIME in the cloud
  - KNIME on Spark

- A Guided Analytics use case: patent analysis
- Quick advertisement for a workshop tomorrow

KNIME in the cloud

- KNIME Analytics platform on AWS/Azure
- KNIME Server, Server + Big Data on AWS/Azure
- Launch Hadoop/Spark clusters on AWS/Azure
- Connectors for S3, Blob Store, Azure SQL DB, Redshift

- What else should we be thinking about/working on?

KNIME on Spark

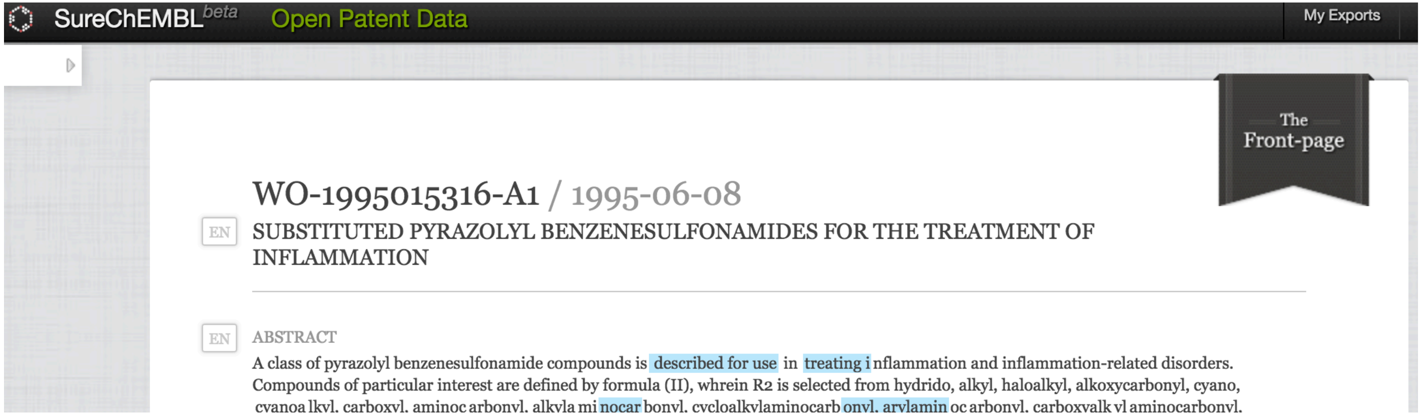
Demo

The case study: working with patent data

- Start with the PDF for a med-chem patent
- Extract the structures from the patent
- Attempt to identify the key compound

Inspired by:
http://chembl.blogspot.ch/2014/11/finding-key-compounds-in-med-chemistry.html

The patent



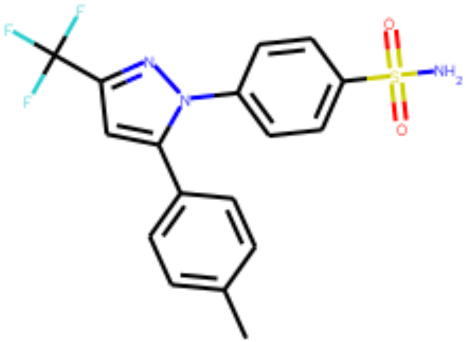
WO-1995015316-A1 / 1995-06-08

EN SUBSTITUTED PYRAZOLYL BENZENESULFONAMIDES FOR THE TREATMENT OF INFLAMMATION

EN ABSTRACT

A class of pyrazolyl benzenesulfonamide compounds is described for use in treating inflammation and inflammation-related disorders. Compounds of particular interest are defined by formula (II), whrein R2 is selected from hydrido, alkyl, haloalkyl, alkoxycarbonyl, cyano, cyanoalkyl, carboxyl, aminocarbonyl, alkylaminocarbonyl, cycloalkylaminocarbonyl, arylaminocarbonyl, carboxyalkylaminocarbonyl,

This is the patent “for Celebrex” :

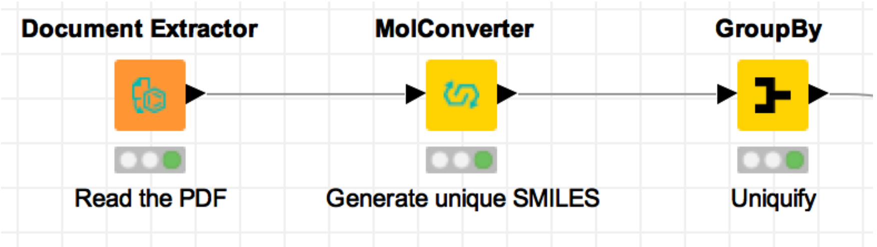


One of the validation patents from:

Hattori, K., Wakabayashi, H. & Tamaki, K. Predicting Key Example Compounds in Competitors’ Patent Applications Using Structural Information Alone. *J. Chem. Inf. Model.* **48,** 135–142 (2008).

Extracting structures

Grab the PDF from SureChEMBL and the use the ChemAxon Document Extractor:

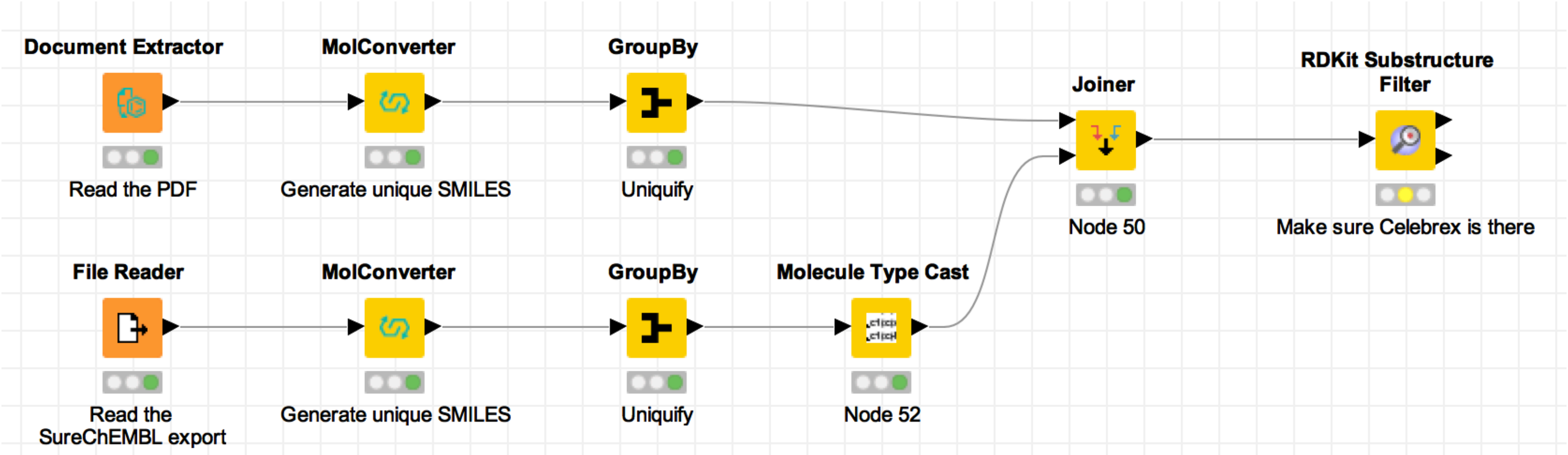


Document Extractor → MolConverter → GroupBy

1. Document Extractor: Read the PDF
2. MolConverter: Generate unique SMILES
3. GroupBy: Uniquify

Let’s start with a bit of validation:

Compare the Document Extractor to the structures downloadable from SureChEMBL:



Document Extractor – Read the PDF

MolConverter – Generate unique SMILES

GroupBy – Uniquify

File Reader – Read the SureChEMBL export

MolConverter – Generate unique SMILES

GroupBy – Uniquify

Molecule Type Cast – Node 52

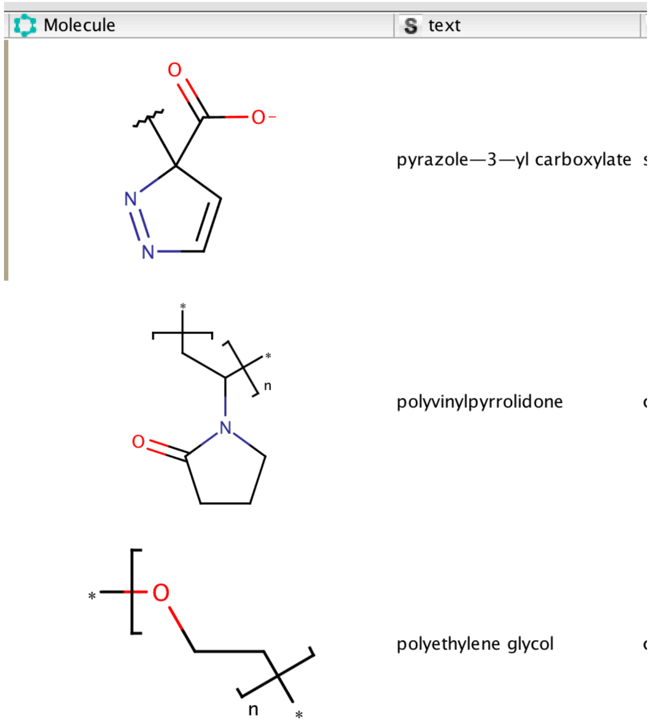
Joiner – Node 50

RDKit Substructure Filter – Make sure Celebrex is there

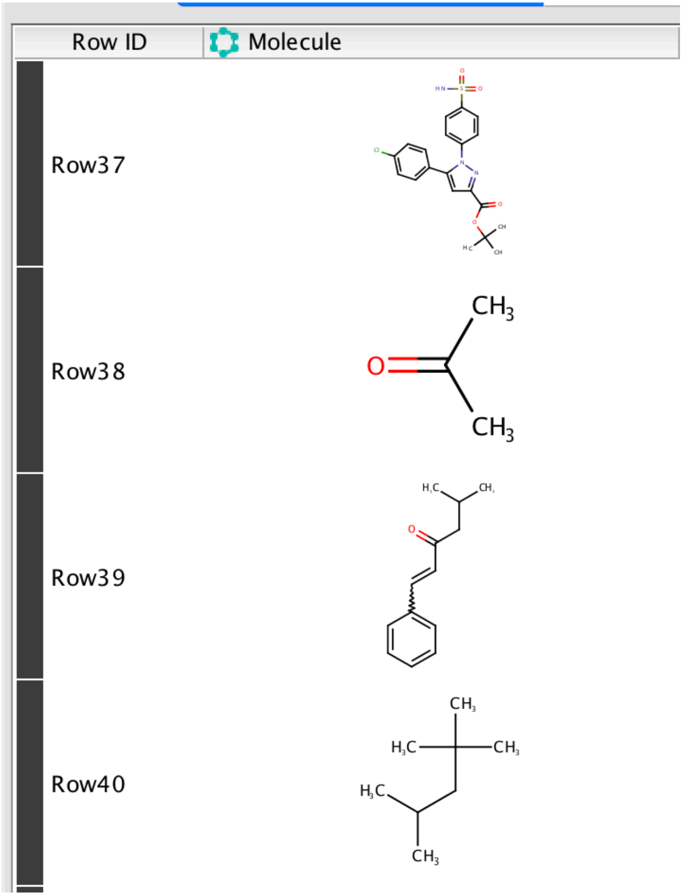
Results:

- 451 structures from Document Extractor
- 548 structures from ChEMBL
- 192 are in common, and Celebrex is one of them.

Cleaning up the data



Molecule	text
	pyrazole—3—yl carboxylate
	polyvinylpyrrolidone
	polyethylene glycol



Row ID	Molecule
Row37	
Row38	
Row39	
Row40	

The interesting structures are mixed in with a bunch of stuff we'd rather not see

Cleaning up the data, but with which rules?

- Some things, like removing molecules with attachment points, are obvious. But what about the other properties?
- There are multiple different ways of filtering these down to the “interesting” ones.
- Instead of picking some hard-and-fast rule, let’s do it interactively using a set of reasonable properties:
  - # of heavy atoms
  - # of rotatable bonds
  - # of rings

Finding the key compound in the patent

- Idea: the key compound is likely to have a lot of similar compounds
- Classic approach: find the compounds in the patent with the most near neighbors
- Alternate approach: use metrics from network analysis

Method from:

Hattori, K., Wakabayashi, H. & Tamaki, K. Predicting Key Example Compounds in Competitors' Patent Applications Using Structural Information Alone. *J. Chem. Inf. Model.* **48,** 135–142 (2008).

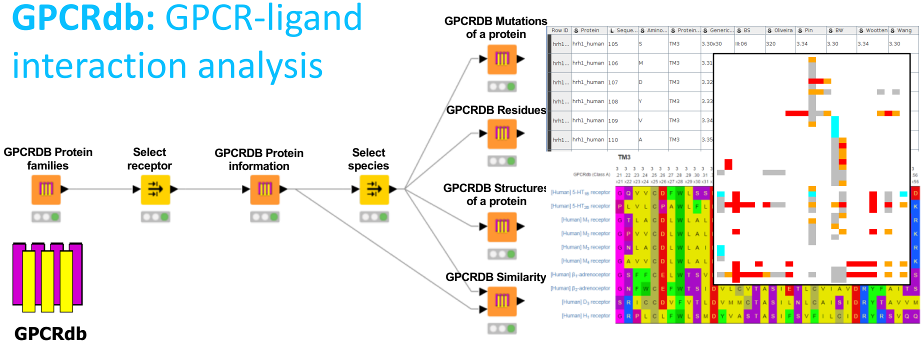
Advertising!

3D-e-Chem KNIME Nodes for Integrated Structural Cheminformatics Analyses and Computer-Aided Drug Discovery

Information: https://tech.knime.org/3d-e-chem-nodes-for-knime

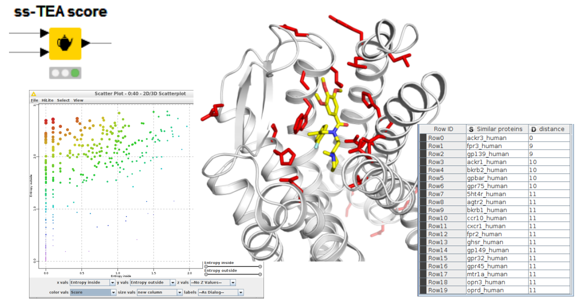
Workshop: https://github.com/3D-e-Chem/workflows (Clone or download - Download **ZIP**)

GPCRdb: GPCR-ligand interaction analysis



Row ID	Protein	Seque...	Amino...	Protein...	Generic...	BS	Oliveira	Pin	BW	Wootten	Wang
hrh1...	hrh1_human	105	S	TM3	3.30x30	III:06	320	3.34	3.30	3.34	3.30
hrh1...	hrh1_human	106	M	TM3	3.31						
hrh1...	hrh1_human	107	D	TM3	3.32						
hrh1...	hrh1_human	108	Y	TM3	3.33						
hrh1...	hrh1_human	109	V	TM3	3.34						
hrh1...	hrh1_human	110	A	TM3	3.35						

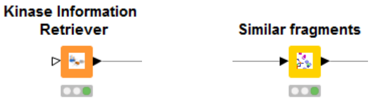
ss-TEA: GPCR binding site prediction

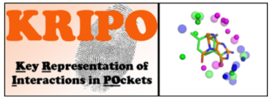


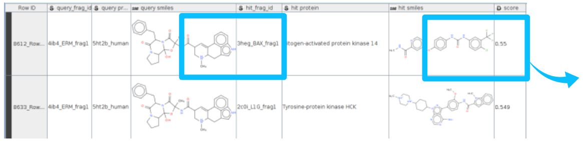
Row ID	Similar proteins	distance
Row0	ackr3_human	0
Row1	fpr3_human	9
Row2	gp139_human	9
Row3	ackr1_human	10
Row4	bkrb2_human	10
Row5	gpbar_human	10
Row6	gpr75_human	10
Row7	5ht4r_human	11
Row8	agtr2_human	11
Row9	bkrb1_human	11
Row10	ccr10_human	11
Row11	cxcr1_human	11
Row12	fpr2_human	11
Row13	ghsr_human	11
Row14	gp149_human	11
Row15	gpr32_human	11
Row16	gpr45_human	11
Row17	mtr1a_human	11
Row18	opn3_human	11
Row19	oprd_human	11

KLIFS: kinase-ligand interaction analysis

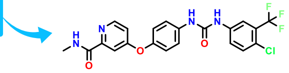
KRIPO: Pharmacophore-based binding site similarity



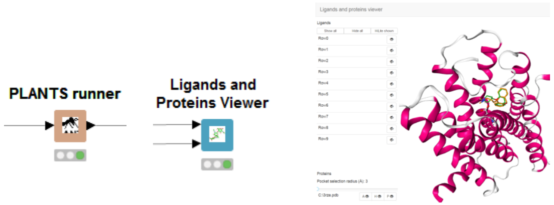




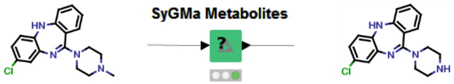
Row ID	query_frag_id	query pr...	query smiles	hit_frag_id	hit protein	hit smiles	score
8612_Row...	4ib4_ERM_frag1	5ht2b_human		3heg_BAX_frag1	itogen-activated protein kinase 14		0.55
8633_Row...	4ib4_ERM_frag1	5ht2b_human		2c0i_L1G_frag1	Tyrosine-protein kinase HCK		0.549



Molviewer: Protein-ligand complexes



SygMA: metabolite prediction



McGuire, Verhoeven, Vass, (...), De Graaf. *J Chem Info Model* 2017, 57: 115