A challenging dataset to validate pharmacophore programs – Automated protocol to select and overlay structures from the RCSB Protein Data Bank.

Ilenia Giangreco

6th Joint Sheffield Conference on Chemoinformatics

22nd - 24th July 2013

Overview

- Introduction and background
  - How to validate overlay programs?
  - Already available datasets
- Methods
  - Selection and filtering of complexes
  - New approach to overlay the structures
    - How and why?
  - Sub setting highly populated set
    - Contact analysis
    - CDK2 as an example
- Results
  - Comparison with standard protocols
  - Scoring overlays
    - Examples of good and poor overlays for pharmacophore validation
- Conclusions

How to validate overlay programs?

Pharmacophore elucidation: a molecular alignment problem

i. Select an enzyme for which multiple structures complexed with different ligands are available

ii. Overlay the proteins into a common reference frame

iii. Extract the ligands from the overlaid proteins and denote this as the target overlay

iv. Compare results obtained from overlay programs with the target overlay

Available datasets

Structures retrieved from the Protein Data Bank

- Patel Y. et al. *J. Comput. Aided Mol. Des.* **2002**

  35 ligands for 5 drug targets

  Catalyst/HipHop
  GASP
  DISCO
  PHASE
  GALAHAD

- Jones G. *J. Chem. Inf. Model.* **2010**

  80 ligands for 9 drug targets

  GAPE
  MARS

- Taylor R. et al. *J. Comput. Aided Mol. Des.* **2012**

  87 ligands for 10 drug targets

  An improved MOGA-based overlay program

- Cross S. et al. *J. Chem. Inf. Model.* **2012**

  960 ligands for 81 targets,
  but 5 taken from Patel et al.

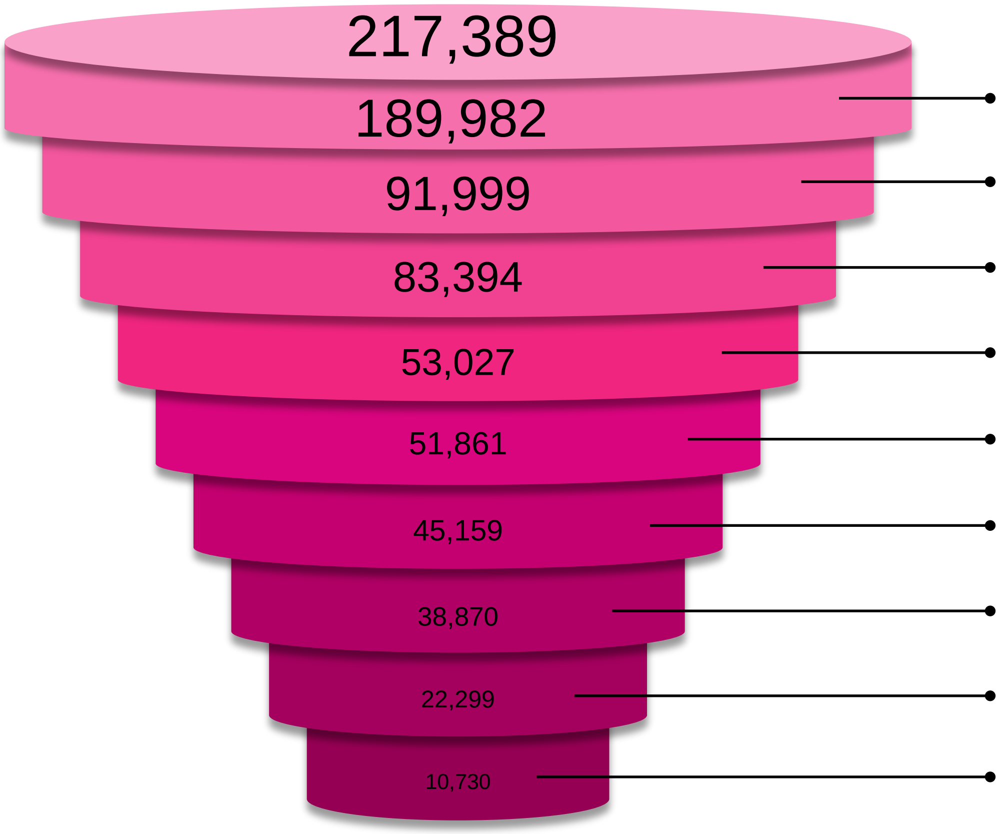
  FLAPpharm

We present a dataset of 1445 ligands for 119 targets

Giangreco I. et al. *J. Chem. Inf. Model.* **2013**

Searching and filtering criteria

- Only X-ray structures with resolution ≤ 2.5 Å
- Release date between 01/01/2000 and 03/05/2012



217,389

189,982 — Excluding apo-proteins

91,999 — Excluding solvent and small ions

83,394 — Excluding cofactors

53,027 — Excluding compounds with HAC ≤ 10

51,861 — Excluding compounds with non-organic elements

45,159 — Excluding simple saccharides

38,870 — Excluding compounds with $(CH_2)_4$ or $CH_2O(CH_2CH_2O)_2$ linkers

22,299 — Excluding compounds that fail the rule of five

10,730 — Excluding compounds that occur more than 20 times

Protein grouping

- EC number used for the top level classification
- UniProt ID used to group structures

  Proteins from the same gene but different species have different entries (e.g. DHFR, beta-lactamase)
- Unique ligands selected within a group of structures

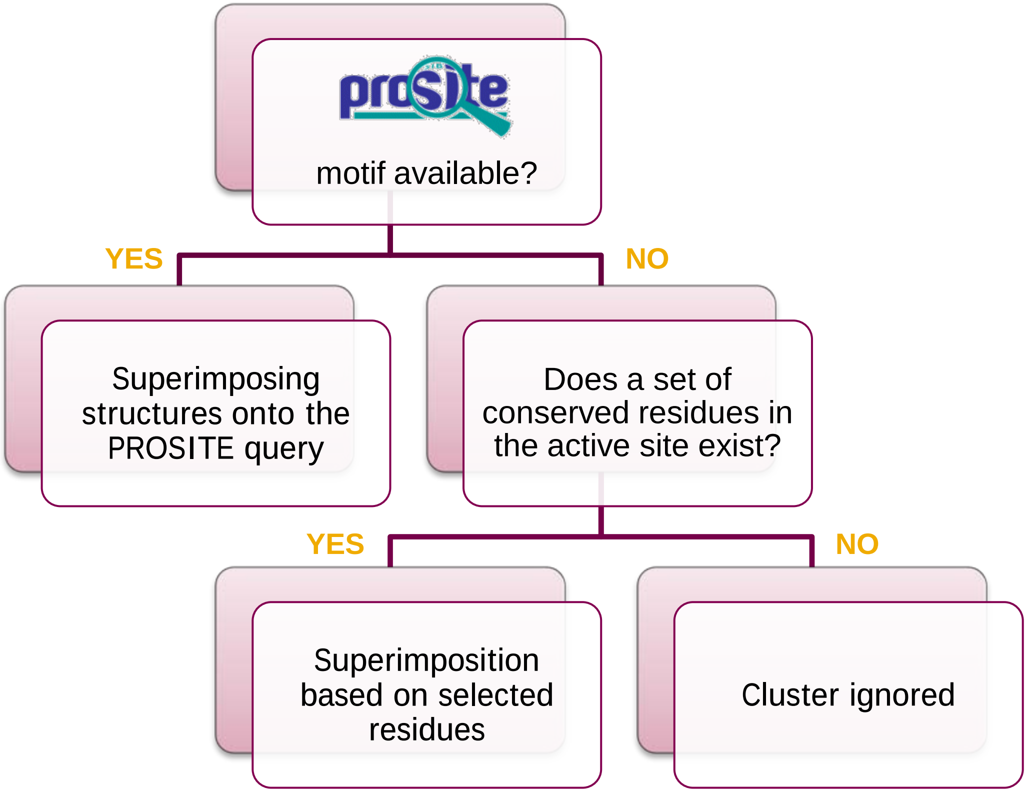
  Where multiple structures of the same protein-ligand complex are available, the structure with the best resolution has been chosen
- Analysis of 183 clusters of proteins out of 2365 total

  At least 5 ligands in each cluster
- Further reduction of targets based on PROSITE motif availability

Protein overlay: how?

A new approach not biased by protein flexibility

Focusing on protein domains or functional sites, if available


prosite
motif available?
YES
NO
Superimposing structures onto the PROSITE query
Does a set of conserved residues in the active site exist?
YES
NO
Superimposition based on selected residues
Cluster ignored

- All-by-all comparison to find a reference structure. Loop over all sequences and calculation of RMSD value. For each pair of proteins, the combination of chains resulting in the lowest RMSD is considered
- Lowest mean RMSD as criterion of selection. For the selected protein, take the most commonly used chain as the reference
- Superimposition of protein structures using the backbone atom coordinates of selected residues

The prosite database: why?

- PROSITE is a resource for the identification and annotation of conserved regions in protein sequences

- Large collection of biologically meaningful signatures:

  - Generalised profiles (weight matrices) describing protein families and modular protein domains
  - Patterns (regular expressions) describing short sequence motifs often corresponding to functionally or structurally important residues

- All signatures are built from manually derived alignments and are provided with extensive manually curated documentation

- Each PROSITE profile is associated with a manually curated annotation template called ProRule

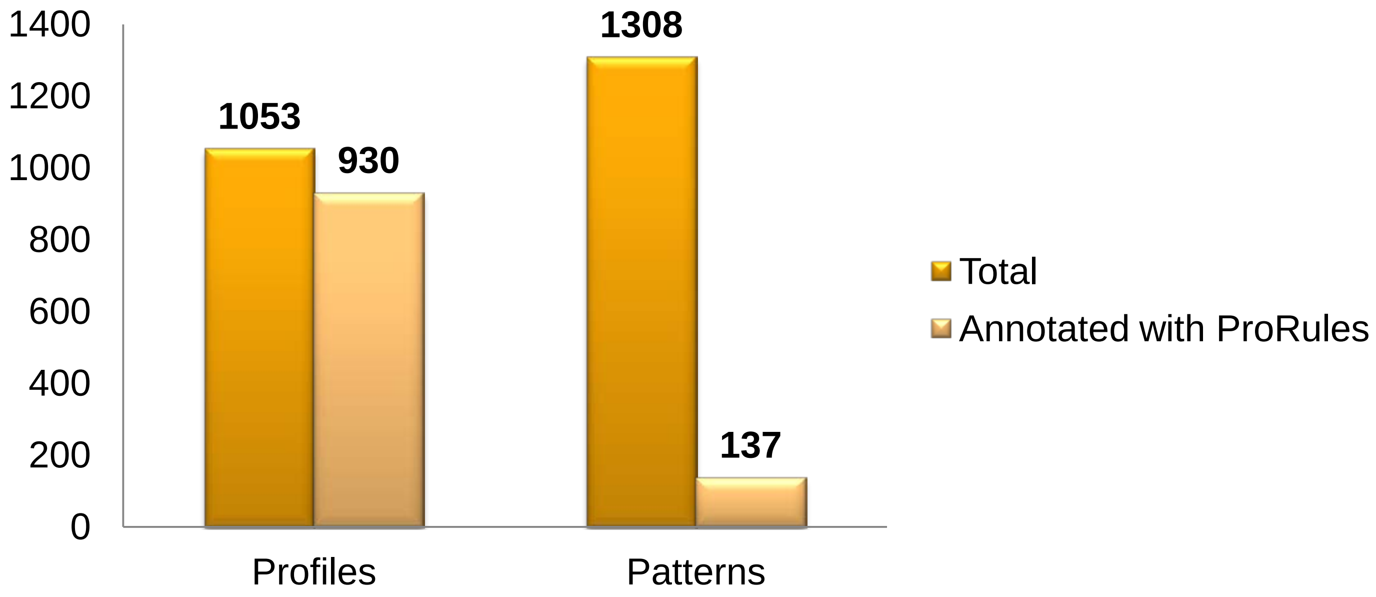
  - ProRules add motif-specific information to their associated profile, allowing the detection of intra-domain features (e.g. active sites, binding sites, disulfide bridges)

Sigrist et al. Nucleic Acids Research, 2012, 1–4

PROSITE stats

As of 14th March 2013

2361 motifs distributed as follows:



PROSITE patterns and profiles

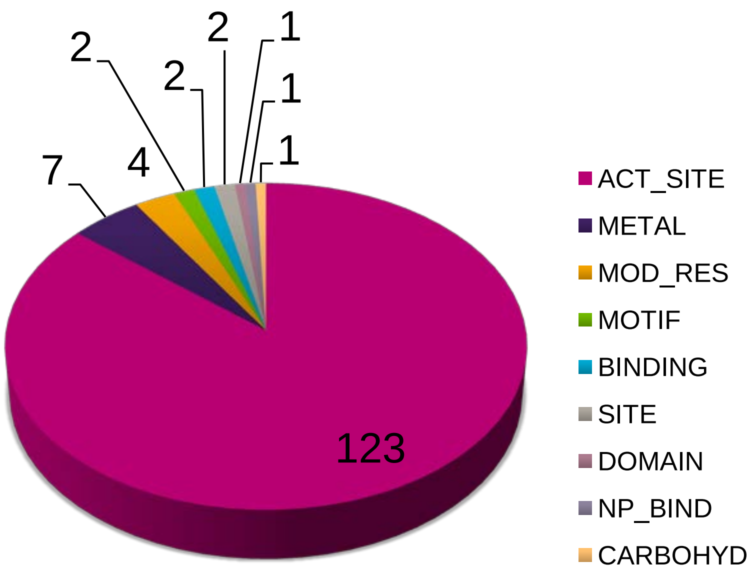
- Patterns are regular expressions matching short sequence motifs usually of biological meaning
  - ~10 to 20 amino acids in length
  - thought or proved to be important to the biological function
  - conserved in both structure and sequence during evolution
- Patterns are qualitative motif descriptors
- Profiles are more sensitive than patterns
  - Patterns have intrinsic limitations in identifying distant homologues as they do not accept any mismatch
- Profiles usually correspond to protein domains
- Profiles are quantitative motif descriptors
  - Numerical weights for each possible match or mismatch between a sequence residue and a profile position

Sigrist et al. BRIEFINGS IN BIOINFORMATICS, 2002, 3, 265–274

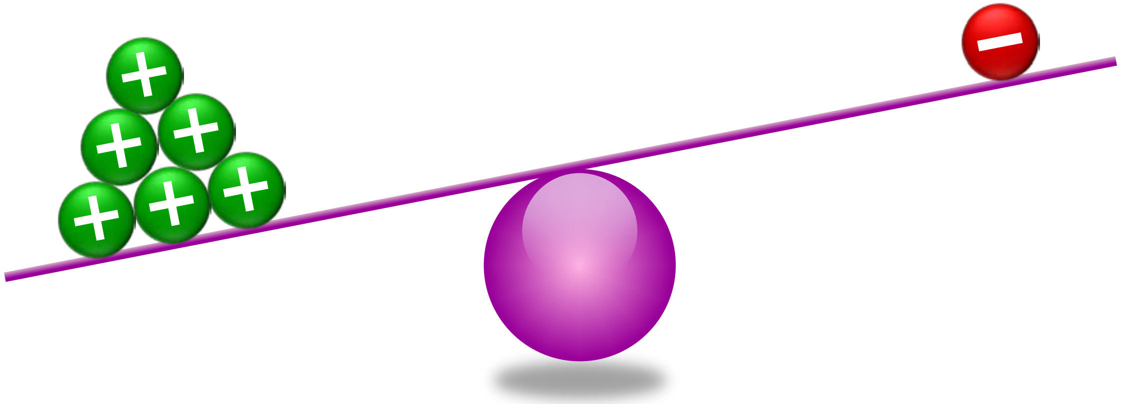
PROSITE patterns

These biologically significant regions or residues are generally:

- Enzyme catalytic sites – ACT_SITE
- Binding site for any chemical group (co-enzyme, prosthetic group, etc.) – BINDING
- Amino acids involved in binding a metal ion – METAL
- Cysteines involved in disulphide bonds – DISULFID
- Interesting single amino acid site on the sequence – SITE
- Modified residues excluding lipids, glycans and protein cross-links – MOD_RES



Pros & cons



✓ A clearly defined set of residues used for the superimposition

- No subjective choice
- No variability based on the size of ligands

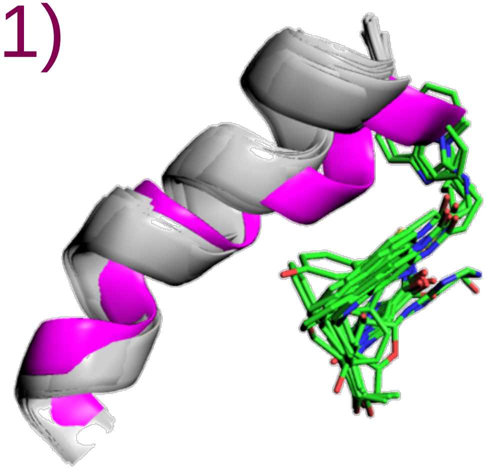
✓ Reproducibility

✓ Better results in case of protein flexibility

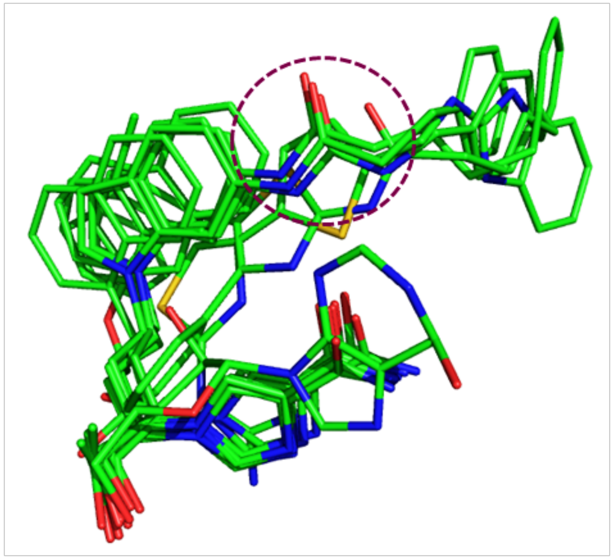
✘ Motifs not available for some targets of interest, or if available, may not be located close to the binding site

Comparison with previous approaches (1)

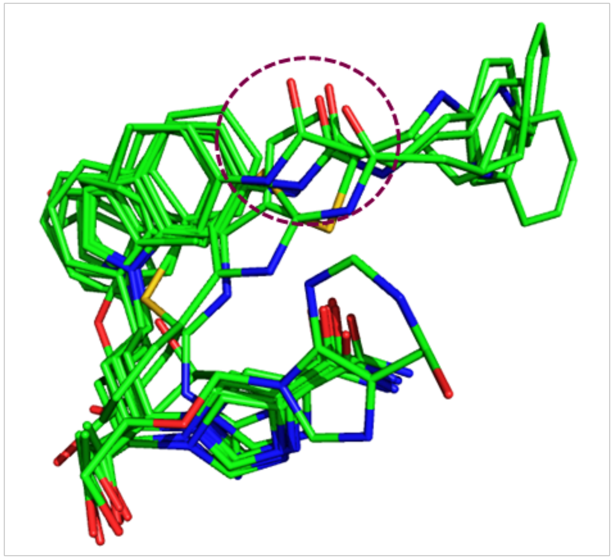
Adenosine deaminase (ADA) - example



- 11 protein-ligand complexes from Taylor et al. *J. Comput. Aided Mol. Des.* **2012**
- All but one (1krm – magenta helix) structures in open form



Overlay onto the PROSITE motif PS00485

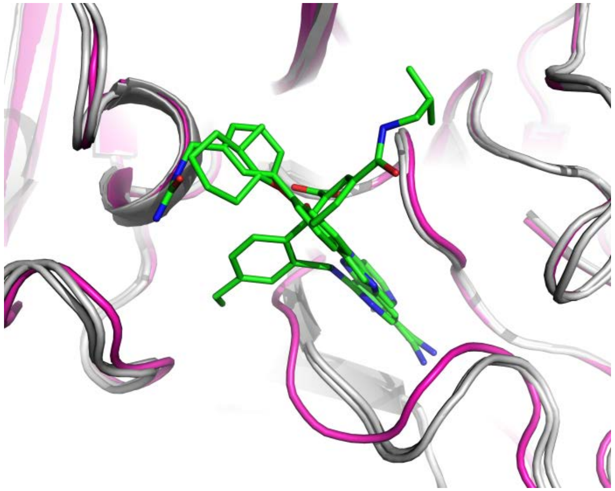


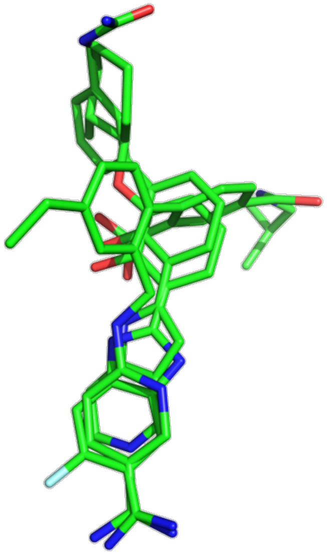
Overlay by least-square fitting of the binding site atoms (Relibase+)

Comparison with previous approaches (2)

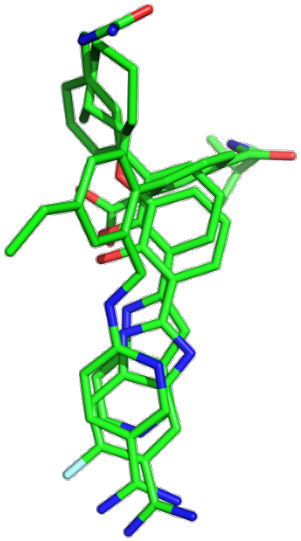
Coagulation factor VII (fVII) - example

- 3 protein-ligand complexes common to PharmBench
- A flexible loop in the binding site





Overlay onto the PROSITE motif PS00135

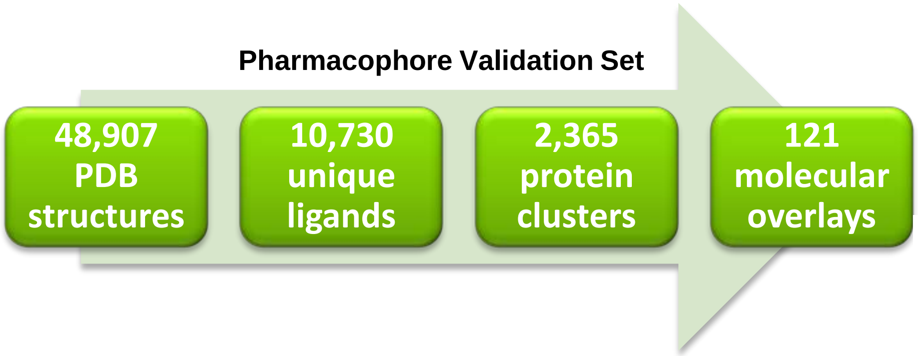


Overlay obtained using the CE algorithm as implemented in PyMOL

Final dataset

121 sets of molecular overlays for 119 targets

- 2 targets with an additional overlay of allosteric ligands
  - PDPK1 and FPS
- 9 targets with more than 40 ligands reduced through a contact analysis
  - Carbonic anhydrase II, CDK2, Thrombin, p38MAPK, HSP90, Trypsin, BACE, CHK1 and Glycogen phosphorylase
- Molecules put in a sensible charge and tautomer state
- Visual inspection of each molecule to guarantee a high quality set
  - Structures with bad conformations flag up as poor, but still included. We assume that they will not affect the feature assignment

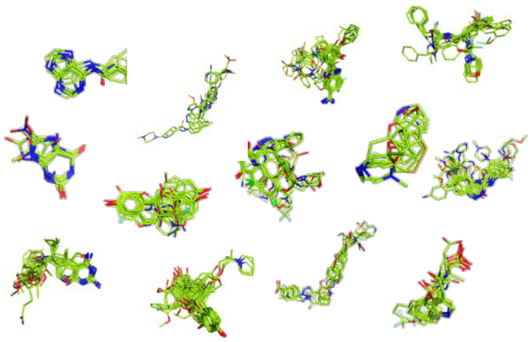

Pharmacophore Validation Set

48,907 PDB structures

10,730 unique ligands

2,365 protein clusters

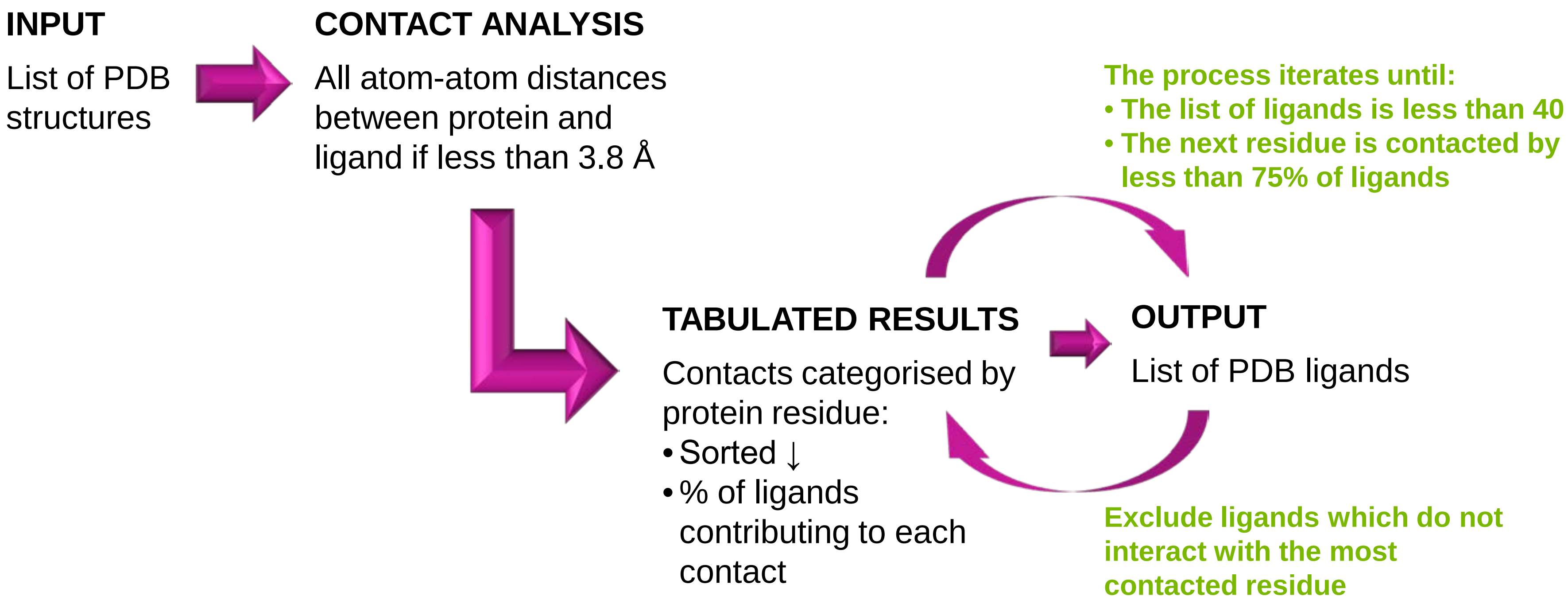
121 molecular overlays



Contact analysis

A rational way to subset clusters

Pharmacophore elucidation is a combinatorial problem - large data sets provide a barrier to validation.



INPUT

List of PDB structures

CONTACT ANALYSIS

All atom-atom distances between protein and ligand if less than 3.8 Å

TABULATED RESULTS

Contacts categorised by protein residue:

- Sorted ↓
- % of ligands contributing to each contact

OUTPUT

List of PDB ligands

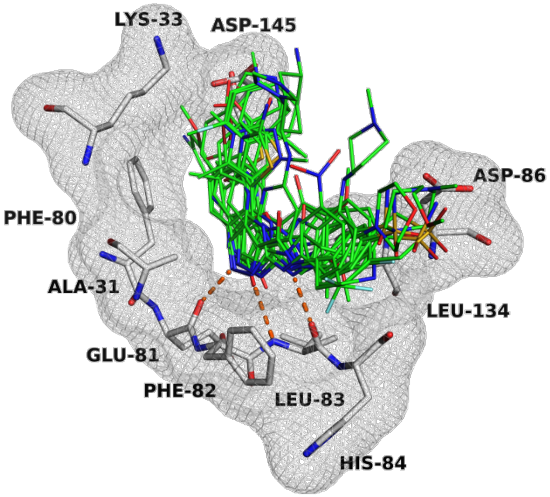
The process iterates until:

- **The list of ligands is less than 40**
- **The next residue is contacted by less than 75% of ligands**

Exclude ligands which do not interact with the most contacted residue

CDK2

From **105** ligands in the original set to **24** in the final set



Residue	Ligands with contact	%	BB contacts (%)	SC contacts (%)	Polar contacts (%)	Hphobe contacts (%)	HB contacts (%)
ALA31	24	100	0	100	0	100	0
ASP145	24	100	20	80	42	37	21
ASP86	24	100	17	83	64	13	23
GLU81	24	100	100	0	50	90	41
HIS84	24	100	100	0	45	36	18
ILE10	24	100	15	85	12	87	1
LEU134	24	100	0	100	0	100	0
LEU83	24	100	99	1	54	13	33
PHE80	24	100	0	100	0	100	0
PHE82	24	100	51	49	0	100	0

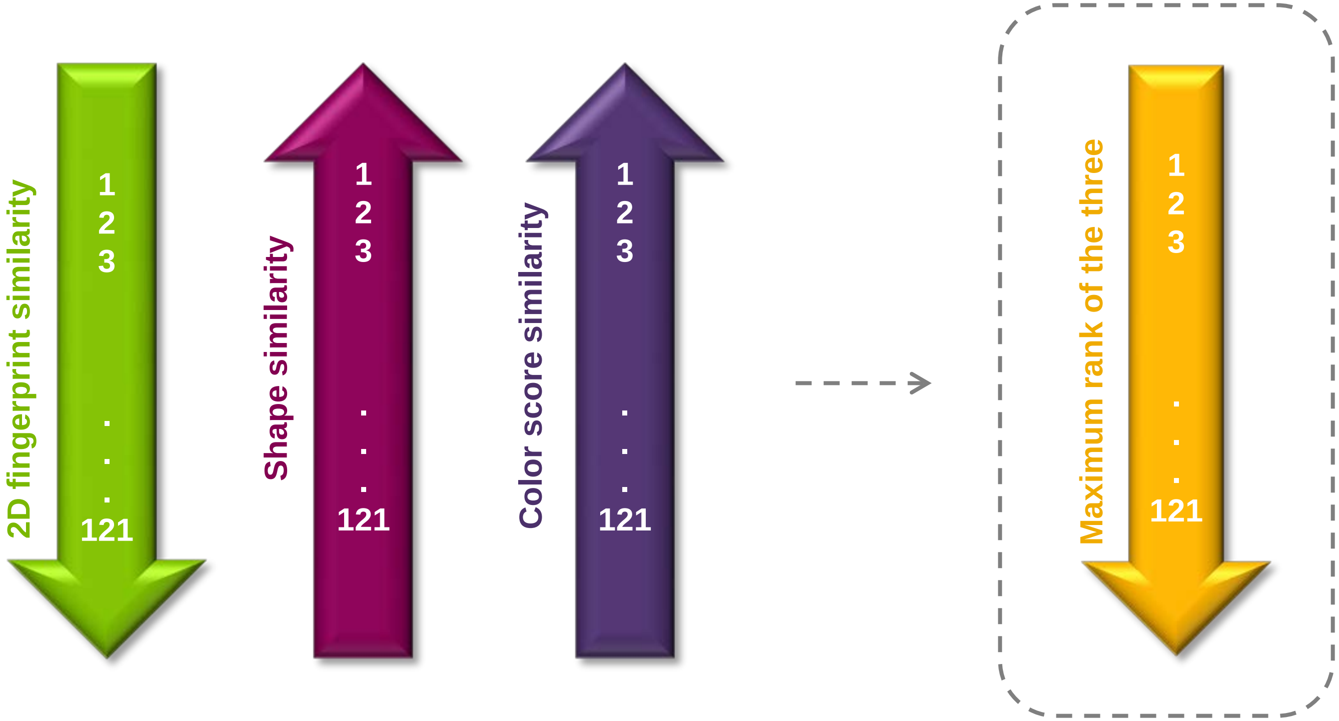
No.	Average 2D fingerprint similarity	Average shape similarity	Average color score
105	0.473	0.464	0.101
24	0.496	0.533	0.150

Scoring overlays

Good or poor for pharmacophore validation?

Consensus approach based on the maximum value of three parameters:

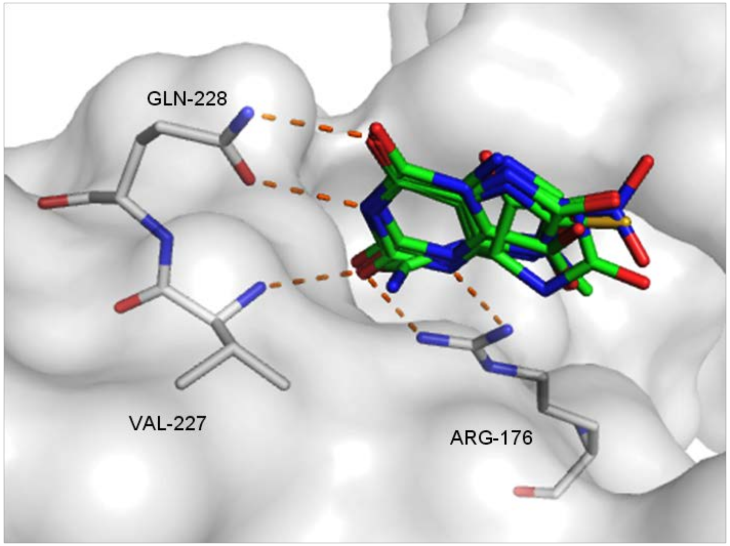
1. Average 2D similarity (in-house fingerprint)
2. Average shape similarity (OEShape toolkit)
3. Average color score (OEShape toolkit)

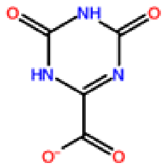


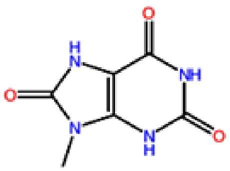
The best overlay

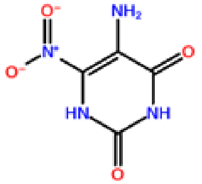
Uricase – 8 ligands

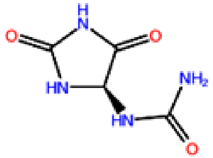
Good shape
Good color score
Good 2D similarity

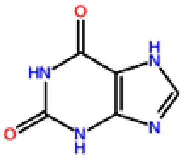


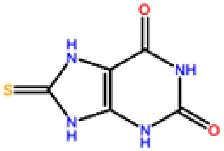

1r4u - OXC

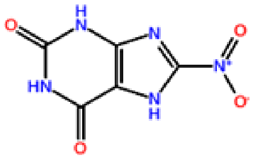

1r4s - MUA

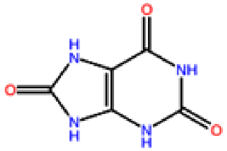

1wrr - UNC


2fxl - 2AL


3l8w - XAN


3lbg - 8TX

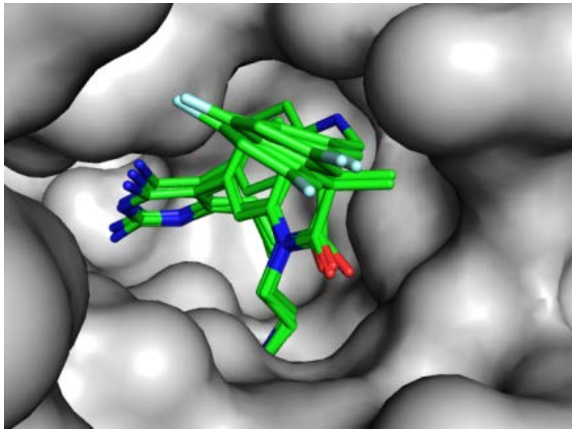

3ld4 - 8NX


3obp - URC

The poorest overlays

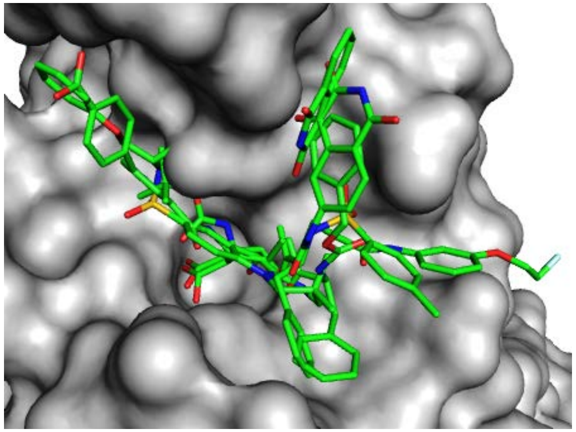
Renin – 5 ligands

Good shape
Good color score
Poor 2D similarity



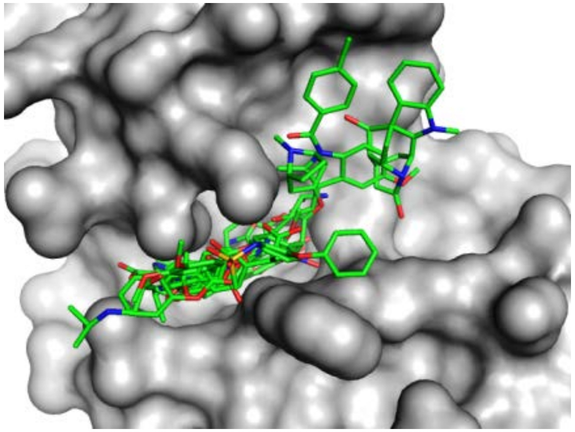
Caspase 3 – 7 ligands

Poor shape
Poor color score
Good 2D similarity



Phospholipase A2 – 16 ligands

Poor shape
Poor color score
Good 2D similarity



Conclusions

- The biggest and most diverse set ever published for pharmacophore validation – automated protocol
- A different and sensible approach to superimpose protein-ligand complexes, with better performance in cases of protein flexibility
- A rational way to reduce the number of ligands within a set, if higher than 40
- Overlays scored with a max-consensus based approach to distinguish between good and poor sets for pharmacophore validation

Acknowledgements

David A. Cosgrove

Martin J. Packer

AZ CompChem UK, Sweden, US

AstraZeneca

AZ global Post-doc program for funding

Jason Cole

Oliver Korb

John W. Liebeschuetz

Juliette Pradon

www.ccdc.cam.ac.uk

Robin Taylor

Confidentiality Notice