The Druggable Genome

Anna Gaulton
EMBL - European Bioinformatics Institute, ChEMBL Group
Overview

• What is druggability?
• Importance in drug discovery
• Overview of different methods
  • Precedence-based
  • Structure-based
  • Feature-based
• Beyond small-molecules
• The druggable genome
What is druggability?

Definitions of ‘druggability’ vary:

‘the ability of a protein target to bind small molecules with high affinity’
  - sometimes (perhaps more appropriately) called ‘ligandability’

Easier to assess

More predictive of success

‘the likelihood of finding orally bioavailable small molecules that bind to a particular target in a disease-modifying way’
  - includes consideration of likely pharmacokinetic and pharmacodynamic properties of compounds

(from Edfeldt, Folmer & Breeze, 2011)
What is druggability?

Definitions of ‘druggability’ vary:

‘the ability of a protein target to bind small molecules with high affinity’
- sometimes (perhaps more appropriately) called ‘ligandability’

Ability of a protein to be modulated by a drug-like small molecule
Easier to assess
More predictive of success

‘the likelihood of finding orally bioavailable small molecules that bind to a particular target in a disease-modifying way’
- includes consideration of likely pharmacokinetic and pharmacodynamic properties of compounds

(from Edfeldt, Folmer & Breeze, 2011)
Beautiful binding sites

• Orally bioavailable (drug-like) small molecules tend to have properties within certain parameters e.g.,
  • Mwt <= 500 Da,
  • LogP <= 5
  • H-bond Acceptors <= 10
  • H-bond Donors <= 5

  (Lipinski et al, 1997)

• In order to bind such compounds, a protein should have a binding site with complementary properties e.g.,
  • Appropriate size to accommodate a drug-like ligand
  • Buried, to increase interaction surface,
  • Not too polar
Beautiful binding sites

Beautiful PDE-5

Ugly CMV protease
Importance of assessing druggability

- Omics techniques are producing large amounts of data relating genes/proteins to disease e.g.,
  - Sequencing/GWAS studies to identify mutations
  - Expression studies in diseased vs normal tissue
  - Proteomics/metabolomics studies to find biomarkers

- Estimates suggest around 10-15% of human genome may be druggable (with small molecule approach)
Importance of assessing druggability

- Important to prioritise potential targets and pursue those that are most likely to be amenable to a small molecule approach
- Choosing a target that is unlikely to bind small molecules with high affinity is likely to result in failure of costly screening experiments
- Choosing a target that is unlikely to bind ‘drug-like’ small molecules may lead to later (even more costly) failure due to poor pharmacokinetic properties of compounds
- >60% projects may fail at lead identification/optimisation stages  (Brown & Superti-Furga, 2003)
Historical success

- Currently approved small molecule drugs target <400 human proteins

- Many of these are members of the same key families:
  - G protein-coupled receptors
  - Nuclear hormone receptors
  - Ion channels
  - Kinases

- Therefore, membership of one of these families may increase likelihood of success

- However, even within families, druggability can differ (e.g., peptide vs aminergic GPCRs)
Methods for assessing druggability

Sequence-based:
- Predicted druggable based on sequence features
  - Sequence analysis
    - e.g., Machine-learning algorithms

Structure-based:
- Protein structure contains drug-like pockets
  - Structural analysis
    - Protein Data Bank

Ligand-based:
- Binds endogenous drug-like ligands
  - Metabolite/ligand databases
    - ChEBI, HMDB, KEGG, IUPHAR

Precedence-based:
- High affinity drug-like compounds available
  - Pharmacology databases
    - ChEMBL, PubChem, BindingDB
- Compounds in clinical trials for the protein
  - Clinical trial databases
    - ClinicalTrials.gov
- Protein is an established small molecule drug target
  - Drug databases
    - DrugBank, TTD, DailyMed, ChEMBL

Increasing confidence in druggability

(from Campbell et al, 2010)
Precedence-based assessment

- If a protein is the target of an approved small molecule drug this gives a high degree of confidence in druggability.
- Also implies that targeting this protein is (relatively) safe.
- Caveats - still not an absolute guarantee of success for a different disease/product profile e.g.,
  - Agonist vs antagonist,
  - CNS-penetrant vs systemic
  - Long vs short-acting
  - Acceptability of side-effects
DrugBank

- Freely-available database of approved drugs and their targets: [http://www.drugbank.ca](http://www.drugbank.ca)
Abiraterone is a derivative of steroidal progesterone and is an innovative drug that offers clinical benefit to patients with hormone refractory prostate cancer. Abiraterone is administered as an acetate salt prodrug because it has a higher bioavailability and less susceptible to hydrolysis than abiraterone itself. FDA approved on April 28, 2011.

### Synonyms
- CB7630

### Salts
- Abiraterone Acetate

### Brand names
<table>
<thead>
<tr>
<th>Name</th>
<th>Company</th>
</tr>
</thead>
<tbody>
<tr>
<td>Zyliga</td>
<td>Janssen Biotech</td>
</tr>
</tbody>
</table>
### DrugBank: Abiraterone (DB05812)

**Categories**
- Antineoplastic Agents

**CAS number**
154229-19-3

**Weight**
- Average: 391.5457
- Monoisotopic: 391.251129305

**Chemical Formula**
C_{26}H_{33}NO_{2}

**InChI Key**
InChIKey:=UUVG8jCZGCLX3Z-3WZ79N

**InChI**
InChI=1S/C26H33NO2/c1-17(28)29-20-10-12-25(21)19(15-20)8-7-21-23-9-8-22(18-5-4-14-27-16-18)26(23,3)13-11-24(21)25/h4-6,8,14,16,20-21,23-24H,7,9-13,15H2,1-3H3/t20-21\_27,28\_29-25-26+\_m0/s1

**IUPAC Name**
(2'R,5S,15S)-2,15-dimethyl-14-(pyridin-3-yl)tetracyclo[8.7.0.0^2.7,0^11.15]heptadeca-7,13-dien-5-yl acetate

**SMILES**
Cc1o[c{O}][C@H]1Cc[c@H]2Cc2C(c(C)c[C@H]3Cc3Cc4Cc4Cc(C)cC(C)CCCN=Cc[C@H]4Cc4Cc3Cc2Cc1

**Mass Spec**
- Not Available

### Taxonomy

**Kingdom**
- Not Available

**Classes**
- Not Available

**Substructures**
- Not Available

### Pharmacology

**Indication**
Used in combination with prednisone for the treatment of metastatic, castration-resistant prostate cancer.

**Pharmacodynamics**
Abiraterone is associated with decreases in PSA levels, tumor shrinkage (as evaluated by RECIST criteria), radiographic regression of bone metastases and improvement in pain. Levels of adrenocorticotropic hormones increased up to 6-fold but this can be suppressed by dexamethasone.

**Mechanism of action**
Abiraterone is an orally active inhibitor of the steroidal enzyme CYP17A1 (17 alpha-hydroxylase/C17,20 lyase). It inhibits CYP17A1 in a selective and irreversible manner via covalent binding mechanism. CYP17A1 is an enzyme that catalyzes the biosynthesis of androgen and is highly expressed in testicular, adrenal, and prostatic tumor tissue. More specifically, abiraterone inhibits the conversion of 17-hydroxyprogrenolone to dehydroepiandrosterone (DHEA) by the enzyme CYP17A1 to decrease serum levels of testosterone and other androgens.

**Absorption**
Abiraterone itself is poorly absorbed and is susceptible to hydrolysis by esterases. The salt form, abiraterone acetate, is a prodrug which has a much higher oral bioavailability and is also esterase resistant. Peak drug concentrations of abiraterone were reached in 1.5 - 4 hours. Abiraterone acetate was rapidly and completely deacetylated into abiraterone-the parent salt form was not detectable in early pharmacokinetic studies. Food and high fat meals increases absorption 4.4-fold.

**Volume of distribution**
V_{dss}= 19,699 ± 13,358 L

**Protein binding**
>99% protein bound to alpha-1-acid glycoprotein and albumin.
1. **Cytochrome P450 17A1**

**Pharmacological action:** yes  
**Actions:** inhibitor

Conversion of pregnenolone and progesterone to their 17α,α-hydroxylated products and subsequently to dehydroepiandrosterone (DHEA) and androstenedione. Catalyzes both the 17α-hydroxylation and the 17,20-lyase reaction. Involved in sexual development during fetal life and at puberty.

**Organism class:** human  
**UniProt ID:** P05083  
**Gene:** CYP17A1  
**Protein Sequence:** FASTA  
**Gene Sequence:** FASTA  
**SNPs:** SNPJam Report

**References:**

2. **Cytochrome P450 3A4**

**Actions:** substrate

Cytochromes P450 are a group of heme-thiolate monoxygenases. In liver microsomes, this enzyme is involved in an NADPH-dependent electron transport pathway. It performs a variety of oxidation reactions (e.g. caffeine 8-oxidation, omeprazole sulfoxidation, midazolam 1′-hydroxylation and midazolam 4′-hydroxylation) of structurally unrelated compounds, including steroids, fatty acids, and xenobiotics. The enzyme also hydroxylates etoposide.

**UniProt ID:** P08684  
**Gene:** CYP3A4  
**Protein Sequence:** FASTA  
**Gene Sequence:** FASTA  
**SNPs:** SNPJam Report

**References:**

3. **Bile salt sulfotransferase**
Therapeutic Targets Database

- Non-commercial database of drug/candidate, target and indication information:
<table>
<thead>
<tr>
<th>Name</th>
<th>Infliximab</th>
</tr>
</thead>
<tbody>
<tr>
<td>Trade Name</td>
<td>Remicade</td>
</tr>
<tr>
<td>Company</td>
<td>Johnson &amp; Johnson</td>
</tr>
</tbody>
</table>

**Indication**
- Asthma [ICD9: 493, ICD10: J45]
- Psoriasis, Crohn's disease, ankylosing spondylitis, psoriatic arthritis, rheumatoid arthritis and ulcerative colitis [ICD9: 555, 556, 556.9, 696, 696.0, 710-719, 714, 720.0 ICD10: K50, K50-K52, K31, L40, L40.5, M00-M25, M05-M06, M07, M08.1, M45]

**Phase:** III

**Approved:** 1

**Therapeutic Class:** Immunosuppressive Agents

**CAS Number:** CAS: 170277-31-3

**Formular:** C642H9912N1694O1987546

**SuperDrug ATC ID:** L04AA12;

**Target**
- Tumor necrosis factor
- Tumor necrosis factor

**Ref 1**

**Ref 2**

**Ref 3**
ChEMBL

- EBI database of drug, target and bioactivity information: www.ebi.ac.uk/chembl
ChEMBL

- Manually curated efficacy targets for FDA approved drugs and WHO antimalarials
  - Targets with which drug interacts directly
  - Targets responsible for efficacy in approved indication
  - NOT targets responsible for adverse-effects or non-approved indications
  - NOT targets assigned purely on basis of pharmacology data

- Drug type (small molecule, antibody etc), action type (e.g., agonist, antagonist), and binding site/subunit information where available

- Deal with non-specific drugs and targets that are protein complexes

(Bento et al, 2013)
Clinical candidates

• If a target has compounds that have reached the clinic (e.g., phase I/II/III trials) this also provides good confidence in druggability

• Particularly, late-phase development candidates are likely to have shown a degree of safety and an adequate pharmacokinetic profile

• But many databases providing candidate information are commercial (and expensive)
ClinicalTrials.gov

- Freely accessible database of clinical trials: http://clinicaltrials.gov
### ChEMBL USAN Information

<table>
<thead>
<tr>
<th>Parent Molecule</th>
<th>Synonyms</th>
<th>Phase</th>
<th>Research Codes</th>
<th>Applicants</th>
<th>USAN Stem</th>
<th>USAN Year</th>
<th>First Approval</th>
<th>ATC Code</th>
<th>Icon</th>
</tr>
</thead>
<tbody>
<tr>
<td>CHEMBL2364642</td>
<td>Dasiprotimut-T (USAN)</td>
<td></td>
<td></td>
<td>Biovest International, Inc.</td>
<td></td>
<td>2013</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CHEMBL2017974</td>
<td>Buparlisib (INN, USAN)</td>
<td></td>
<td>BKM120-NX</td>
<td>Novartis</td>
<td></td>
<td>2013</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Buparlisib Hydrochloride (USAN)</td>
<td></td>
<td>BKM120-AAA</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CHEMBL2364626</td>
<td>Ralimetinib (USAN)</td>
<td></td>
<td>LY-22288220</td>
<td>Eli Lilly And Company</td>
<td>-ralib</td>
<td>2013</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Ralimetinib Mesylate (USAN)</td>
<td></td>
<td>LSN-2322600</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Rimegepant (USAN)</td>
<td></td>
<td>BMS-927711</td>
<td>Bristol-Myers Squibb</td>
<td>-gepant</td>
<td>2013</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Rimegepant Sulfate (USAN)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Candidate target information

- More difficult to identify targets for clinical candidates
- TTD contains some information
- May require searching of literature, company pipeline information etc
- Useful paper (lists targets but no drug links):
- Can also use bioactivity data to identify potential targets
Pharmacology data

- A number of public resources exist that capture pharmacology data e.g.,
  - ChEMBL – https://www.ebi.ac.uk/chembldb
  - BindingDB – http://www.bindingdb.org

- Existence of compounds that bind with high-affinity to the protein implies it is druggable

- But important to consider whether the compounds are ‘drug-like’, also whether there are any selectivity issues
• Many potent inhibitors for caspase-9 but largely peptidic
Endogenous ligands

- Even in the absence of ‘med-chem’ compounds, knowledge of the endogenous ligand/substrate of a protein can be useful in assessing druggability.

- Databases containing ligands can be used to identify proteins that contain small-molecule binding sites e.g.,
  - IUPHAR db: http://www.iuphar-db.org/
  - PDBe (crystal structures): http://www.ebi.ac.uk/pdbe/

- Proteins whose endogenous ligands are peptides/proteins (e.g., proteases), or which are involved in protein-protein interactions are less likely to be druggable.
IUPHAR db: a receptor/ion channel resource

Welcome to the official database of the IUPHAR Committee on Receptor Nomer

Detailed, peer-reviewed pharmacological, functional and pathophysiological information on human, mouse and rat Nuclear Hormone Receptors and selected Enzymes, including all Protein Kinases.

For a wider range of targets and overviews of their key properties visit the IUPHAR/BSG Guide to PHARMACOLOGY

**Natural/Endogenous Ligand(s)**

- melanin-concentrating hormone (Sp: Human, Mouse, Rat)

### Agonists

<table>
<thead>
<tr>
<th>Ligand</th>
<th>Sp.</th>
<th>Action</th>
<th>Affinity</th>
<th>Units</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>[Phe\textsuperscript{13},Tyr\textsuperscript{17}] - MCH</td>
<td>Hs</td>
<td>Full agonist</td>
<td>8.4 – 9.2</td>
<td>pIC\textsubscript{50}</td>
<td>14,37</td>
</tr>
<tr>
<td>melanin-concentrating hormone</td>
<td>Hs</td>
<td>Full agonist</td>
<td>8.2 – 9.2</td>
<td>pIC\textsubscript{50}</td>
<td>37</td>
</tr>
<tr>
<td>Ac-hMCH\textsubscript{6-16} - NH\textsubscript{2}</td>
<td>Hs</td>
<td>Full agonist</td>
<td>8.4 – 8.8</td>
<td>pIC\textsubscript{50}</td>
<td>34</td>
</tr>
<tr>
<td>salmon MCH</td>
<td>Hs</td>
<td>Full agonist</td>
<td>6.2 – 7.1</td>
<td>pIC\textsubscript{50}</td>
<td>14,37</td>
</tr>
</tbody>
</table>

### Antagonists

- compound 30
  - Hs
  - Antagonist
  - Affinity: 7.3
  - Units: pIC\textsubscript{50}
  - Reference: 34
- ATC0055
  - Hs
  - Antagonist
  - Affinity: 5.8 – 5.9
  - Units: pIC\textsubscript{50}
  - Reference: 26
Structure-based prediction methods

• Rely on identifying cavities in protein crystal structures and assessing the properties of these cavities

• Rules for properties that indicate a druggable cavity learnt from analysis of co-crystal complexes with drug-like ligands

• Can be applied to apo-protein structures to predict/score druggability
Algorithms

• PocketFinder – An, Totrov & Abagyan, 2005
  • Trained on a set of 5616 binding sites from PDB
  • Tested on 11,510 corresponding apo-protein structures – high agreement with liganded sites
  • Known ligand-binding site is largest site 80% of time (and in top 2 sites 92.7% of time)
  • Average volume of binding sites 610.8Å

• Druggability Indices – Hajduk, Huth & Fesik, 2005
  • Used NMR-based screening results against 23 proteins with diverse hit-rates to derive druggability rules
  • Properties of pockets (e.g., polar/apolar surface area, roughness, charged residues, shape and compactness) calculated and assessed for correlation with screening hit rate using regression analysis
  • No single factor predictive, but higher total surface area, apolar contacts, compactness and surface roughness generally found in druggable pockets
Algorithms

- **Volkamer, Kuhn, Grombacher, et al., 2012**
  - Trained on dataset of 1069 targets scored as druggable/difficult/undruggable
  - Support Vector Machine model developed based on pocket descriptors
  - In addition to global properties of binding pocket, local properties also taken into account using nearest neighbour method
  - Undruggable had more short-range hydrophilic-hydrophilic interactions and less short-range lipophilic-lipophilic interactions
  - Webserver available: [http://dogsite.zbh.uni-hamburg.de/](http://dogsite.zbh.uni-hamburg.de/)
Other algorithms

DrugEBIility

- [https://www.ebi.ac.uk/chembl/drugability/structure](https://www.ebi.ac.uk/chembl/drugability/structure)
- All potential pockets in crystal structures from PDB predicted using a pocket-finding algorithm (based on SurfNet, Laskowski 1995)
- Wide range of properties calculated for each cavity
- Decision tree algorithm trained on known binding pockets for drug-like ligands (e.g., rule-of-five)
- Decision tree used to classify unknown pockets into druggable/undruggable
- Second ‘tractability’ algorithm also trained with more relaxed ligand criteria (e.g., Mwt < 800)
e.g., PDE-5
e.g., PTP-1B
Details of sites identified

View cavities (and ligands) on structure
Issues with ligand/structure-based methods

- Ligand/structure-based druggability assessment gives highest degree of confidence, but not useful for novel targets or those that don’t have crystal structures available.

- Methods are needed to help prioritise the remainder of the proteins in the human genome, some of which may be ‘druggable’ but have not yet been investigated.

- Homology may help in some cases (e.g., member of known drug target family), but this is only based on ‘past success’ and not very useful in assessing novel families.
Feature-based druggability

- Various machine-learning/modelling methods can be used to identify more general ‘features’ of what makes a good drug target.
- Since these are based only on amino-acid sequence, they can be applied to whole genomes.
- Large numbers of different descriptors can be calculated e.g.,
  - Amino acid composition, length, hydrophobicity
  - Transmembrane domains, signal peptide, glycosylation sites
  - Secondary structure, domain composition
  - Subcellular localisation
• Analysed 148 human drug targets and 3573 ‘non targets’

• Wide range of sequence-based descriptors compared - eight rules found to be most discriminative:
  • High hydrophobicity
  • Length > 550 aa
  • Signal Peptide
  • No PEST motif
  • N-glycosylation sites > 2
  • O-glycosylation sites <=1
  • pI < 7.2
  • Membrane located

- These properties (plus others, such as amino-acid composition, secondary structure and low complexity regions) used to create a support vector machine classifier with a reported accuracy of 89%

- Within the non-target set, 668 potential targets identified with drug target-like features (i.e., predicted druggable)
Beyond small molecule druggability

• Based on current estimates, the overlap between small-molecule druggable and disease-modifying targets may be relatively small

• Therefore, other approaches may be necessary to target proteins that do not have ‘beautiful’ small molecule binding sites:
  • Inhibition of protein-protein interactions
  • Protein therapeutics
  • SiRNA etc
Protein therapeutic druggability

• The ability to target a protein with monoclonal antibodies/protein drugs depends largely on extra-cellular location
  • i.e., protein should be secreted or membrane bound

• This can be determined in a variety of ways e.g.,
  • Precedence (e.g., known protein therapeutic drugs)
  • Annotation/experimental evidence (e.g., known membrane protein, isolated from plasma)
  • Predictive methods (e.g., transmembrane domains, signal peptide, subcellular localisation prediction algorithms)
Protein-protein interactions

- Most proteins participate in protein-protein interactions, even if they don’t have a small molecule binding site, so modulating this interaction may be a highly effective means of modulating a target.

- However, protein-protein interaction interfaces are often large, flat surfaces (i.e., not beautiful binding sites).

- Targeting these is difficult, particularly with an oral small molecule drug, peptidomimetics more likely.

- Surfaces may have ‘hot-spots’ that contribute most of the binding affinity, allowing smaller inhibitors to target these.

- Many algorithms now being developed to predict such hot-spots and assess druggability.
Protein-protein interaction druggability


- Sugaya N, Furuya T. **Dr. PIAS: an integrative system for assessing the druggability of protein-protein interactions.** *BMC Bioinformatics.* 2011, 12, 50.


*This list is definitely not comprehensive!
The druggable genome

• Hopkins & Groom, 2002
The druggable genome

- Sakharkar, Sakharkar & Pervalz 2007
The druggable genome (based on ChEMBL)

- Targets of approved drugs

![Venn diagram showing the overlap between small molecules (384) and biotherapeutics (108) with 23 targets in common.](image-url)
The druggable genome

- Targets of drugs in development

- Small molecules ~600
- Biotherapeutics ~300
The druggable genome

- Targets with small molecules in ChEMBL

Small molecules
~1500
The druggable genome

- Small molecule-druggable

Small molecules
~2500
The druggable genome

- Extracellular (biotherapeutic druggable)

Biotherapeutics

\(~2500-4000\)
The druggable genome

- Total druggable space

Small molecules
~2500

Biotherapeutics
~2500-4000

~4500
References

- Fragment screening to predict druggability (ligandability) and lead discovery success. Edfeldt FN, Folmer RH, Breeze AL. Drug Discov Today. 2011; 16(7-8); 284-7.
- Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. Lipinski CA, Lombardo F, Dominy BW, Feeney PJ. Adv Drug Deliv Rev. 1997; 23(1-3); 3-25.
Acknowledgements

ChEMBL group

DrugEBillity:
• Bissan Al-Lazikani (ICR)
• Mishal Patel (ICR)
• Edith Chan (UCL)

Drug target curation:
• Rita Santos
• Ramesh Donadi
• Anne Hersey
• Patricia Bento
• Anneli Karlsson