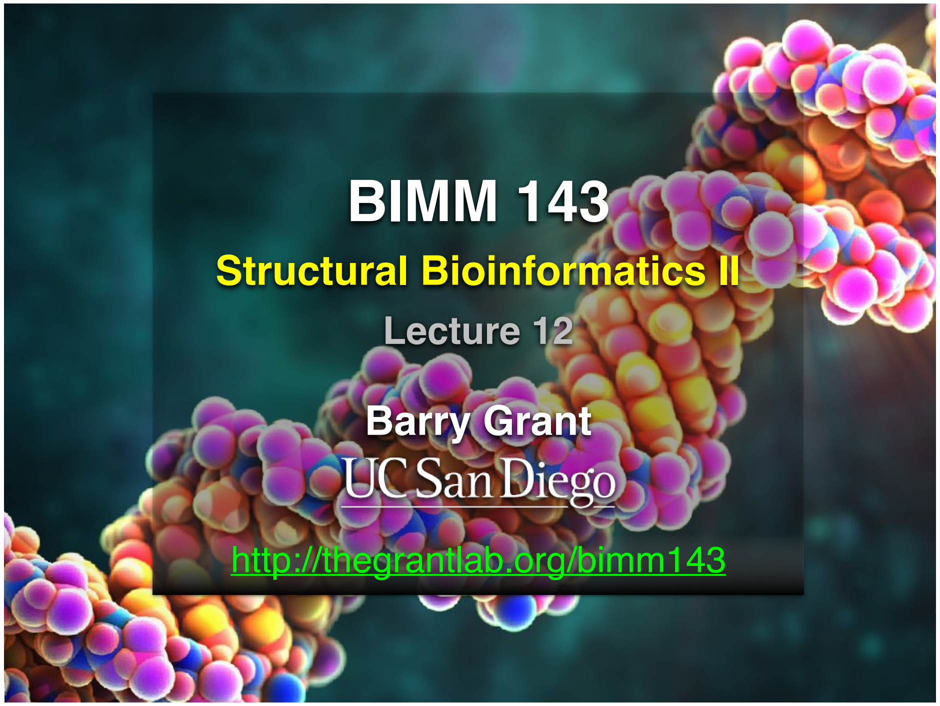


BIMM 143

Structural Bioinformatics II

Lecture 12

Barry Grant

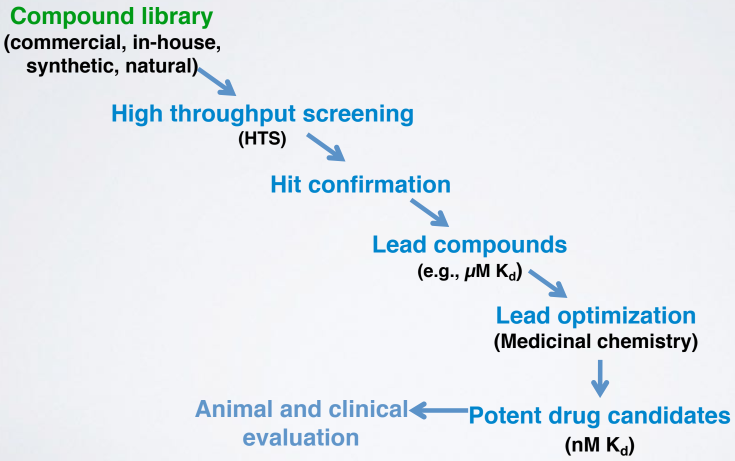
UC San Diego

http://thegrantlab.org/bimm143

NEXT UP:

- **Overview of structural bioinformatics**
  - Major motivations, goals and challenges
- **Fundamentals of protein structure**
  - Composition, form, forces and dynamics
- **Representing and interpreting protein structure**
  - Modeling energy as a function of structure
- **Example application areas**
  - **drug discovery** & Predicting functional dynamics

THE TRADITIONAL EMPIRICAL PATH TO DRUG DISCOVERY



Compound library (commercial, in-house, synthetic, natural)

High throughput screening (HTS)

Hit confirmation

Lead compounds (e.g., μM K_d)

Lead optimization (Medicinal chemistry)

Potent drug candidates (nM K_d)

Animal and clinical evaluation

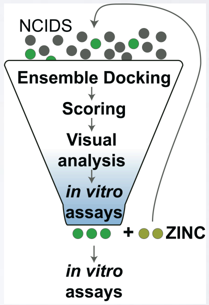
COMPUTER-AIDED LIGAND DESIGN

Aims to reduce number of compounds synthesized and assayed

Lower costs

Reduce chemical waste

Facilitate faster progress



Two main approaches:

(1). **Receptor/Target-Based**

(2). **Ligand/Drug-Based**

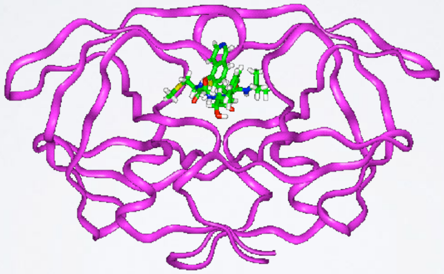
Two main approaches:

(1). **Receptor/Target-Based**

(2). **Ligand/Drug-Based**

SCENARIO 1: RECEPTOR-BASED DRUG DISCOVERY

Structure of Targeted Protein Known: Structure-Based Drug Discovery



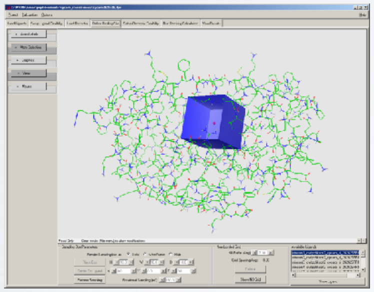
HIV Protease/KNI-272 complex

PROTEIN-LIGAND DOCKING

Structure-Based Ligand Design

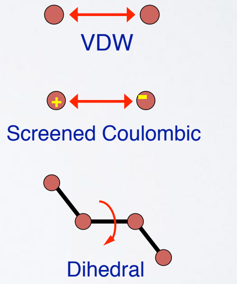
Docking software

Search for structure of lowest energy



Potential function

Energy as function of structure

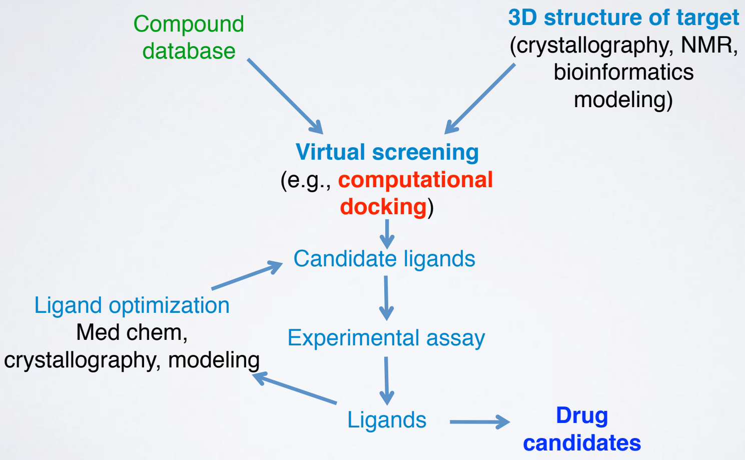


VDW

Screened Coulombic

Dihedral

STRUCTURE-BASED VIRTUAL SCREENING



Compound database

3D structure of target (crystallography, NMR, bioinformatics modeling)

Virtual screening (e.g., computational docking)

Candidate ligands

Experimental assay

Ligands

Drug candidates

Ligand optimization Med chem, crystallography, modeling

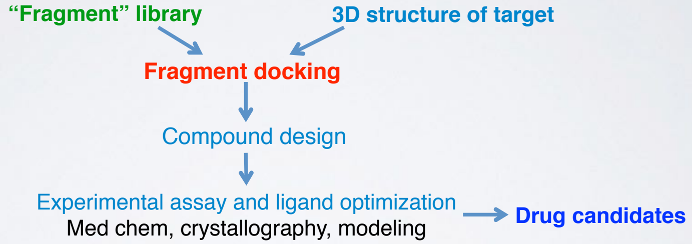
COMPOUND LIBRARIES

Commercial (in-house pharma)

Government (NIH)

Academia

FRAGMENTAL STRUCTURE-BASED SCREENING



“Fragment” library

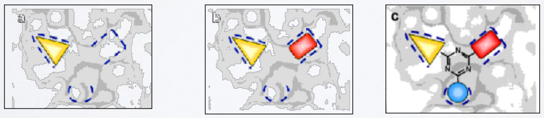
3D structure of target

Fragment docking

Compound design

Experimental assay and ligand optimization Med chem, crystallography, modeling

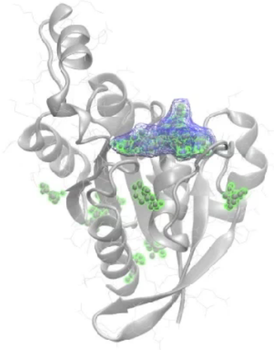
Drug candidates

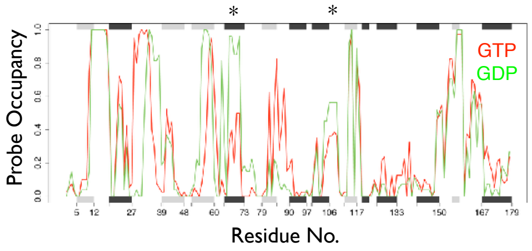


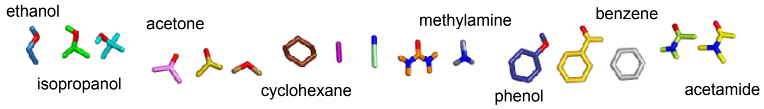
http://www.beilstein-institut.de/bozen2002/proceedings/Jhoti/jhoti.html

Multiple non active-site pockets identified

Small organic probe fragment affinities map multiple potential binding sites across the structural ensemble.



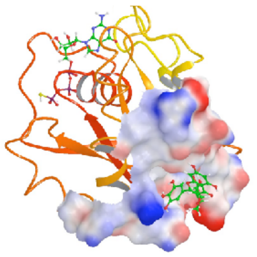


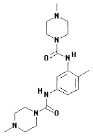


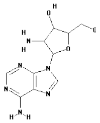
Ensemble docking & candidate inhibitor testing

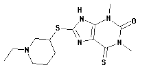
Top hits from ensemble docking against distal pockets were tested for inhibitory effects on basal ERK activity in glioblastoma cell lines.

Ensemble computational docking

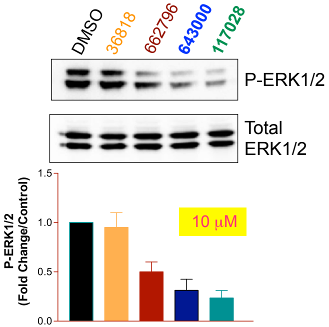








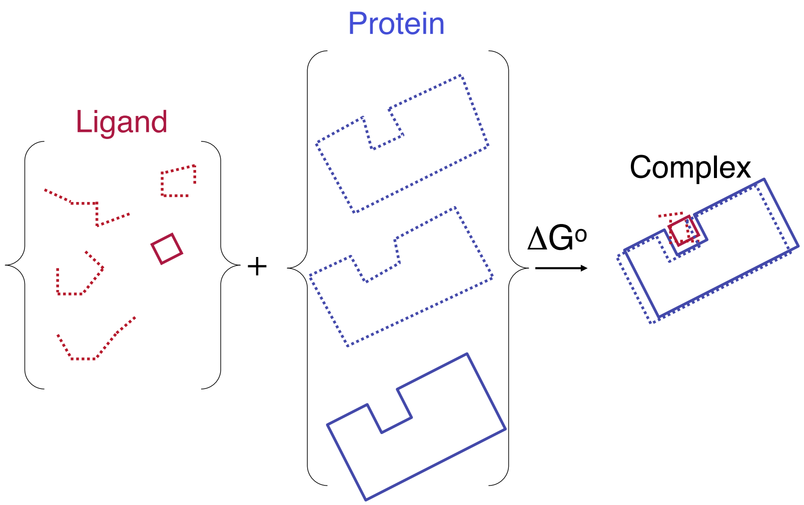
Compound effect on U251 cell line



Compound testing in cancer cell lines

PLoS One (2011, 2012)

Proteins and Ligand are Flexible



COMMON SIMPLIFICATIONS USED IN PHYSICS-BASED DOCKING

Quantum effects approximated classically

Protein often held rigid

Configurational entropy neglected

Influence of water treated crudely

Two main approaches:

(1). **Receptor/Target-Based**

(2). **Ligand/Drug-Based**

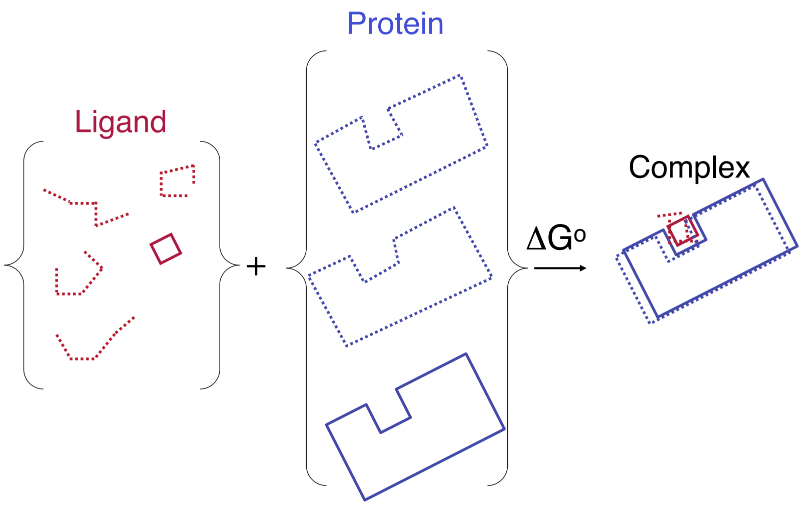
Do it Yourself!

Hand-on time!

https://bioboot.github.io/bimm143_W18/lectures/#12

You can use the classroom computers or your own laptops. If you are using your laptops then you will need to install **VMD** and **MGLTools**

Proteins and Ligand are Flexible



HTTP://129.177.232.111:3848/PCA-APP/

HTTPS://DCMB-GRANT-SHINY.UMMS.MED.UMICH.EDU/PCA-APP/

HTTP://BIO3D.UCSD.EDU/PCA-APP/

Two main approaches:

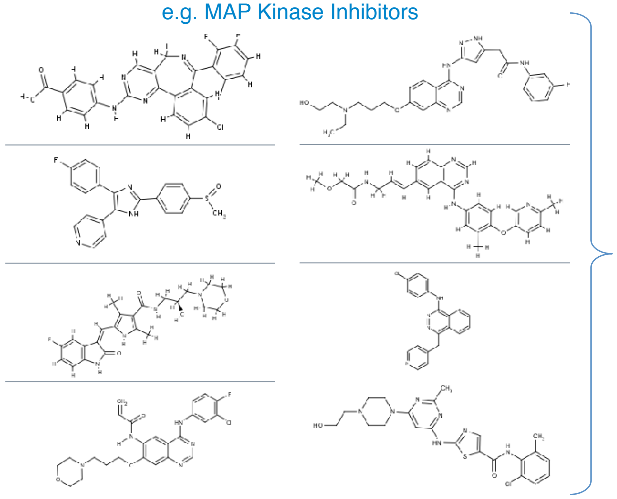
(1). **Receptor/Target-Based**

(2). **Ligand/Drug-Based**

Scenario 2

Structure of Targeted Protein Unknown:
Ligand-Based Drug Discovery

e.g. MAP Kinase Inhibitors



Using knowledge of existing inhibitors to discover more

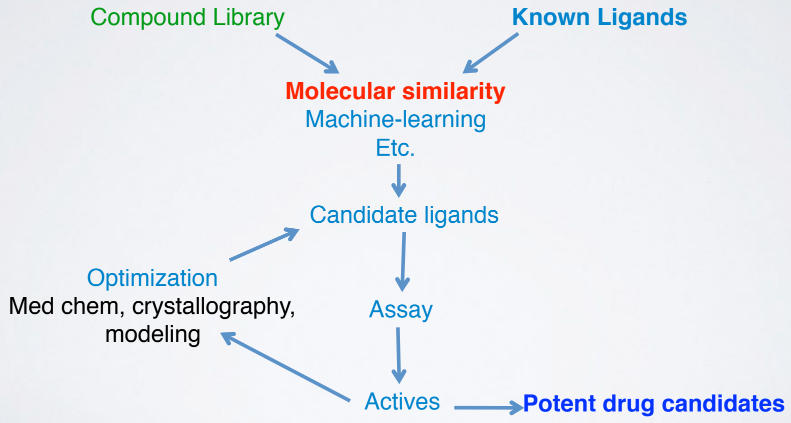
Why Look for Another Ligand if You Already Have Some?

Experimental screening generated some ligands, but they don’t bind tightly enough

A company wants to work around another company’s chemical patents

An high-affinity ligand is toxic, is not well-absorbed, difficult to synthesize etc.

LIGAND-BASED VIRTUAL SCREENING



Compound Library

Known Ligands

Molecular similarity
Machine-learning
Etc.

Candidate ligands

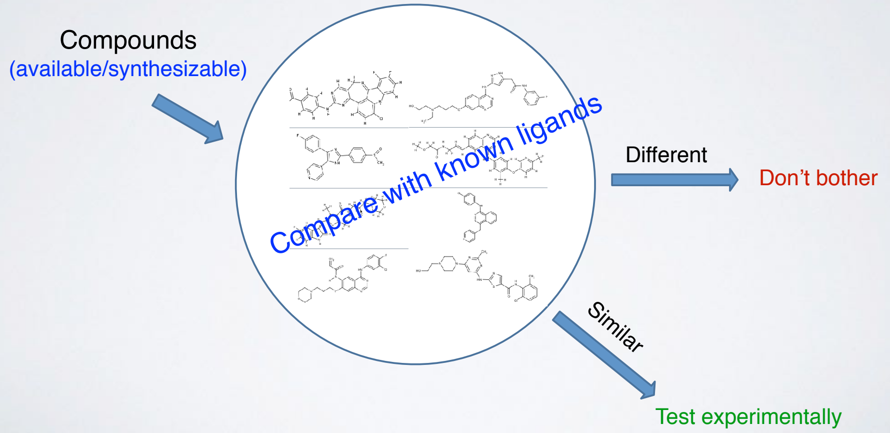
Assay

Actives

Potent drug candidates

Optimization
Med chem, crystallography,
modeling

CHEMICAL SIMILARITY
LIGAND-BASED DRUG-DISCOVERY



Compounds
(available/synthesizable)

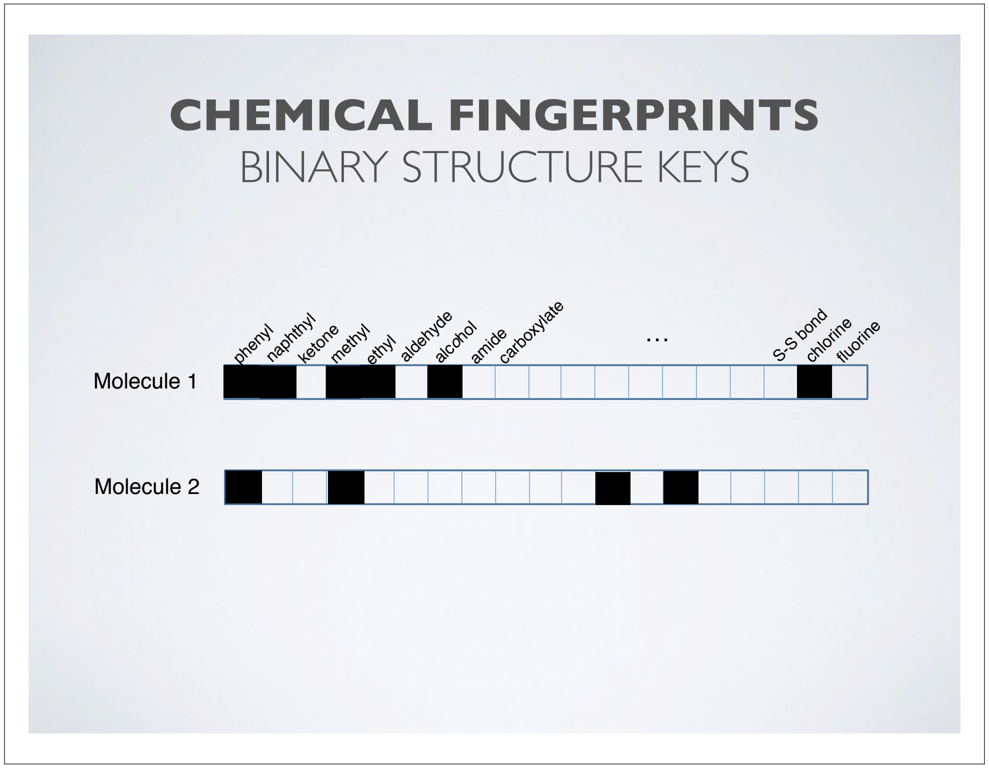
Compare with known ligands

Different

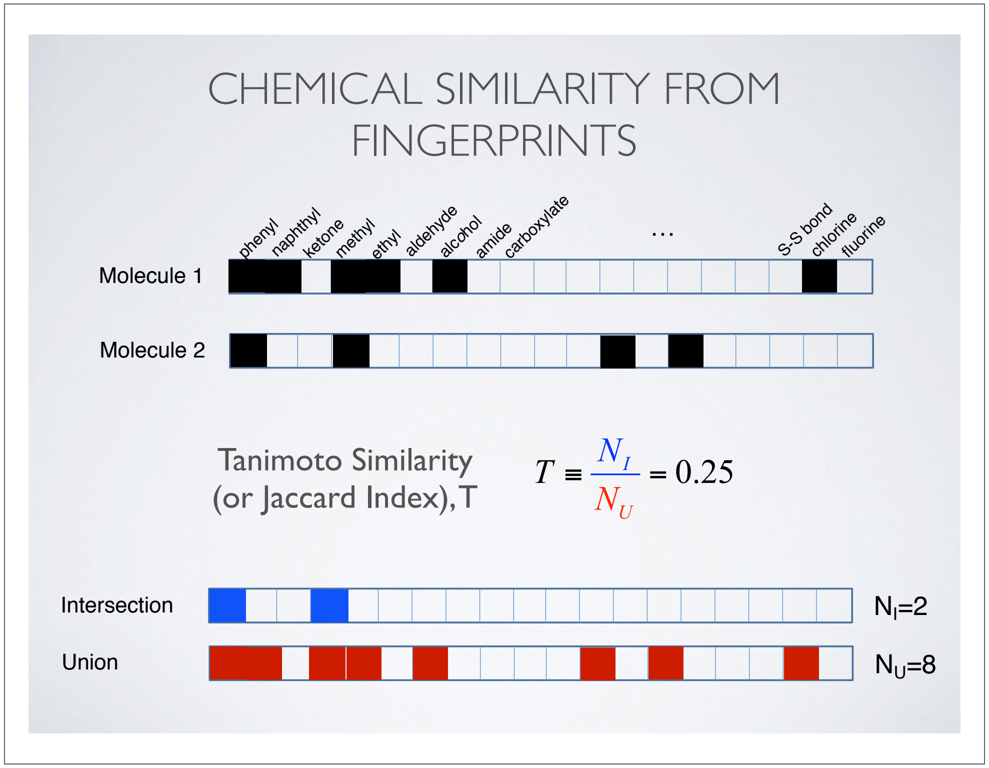
Don’t bother

Similar

Test experimentally



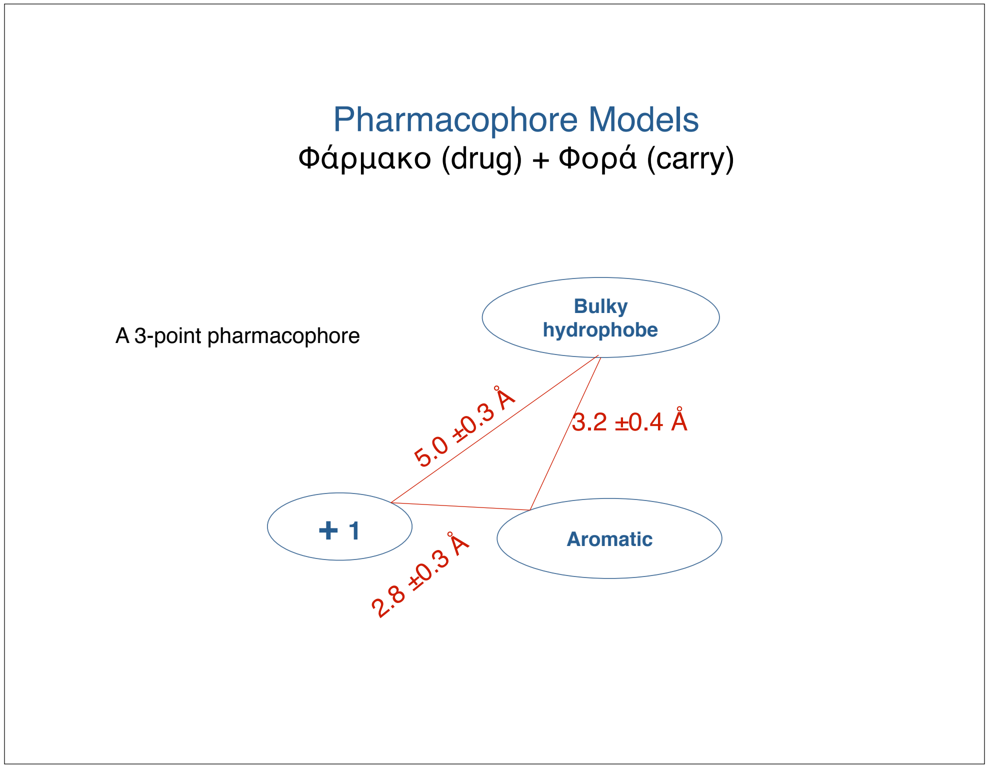
CHEMICAL FINGERPRINTS
BINARY STRUCTURE KEYS



CHEMICAL SIMILARITY FROM FINGERPRINTS

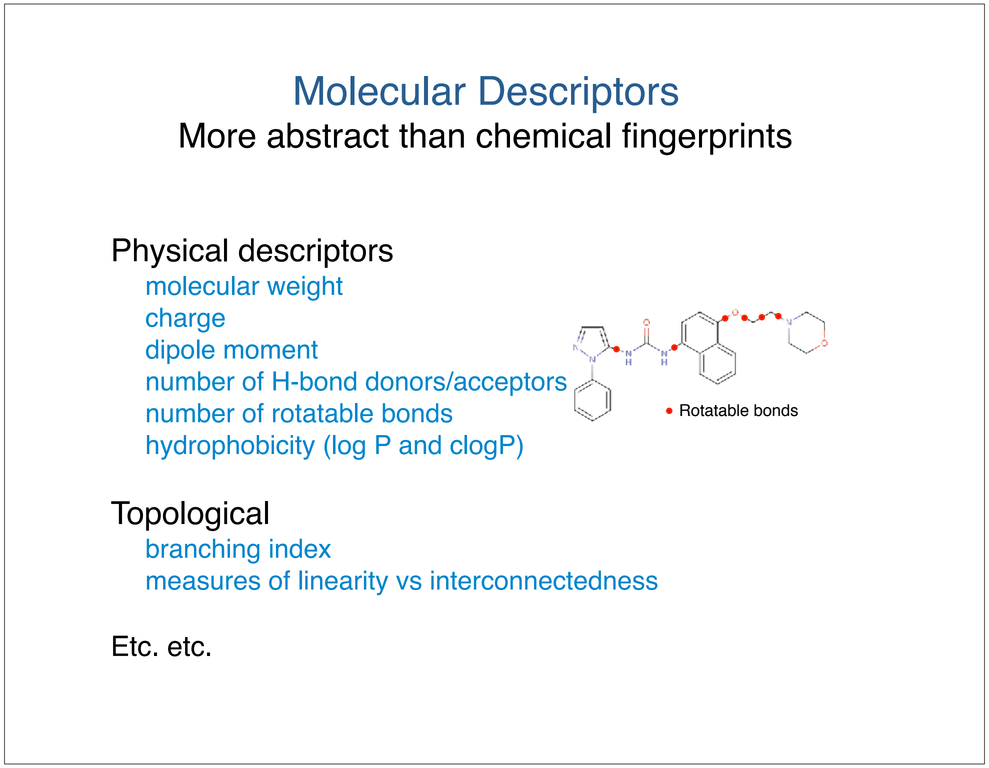
Tanimoto Similarity (or Jaccard Index), T

$$T \equiv \frac{N_I}{N_U} = 0.25$$



Pharmacophore Models
Φάρμακο (drug) + Φορά (carry)

A 3-point pharmacophore



Molecular Descriptors
More abstract than chemical fingerprints

Physical descriptors
- molecular weight
- charge
- dipole moment
- number of H-bond donors/acceptors
- number of rotatable bonds
- hydrophobicity (log P and clogP)

● Rotatable bonds

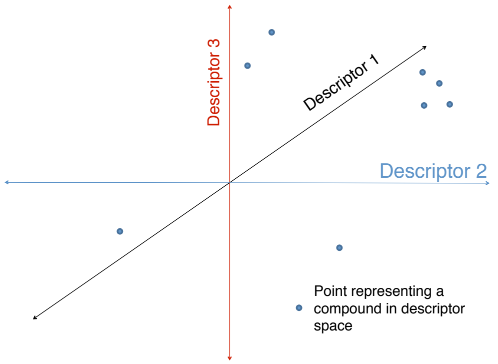
Topological
- branching index
- measures of linearity vs interconnectedness

Etc. etc.

A High-Dimensional “Chemical Space”

Each compound is at a point in an n-dimensional space

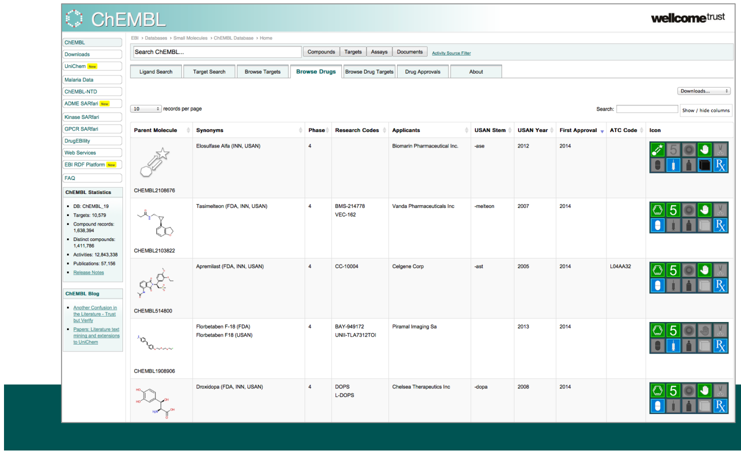
Compounds with similar properties are near each other



Apply **multivariate statistics** and **machine learning** for descriptor-selection. (e.g. partial least squares, PCA, support vector machines, random forest, deep learning etc.)

Approved drugs and clinical candidates

- Catalogue approved drugs and clinical candidates from FDA Orange Book, and USAN applications
- Small molecules and biotherapeutics



ChEMBL Statistics

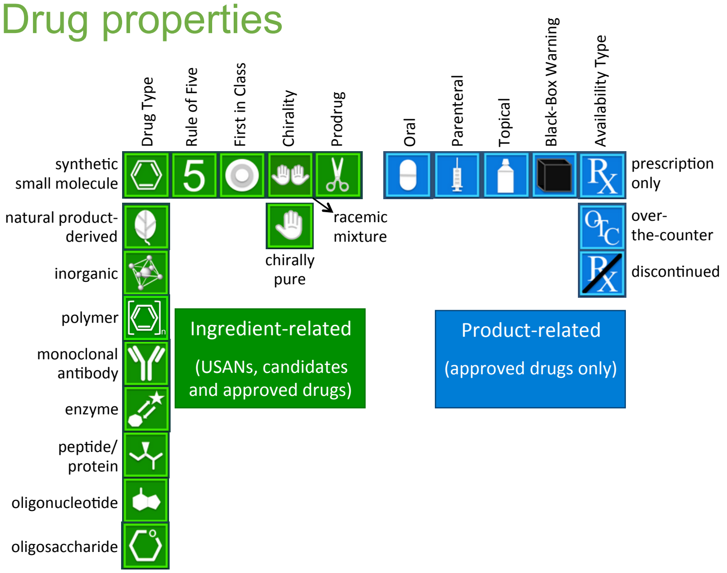
- DB: ChEMBL_19
- Targets: 10,579
- Compound records: 1,638,394
- Distinct compounds: 1,411,786
- Activities: 12,843,338
- Publications: 57,156
- Release Notes

ChEMBL Blog

- Another Confusion in the Literature - Trust but Verify
- Papers: Literature text mining and extensions to UniChem

Parent Molecule	Synonyms	Phase	Research Codes	Applicants	USAN Stem	USAN Year	First Approval	ATC Code	Icon
CHEMBL2108676	Elosulfase Alfa (INN, USAN)	4		Biomarin Pharmaceutical Inc.	-ase	2012	2014		
CHEMBL2103822	Tasimelteon (FDA, INN, USAN)	4	BMS-214778 VEC-162	Vanda Pharmaceuticals Inc	-melteon	2007	2014		
CHEMBL514800	Apremilast (FDA, INN, USAN)	4	CC-10004	Celgene Corp	-ast	2005	2014	L04AA32	
CHEMBL1908906	Florbetaben F-18 (FDA) Florbetaben F18 (USAN)	4	BAY-949172 UNII-TLA7312TOI	Piramal Imaging Sa		2013	2014		
	Droxidopa (FDA, INN, USAN)	4	DOPS L-DOPS	Chelsea Therapeutics Inc	-dopa	2008	2014		

Drug properties



Ingredient-related

(USANs, candidates and approved drugs)

Product-related

(approved drugs only)

LIPINSKI’S RULE OF FIVE

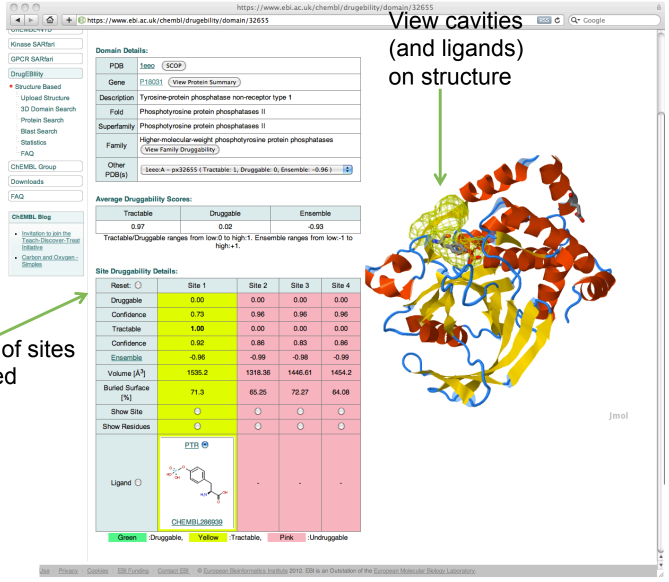
Lipinski’s rule of five states that, in general, an orally active drug has no more than one violation of the following criteria:

- Not more than 5 hydrogen bond donors (nitrogen or oxygen atoms with one or more hydrogen atoms)
- Not more than 10 hydrogen bond acceptors (nitrogen or oxygen atoms)
- A molecular mass less than 500 daltons
- An octanol-water partition coefficient log P not greater than 5

Rules for drug discovery success

- Set of approved drugs or medicinal chemistry compounds and their targets can be used to derive rules for drug discovery success (or failure):
  - What features make a successful drug target?
  - What features make a protein druggable by small molecules?
  - What features of a compound contribute to good oral bioavailability?
  - What chemical groups may be associated with toxicity?

Druggability prediction



https://www.ebi.ac.uk/chembl/drugebility/domain/32655

Kinase SARfari
GPCR SARfari
DrugEBIlity
- Structure Based
  - Upload Structure
  - 3D Domain Search
  - Protein Search
  - Blast Search
  - Statistics
  - FAQ

ChEMBL Group
Downloads
FAQ

ChEMBL Blog
- Invitation to join the Teach-Discover-Treat Initiative
- Carbon and Oxygen - Simples

Domain Details:

PDB	1eeo SCOP
Gene	P18031 View Protein Summary
Description	Tyrosine-protein phosphatase non-receptor type 1
Fold	Phosphotyrosine protein phosphatases II
Superfamily	Phosphotyrosine protein phosphatases II
Family	Higher-molecular-weight phosphotyrosine protein phosphatases View Family Druggability
Other PDB(s)	1eeo:A – px32655 (Tractable: 1, Druggable: 0, Ensemble: –0.96)

Average Druggability Scores:

Tractable	Druggable	Ensemble
0.97	0.02	-0.93

Tractable/Druggable ranges from low:0 to high:1. Ensemble ranges from low:-1 to high:+1.

Site Druggability Details:

Reset: ○	Site 1	Site 2	Site 3	Site 4
Druggable	0.00	0.00	0.00	0.00
Confidence	0.73	0.96	0.96	0.96
Tractable	**1.00**	0.00	0.00	0.00
Confidence	0.92	0.86	0.83	0.86
Ensemble	-0.96	-0.99	-0.98	-0.99
Volume [Å³]	1535.2	1318.36	1446.61	1454.2
Buried Surface [%]	71.3	65.25	72.27	64.08
Show Site	○	○	○	○
Show Residues	○	○	○	○
Ligand ○	PTR ● CHEMBL286939	-	-	-

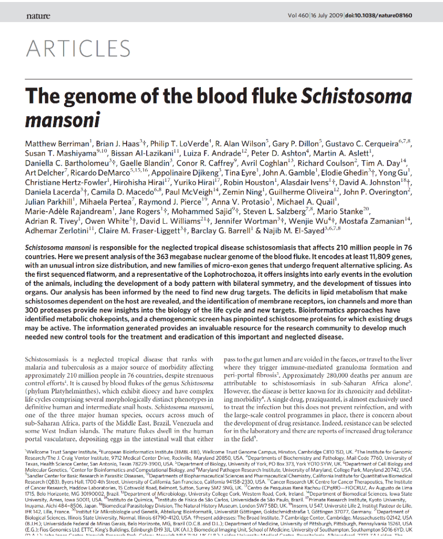
🟩 Green :Druggable, 🟨 Yellow :Tractable, 🟥 Pink :Undruggable

Use | Privacy | Cookies | EBI Funding | Contact EBI | © European Bioinformatics Institute 2012. EBI is an Outstation of the European Molecular Biology Laboratory.

View cavities (and ligands) on structure

Details of sites identified

Examples



nature Vol 460|16 July 2009|doi:10.1038/nature08160

ARTICLES

The genome of the blood fluke *Schistosoma mansoni*

Matthew Berriman[1], Brian J. Haas[3]†, Philip T. LoVerde[4], R. Alan Wilson[5], Gary P. Dillon[5], Gustavo C. Cerqueira[6,7,8], Susan T. Mashiyama[9,10], Bissan Al-Lazikani[11], Luiza F. Andrade[12], Peter D. Ashton[4], Martin A. Aslett[1], Daniella C. Bartholomeu[3]†, Gaelle Blandin[3], Conor R. Caffrey[9], Avril Coghlan[13], Richard Coulson[2], Tim A. Day[14], Art Delcher[7], Ricardo DeMarco[5,15,16], Appolinaire Djikeng[3], Tina Eyre[1], John A. Gamble[1], Elodie Ghedin[3]†, Yong Gu[1], Christiane Hertz-Fowler[1], Hirohisha Hirai[17], Yuriko Hirai[17], Robin Houston[1], Alasdair Ivens[1]†, David A. Johnston[18]†, Daniela Lacerda[3]†, Camila D. Macedo[6,8], Paul McVeigh[14], Zemin Ning[1], Guilherme Oliveira[12], John P. Overington[2], Julian Parkhill[1], Mihaela Pertea[7], Raymond J. Pierce[19], Anna V. Protasio[1], Michael A. Quail[1], Marie-Adèle Rajandream[1], Jane Rogers[1]†, Mohammed Sajid[9]†, Steven L. Salzberg[7,8], Mario Stanke[20], Adrian R. Tivey[1], Owen White[3]†, David L. Williams[21]†, Jennifer Wortman[3]†, Wenjie Wu[4]†, Mostafa Zamanian[14], Adhemar Zerlotini[11], Claire M. Fraser-Liggett[3]†, Barclay G. Barrell[1] & Najib M. El-Sayed[3,6,7,8]

Schistosoma mansoni is responsible for the neglected tropical disease schistosomiasis that affects 210 million people in 76 countries. Here we present analysis of the 363 megabase nuclear genome of the blood fluke. It encodes at least 11,809 genes, with an unusual intron size distribution, and new families of micro-exon genes that undergo frequent alternative splicing. As the first sequenced flatworm, and a representative of the Lophotrochozoa, it offers insights into early events in the evolution of the animals, including the development of a body pattern with bilateral symmetry, and the development of tissues into organs. Our analysis has been informed by the need to find new drug targets. The deficits in lipid metabolism that make schistosomes dependent on the host are revealed, and the identification of membrane receptors, ion channels and more than 300 proteases provide new insights into the biology of the life cycle and new targets. Bioinformatics approaches have identified metabolic chokepoints, and a chemogenomic screen has pinpointed schistosome proteins for which existing drugs may be active. The information generated provides an invaluable resource for the research community to develop much needed new control tools for the treatment and eradication of this important and neglected disease.

Schistosomiasis is a neglected tropical disease that ranks with malaria and tuberculosis as a major source of morbidity affecting approximately 210 million people in 76 countries, despite strenuous control efforts[1]. It is caused by blood flukes of the genus *Schistosoma* (phylum Platyhelminthes), which exhibit dioecy and have complex life cycles comprising several morphologically distinct phenotypes in definitive human and intermediate snail hosts. *Schistosoma mansoni*, one of the three major human species, occurs across much of sub-Saharan Africa, parts of the Middle East, Brazil, Venezuela and some West Indian islands. The mature flukes dwell in the human portal vasculature, depositing eggs in the intestinal wall that either pass to the gut lumen and are voided in the faeces, or travel to the liver where they trigger immune-mediated granuloma formation and peri-portal fibrosis[2]. Approximately 280,000 deaths per annum are attributable to schistosomiasis in sub-Saharan Africa alone[3]. However, the disease is better known for its chronicity and debilitating morbidity[4]. A single drug, praziquantel, is almost exclusively used to treat the infection but this does not prevent reinfection, and with the large-scale control programmes in place, there is concern about the development of drug resistance. Indeed, resistance can be selected for in the laboratory and there are reports of increased drug tolerance in the field[5].

[1]Wellcome Trust Sanger Institute, [2]European Bioinformatics Institute (EMBL-EBI), Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SD, UK. [3]The Institute for Genomic Research/The J. Craig Venter Institute, 9712 Medical Center Drive, Rockville, Maryland 20850, USA. [4]Departments of Biochemistry and Pathology, Mail Code 7760, University of Texas, Health Science Center, San Antonio, Texas 78229-3900, USA. [5]Department of Biology, University of York, PO Box 373, York YO10 5YW, UK. [6]Department of Cell Biology and Molecular Genetics, [7]Center for Bioinformatics and Computational Biology, and [8]Maryland Pathogen Research Institute, University of Maryland, College Park, Maryland 20742, USA. [9]Sandler Center for Basic Research in Parasitic Diseases, [10]Departments of Biopharmaceutical Sciences and Pharmaceutical Chemistry, California Institute for Quantitative Biomedical Research (QB3), Byers Hall, 1700 4th Street, University of California, San Francisco, California 94158-2330, USA. [11]Cancer Research UK Centre for Cancer Therapeutics, The Institute of Cancer Research, Haddow Laboratories, 15 Cotswold Road, Belmont, Sutton, Surrey SM2 5NG, UK. [12]Centro de Pesquisas René Rachou (CPqRR)—FIOCRUZ, Av Augusto de Lima 1715, Belo Horizonte, MG 30190002, Brazil. [13]Department of Microbiology, University College Cork, Western Road, Cork, Ireland. [14]Department of Biomedical Sciences, Iowa State University, Ames, Iowa 50011, USA. [15]Instituto de Quimica, [16]Instituto de Fisica de São Carlos, Universidade de São Paulo, Brazil. [17]Primate Research Institute, Kyoto University, Inuyama, Aichi 484–8506, Japan. [18]Biomedical Parasitology Division, The Natural History Museum, London SW7 5BD, UK. [19]Inserm, U 547, Université Lille 2, Institut Pasteur de Lille, IFR 142, Lille, France. [20]Institut für Mikrobiologie und Genetik, Abteilung Bioinformatik, Universität Göttingen, Goldschmidtstraße 1, Göttingen 37077, Germany. [21]Department of Biological Sciences, Illinois State University, Normal, Illinois 61790-4120, USA. †Present addresses: The Broad Institute, 7 Cambridge Center, Cambridge, Massachusetts 02142, USA (B.J.H.); Universidade Federal de Minas Gerais, Belo Horizonte, MG, Brazil (D.C.B. and D.L.); Department of Medicine, University of Pittsburgh, Pittsburgh, Pennsylvania 15261, USA (E.G.); Fios Genomics Ltd, ETTC, King's Buildings, Edinburgh EH9 3JL, UK (A.I.); Biomedical Imaging Unit, School of Medicine, University of Southampton, Southampton SO16 6YD, UK

NATURE CHEMISTRY | ARTICLE

Quantifying the chemical beauty of drugs

G. Richard Bickerton, Gaia V. Paolini, Jérémy Besnard, Sorel Muresan & Andrew L. Hopkins

Affiliations | Contributions | Corresponding author

Nature Chemistry **4**, 90–98 (2012) | doi:10.1038/nchem.1243
Received 01 September 2011 | Accepted 02 December 2011 | Published online 24 January 2012

Abstract

Abstract · References · Author information · Supplementary information

Drug-likeness is a key consideration when selecting compounds during the early stages of drug discovery. However, evaluation of drug-likeness in absolute terms does not reflect adequately the whole spectrum of compound quality. More worryingly, widely used rules may inadvertently foster undesirable molecular property inflation as they permit the encroachment of rule-compliant compounds towards their boundaries. We propose a measure of drug-likeness based on the concept of desirability called the quantitative estimate of drug-likeness (QED). The empirical rationale of QED reflects the underlying distribution of molecular properties. QED is intuitive, transparent, straightforward to implement in many practical settings and allows compounds to be ranked by their relative merit. We extended the utility of QED by applying it to the problem of molecular target druggability assessment by prioritizing a large set of published bioactive compounds. The measure may also capture the abstract notion of aesthetics in medicinal chemistry.

Subject terms: Pharmacology · Theoretical chemistry

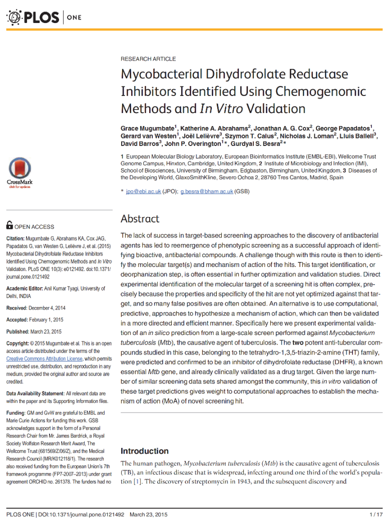
At a glance

Target prediction models

- Active compounds from ChEMBL can be used to train target prediction models
- Variety of methods used
  - Multi-Category Naïve Bayesian Classifier (e.g., ChEMBL)
  - Chemical similarity between ligand sets (e.g., SEA)
  - 3D similarity between ligands (e.g., SwissTargetPrediction)
  - Protein and ligand descriptors (e.g., Proteochemometric models)
- Open source tools available for many methods
  - E.g., Scikit-learn with RDKit

  Examples at: https://github.com/chembl/mychembl/blob/master/ipython_notebooks

Examples



PLOS | ONE

RESEARCH ARTICLE

Mycobacterial Dihydrofolate Reductase Inhibitors Identified Using Chemogenomic Methods and *In Vitro* Validation

Grace Mugumbate[1], Katherine A. Abrahams[2], Jonathan A. G. Cox[2], George Papadatos[1], Gerard van Westen[1], Joël Lelièvre[3], Szymon T. Calus[2], Nicholas J. Loman[2], Lluis Ballell[3], David Barros[3], John P. Overington[1]*, Gurdyal S. Besra[2]*

1 European Molecular Biology Laboratory, European Bioinformatics Institute (EMBL-EBI), Wellcome Trust Genome Campus, Hinxton, Cambridge, United Kingdom, 2 Institute of Microbiology and Infection (IMI), School of Biosciences, University of Birmingham, Edgbaston, Birmingham, United Kingdom, 3 Diseases of the Developing World, GlaxoSmithKline, Severo Ochoa 2, 28760 Tres Cantos, Madrid, Spain

* jpo@ebi.ac.uk (JPO); g.besra@bham.ac.uk (GSB)

CrossMark

OPEN ACCESS

Citation: Mugumbate G, Abrahams KA, Cox JAG, Papadatos G, van Westen G, Lelièvre J, et al. (2015) Mycobacterial Dihydrofolate Reductase Inhibitors Identified Using Chemogenomic Methods and *In Vitro* Validation. PLoS ONE 10(3): e0121492. doi:10.1371/journal.pone.0121492

Academic Editor: Anil Kumar Tyagi, University of Delhi, INDIA

Received: December 4, 2014

Accepted: February 1, 2015

Published: March 23, 2015

Copyright: © 2015 Mugumbate et al. This is an open access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Data Availability Statement: All relevant data are within the paper and its Supporting Information files.

Funding: GM and GvW are grateful to EMBL and Marie Curie Actions for funding this work. GSB acknowledges support in the form of a Personal Research Chair from Mr. James Bardrick, a Royal Society Wolfston Research Merit Award, The Wellcome Trust (681569/Z/06/Z), and the Medical Research Council (MR/K012118/1). The research also received funding from the European Union's 7th framework programme (FP7-2007–2013) under grant agreement ORCHID no. 261378. The funders had no

Abstract

The lack of success in target-based screening approaches to the discovery of antibacterial agents has led to reemergence of phenotypic screening as a successful approach of identifying bioactive, antibacterial compounds. A challenge though with this route is then to identify the molecular target(s) and mechanism of action of the hits. This target identification, or deorphanization step, is often essential in further optimization and validation studies. Direct experimental identification of the molecular target of a screening hit is often complex, precisely because the properties and specificity of the hit are not yet optimized against that target, and so many false positives are often obtained. An alternative is to use computational, predictive, approaches to hypothesize a mechanism of action, which can then be validated in a more directed and efficient manner. Specifically here we present experimental validation of an *in silico* prediction from a large-scale screen performed against *Mycobacterium tuberculosis* (*Mtb*), the causative agent of tuberculosis. The **two** potent anti-tubercular compounds studied in this case, belonging to the tetrahydro-1,3,5-triazin-2-amine (THT) family, were predicted and confirmed to be an inhibitor of dihydrofolate reductase (DHFR), a known essential *Mtb* gene, and already clinically validated as a drug target. Given the large number of similar screening data sets shared amongst the community, this *in vitro* validation of these target predictions gives weight to computational approaches to establish the mechanism of action (MoA) of novel screening hit.

Introduction

The human pathogen, *Mycobacterium tuberculosis* (*Mtb*) is the causative agent of tuberculosis (TB), an infectious disease that is widespread, infecting around one third of the world's population [1]. The discovery of streptomycin in 1943, and the subsequent discovery and

PLOS ONE | DOI:10.1371/journal.pone.0121492 March 23, 2015 1 / 17

ARTICLE

doi:10.1038/nature11159

Large-scale prediction and testing of drug activity on side-effect targets

Eugen Lounkine[1]*, Michael J. Keiser[2,3]*, Steven Whitebread[1], Dmitri Mikhailov[1], Jacques Hamon[4], Jeremy L. Jenkins[1], Paul Lavan[4], Eckhard Weber[4], Allison K. Doak[3], Serge Côté[4], Brian K. Shoichet[3] & Laszlo Urban[1]

Discovering the unintended 'off-targets' that predict adverse drug reactions is daunting by empirical methods alone. Drugs can act on several protein targets, some of which can be unrelated by conventional molecular metrics, and hundreds of proteins have been implicated in side effects. Here we use a computational strategy to predict the activity of 656 marketed drugs on 73 unintended 'side-effect' targets. Approximately half of the predictions were confirmed, either from proprietary databases unknown to the method or by new experimental assays. Affinities for these new off-targets ranged from 1 nM to 30 μM. To explore relevance, we developed an association metric to prioritize those new off-targets that explained side effects better than any known target of a given drug, creating a drug–target–adverse drug reaction network. Among these new associations was the prediction that the abdominal pain side effect of the synthetic oestrogen chlorotrianisene was mediated through its newly discovered inhibition of the enzyme cyclooxygenase-1. The clinical relevance of this inhibition was borne out in whole human blood platelet aggregation assays. This approach may have wide application to de-risking toxicological liabilities in drug discovery.

Adverse drug reactions (ADRs) can limit the use of otherwise effective drugs. Next to lack of efficacy, they are the leading cause for attrition in clinical trials of new drugs[1–3] and are more prominent still in the failure of molecules to advance from pre-clinical research into human trials[4]. Some ADRs are caused by modulation of the primary target of a drug[5], others result from non-specific interactions of reactive metabolites[6]. In many cases, however, ADRs are caused by unintended activity at off-targets. Notorious examples of off-target toxicity include that of the appetite suppressant fenfluramine–phentermine (fen–phen), which was withdrawn from the market after numerous patient deaths. These owed to the activation of the 5-hydroxytryptamine-2B (5-HT$_{2B}$) receptor by one of its metabolites, norfenfluramine, leading to proliferative valvular heart disease[7]. Similarly, well-known drugs, such as the antihistamine terfenadine, have been withdrawn because they caused arrhythmias and death, which have been attributed to their off-target inhibition of the human *ether-à-go-go*-related gene potassium channel (hERG, also known as KCNH2)[8,9]. Prediction of unknown off-target drug interactions might prevent such disastrous drug toxicities, which are often detected only after fatalities in the clinic, and might allow safer molecules to be prioritized for pre-clinical development. Methods to systematically predict off-targets, and associate these with side effects, have thus attracted intense interest[10–16], frequently in the form of either chemical genomics[17,18] or informatics[19–26] approaches.

Whereas the informatics methods have never been tested systematically on a large scale, in principle they can be deployed against thousands of targets. Here we present a large-scale, prospective evaluation of safety target prediction using one such method, the similarity ensemble approach (SEA)[25–27]. SEA calculates whether a molecule will bind to a target based on the chemical features it shares with those of known ligands, using a statistical model to control for random similarity. Because SEA relies only on chemical similarity, it can be applied systematically and, for those targets that have known ligands, comprehensively. For 656 drugs approved for human use (Supplementary Table 1), targets were predicted from among 73 proteins (Supplementary Table 2 and Methods) with established association of ADRs[22,28], for which assays were available at Novartis. Encouragingly, many of the predictions were confirmed, often at pharmacologically relevant concentrations. This motivated us to develop a guilt-by-association metric that linked the new targets to the ADRs of those drugs for which they are the primary or well-known off-targets, creating a drug–target–ADR network. The applicability and the limitations of this approach will be considered.

Testing the predictions

The 656 drugs were computationally screened for their likelihood to bind to 73 targets (Supplementary Table 2) using SEA[25–27]. The targets belong to the Novartis *in vitro* safety panels based on their association with ADRs[22,28]. Here we insisted that they also be described in the ChEMBL database[29], enabling correspondence with SEA predictions (Supplementary Table 2). ChEMBL annotates more than 285,000 ligands modulating more than 1,500 different human targets with affinities better than 30 μM. SEA calculated the similarity of each drug versus each set of ligands for the 73 targets, comparing the overall set similarity to a model of such expected at random. For instance, the sodium channel blocker aprindine loosely resembled the set of histamine H_1 ligands; although no single H_1 ligand was strongly similar to the drug (Table 1), the overall similarity of the set was much greater than expected at random, leading to a highly significant SEA expectation value (*E* value) of 5×10^{-26} between aprinidine and H_1 receptor ligands. Only 1,644 of the more than 47,000 possible drug–target pairs had significant *E* values. Of these, 403 were already known in ChEMBL and so were trivially confirmed; we do not consider these further. Of the remaining 1,241 predictions, 348 (28%) were unknown to ChEMBL, but could be found in proprietary ligand–target databases that were unavailable to SEA (see Methods). The remaining

[1]Novartis Institutes for Biomedical Research, Cambridge, Massachusetts 02139, USA. [2]SeaChange Pharmaceuticals Inc. 409 Illinois Street, San Francisco, California 94158, USA. [3]Department of Pharmaceutical Chemistry, University of California, San Francisco, 1700 4th Street, Byers Hall Suite 508D, California 94158-2550, USA. [4]Novartis Institutes for Biomedical Research, 4056 Basel, Switzerland.

*These authors contributed equally to this work.

21 JUNE 2012 | VOL 486 | NATURE | 361

©2012 Macmillan Publishers Limited. All rights reserved

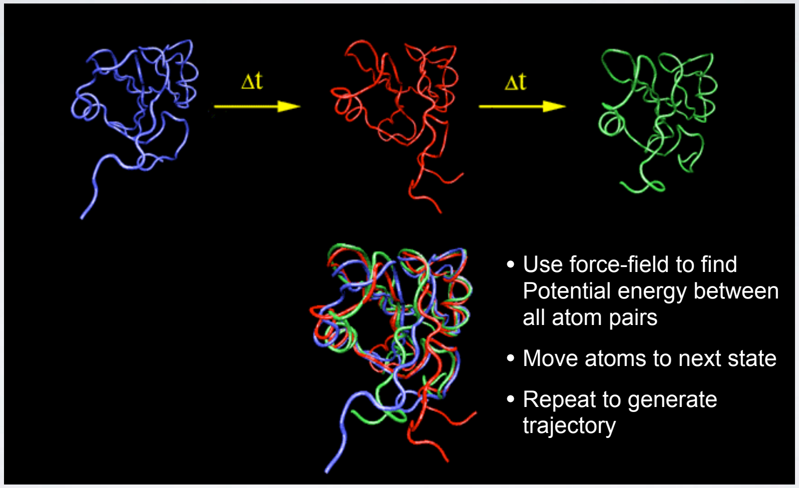
NEXT UP:

- **Overview of structural bioinformatics**
  - Major motivations, goals and challenges
- **Fundamentals of protein structure**
  - Composition, form, forces and dynamics
- **Representing and interpreting protein structure**
  - Modeling energy as a function of structure
- **Example application areas**
  - Drug discovery & predicting **functional dynamics**

PREDICTING FUNCTIONAL DYNAMICS

- **Proteins are intrinsically flexible molecules with internal motions that are often intimately coupled to their biochemical function**
  - E.g. ligand and substrate binding, conformational activation, allosteric regulation, etc.
- **Thus knowledge of dynamics can provide a deeper understanding of the mapping of structure to function**
  - **Molecular dynamics** (MD) and **normal mode analysis** (NMA) are two major methods for predicting and characterizing molecular motions and their properties

MOLECULAR DYNAMICS SIMULATION



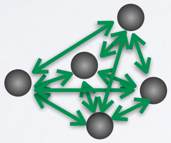
- Use force-field to find Potential energy between all atom pairs
- Move atoms to next state
- Repeat to generate trajectory

McCammon, Gelin & Karplus, *Nature* (1977)
[See: https://www.youtube.com/watch?v=ui1ZysMFcKk]

- Divide **time** into discrete (~1fs) **time steps** (**Δt**)
(for integrating equations of motion, see below)

- Divide **time** into discrete (~1fs) **time steps** (**Δt**)
(for integrating equations of motion, see below)

- At each time step calculate pair-wise atomic **forces** (***F(t)***)
(by evaluating **force-field** gradient)



Nucleic motion described classically

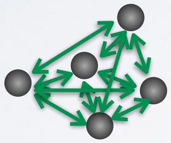
$$m_i \frac{d^2}{dt^2}\vec{R}_i = -\vec{\nabla}_i E(\vec{R})$$

Empirical force field

$$E(\vec{R}) = \sum_{\mathrm{bonded}} E_i(\vec{R}) + \sum_{\mathrm{non-bonded}} E_i(\vec{R})$$

- Divide **time** into discrete (~1fs) **time steps** (**Δt**)
(for integrating equations of motion, see below)

- At each time step calculate pair-wise atomic **forces** (***F(t)***)
(by evaluating **force-field** gradient)



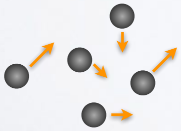
Nucleic motion described classically

$$m_i \frac{d^2}{dt^2}\vec{R}_i = -\vec{\nabla}_i E(\vec{R})$$

Empirical force field

$$E(\vec{R}) = \sum_{\mathrm{bonded}} E_i(\vec{R}) + \sum_{\mathrm{non-bonded}} E_i(\vec{R})$$

- Use the forces to calculate **velocities** and move atoms to new **positions**
(by integrating numerically via the “leapfrog” scheme)

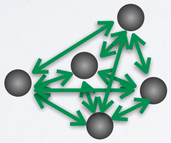


$$\begin{aligned} \boldsymbol{v}(t+\frac{\Delta t}{2}) &= \boldsymbol{v}(t-\frac{\Delta t}{2}) + \frac{\boldsymbol{F}(t)}{m}\Delta t \\ \boldsymbol{r}(t+\Delta t) &= \boldsymbol{r}(t) + \boldsymbol{v}(t+\frac{\Delta t}{2})\Delta t \end{aligned}$$

BASIC ANATOMY OF A MD SIMULATION

- Divide **time** into discrete (~1fs) **time steps** (**Δt**)
(for integrating equations of motion, see below)

- At each time step calculate pair-wise atomic **forces** (***F(t)***)
(by evaluating **force-field** gradient)



Nucleic motion described classically

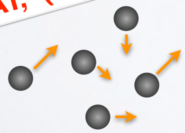
$$m_i \frac{d^2}{dt^2}\vec{R}_i = -\vec{\nabla}_i E(\vec{R})$$

Empirical force field

$$E(\vec{R}) = \sum_{\mathrm{bonded}} E_i(\vec{R}) + \sum_{\mathrm{non-bonded}} E_i(\vec{R})$$

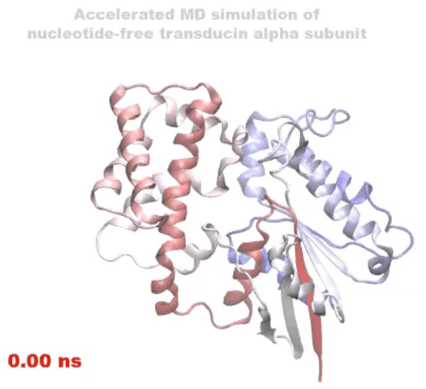
- Use the forces to calculate **velocities** and move atoms to new **positions**
(by integrating numerically via the “leapfrog” scheme)

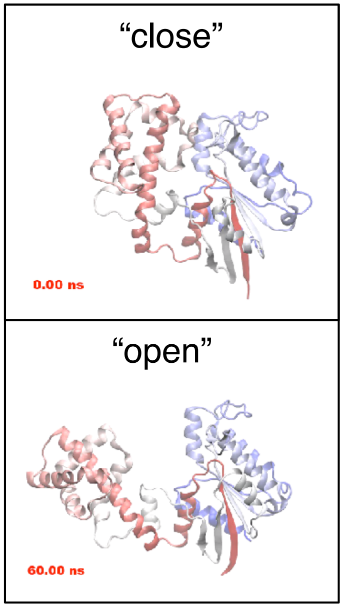
- **REPEAT, (iterate many, many times... 1ms = 10^{12} time steps)**



$$\begin{aligned} \boldsymbol{v}(t+\frac{\Delta t}{2}) &= \boldsymbol{v}(t-\frac{\Delta t}{2}) + \frac{\boldsymbol{F}(t)}{m}\Delta t \\ \boldsymbol{r}(t+\Delta t) &= \boldsymbol{r}(t) + \boldsymbol{v}(t+\frac{\Delta t}{2})\Delta t \end{aligned}$$

MD Prediction of Functional Motions

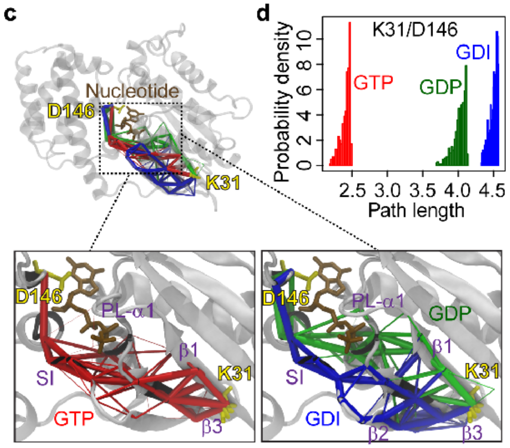

Accelerated MD simulation of
nucleotide-free transducin alpha subunit
0.00 ns


“close”

“open”

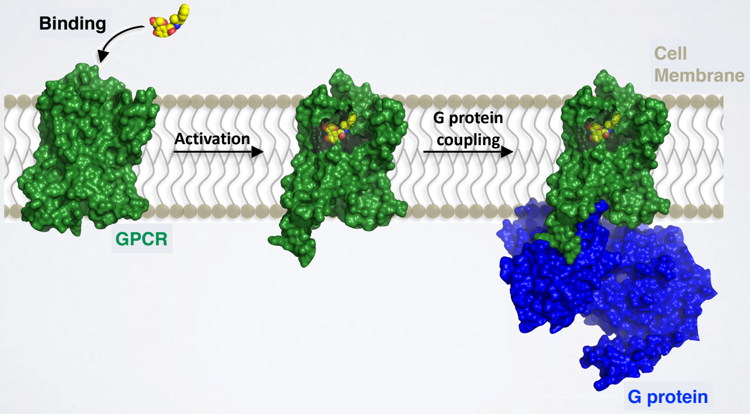
Yao and Grant, Biophys J. (2013)

Simulations Identify Key Residues Mediating Dynamic Activation

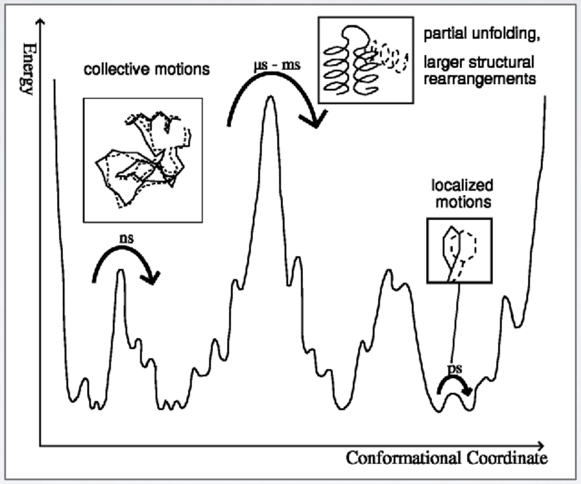


Yao … Grant, Journal of Biological Chemistry (2016)

EXAMPLE APPLICATION OF MOLECULAR SIMULATIONS TO GPCRS



PROTEINS JUMP BETWEEN MANY, HIERARCHICALLY ORDERED “CONFORMATIONAL SUBSTATES”



H. Frauenfelder et al., *Science* **229** (1985) 337

MOLECULAR DYNAMICS IS VERY EXPENSIVE

Example: F_1-ATPase in water (183,674 atoms) for 1 nanosecond:

=> 10^6 integration steps

=> 8.4 * 10^{11} floating point operations/step
[n(n-1)/2 interactions]

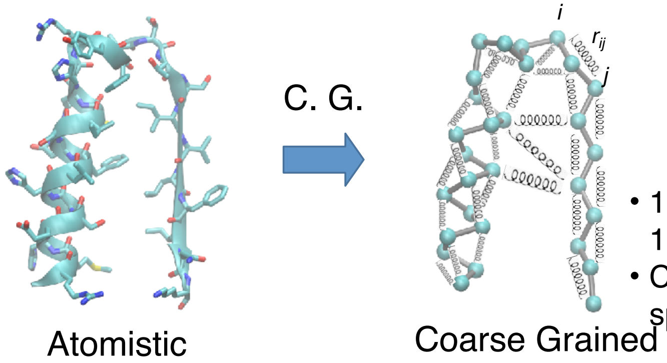
Total: 8.4 * 10^{17} flop
(on a 100 Gflop/s cpu: **ca 25 years!**)

… but performance has been improved by use of:

multiple time stepping	ca. 2.5 years
fast multipole methods	ca. 1 year
parallel computers	ca. 5 days
modern GPUs	**ca. 1 day**
(Anton supercomputer	**ca. minutes)**

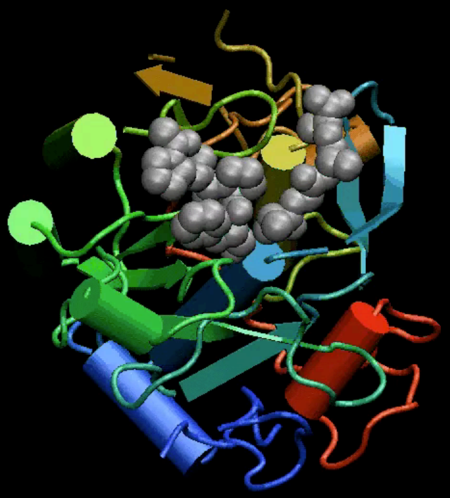
COARSE GRAINING: **NORMAL MODE ANALYSIS** (NMA)

- MD is still time-consuming for large systems
- Elastic network model NMA (ENM-NMA) is an example of a lower resolution approach that finishes in seconds even for large systems.



- 1 bead / 1 amino acid
- Connected by springs

NMA models the protein as a network of elastic strings



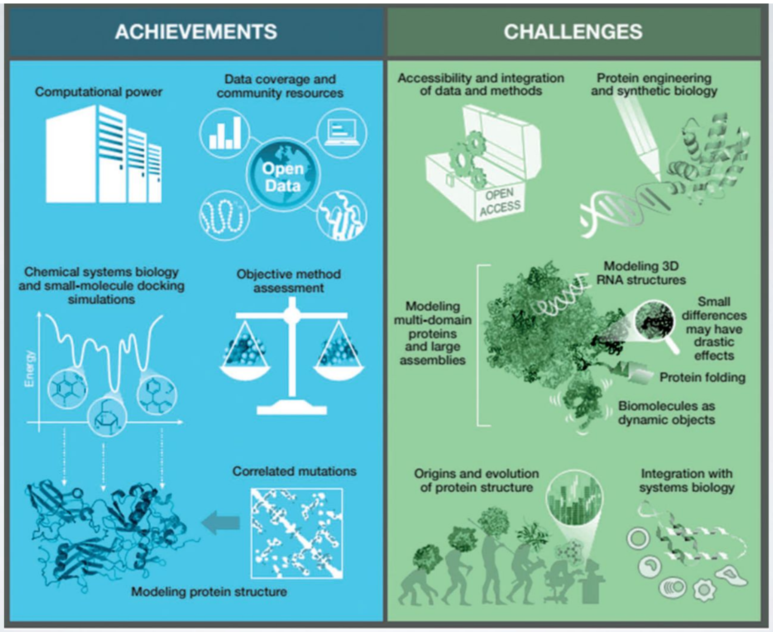
Proteinase K

Do it Yourself!

Hand-on time!

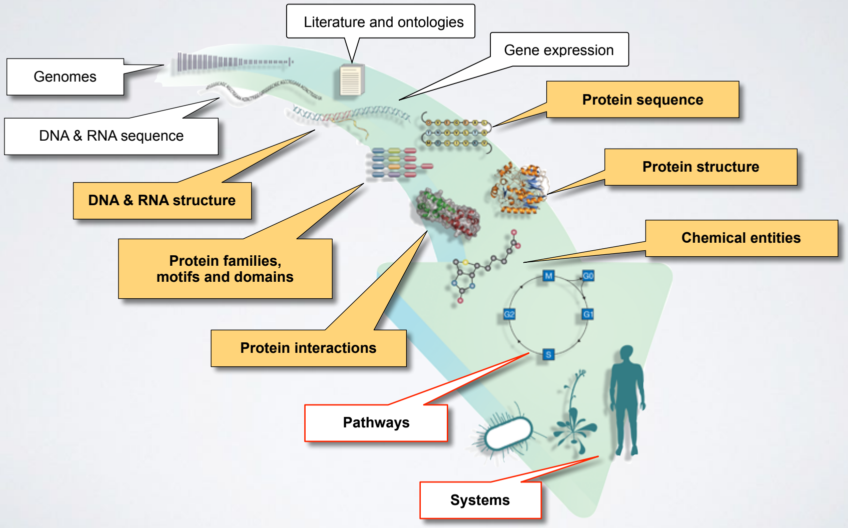
https://bioboot.github.io/bimm143_W18/lectures/#12

Focus on **section 3** & **4** exploring **PCA** and **NMA apps**


ACHIEVEMENTS
CHALLENGES
Computational power
Data coverage and community resources
Open Data
Chemical systems biology and small-molecule docking simulations
Energy
Objective method assessment
Correlated mutations
Modeling protein structure
Accessibility and integration of data and methods
OPEN ACCESS
Protein engineering and synthetic biology
Modeling multi-domain proteins and large assemblies
Modeling 3D RNA structures
Small differences may have drastic effects
Protein folding
Biomolecules as dynamic objects
Origins and evolution of protein structure
Integration with systems biology

Ilan Samish et al. Bioinformatics 2015;31:146-150

INFORMING SYSTEMS BIOLOGY?



SUMMARY

- Structural bioinformatics is computer aided structural biology
- Described major motivations, goals and challenges of structural bioinformatics
- Reviewed the fundamentals of protein structure
- Introduced both physics and knowledge based modeling approaches for describing the structure, energetics and dynamics of proteins computationally
- Introduced both structure and ligand based bioinformatics approaches for drug discovery and design