

A genome-wide quest for drug discovery targets against tropical diseases.

Marc A. Marti-Renom

Genome Biology Group (CNAG)
Structural Genomics Group (CRG)

***iCrea**
INSTITUCIÓ CATALANA DE
RECERCA I ESTUDIS AVANÇATS

***iCrea**
INSTITUCIÓ CATALANA DE
RECERCA I ESTUDIS AVANÇATS

cnag  **CRG**
Centre
for Genomic
Regulation

Outline...

COMPARATIVE MODELING

EXAMPLES

THE TROPICAL DISEASE INITIATIVE

Nomenclature

Homology: Sharing a common ancestor, may have similar or dissimilar functions

Similarity: Score that quantifies the degree of relationship between two sequences.

Identity: Fraction of identical aminoacids between two aligned sequences (case of similarity).

Target: Sequence corresponding to the protein to be modeled.

Template: 3D structure/s to be used during protein structure prediction.

Model: Predicted 3D structure of the target sequence.

Nomenclature

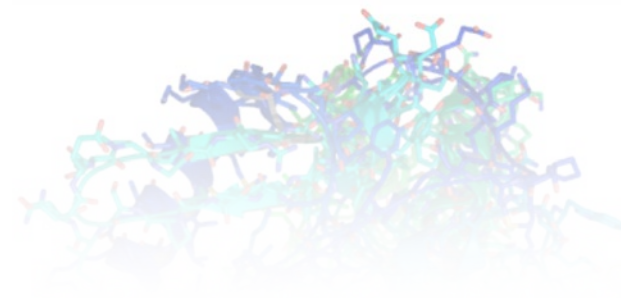
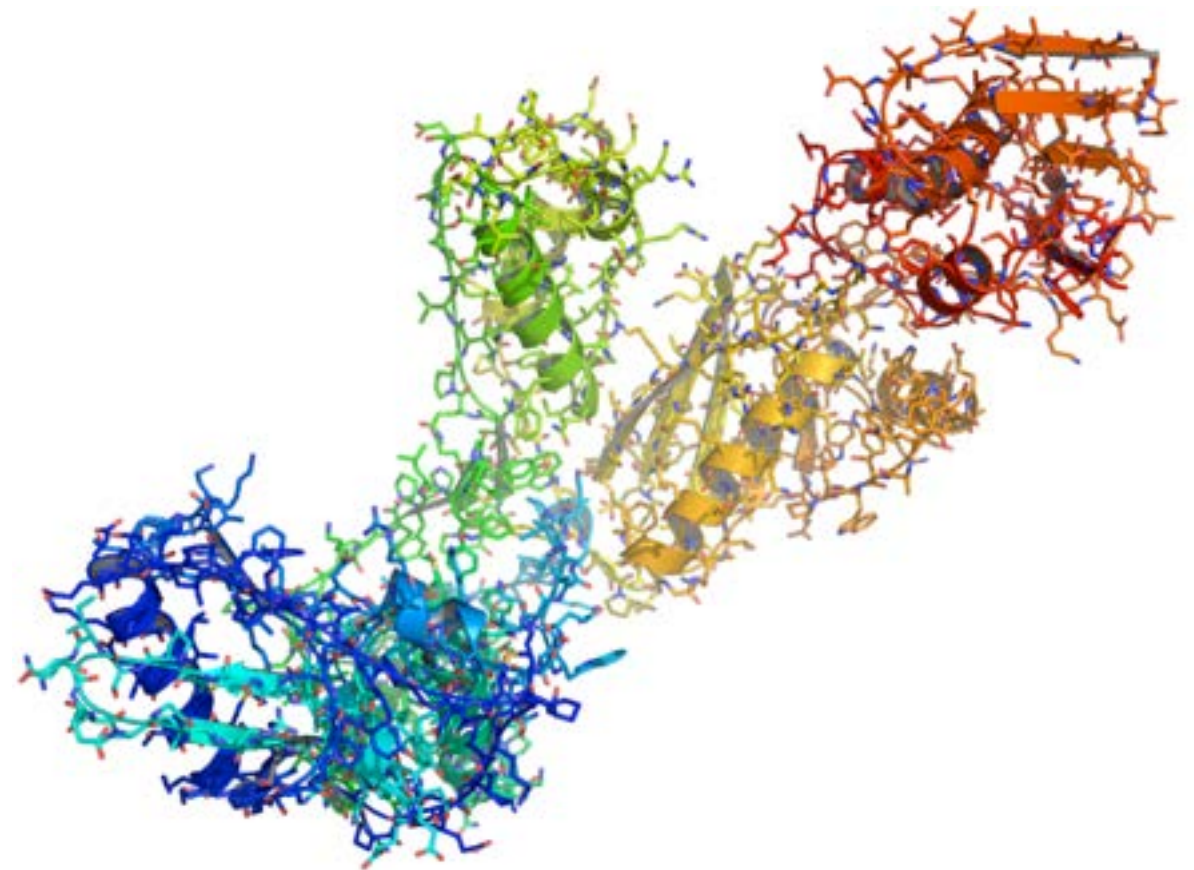
Fold: Three dimensional conformation of a protein sequence (usually at domain level).

Domain: Structurally globular part of a protein, which may independently fold.

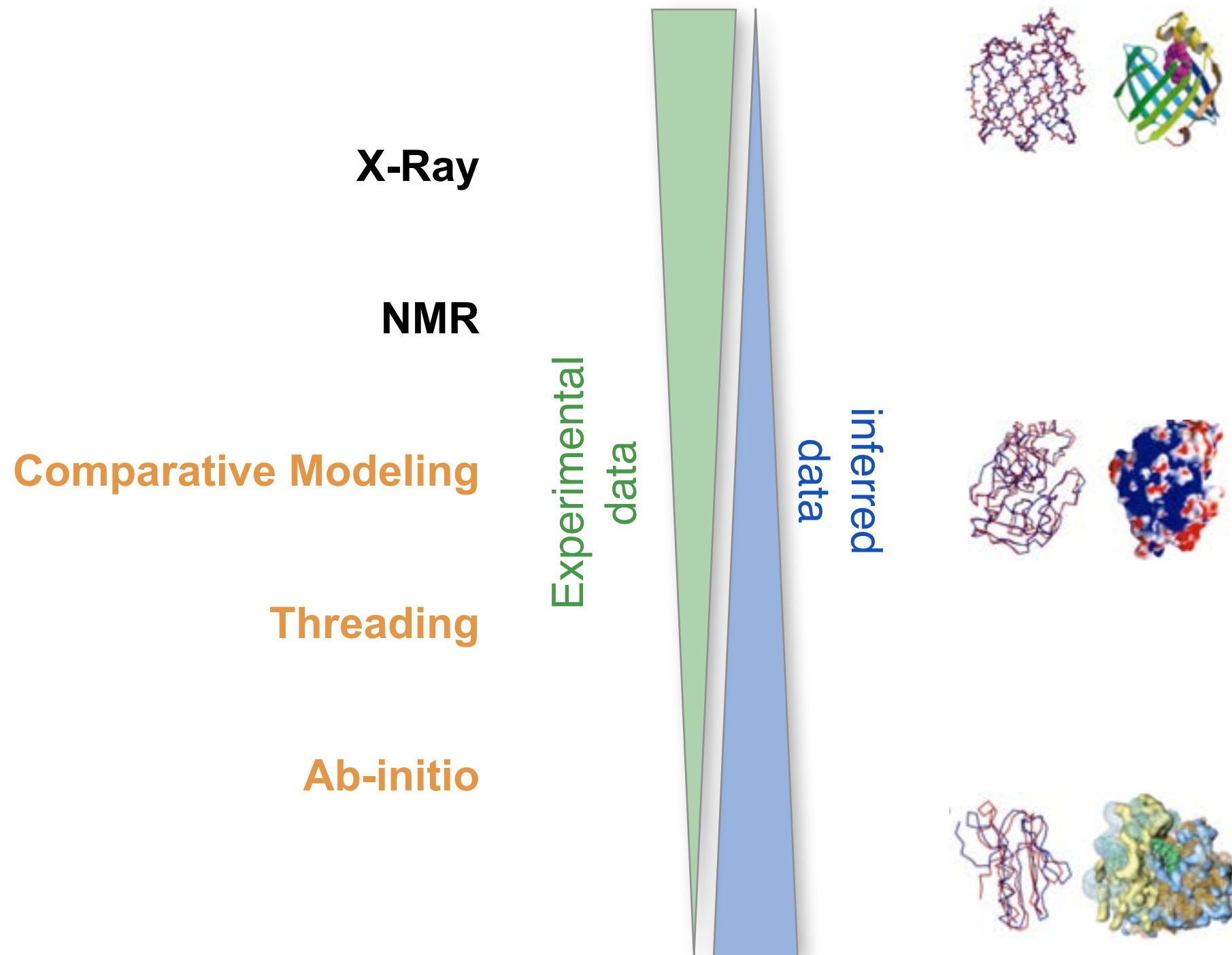
Secondary Structure: Regular sub-domain structures composed by alpha-helices, beta-sheets and coils (or loops).

Backbone: Protein structure skeleton composed by the carbon, nitrogen and oxygen atoms.

Side-Chain: Specific atoms identifying each of the 20 residues types.



protein prediction .vs. protein determination



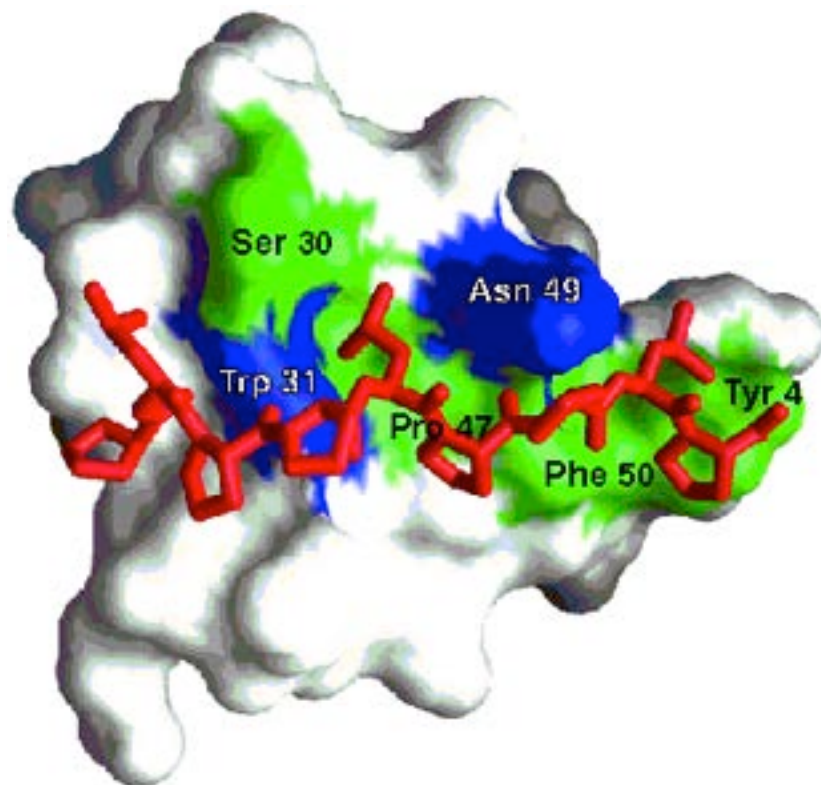
Why is it useful to know the **structure** of a protein, not only its sequence?

- ◆ The biochemical function (activity) of a protein is defined by its interactions with other molecules.
- ◆ The biological function is in large part a consequence of these interactions.
- ◆ The 3D structure is more informative than sequence because interactions are determined by residues that are close in space but are frequently distant in sequence.

YDL117W
(15-64)

10 20 30 40 50

K A R Y G W S G Q T K G D L G F L E G D I M E V T R I A G S W F Y G K L L R N K K C S G Y F P H I L F

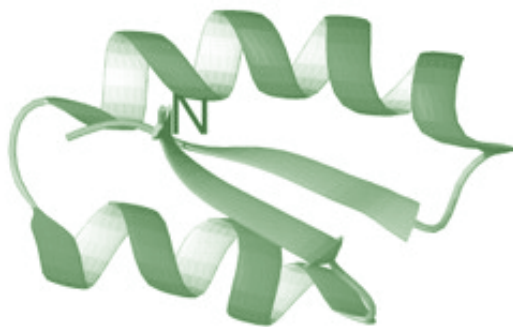


In addition, since evolution tends to conserve function and function depends more directly on structure than on sequence, **structure is more conserved in evolution than sequence.**

The net result is that **patterns in space are frequently more recognizable than patterns in sequence.**

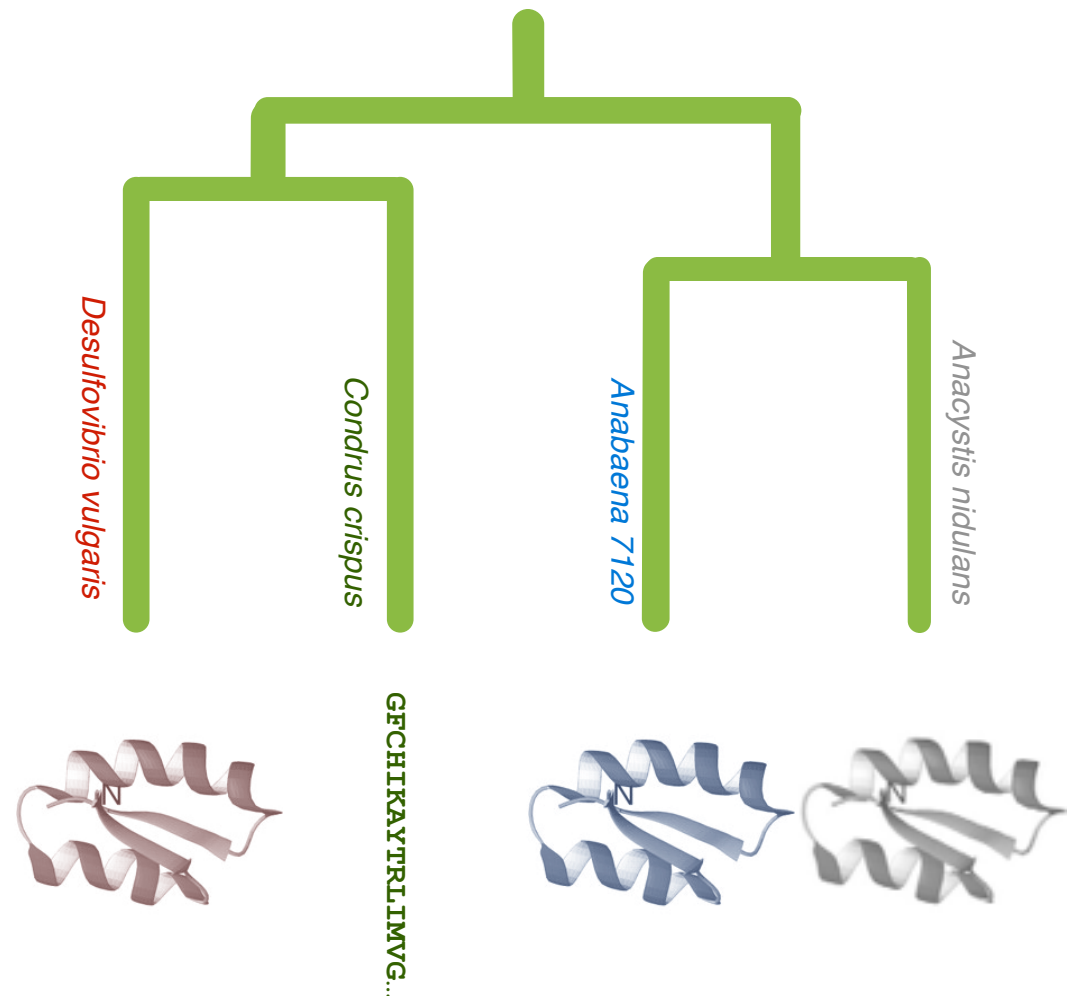
Principles of protein structure

GFCHIKAYTRLIMVG...



Folding (physics)

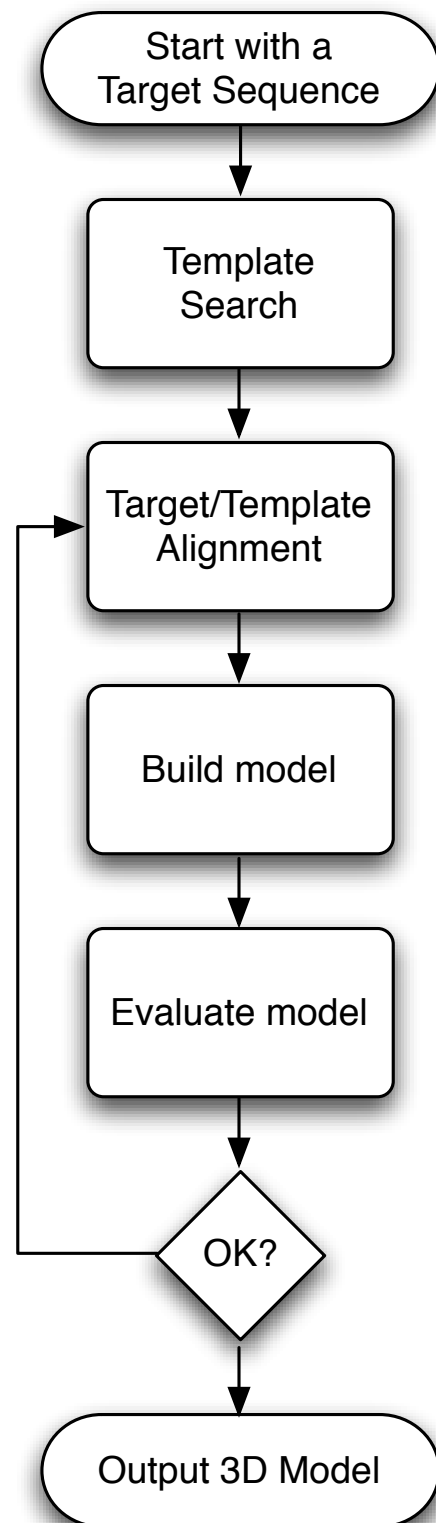
Ab initio prediction



Evolution (rules)

Threading
Comparative Modeling

Comparative modeling by satisfaction of spatial restraints



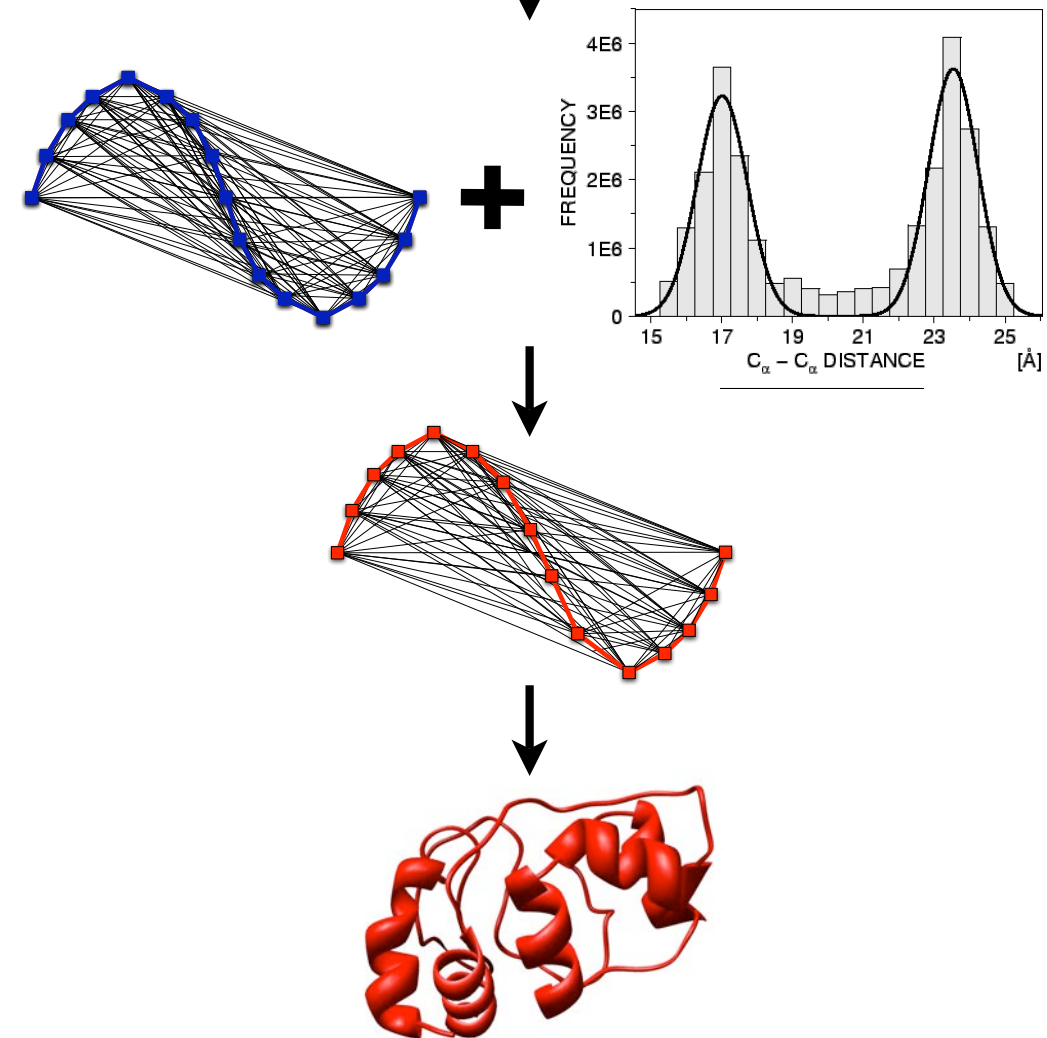
Given an alignment...

extract spatial features from the template(s) and statistics from known structures

apply these features as restraints on your target sequence

optimize to find the best solution for the restraints to produce your 3D model

MSVIPKR--GNCEQTSE
ASILPKRLFGNCEQTS



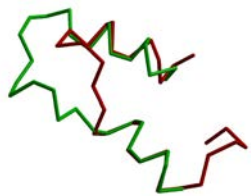
A. Šali & T. Blundell, *J. Mol. Biol.* 234, 779, 1993.
J.P. Overington & A. Šali, *Prot. Sci.* 3, 1582, 1994.
A. Fiser, R. Do & A. Šali, *Prot. Sci.*, 9, 1753, 2000.

Comparative modeling by satisfaction of spatial restraints

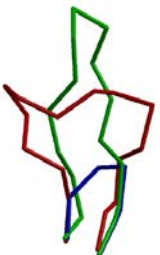
Types of errors and their impact



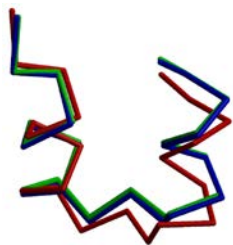
Wrong fold



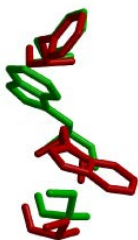
Miss alignments



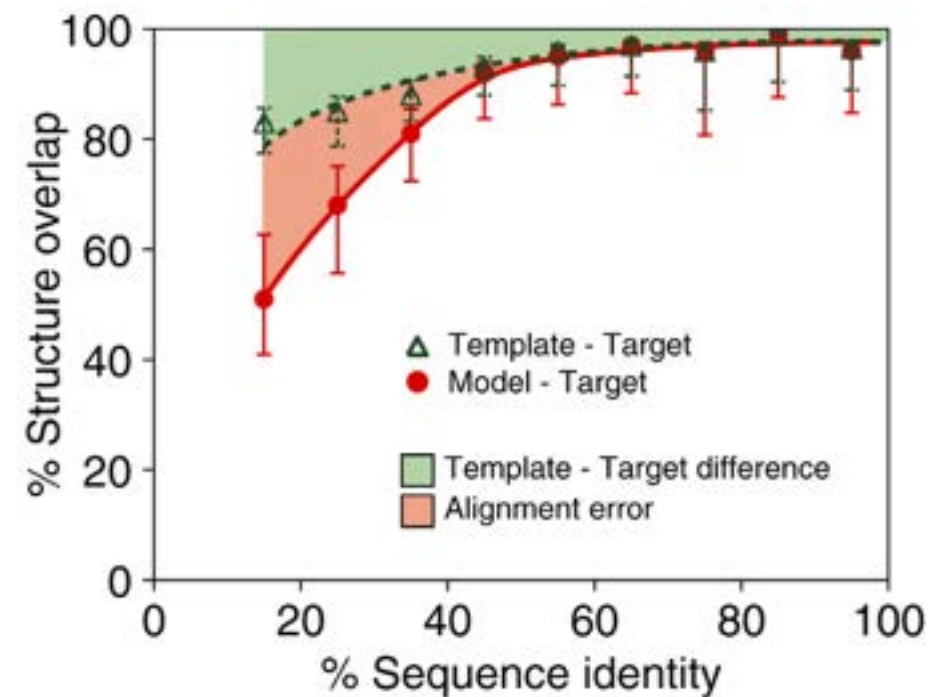
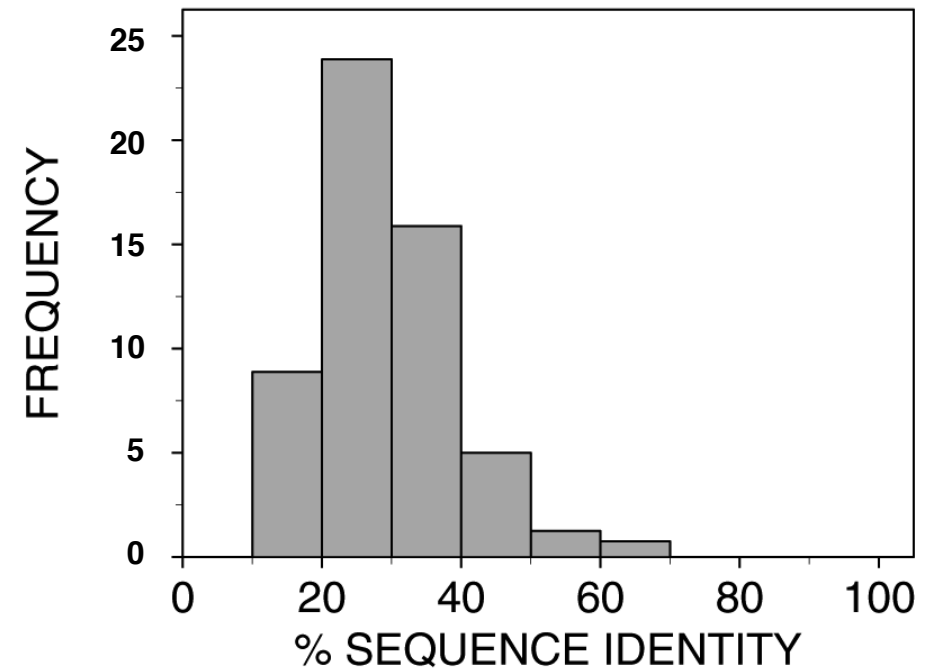
Loop regions



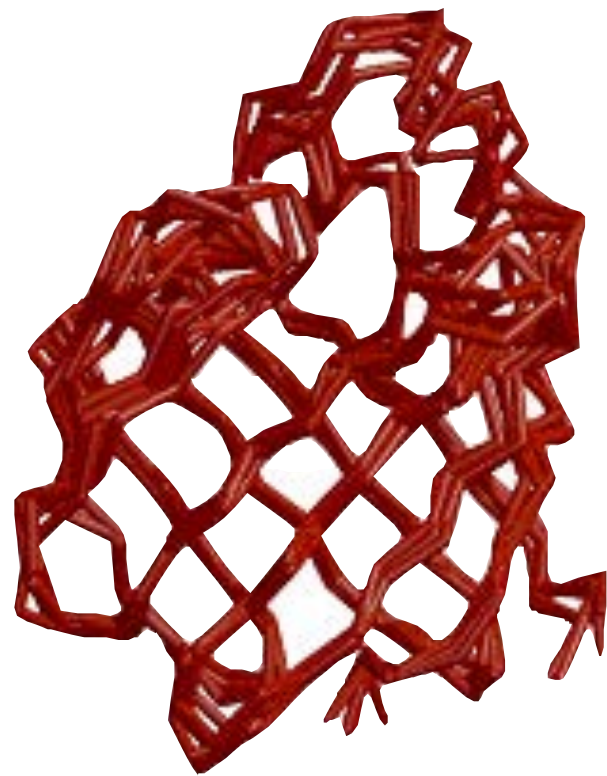
Rigid body distortions



Side-chain packing

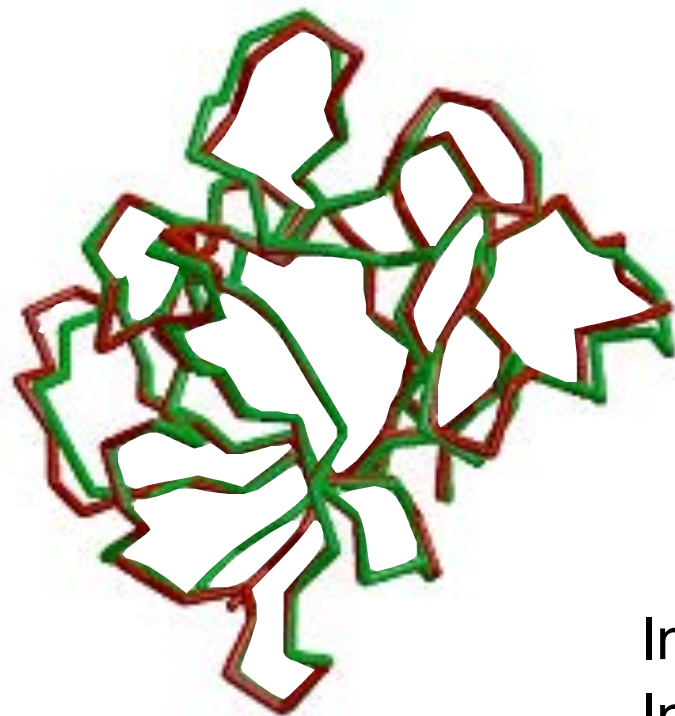
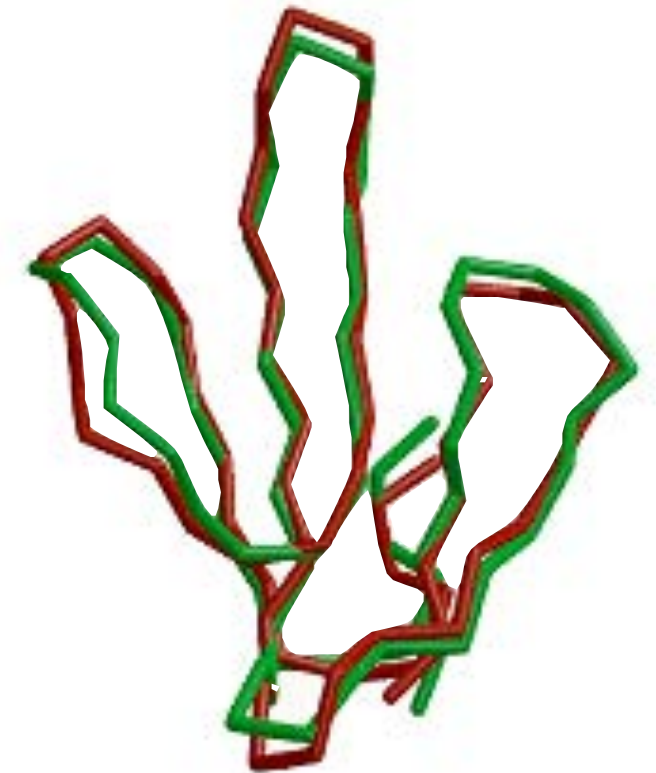


“Biological” significance of modeling errors



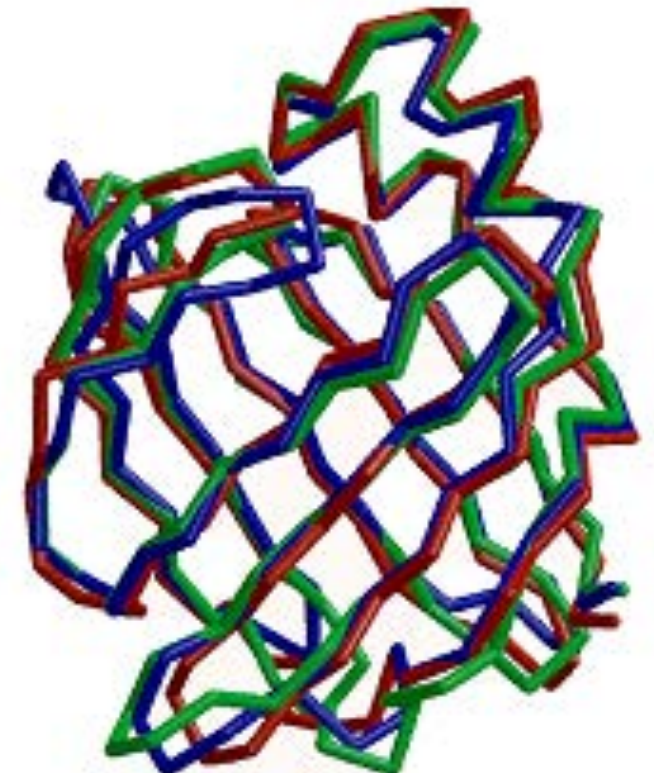
NMR
Ileal lipid-binding protein
1eal

NMR – X-RAY
Erabutoxin 3ebx
Erabutoxin 1era



X-RAY
Interleukin 1 β 41bi (2.9Å)
Interleukin 1 β 2mib (2.8Å)

CRABPII 1opbB
FABP 1ftpA
ALBP 1lib
40% seq. id.

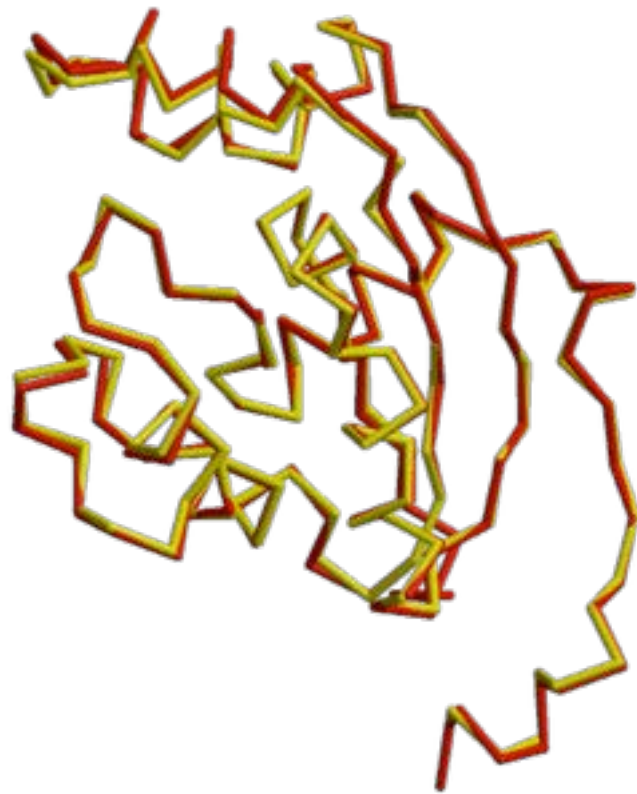


Model Accuracy

HIGH ACCURACY

NM23
Seq id 77%

C α equiv 147/148
RMSD 0.41Å

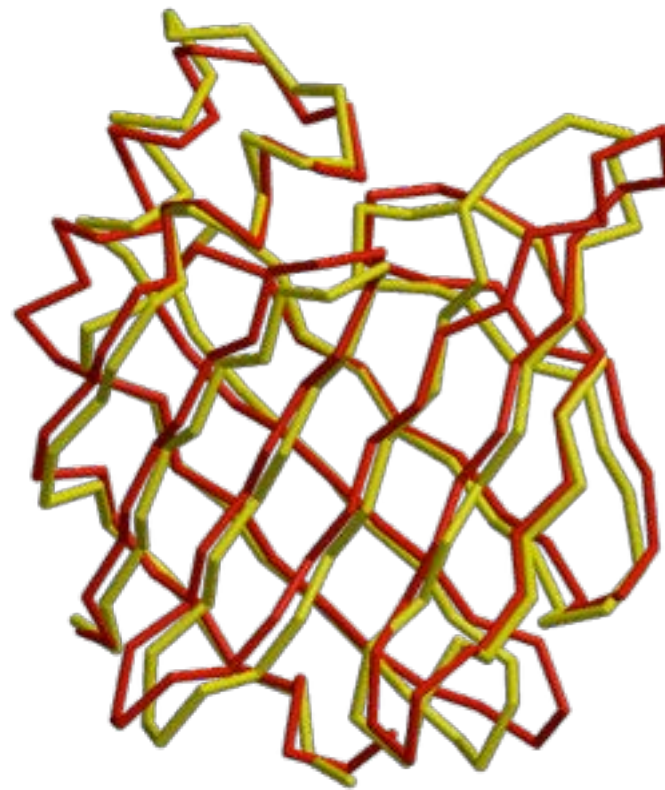


Sidechains
Core backbone
Loops

MEDIUM ACCURACY

CRABP
Seq id 41%

C α equiv 122/137
RMSD 1.34Å



Sidechains
Core backbone
Loops
Alignment

LOW ACCURACY

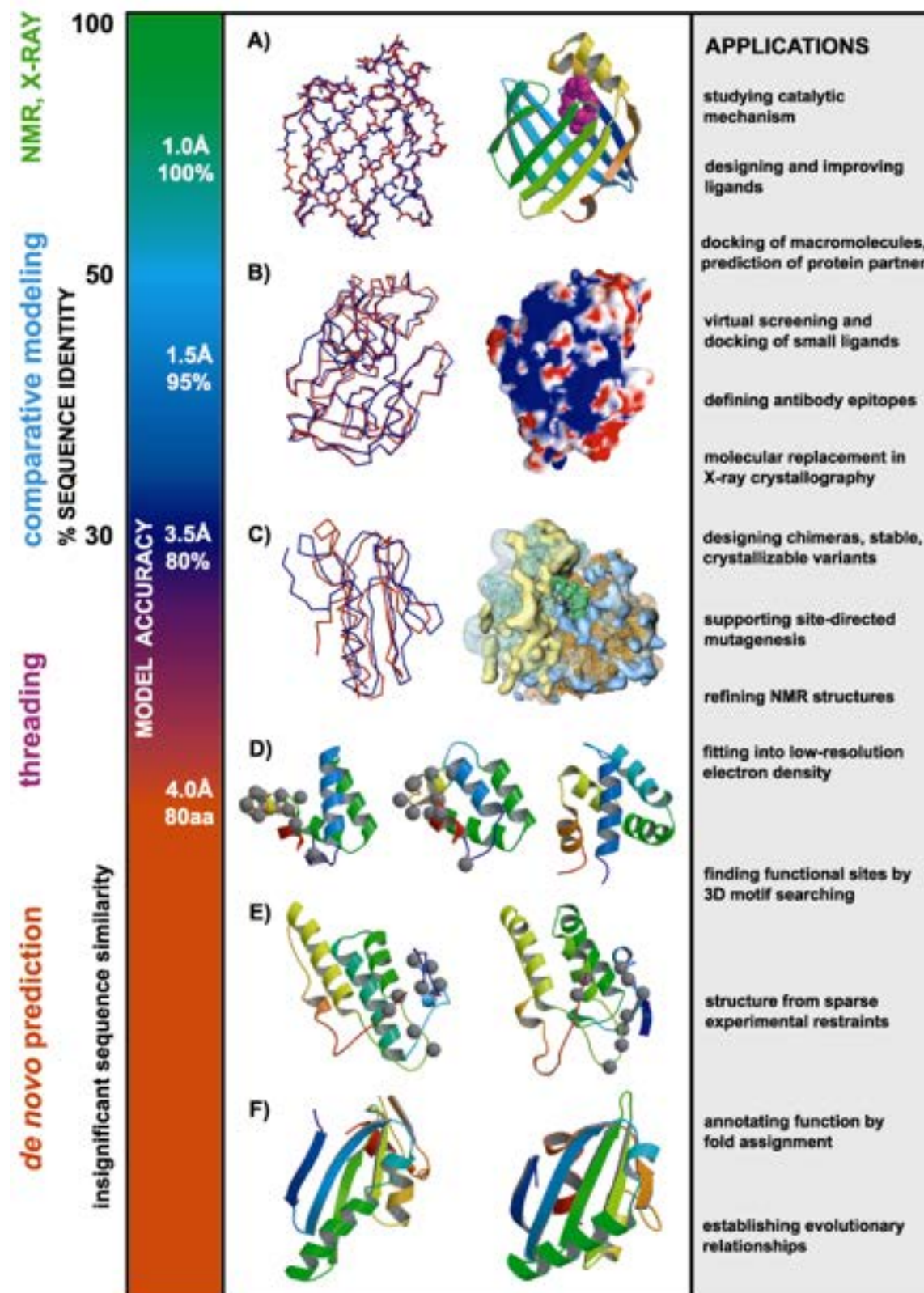
EDN
Seq id 33%

C α equiv 90/134
RMSD 1.17Å



Sidechains
Core backbone
Loops
Alignment
Fold assignment

Utility of protein structure models, despite errors



Structural analysis of missense mutations in human BRCA1 BRCT domains

Mirkovic et al. Structure-based assessment of missense mutations in human BRCA1: implications for breast and ovarian cancer predisposition. Cancer Res (2004) vol. 64 (11) pp. 3790-7

CANCER RESEARCH 64, 3790–3797, June 1, 2004

Structure-Based Assessment of Missense Mutations in Human BRCA1: Implications for Breast and Ovarian Cancer Predisposition

Nebojsa Mirkovic,¹ Marc A. Marti-Renom,² Barbara L. Weber,³ Andrej Sali,² and Alvaro N. A. Monteiro^{4,5}

¹Laboratory of Molecular Biophysics, Pels Family Center for Biochemistry and Structural Biology, Rockefeller University, New York, New York; ²Departments of Biopharmaceutical Sciences and Pharmaceutical Chemistry, and California Institute for Quantitative Biomedical Research, University of California at San Francisco, San Francisco, California; ³Abramson Family Cancer Research Institute, University of Pennsylvania, Philadelphia, Pennsylvania; ⁴Strang Cancer Prevention Center, New York, New York; and ⁵Department of Cell and Developmental Biology, Weill Medical College of Cornell University, New York, New York

ABSTRACT

The *BRCA1* gene from individuals at risk of breast and ovarian cancers can be screened for the presence of mutations. However, the cancer association of most alleles carrying missense mutations is unknown, thus creating significant problems for genetic counseling. To increase our ability to identify cancer-associated mutations in *BRCA1*, we set out to use the principles of protein three-dimensional structure as well as the correlation between the cancer-associated mutations and those that abolish transcriptional activation. Thirty-one of 37 missense mutations of known impact on the transcriptional activation function of BRCA1 are readily rationalized in structural terms. Loss-of-function mutations involve non-conservative changes in the core of the BRCA1 C-terminus (BRCT) fold or are localized in a groove that presumably forms a binding site involved in the transcriptional activation by BRCA1; mutations that do not abolish transcriptional activation are either conservative changes in the core or are on the surface outside of the putative binding site. Next, structure-based rules for predicting functional consequences of a given missense mutation were applied to 57 germ-line BRCA1 variants of unknown cancer association. Such a structure-based approach may be helpful in an integrated effort to identify mutations that predispose individuals to cancer.

INTRODUCTION

Many germ-line mutations in the human *BRCA1* gene are associated with inherited breast and ovarian cancers (1, 2). This information has allowed clinicians and genetic counselors to identify individuals at high risk for developing cancer. However, the disease association of over 350 missense mutations remains unclear, primarily because their relatively low frequency and ethnic specificity limit the usefulness of the population-based statistical approaches to identifying cancer-causing mutations. To address this problem, we use here the three-dimensional structure of the human BRCA1 BRCT domains to assess the transcriptional activation functions of BRCA1 mutants. Our study is made possible by the recently determined sequences (3–6) and three-dimensional structures of the BRCA1 homologs (7, 8). In addition, we benefited from prior studies that attempted to rationalize and predict functional effects of mutations in various proteins (9–12), including those of BRCA1 (13, 14).

BRCA1 is a nuclear protein that activates transcription and facilitates DNA damage repair (15, 16). The tandem BRCT domains at the

COOH-terminus of BRCA1 are involved in several of its functions, including modulation of the activity of several transcription factors (15), binding to the RNA polymerase II holoenzyme (17), and activating transcription of a reporter gene when fused to a heterologous DNA-binding domain (18, 19). Importantly, cancer-associated mutations in the BRCT domains, but not benign polymorphisms, inactivate transcriptional activation and binding to RNA polymerase II (18–21). These observations suggest that abolishing the transcriptional activation function of BRCA1 leads to tumor development and provides a genetic framework for characterization of BRCA1 BRCT variants.

MATERIALS AND METHODS

The multiple sequence alignment (MSA) of orthologous BRCA1 BRCT domains from seven species, including *Homo sapiens* (GenBank accession number U14680), *Pan troglodytes* (AF207822), *Mus musculus* (U68174), *Rattus norvegicus* (AF036760), *Gallus gallus* (AF355273), *Canis familiaris* (U50709), and *Xenopus laevis* (AF416868), was obtained by using program ClustalW (22) and contains only one gapped position (Supplementary Fig. 1). According to PSI-BLAST (23), the latter six sequences are the only sequences in the nonredundant protein sequence database at National Center for Biotechnology Information that have between 30% and 90% sequence identity to the human BRCA1 BRCT domains (residues 1649–1859).

The multiple structure-based alignment of the native structures of the BRCT-like domains was obtained by the SALIGN command in MODELLER (Supplementary Fig. 2). It included the experimentally determined structures of the two human BRCA1 BRCT domains (Protein Data Bank code 1JNX; Refs. 8, 24), rat BRCA1 BRCT domains (1L0B; Ref. 7), human p53-binding protein (1KZY; Ref. 7), human DNA-ligase IIIα (1IMO; Ref. 25), and human XRCC1 protein (1CDZ; Ref. 13). Structure variability was defined by the root-mean-square deviation among the superposed Cα positions, as calculated by the COMPARE command of MODELLER. The purpose of these calculations was to gain insight into the variability of surface-exposed residues (left panel in Fig. 2). In conjunction with observed mutation clustering, these data may point to putative functional site(s) on the surface of BRCT repeats.

Comparative protein structure modeling by satisfaction of spatial restraints, implemented in the program MODELLER-6 (26), was used to produce a three-dimensional model for each of the 94 mutants. The crystallographic structure of the human wild-type BRCA1 BRCT domains was used as the template for modeling (8). The four residues missing in the crystallographic structure (1694 and 1817–1819) were modeled *de novo* (27). All of the models are available in the BRCA1 model set deposited in our ModBase database of comparative protein structure models (28).⁶

For the native structure of the human BRCT tandem repeat and each of the 94 mutant models, a number of sequence and structure features were calculated. These features were used in the classification tree in Fig. 3 (values for all 94 mutations are given in Supplementary Tables 1 and 2).

Buriedness. Accessible surface area of an amino acid residue was calculated by the program DSSP (29) and normalized by the maximum accessible surface area for the corresponding amino acid residue type. A residue was considered exposed if its accessible surface area was larger than 40Å² and if its relative accessible surface area was larger than 9% and buried otherwise. A mutation of a more exposed residue is less likely to change the structure and therefore its function.

Received 9/24/03; revised 1/30/04; accepted 3/15/04.

Grant support: This work was supported by Lee Kaplan Foundation, the Fashion Footwear Association of New York/QVC; United States Army award DAMD17-99-1-9389 and NIH CA92309 (A. N. A. M.); the Mathers Foundation, Sandler Family Supporting Foundation, Sun, IBM and Intel (A. S.); and NIH GM 54762 GM61390 (A. S.); and the Breast Cancer Research Foundation (B. L. W.). M. A. M.-R. is a Rockefeller University Presidential Postdoctoral Fellow; A. S. is an Irma T. Hirschl Trust Cancer Scientist; and B. L. W. is an Abramson Investigator.

The costs of publication of this article were defrayed in part by the payment of page charges. This article must therefore be hereby marked *advertisement* in accordance with 18 U.S.C. Section 1734 solely to indicate this fact.

Note: The authors declare that they have no competing financial interests. Supplemental data for this article are available at Cancer Research Online (<http://cancerres.aacrjournals.org>).

Requests for reprints: Alvaro N. A. Monteiro, H. Lee Moffitt Cancer Center and Research Institute, MRC 3 West, 12902 Magnolia Drive, Tampa, FL 33612. Phone: (813) 745-6321; Fax: (813) 903-6847; E-mail: monteian@moffitt.usf.edu.

3790

(813) 745-6321; Fax: (813) 903-6847; E-mail: monteian@moffitt.usf.edu

Received 9/24/03; revised 1/30/04; accepted 3/15/04.

Grant support: This work was supported by Lee Kaplan Foundation, the Fashion Footwear Association of New York/QVC; United States Army award DAMD17-99-1-9389 and NIH CA92309 (A. N. A. M.); the Mathers Foundation, Sandler Family Supporting Foundation, Sun, IBM and Intel (A. S.); and NIH GM 54762 GM61390 (A. S.); and the Breast Cancer Research Foundation (B. L. W.). M. A. M.-R. is a Rockefeller University Presidential Postdoctoral Fellow; A. S. is an Irma T. Hirschl Trust Cancer Scientist; and B. L. W. is an Abramson Investigator.

The costs of publication of this article were defrayed in part by the payment of page charges. This article must therefore be hereby marked *advertisement* in accordance with 18 U.S.C. Section 1734 solely to indicate this fact.

Note: The authors declare that they have no competing financial interests. Supplemental data for this article are available at Cancer Research Online (<http://cancerres.aacrjournals.org>).

Requests for reprints: Alvaro N. A. Monteiro, H. Lee Moffitt Cancer Center and Research Institute, MRC 3 West, 12902 Magnolia Drive, Tampa, FL 33612. Phone: (813) 745-6321; Fax: (813) 903-6847; E-mail: monteian@moffitt.usf.edu.

3790

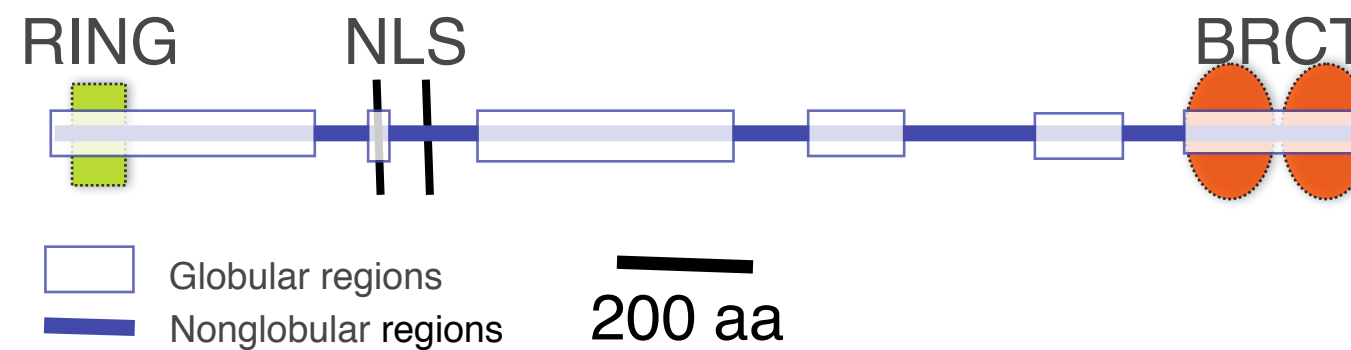
© 2004 American Association for Cancer Research

0732-183X/04/6411-3790/\$20.00

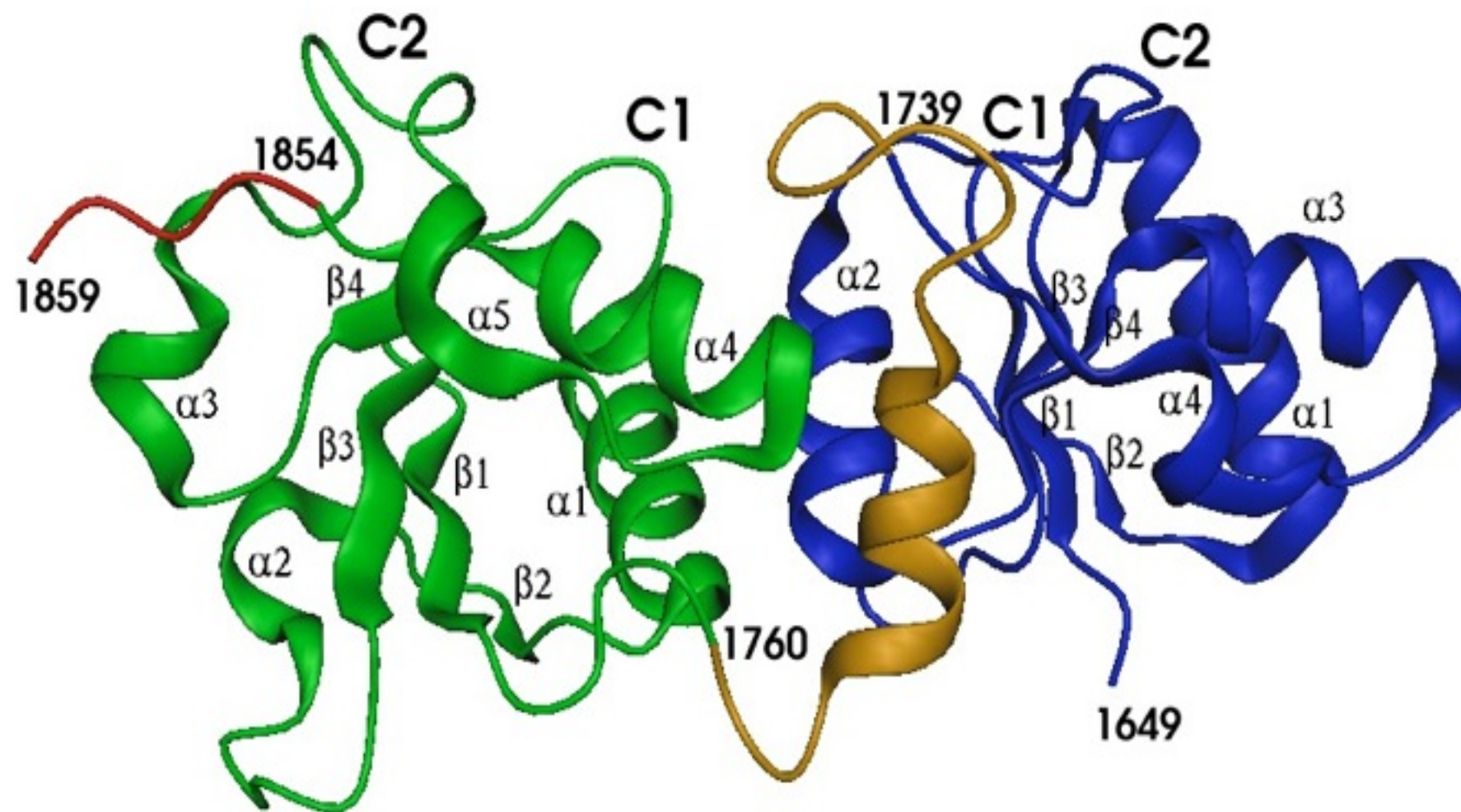
DOI: 10.1158/0732-183X.3790-04

13

Human BRCA1 and its two BRCT domains



BRCA1 BRCT repeats, 1jnx



CONFIDENTIAL



MYRIAD

BRCAAnalysis™

Comprehensive BRCA1-BRCA2 Gene Sequence Analysis Result

Niecee Singer, MS
Strang Cancer Prevention Center
428 E 72nd St
New York, NY 10021

Physician: Fred Gilbert, MD

SPECIMEN
Specimen Type: Blood
Draw Date: n/a
Accession Date: Oct 27, 2000
Report Date: Nov 17, 2000

PATIENT
Name:
Date of Birth: Feb 02, 1953
Patient ID:
Gender: Female
Accession #: 00019998
Requisition #: 56694

Test Result

Gene Analyzed	Specific Genetic Variant
BRCA2	H2116R
BRCA1	None Detected

Interpretation

GENETIC VARIANT OF UNCERTAIN SIGNIFICANCE

The BRCA2 variant H2116R results in the substitution of arginine for histidine at amino acid position 2116 of the BRCA2 protein. Variants of this type **may or may not** affect BRCA2 protein function. Therefore, the contribution of this variant to the relative risk of breast or ovarian cancer cannot be established solely from this analysis. The observation by Myriad Genetic Laboratories of this particular variant in an individual with a deleterious truncating mutation in BRCA2, however, reduces the likelihood that H2116R is itself deleterious.

Authorized Signature:

Brian E. Ward, Ph.D.
Laboratory Director

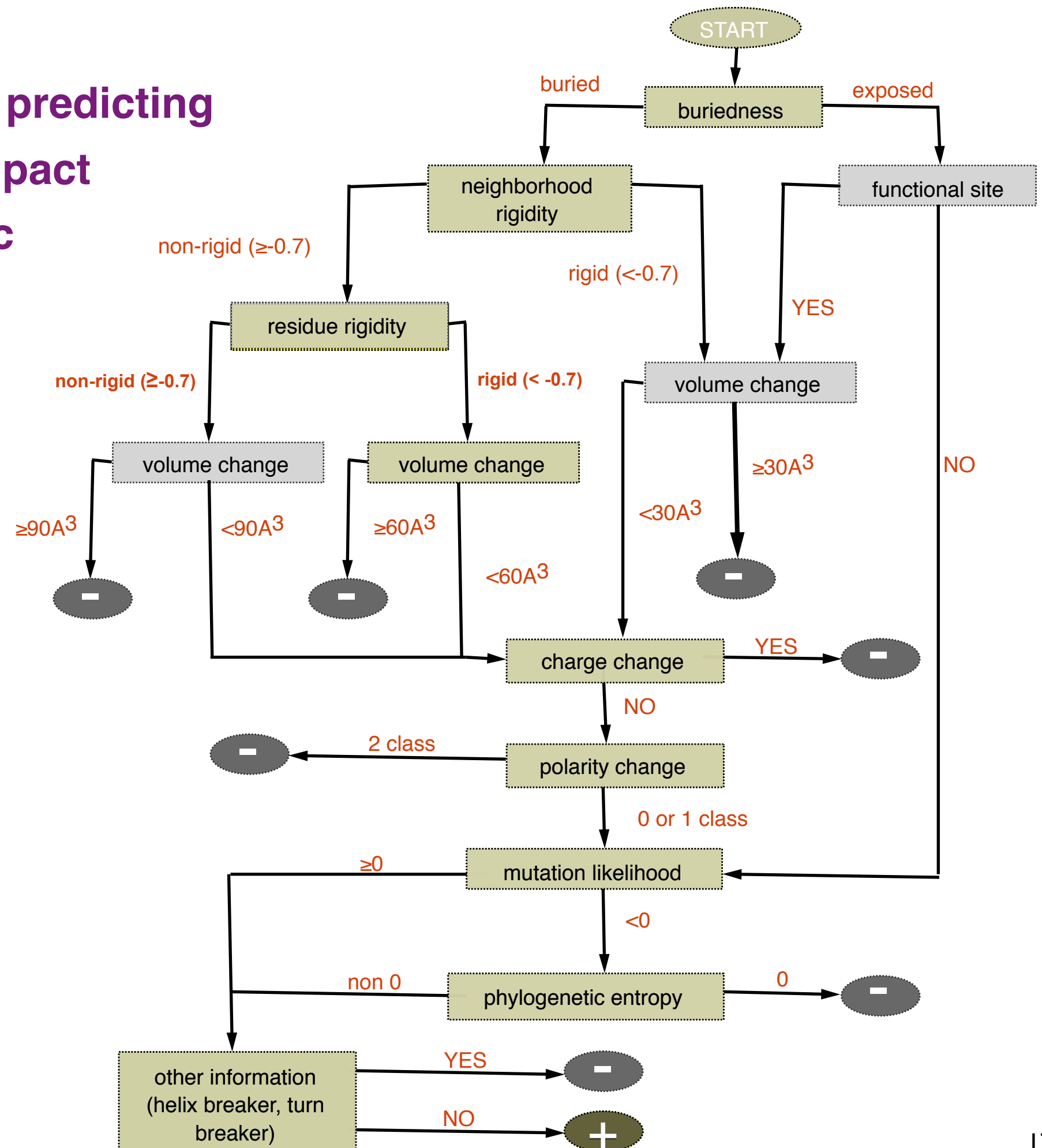

Thomas S. Frank, M.D.
Medical Director

These test results should only be used in conjunction with the patient's clinical history and any previous analysis of appropriate family members. It is strongly recommended that these results be communicated to the patient in a setting that includes appropriate counseling. The accompanying Technical Specifications summary describes the analysis, method, performance characteristics, nomenclature, and interpretive criteria of this test. This test may be considered investigational by some states. This test was developed and its performance characteristics determined by Myriad Genetic Laboratories. It has not been reviewed by the U.S. Food and Drug Administration. The FDA has determined that such clearance or approval is not necessary.

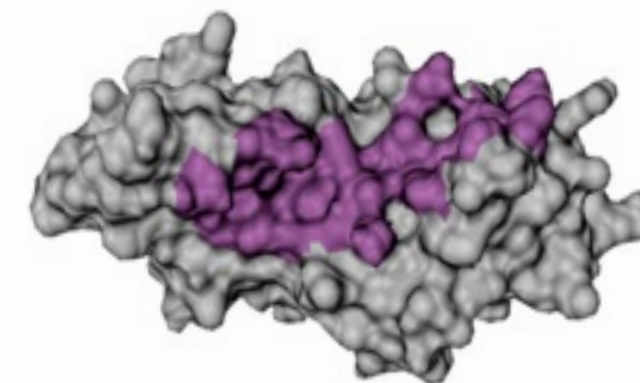
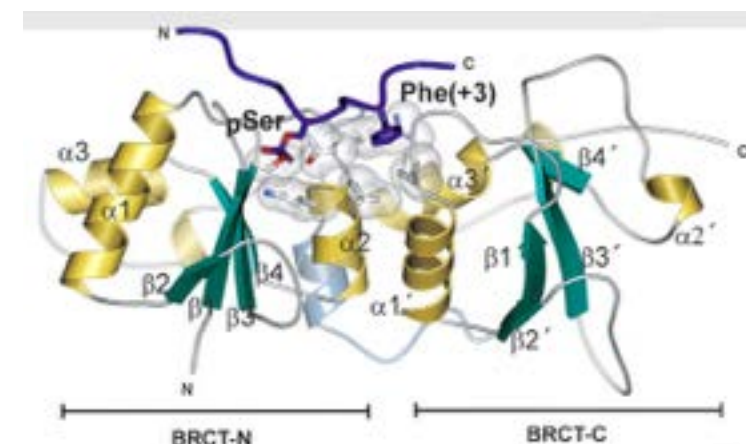
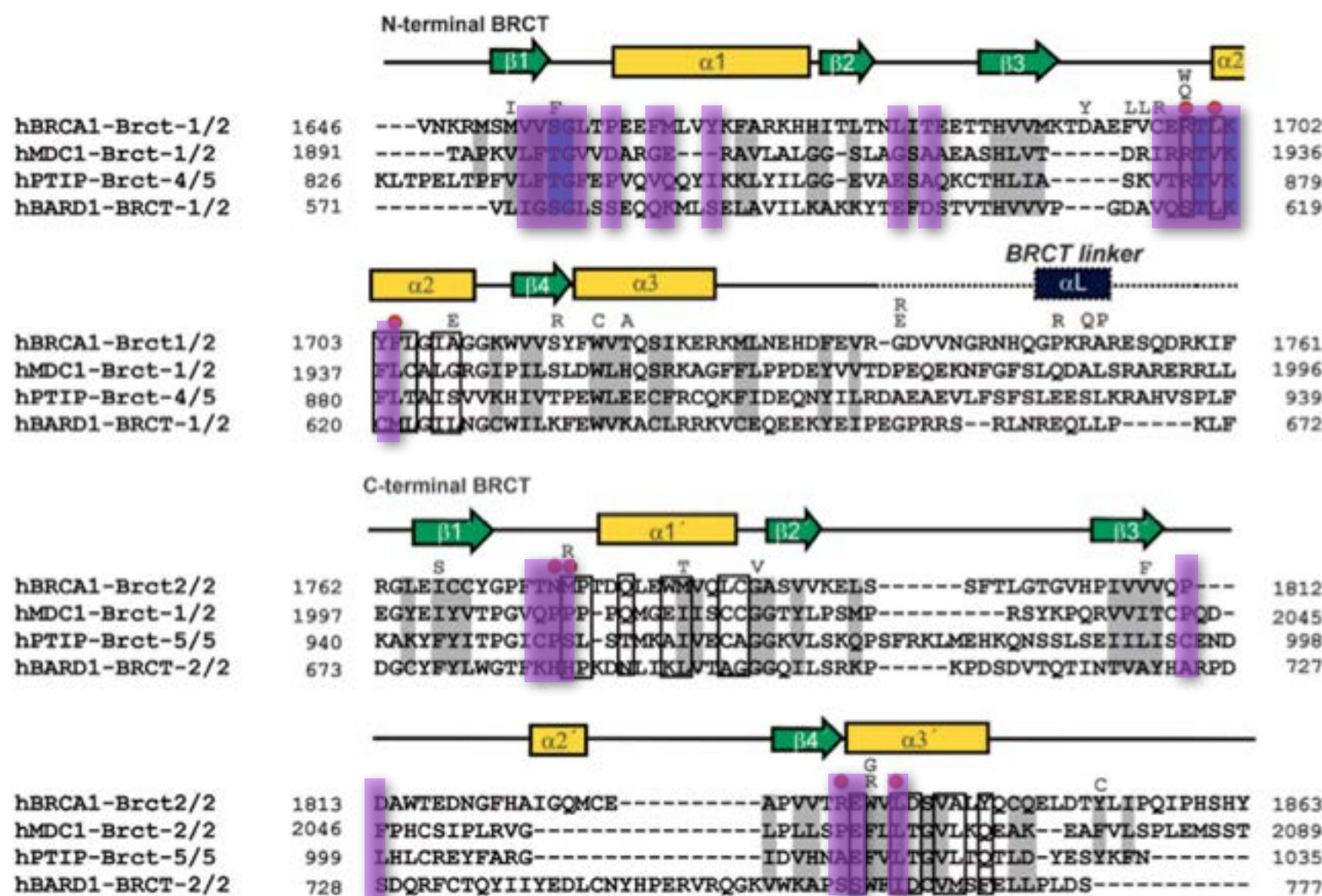
Missense mutations in BRCT domains by function

	cancer associated	not cancer associated	?				
no transcription activation	C1697R R1699W A1708E S1715R P1749R M1775R		M1652K L1657P E1660G H1686Q R1699Q K1702E Y1703HF 1704S	L1705PS 1715NS1 722FF17 34LG173 8EG1743 RA1752P F1761I	F1761S M1775E M1775K L1780P I1807S V1833E A1843T		
transcription activation		M1652I A1669S	V1665M D1692N G1706A D1733G M1775V P1806A				
?			M1652T V1653M L1664P T1685A T1685I M1689R D1692Y F1695L V1696L R1699L G1706E W1718C	W1718S T1720A W1730S F1734S E1735K V1736A G1738R D1739E D1739G D1739Y V1741G H1746N	R1751P R1751Q R1758G L1764P I1766S P1771L T1773S P1776S D1778N D1778G D1778H M1783T	C1787S G1788D G1788V G1803A V1804D V1808A V1809A V1809F V1810G Q1811R P1812S N1819S	A1823T V1833M W1837R W1837G S1841N A1843P T1852S P1856T P1859R

“Decision” tree for predicting functional impact of genetic variants



Putative binding site on BRCA1



Putative binding site predicted in 2003
and accepted for publication on March 2004.

Williams *et al.* 2004 Nature Structure Biology. June 2004 11:519

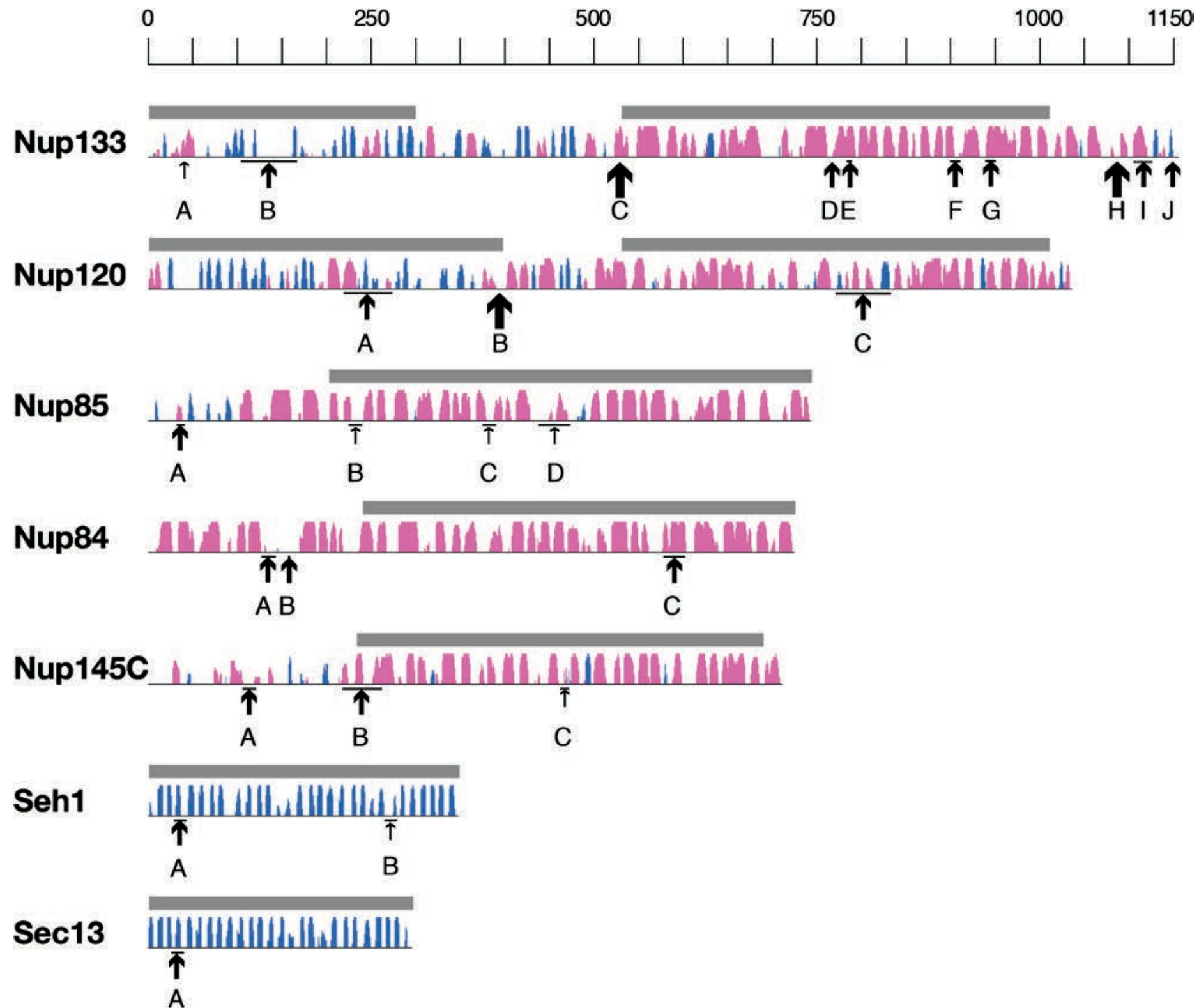
Mirkovic *et al.* 2004 Cancer Research. June 2004 64:3790

Common Evolutionary Origin of Coated Vesicles and Nuclear Pore Complexes

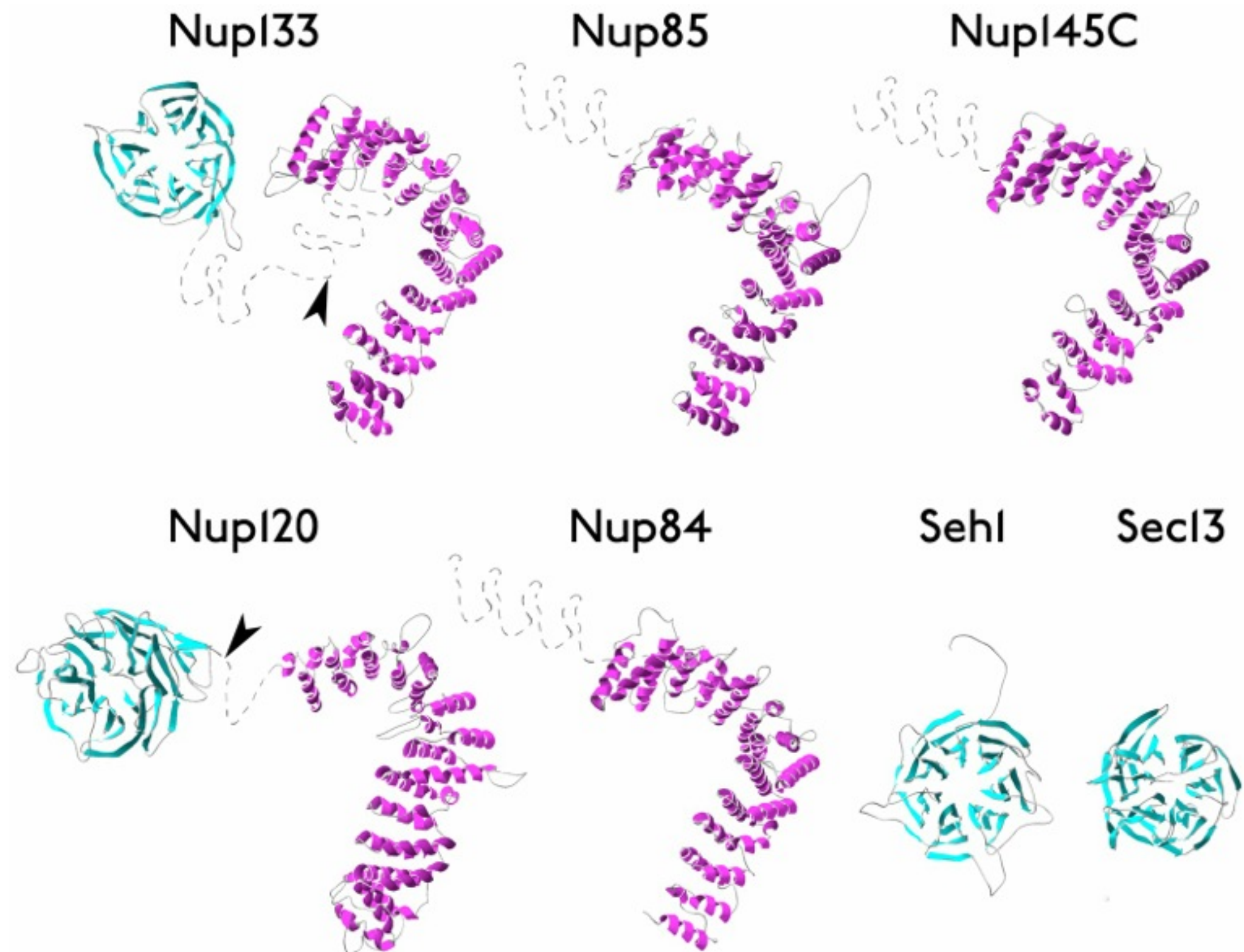
mGenThreader + SALIGN + MOULDER

D. Devos, S. Dokudovskaya, F. Alber, R. Williams, B.T. Chait, A. Sali, M.P. Rout.
Components of Coated Vesicles and Nuclear Pore Complexes Share a Common Molecular Architecture.
PLOS Biology **2(12)**:e380, 2004

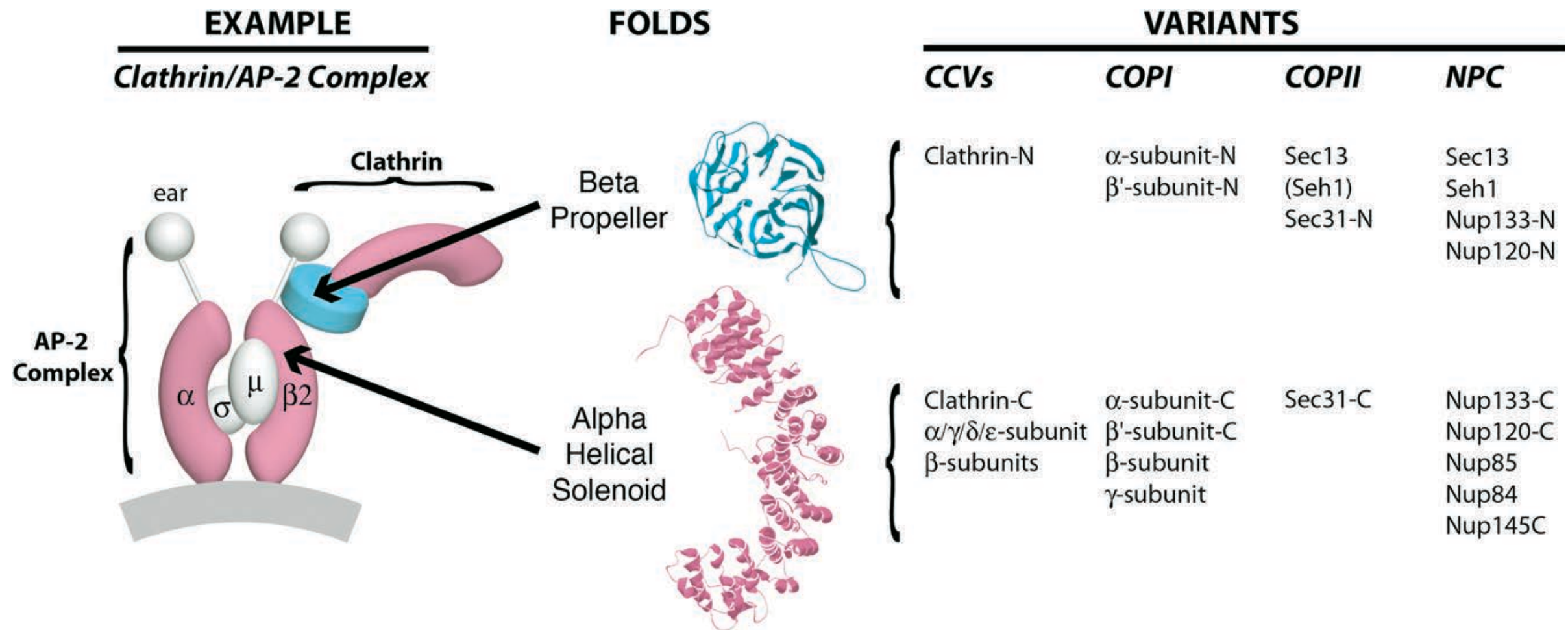
yNup84 complex proteins



All Nucleoporins in the Nup84 Complex are Predicted to Contain β -Propeller and/or α -Solenoid Folds



NPC and Coated Vesicles Share the β -Propeller and α -Solenoid Folds and Associate with Membranes

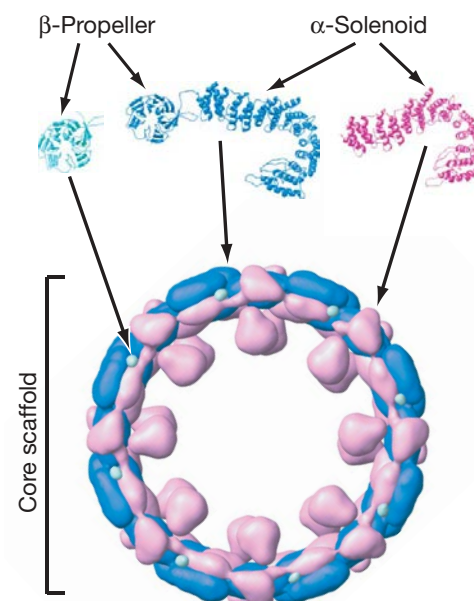
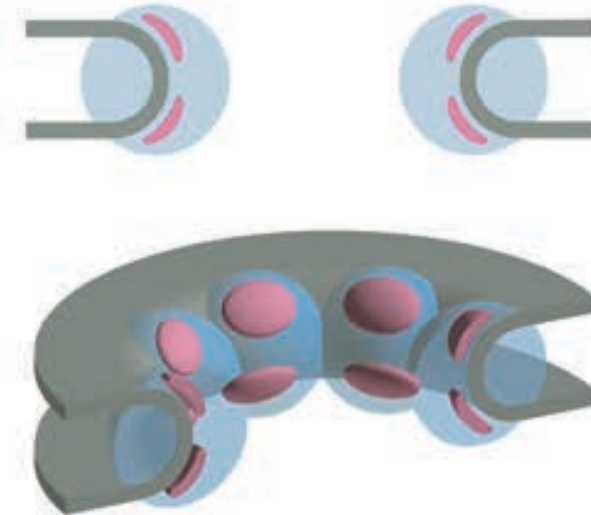


NPC and Coated Vesicles Both Associate with Membranes

Coated Vesicle

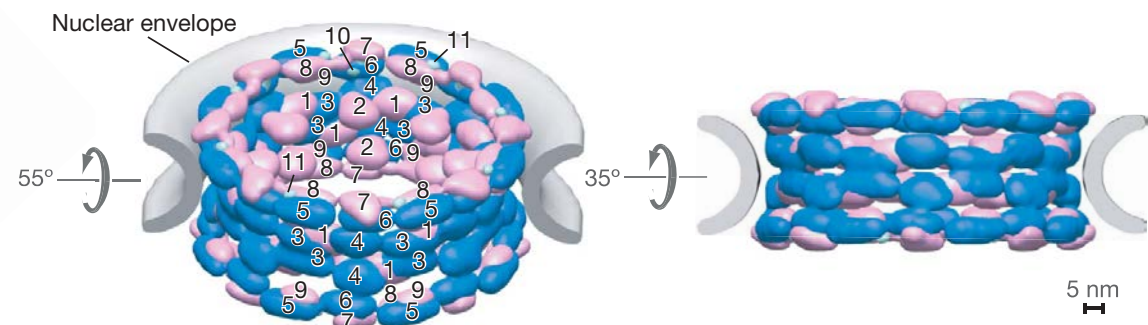


NPC model



Nup 84 complex

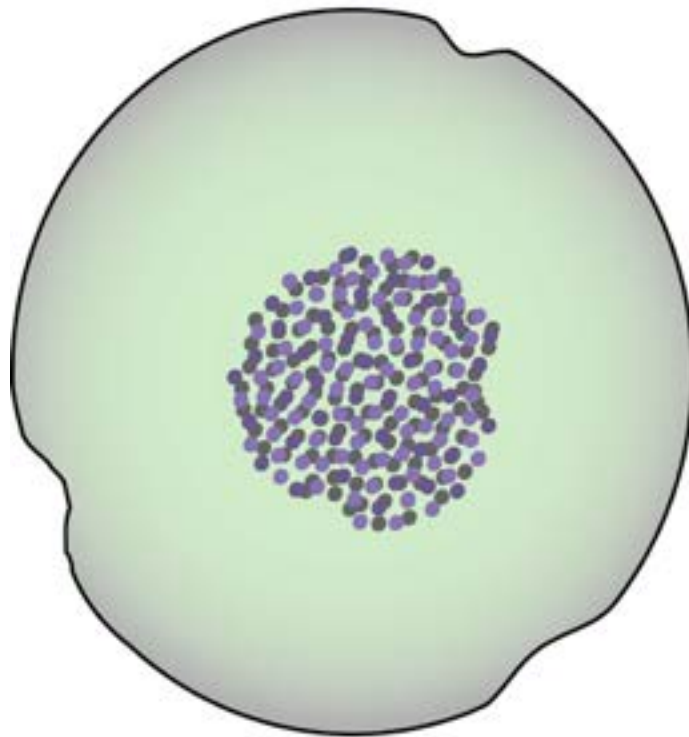
1 Nup192, 2 Nup188, 3 Nup170, 4 Nup157, 5 Nup133,
6 Nup120, 7 Nup85, 8 Nup84, 9 Nup145C, 10 Seh1, 11 Sec13



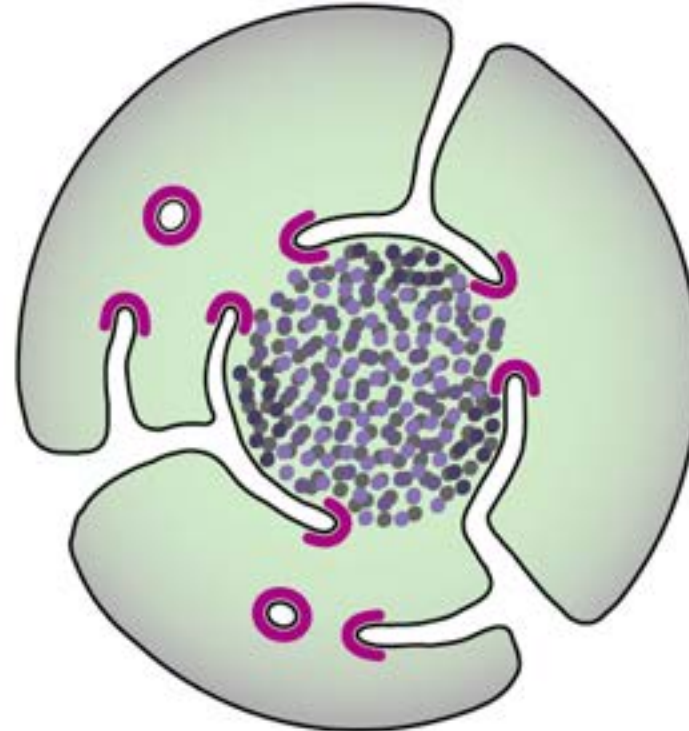
A Common Evolutionary Origin for Nuclear Pore Complexes and Coated Vesicles?

The proto-coatomer hypothesis

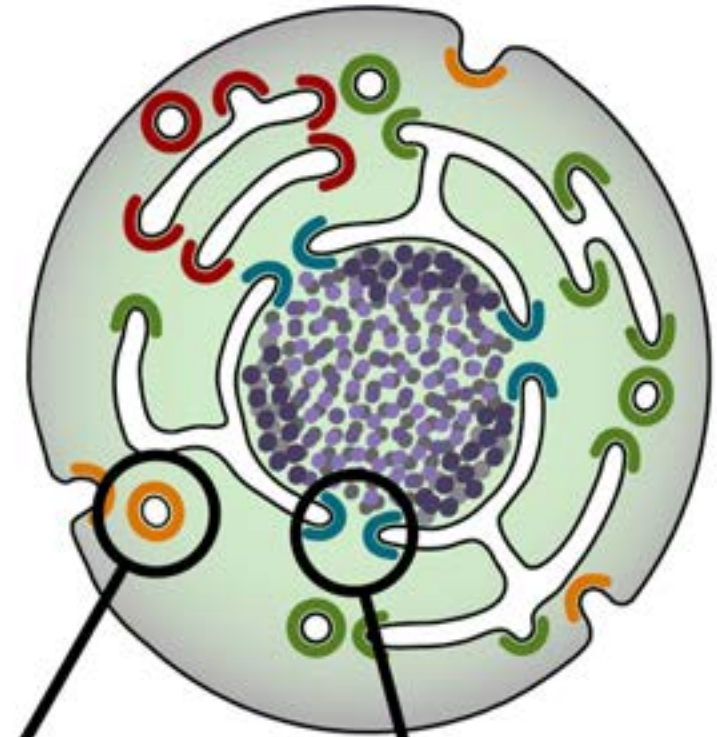
Prokaryote



Early Eukaryote

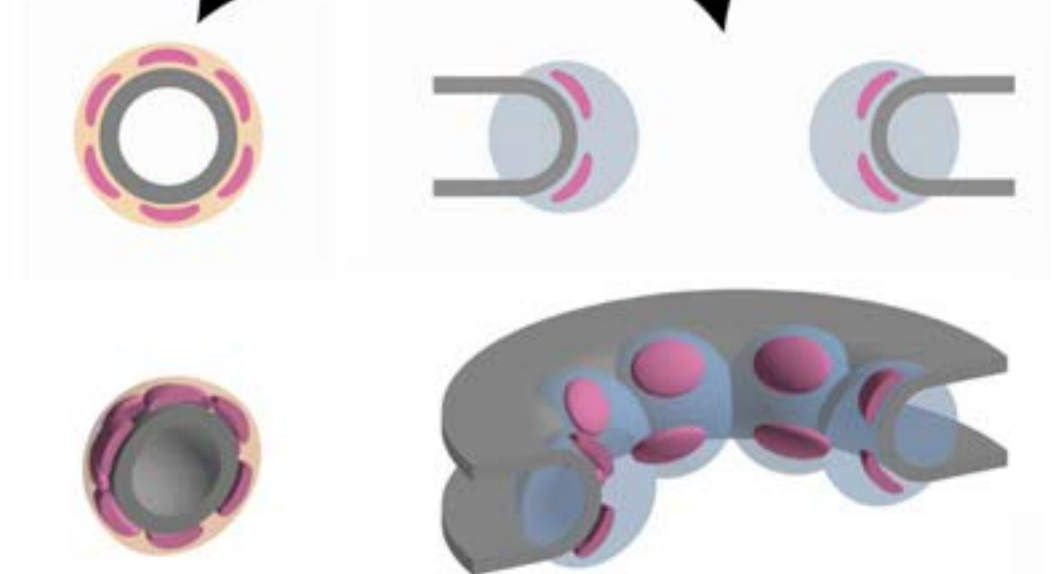


Modern Eukaryote



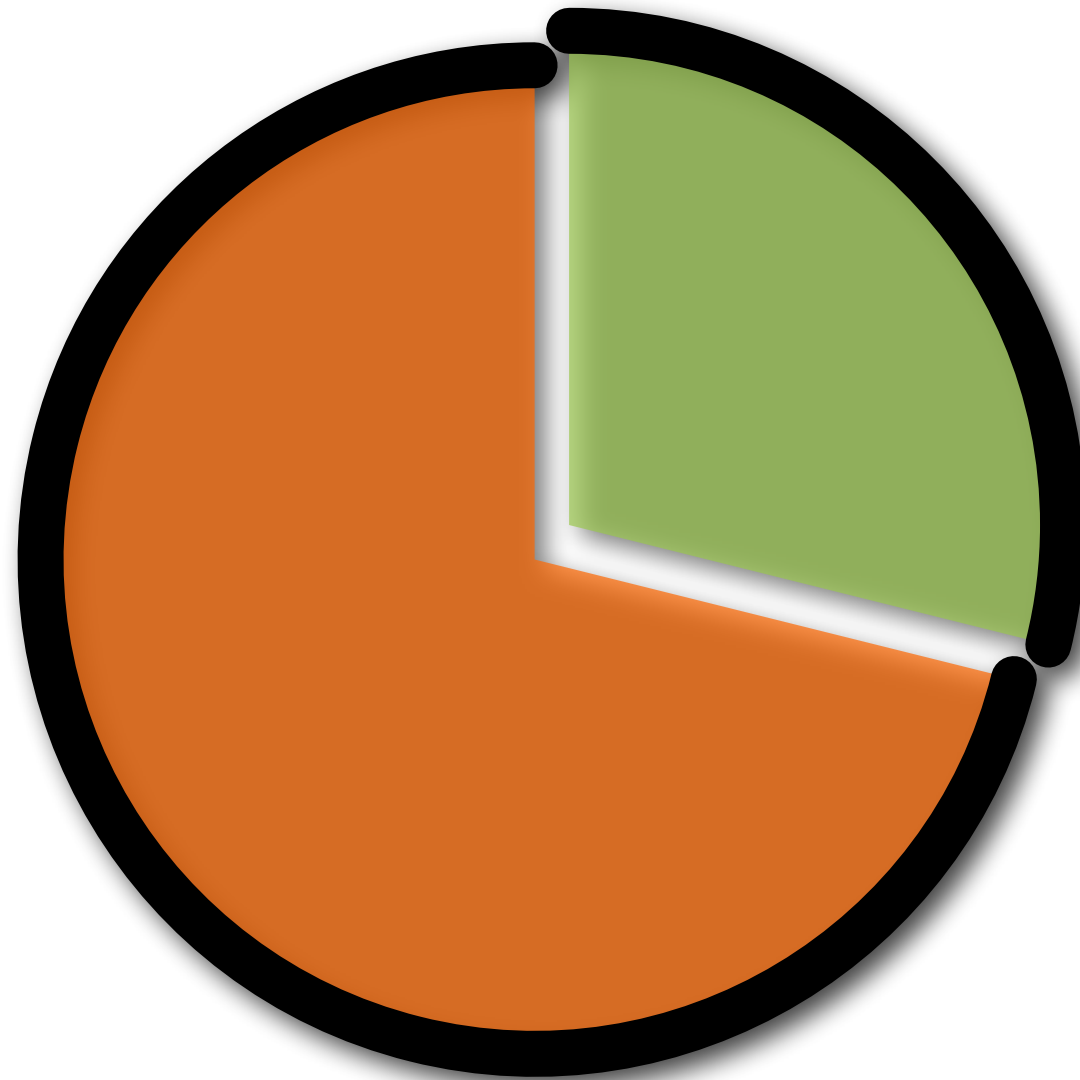
A simple coating module containing minimal copies of the two conserved folds evolved in proto-eukaryotes to bend membranes.

The progenitor of the NPC arose from a membrane-coating module that wrapped extensions of an early ER around the cell's chromatin.



Tropical Disease Initiative (TDI)

Predicting binding sites in protein structure models.

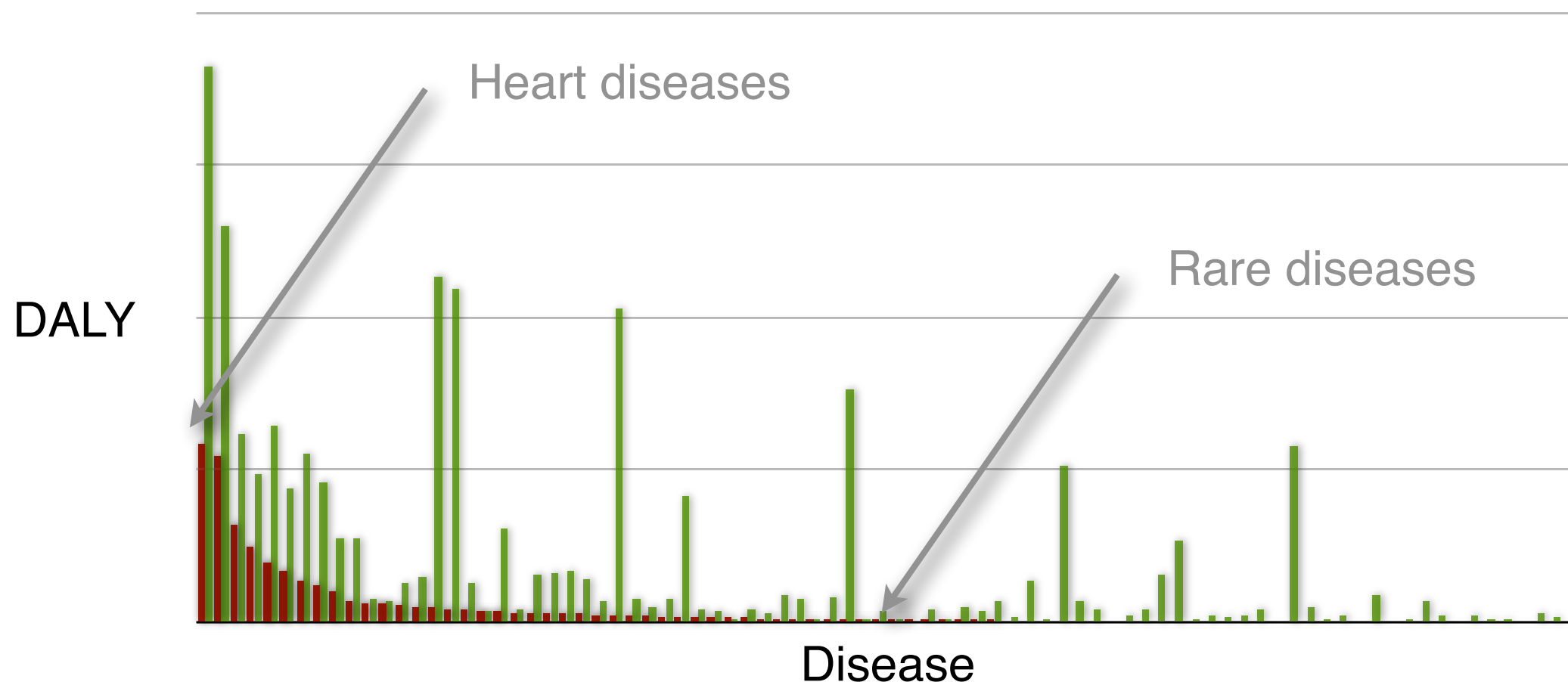


<http://www.tropicaldisease.org>



Need is High in the Tail

- DALY Burden Per Disease in Developed Countries
- DALY Burden Per Disease in Developing Countries



Disease data taken from WHO, *World Health Report 2004*

DALY - Disability adjusted life years

DALY is not a perfect measure of market size, but is certainly a good measure for importance.

DALYs for a disease are the sum of the years of life lost due to premature mortality (YLL) in the population and the years lost due to disability (YLD) for incident cases of the health condition. The DALY is a health gap measure that extends the concept of potential years of life lost due to premature death (PYLL) to include equivalent years of 'healthy' life lost in states of less than full health, broadly termed disability. One DALY represents the loss of one year of equivalent full health.

“Unprofitable” Diseases and Global DALY (in 1000’s)

Malaria*	46,486
Tetanus	7,074
Lymphatic filariasis*	5,777
Syphilis	4,200
Trachoma	2,329
Leishmaniasis*	2,090
Ascariasis	1,817
Schistosomiasis*	1,702
Trypanosomiasis*	1,525

Trichuriasis	1,006
Japanese encephalitis	709
Chagas Disease*	667
Dengue*	616
Onchocerciasis*	484
Leprosy*	199
Diphtheria	185
Poliomyelitis	151
Hookworm disease	59

Disease data taken from WHO, *World Health Report 2004*

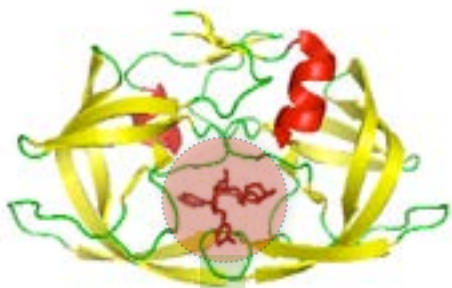
DALY - Disability adjusted life year in 1000’s.

* Officially listed in the WHO Tropical Disease Research [disease portfolio](#).

Comparative docking

Expansion

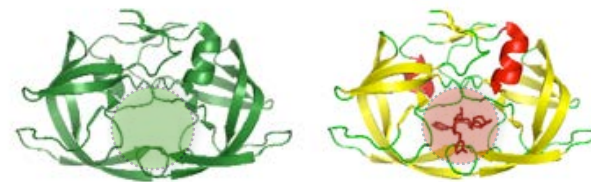
co-crystallized protein/ligand



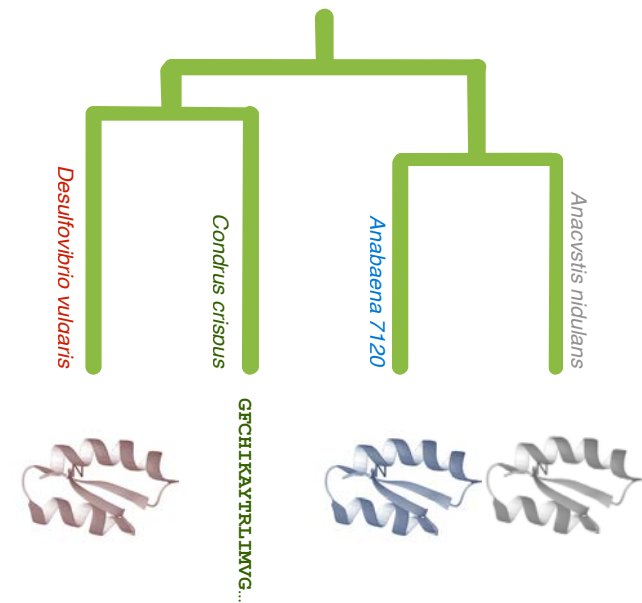
crystallized protein

2. Inheritance

model



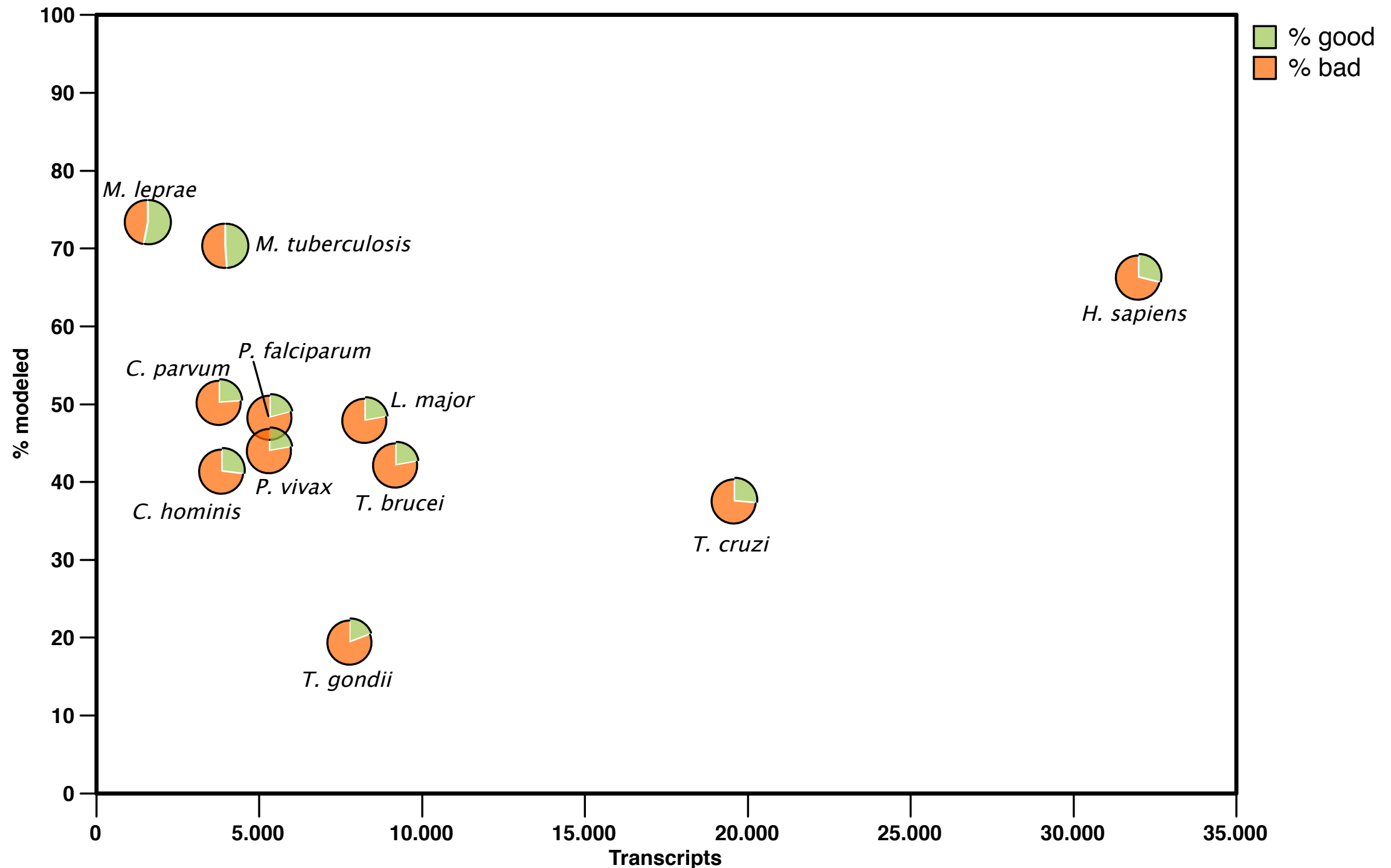
template



1. Modeling

Modeling Genomes

data from models generated by ModPipe (Eswar, Pieper & Sali)



A good model has MPQS of 1.0 or higher

Summary table

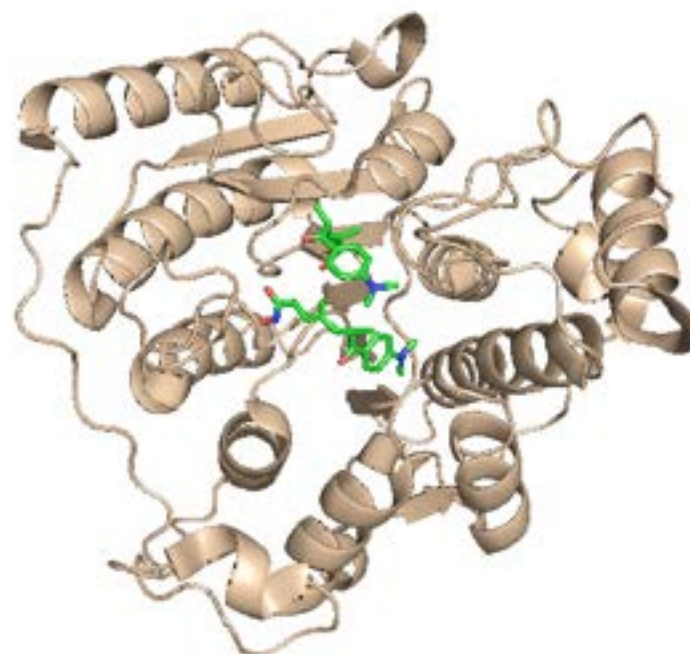
models with inherited ligands

29,271 targets with good models, 297 inherited a ligand/substance similar to a known drug in DrugBank

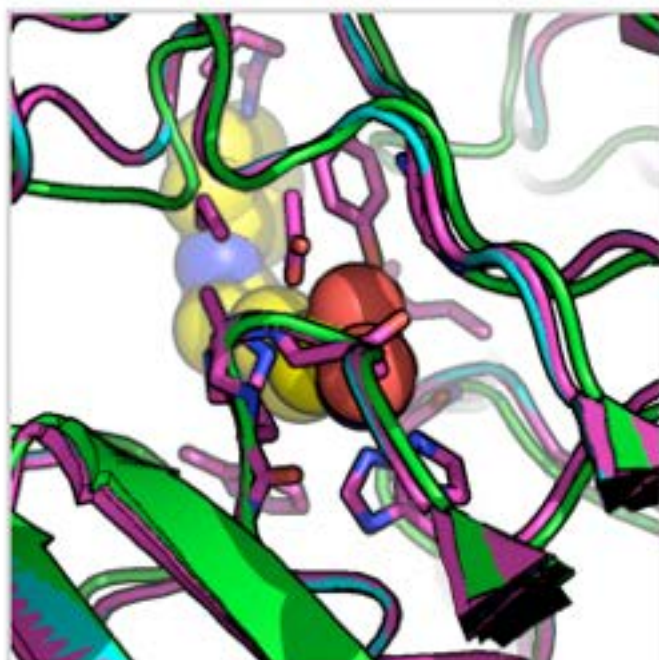
	Transcripts	Modeled targets	Selected models	Inherited ligands	Similar to a drug	Drugs
<i>C. hominis</i>	3,886	1,614	666	197	20	13
<i>C. parvum</i>	3,806	1,918	742	232	24	13
<i>L. major</i>	8,274	3,975	1,409	478	43	20
<i>M. leprae</i>	1,605	1,178	893	310	25	6
<i>M. tuberculosis</i>	3,991	2,808	1,608	365	30	10
<i>P. falciparum</i>	5,363	2,599	818	284	28	13
<i>P. vivax</i>	5,342	2,359	822	268	24	13
<i>T. brucei</i>	7,793	1,530	300	138	13	6
<i>T. cruzi</i>	19,607	7,390	3,070	769	51	28
<i>T. gondii</i>	9,210	3,900	1,386	458	39	21
TOTAL	68,877	29,271	11,714	3,499	297	143

L. major Histone deacetylase 2 + Vorinostat

Template 1t64A a human HDAC8 protein.



PDB	IO	Template	IO	Model	IO	Ligand	Exact	SupStr	SubStr	Similar
1c3sA	83.33/80.00	1t64A	36.00/1.47	LmjF21.0680.1.pdb	90.91/100.00	SHH	DB02546	DB02546	DB02546	DB02546



[DB02546](#) Vorinostat

Small Molecule; Approved; Investigational

Drug categories:

Anti-Inflammatory Agents, Non-Steroidal
Anticarcinogenic Agents
Antineoplastic Agents
Enzyme Inhibitors

Drug indication:

For the treatment of cutaneous manifestations in patients with cutaneous T-cell lymphoma who have progressive, persistent or recurrent disease on or following two systemic therapies.



L. major Histone deacetylase 2 + Vorinostat

Literature

Proc. Natl. Acad. Sci. USA
Vol. 93, pp. 13143–13147, November 1996
Medical Sciences

Apicidin: A novel antiprotozoal agent that inhibits parasite histone deacetylase

(cyclic tetrapeptide/*Apicomplexa*/antiparasitic/malaria/coccidiosis)

SANDRA J. DARKIN-RATTRAY*†, ANNE M. GURNETT*, ROBERT W. MYERS*, PAULA M. DULSKI*,
TAMI M. CRUMLEY*, JOHN J. ALLOCCO*, CHRISTINE CANNOVA*, PETER T. MEINKE‡, STEVEN L. COLLETTI‡,
MARIA A. BEDNAREK‡, SHEO B. SINGH§, MICHAEL A. GOETZ§, ANNE W. DOMBROWSKI§,
JON D. POLISHOOK§, AND DENNIS M. SCHMATZ*

Departments of *Parasite Biochemistry and Cell Biology, †Medicinal Chemistry, and §Natural Products Drug Discovery, Merck Research Laboratories, P.O. Box 2000, Rahway, NJ 07065

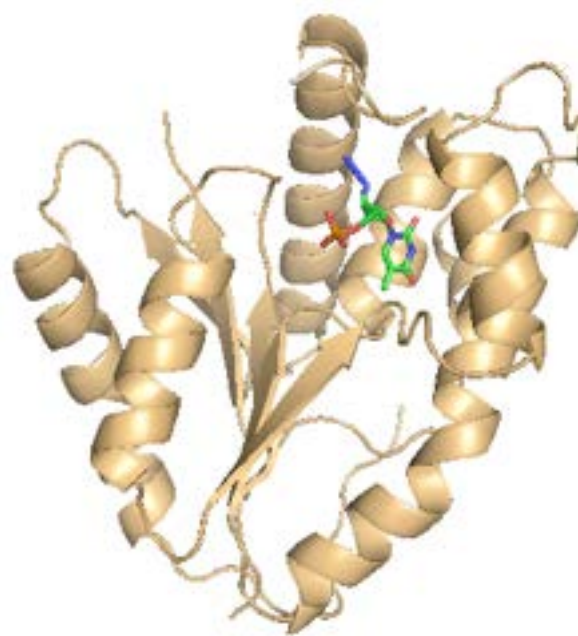
ANTIMICROBIAL AGENTS AND CHEMOTHERAPY, Apr. 2004, p. 1435–1436
0066-4804/04/\$08.00+0 DOI: 10.1128/AAC.48.4.1435–1436.2004
Copyright © 2004, American Society for Microbiology. All Rights Reserved.

Vol. 48, No. 4

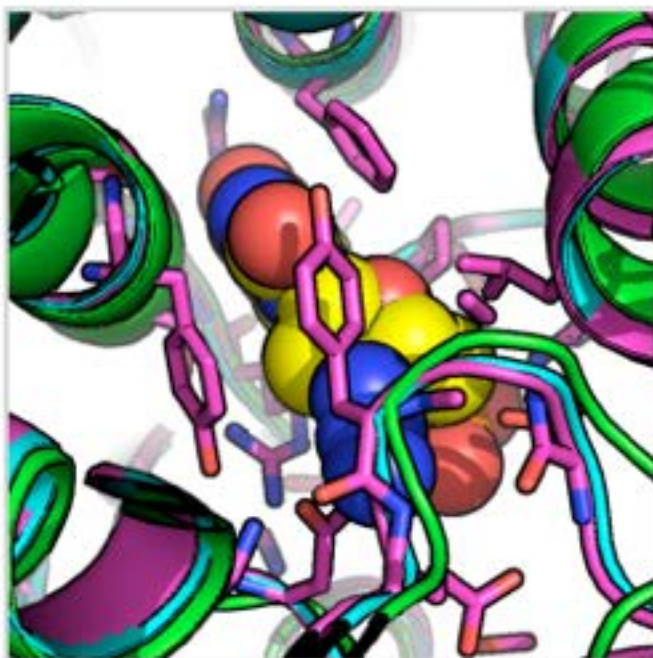
Antimalarial and Antileishmanial Activities of Aroyl-Pyrrolyl-Hydroxyamides, a New Class of Histone Deacetylase Inhibitors

P. falciparum thymidylate kinase + zidovudine

Template 3tmkA a yeast thymidylate kinase.



PDB		Template		Model		Ligand	Exact	SupStr	SubStr	Similar
2tmkB	100.00/100.00	3tmkA	41.00/1.49	PFL2465c.2.pdb	82.61/100.00	ATM		DB00495		DB00495



[DB00495](#) Zidovudine

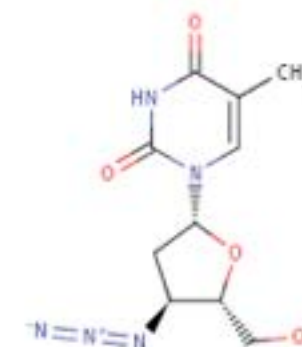
Small Molecule; Approved

Drug categories:

Anti-HIV Agents
Antimetabolites
Nucleoside and Nucleotide Reverse Transcriptase Inhibitors

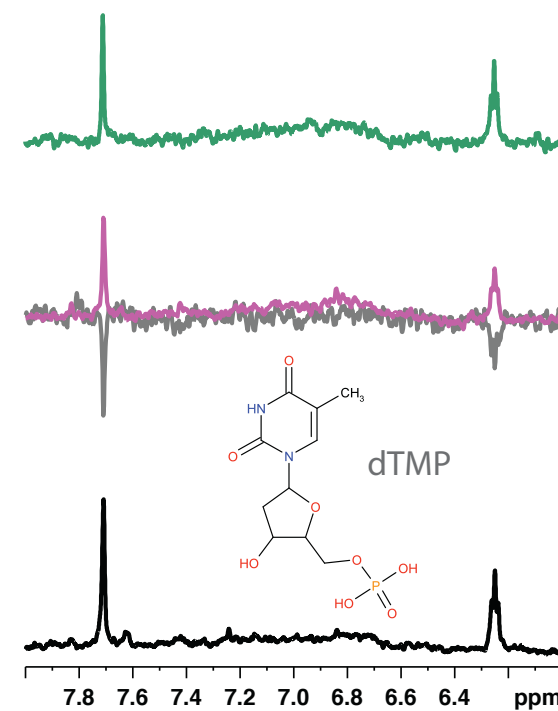
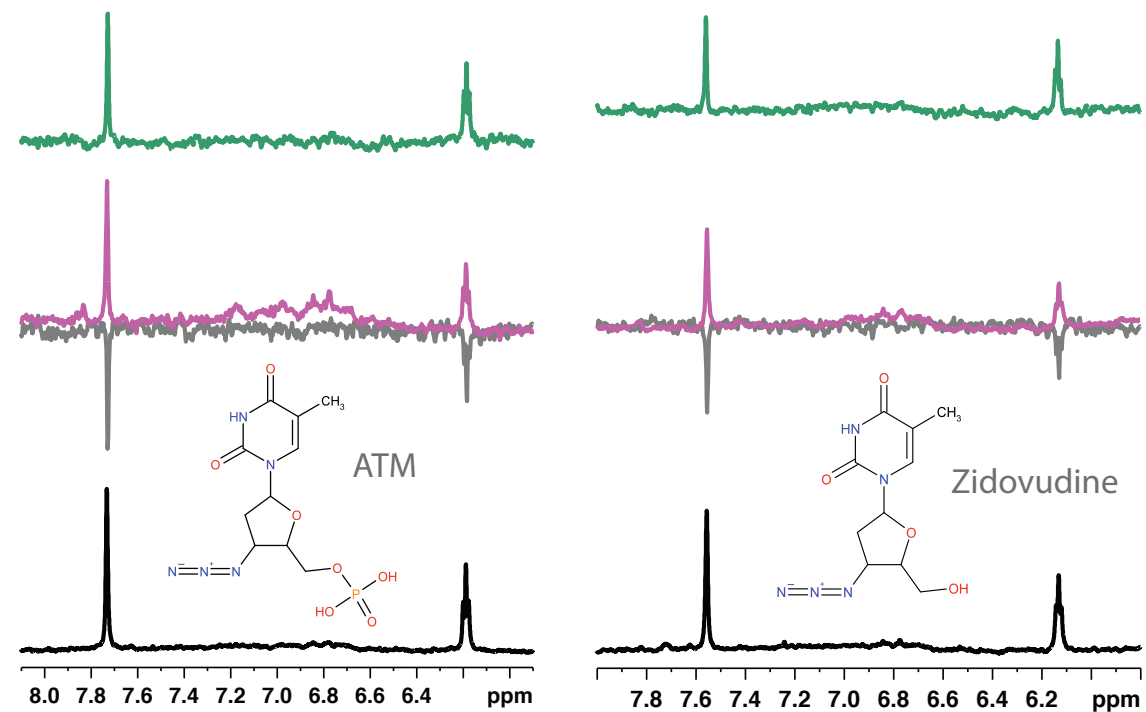
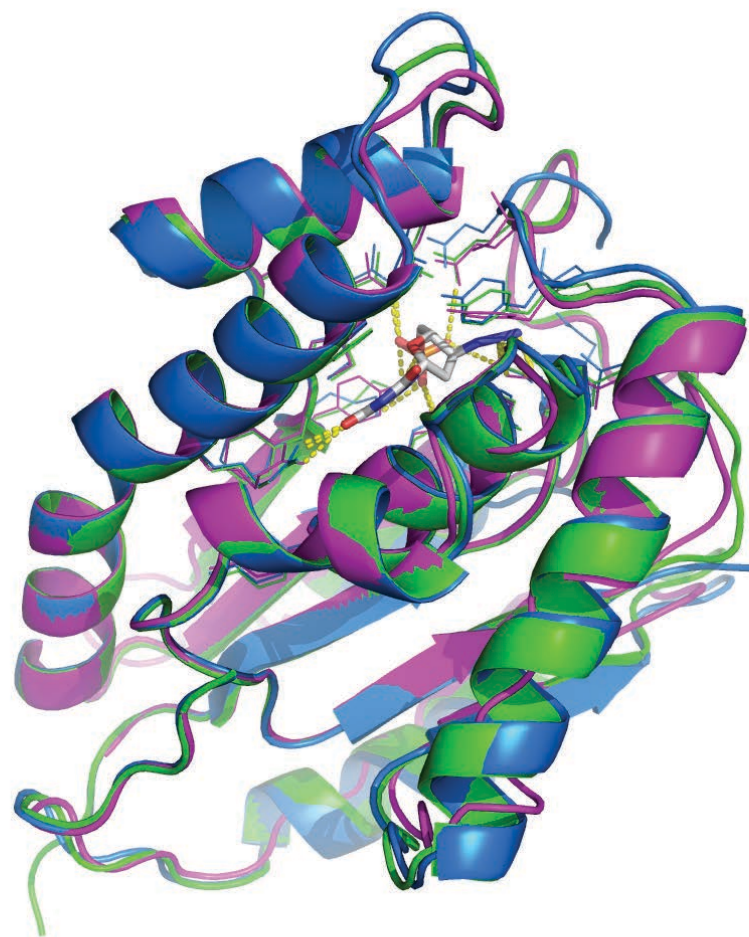
Drug Indication:

For the treatment of human immunovirus (HIV) infections.



P. falciparum thymidilate kinase + zidovudine

NMR Water-LOGSY and STD experiments



TDI's kernel

<http://tropicaldisease.org/kernel>

TDI Kernel database » Q9GU59

http://tropicaldisease.org/kernel/q9gu59/

RSS

Inquisitor

the Tropical Disease Initiative

an open source drug discovery project

You are browsing version 1.0 (2008/05/01) of the TDI Kernel.

Posted on 05.07.08 to Target. Grab the feed. No comments yet. Add your thoughts or trackback from your own site. Edit this entry.

Putative histone deacetylase. predicted to bind 1 ligands [SHH]

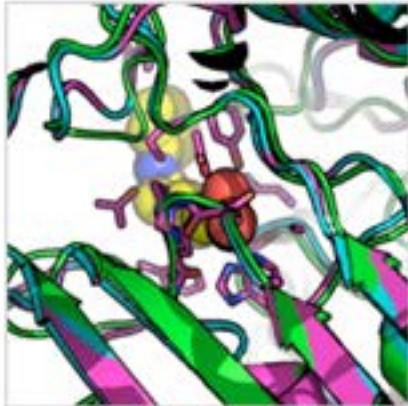
UniPort id: Q9GU59 [C. parvum]

Target keywords: Anticarcinogenic Agents, Antineoplastic Agents, Transcription, Chromatin regulator, Anti-inflammatory Agents, Non-Steroidal, Enzyme Inhibitors, Q9GU59, Transcription regulation, Nucleus

Do you consider this target suitable for drug discovery: ★★★★★ (No Ratings Yet)

Binding site prediction to approved drugs (need help reading this page?):

PDB	id	Template	sim	Model		Ligand	Exact	SupStr	SubStr	Similar
1c3sA	83.33/80.00	1t64A	37.00/1.47	cgd6_1380.1.pdb	90.91/100.00	SHH	DB02546	DB02546	DB02546	DB02546



DB02546 Vorinostat

Small Molecule; Approved; Investigational

Drug categories:

Anti-inflammatory Agents, Non-Steroidal

Anticarcinogenic Agents

Antineoplastic Agents

Enzyme Inhibitors

Drug indication:

For the treatment of cutaneous manifestations in patients with cutaneous T-cell lymphoma who have progressive, persistent or recurrent disease on or following two systemic therapies.

Shown ligand SHH

OCTANEDIOICACIDHYDROXYAMDEPHENYLAMIDE

expanded from 1c3sA to template 1t64A used for building a 3D model of cgd6_1380.1.pdb. Download the coordinates data/Q9GU59/Q9GU59.SHH.952.pdb

Kernel 1.0

SEARCH KERNEL

Q Search...

Advanced Search

Browse the kernel

Download Q9GU59

Login / Register

Batch downloads

Help

Methods

1.73 (100%)

Highest rated target:

A7UD81 (5 out of 5)

2008 : Open Access.

Powered by WordPress.

Theme by Upstart Blogger.

TDI's kernel

<http://tropicaldisease.org/kernel>

L. Orti *et al.*, *Nat Biotechnol* **27**, 320 (Apr, 2009).

L. Orti *et al.*, *PLoS Negl Trop Dis* **3**, e418 (2009).

CORRESPONDENCE

A kernel for the Tropical Disease Initiative

To the Editor:

Identifying proteins that are good drug targets and finding drug leads that bind to them is generally a challenging problem. It is particularly difficult for neglected tropical diseases, such as malaria and tuberculosis, where research resources are relatively scarce¹. Fortunately, several developments improve our ability to deal with drug discovery for neglected diseases: first, the sequencing of many complete genomes of organisms that cause tropical diseases; second, the determination of a large number of protein structures; third, the creation of compound libraries, including already-approved drugs; and fourth, the availability of improved bioinformatics analysis, including methods for comparative protein structure modeling, binding site identification, virtual ligand screening and drug design. Therefore, we are now in a position to increase the odds of identifying high-quality drug targets and drug leads for neglected tropical diseases. Here we encourage a collaboration among scientists to engage in drug discovery for tropical diseases by providing a 'kernel' for the Tropical Disease Initiative (TDI, <http://www.tropicaldisease.org/>). As the Linux kernel did for open source code development, we suggest that the TDI kernel may help overcome a major stumbling block, in this case, for open source drug discovery: the absence of a critical mass of preexisting work that volunteers can build on incrementally. This kernel complements several other initiatives on neglected tropical diseases^{2–5}, including collaborative web portals (e.g., <http://www.thesynapticleap.org/>), public-

private partnerships (e.g., <http://www.mmv.org/>) and private foundations (e.g., <http://www.gatesfoundation.org/>); for an updated list of initiatives, see the TDI website above.

The TDI kernel was derived with our software pipeline^{6,7} for predicting structures of protein sequences by comparative modeling, localizing small-molecule binding sites on the surfaces of the models and predicting ligands that bind to them. Specifically, the pipeline linked 297 proteins from ten pathogen genomes with already approved drugs that were developed for treating other diseases (Table 1). Such links, if proven experimentally, may significantly increase the efficiency of target identification, target validation, lead discovery, lead optimization and clinical trials. Two of the kernel targets were tested for their binding to a known drug by NMR spectroscopy, validating one of our predictions (Fig. 1 and Supplementary Data online).

It is difficult to assess the accuracy of our computational predictions based on this limited experimental testing. Thus, we encourage other investigators to donate their expertise and facilities to test additional predictions. We hope the testing will occur within the

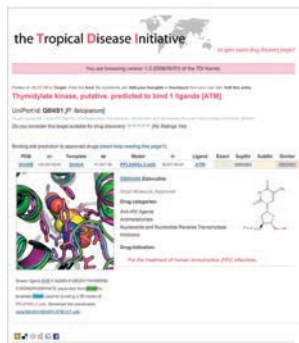


Figure 1 TDI kernel snapshot of the web page for the *Plasmodium falciparum* thymidylate kinase target (<http://tropicaldisease.org/kernel/q8i4s1/>). Our computational pipeline predicted that thymidylate kinase from *P. falciparum* binds ATM (3'-azido-3'-deoxythymidine-5'-monophosphate), a supra-structure of the zidovudine drug approved for the treatment of HIV infection. The binding of this ligand to a site on the kinase was experimentally validated by one-dimensional Water-LOGSY⁹ and saturation transfer difference¹⁰ NMR experiments.

open source context, where results are made available with limited or no restrictions.

A freely downloadable version of the TDI kernel is available in accordance with the Science Commons protocol for implementing open access data (<http://sciencecommons.org/projects/publishing/open-access-data-protocol/>), which prescribes standard academic attribution and facilitates tracking of work but imposes no other restrictions. We do not seek intellectual property rights in the actual discoveries based on the TDI kernel, in the hope of reinvigorating drug discovery for neglected tropical diseases⁸. By minimizing restrictions on the data, including viral terms that would be inherited by all derivative works, we hope to attract as many eyeballs as we possibly can to use and improve the kernel. Although many of the drugs in the kernel are proprietary under diverse types of rights, we believe that the existence of public domain pairs of targets and compounds will reduce the royalties that patent owners can charge and sponsors must pay. This should decrease the large sums of money governments and

Table 1 TDI kernel genomes

Organism ^a	Transcripts ^b	Modeled targets ^c	Similar ^d	Exact ^e
<i>Cryptosporidium hominis</i>	3,886	666	20	13
<i>Cryptosporidium parvum</i>	3,806	742	24	13
<i>Leishmania major</i>	8,274	1,409	43	20
<i>Mycobacterium leprae</i>	1,605	893	25	6
<i>Mycobacterium tuberculosis</i>	3,991	1,608	30	10
<i>Plasmodium falciparum</i>	5,363	818	28	13
<i>Plasmodium vivax</i>	5,342	822	24	13
<i>Toxoplasma gondii</i>	7,793	300	13	6
<i>Trypanosoma cruzi</i>	19,607	3,070	51	28
<i>Trypanosoma brucei</i>	9,210	1,386	39	21
Total	68,877	11,714	297	143

^aOrganisms in bold are included in the World Health Organization (Geneva) Tropical Disease portfolio. ^bNumber of transcripts in each genome. ^cNumber of targets with at least one domain accurately modeled (that is, MODPIPE quality score of at least 1.0). ^dNumber of modeled targets with at least one predicted binding site for a molecule with a Tanimoto score¹¹ of at least 0.9 to a drug in DrugBank¹². ^eNumber of modeled targets with at least one predicted binding site for a molecule in DrugBank.

320

VOLUME 27 NUMBER 4 APRIL 2009 NATURE BIOTECHNOLOGY

330

LOGOS 33, 1000000 4, 1000000 1000000 1000000 1000000

Organism	Transcripts	Modeled targets	Similar	Exact
<i>Cryptosporidium hominis</i>	3,886	666	20	13
<i>Cryptosporidium parvum</i>	3,806	742	24	13
<i>Leishmania major</i>	8,274	1,409	43	20
<i>Mycobacterium leprae</i>	1,605	893	25	6
<i>Mycobacterium tuberculosis</i>	3,991	1,608	30	10
<i>Plasmodium falciparum</i>	5,363	818	28	13
<i>Plasmodium vivax</i>	5,342	822	24	13
<i>Toxoplasma gondii</i>	7,793	300	13	6
<i>Trypanosoma cruzi</i>	19,607	3,070	51	28
<i>Trypanosoma brucei</i>	9,210	1,386	39	21
Total	68,877	11,714	297	143

the kernel. The kernel is available in accordance with the Science Commons protocol for implementing open access data (<http://sciencecommons.org/projects/publishing/open-access-data-protocol/>), which prescribes standard academic attribution and facilitates tracking of work but imposes no other restrictions. We do not seek intellectual property rights in the actual discoveries based on the TDI kernel, in the hope of reinvigorating drug discovery for neglected tropical diseases⁸. By minimizing restrictions on the data, including viral terms that would be inherited by all derivative works, we hope to attract as many eyeballs as we possibly can to use and improve the kernel. Although many of the drugs in the kernel are proprietary under diverse types of rights, we believe that the existence of public domain pairs of targets and compounds will reduce the royalties that patent owners can charge and sponsors must pay. This should decrease the large sums of money governments and

OPEN ACCESS Freely available online



A Kernel for Open Source Drug Discovery in Tropical Diseases

Leticia Orti^{1,2}, Rodrigo J. Carbajo², Ursula Pieper³, Narayanan Eswar^{3a}, Stephen M. Maurer⁴, Arti K. Rai⁵, Ginger Taylor⁶, Matthew H. Todd⁷, Antonio Pineda-Lucena², Andrej Sali^{3a}, Marc A. Marti-Renom^{1*}

1 Structural Genomics Unit, Bioinformatics and Genomics Department, Centro de Investigación Principe Felipe, Valencia, Spain, **2** Structural Biology Laboratory, Medicinal Chemistry Department, Centro de Investigación Principe Felipe, Valencia, Spain, **3** Department of Bioengineering and Therapeutic Sciences, Department of Pharmaceutical Chemistry, and California Institute for Quantitative Biosciences, University of California San Francisco, San Francisco, California, United States of America, **4** Gould School of Law, University of Southern California, Los Angeles, California, United States of America, **5** School of Law, Duke University, Durham, North Carolina, United States of America, **6** The Synaptic Leap, San Ramon, California, United States of America, **7** School of Chemistry, University of Sydney, Sydney, New South Wales, Australia

Abstract

Background: Conventional patent-based drug development incentives work badly for the developing world, where commercial markets are usually small to non-existent. For this reason, the past decade has seen extensive experimentation with alternative R&D institutions ranging from private-public partnerships to development prizes. Despite extensive discussion, however, one of the most promising avenues—open source drug discovery—has remained elusive. We argue that the stumbling block has been the absence of a critical mass of preexisting work that volunteers can improve through a series of granular contributions. Historically, open source software collaborations have almost never succeeded without such “kernels”.

Methodology/Principal Findings: Here, we use a computational pipeline for: (i) comparative structure modeling of target proteins, (ii) predicting the localization of ligand binding sites on their surfaces, and (iii) assessing the similarity of the predicted ligands to known drugs. Our kernel currently contains 143 and 297 protein targets from ten pathogen genomes that are predicted to bind a known drug or a molecule similar to a known drug, respectively. The kernel provides a source of potential drug targets and drug candidates around which an online open source community can nucleate. Using NMR spectroscopy, we have experimentally tested our predictions for two of these targets, confirming one and invalidating the other.

Conclusions/Significance: The TDI kernel, which is being offered under the Creative Commons attribution share-alike license for free and unrestricted use, can be accessed on the World Wide Web at <http://www.tropicaldisease.org>. We think the kernel will facilitate collaborative efforts towards the discovery of new drugs against parasites that cause tropical diseases.

Citation: Orti L, Carbajo RJ, Pieper U, Eswar N, Maurer SM, et al. (2009) A Kernel for Open Source Drug Discovery in Tropical Diseases. *PLoS Negl Trop Dis* 3(4): e418. doi:10.1371/journal.pntd.0000418

Editor: Timothy G. Geary, McGill University, Canada

Received: December 29, 2008; **Accepted:** March 23, 2009; **Published:** April 21, 2009

Copyright: © 2009 Orti 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.

Funding: MAM-R acknowledges the support from a Spanish Ministerio de Educación y Ciencia grant (BIO2007/66670). AS acknowledges the support from the Sandler Family Supporting Foundation and the National Institutes of Health (R01 GM54762, USA GM074945, P01 AI035707, and P01 GM71790). AP-L acknowledges the support from a Spanish Ministerio de Ciencia e Innovación grant (SAF2008-01845). RJC acknowledges the support from the Ramon y Cajal Program of the Spanish Ministerio de Educación y Ciencia. We are also grateful for computer hardware gifts to AS from Ron Conway, Mike Homer, Intel, IBM, Hewlett-Packard, and NetApp. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: sali@salilab.org (AS); mmarti@icpf.es (MAM-R)

† Current address: DuPont Knowledge Center, Hyderabad, India

Introduction

There is a lack of high-quality protein drug targets and drug leads for neglected diseases [1,2]. Fortunately, many genomes of organisms that cause tropical diseases have already been sequenced and published. Therefore, we are now in a position to leverage this information by identifying potential protein targets for drug discovery. Atomic-resolution structures can facilitate this task. In the absence of an experimentally determined structure, comparative modeling can provide useful models for sequences that are detectably related to known protein structures [3,4]. Approximately half of known protein sequences contain domains that can be currently predicted by comparative modeling [5,6]. This coverage

will increase as the number of experimentally determined structures grows and modeling software improves. A protein model can facilitate at least four important tasks in the early stages of drug discovery [7]: prioritizing protein targets for drug discovery [8], identifying binding sites for small molecules [9,10], suggesting drug leads [11,12], and optimizing these leads [13–15].

Here, we address the first three tasks by assembling our computer programs into a software pipeline that automatically and on large-scale predicts protein structures, their ligand binding sites, and known drugs that interact with them. As a proof of principle, we applied the pipeline to the genomes of ten organisms that cause tropical diseases (“target genomes”). We also experimentally tested two predicted drug-target interactions using Nuclear Magnetic

www.plosntds.org

1

April 2009 | Volume 3 | Issue 4 | e418

www.plosntds.org

1

April 2009 | Volume 3 | Issue 4 | e418

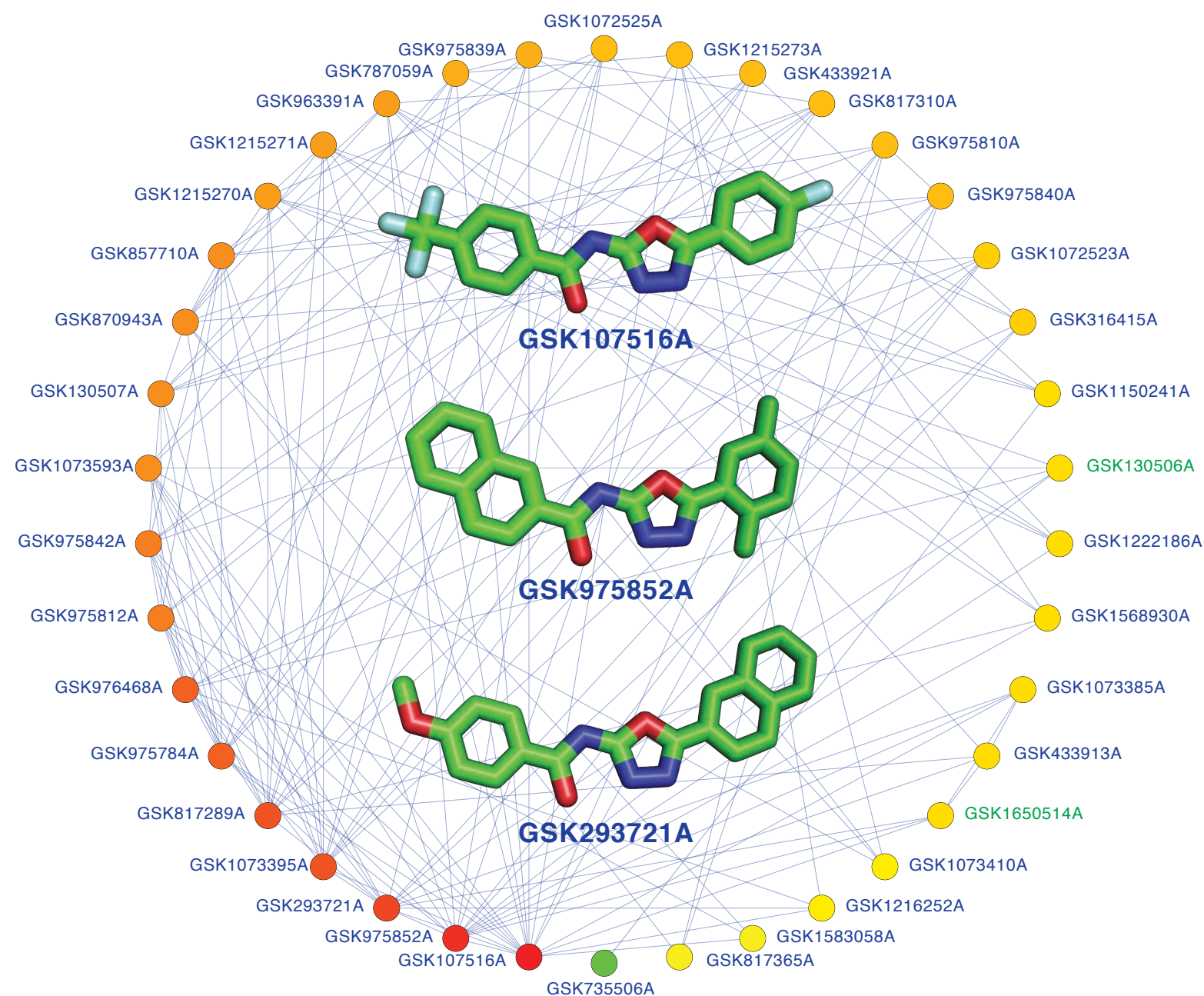
computational pipeline for: (i) comparative structure modeling of target proteins, (ii) predicting the localization of ligand binding sites on their surfaces, and (iii) assessing the similarity of the predicted ligands to known drugs. Our kernel currently contains 143 and 297 protein targets from ten pathogen genomes that are predicted to bind a known drug or a molecule similar to a known drug, respectively. The kernel provides a source of potential drug targets and drug candidates around which an online open source community can nucleate. Using NMR spectroscopy, we have experimentally tested our predictions for two of these targets, confirming one and invalidating the other.

Methodology/Principal Findings: Here, we use a computational pipeline for: (i) comparative structure modeling of target proteins, (ii) predicting the localization of ligand binding sites on their surfaces, and (iii) assessing the similarity of the predicted ligands to known drugs. Our kernel currently contains 143 and 297 protein targets from ten pathogen genomes that are predicted to bind a known drug or a molecule similar to a known drug, respectively. The kernel provides a source of potential drug targets and drug candidates around which an online open source community can nucleate. Using NMR spectroscopy, we have experimentally tested our predictions for two of these targets, confirming one and invalidating the other.

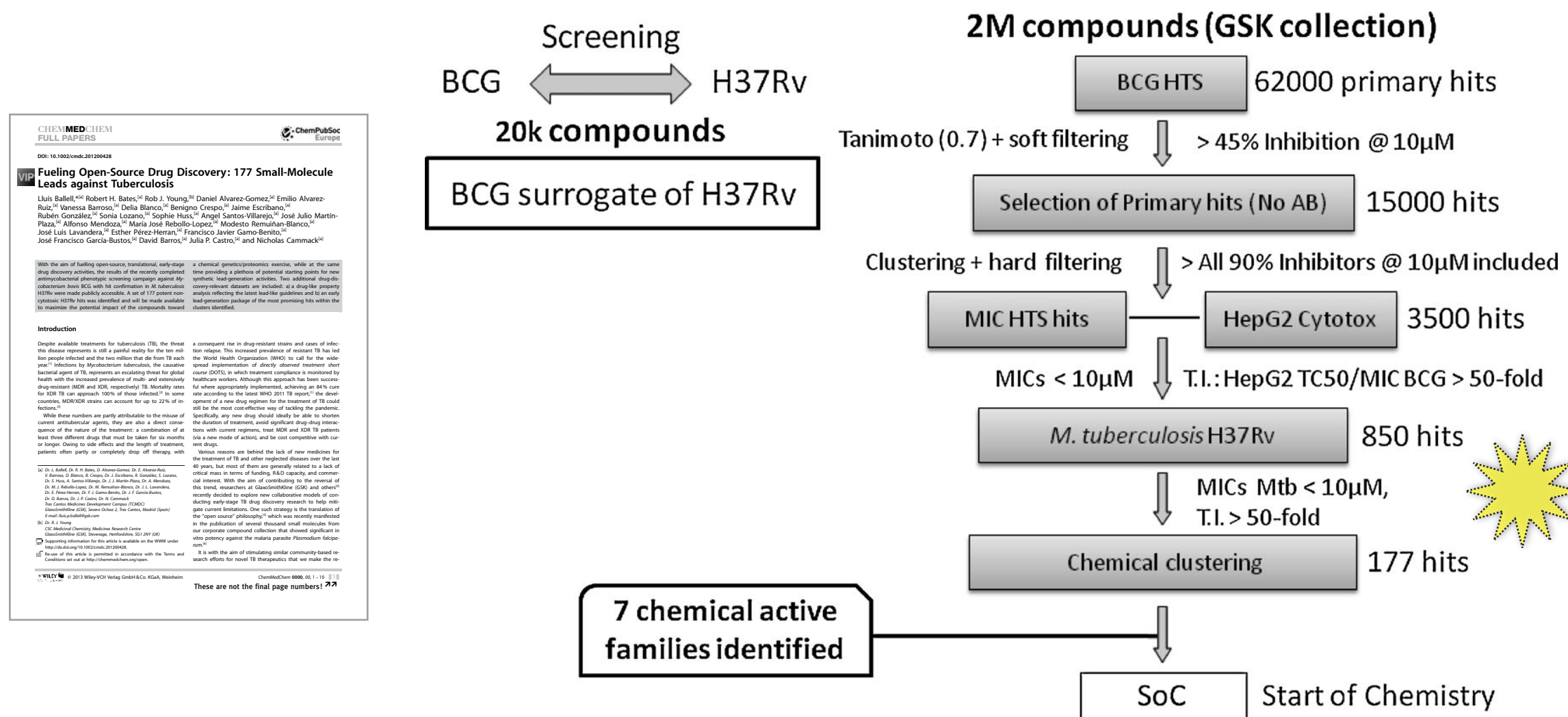
PLoS Negl Trop Dis

TCAMS-TB Target identification

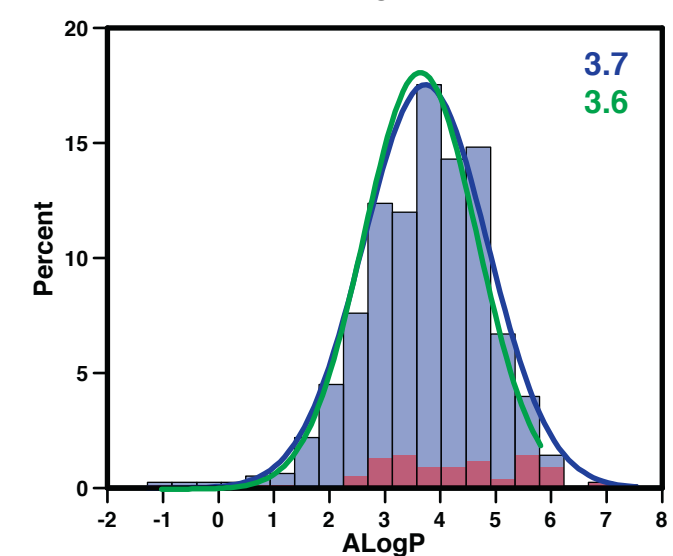
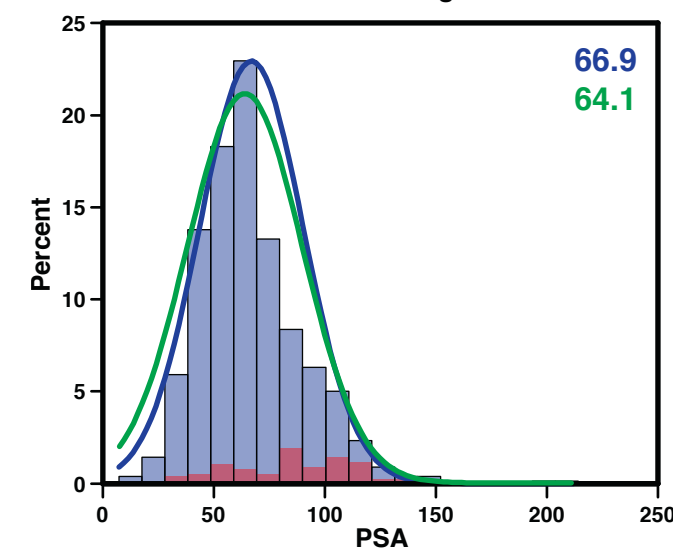
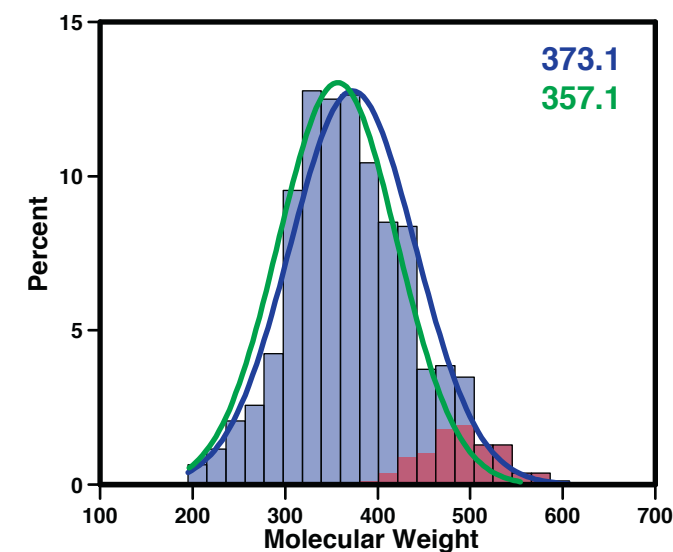
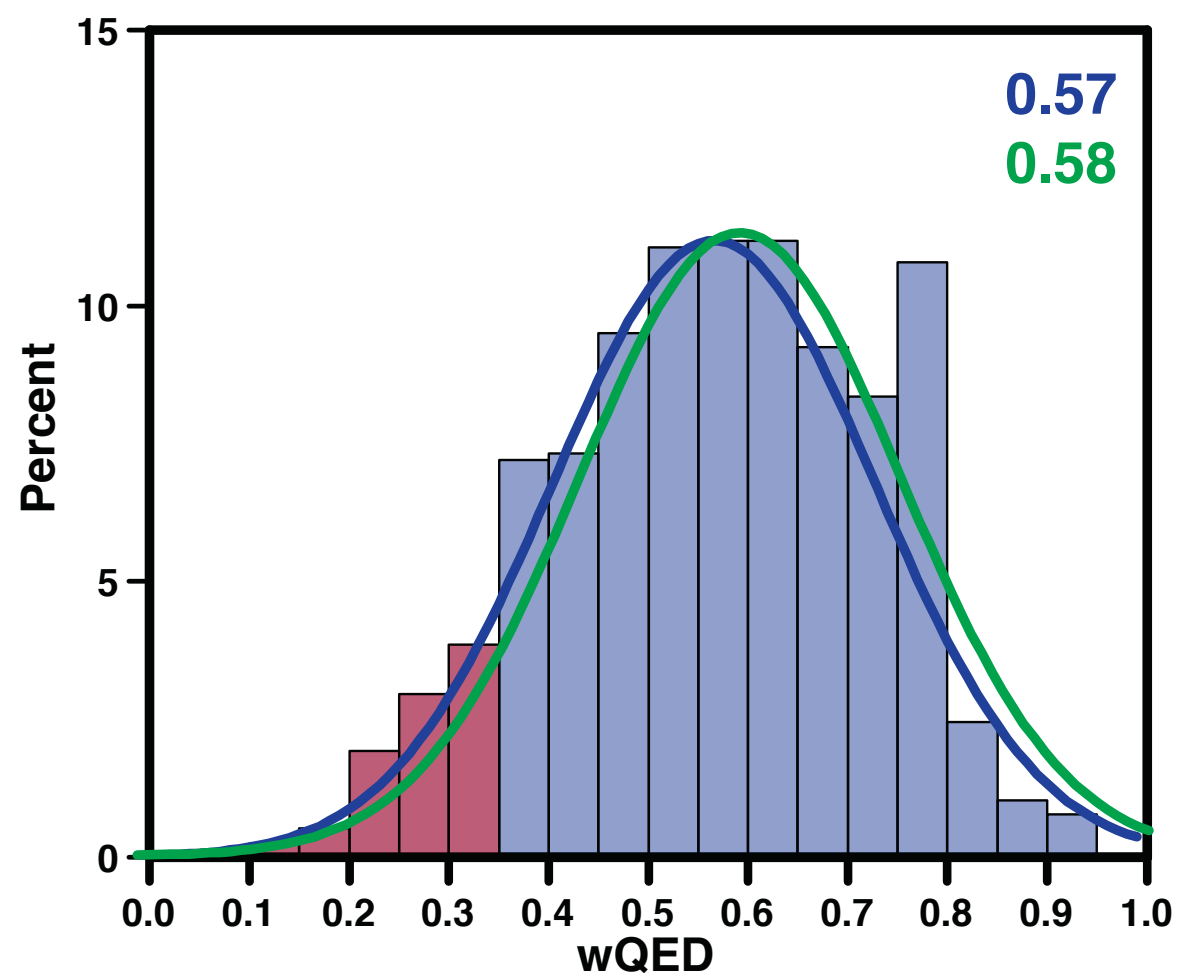
Sali, Brown, Overington & Marti-Renom labs



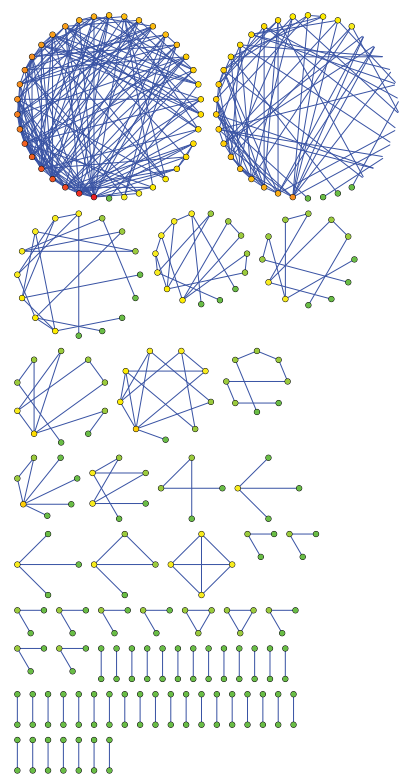
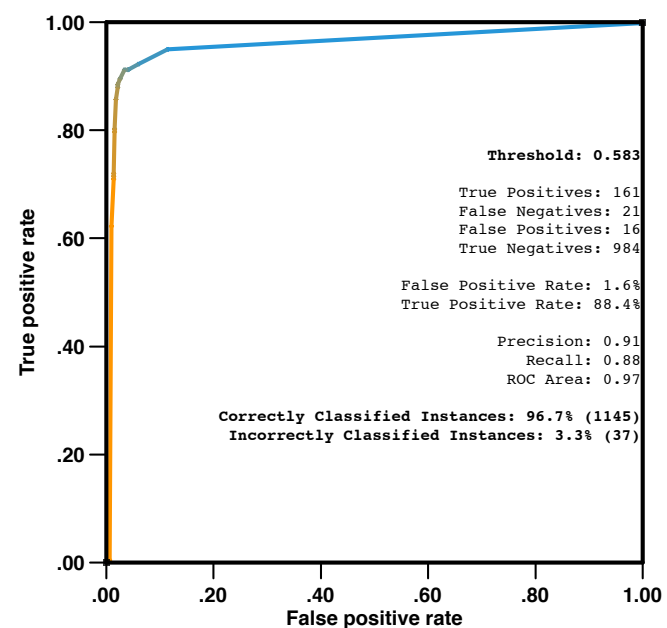
the TCAMS-TB dataset



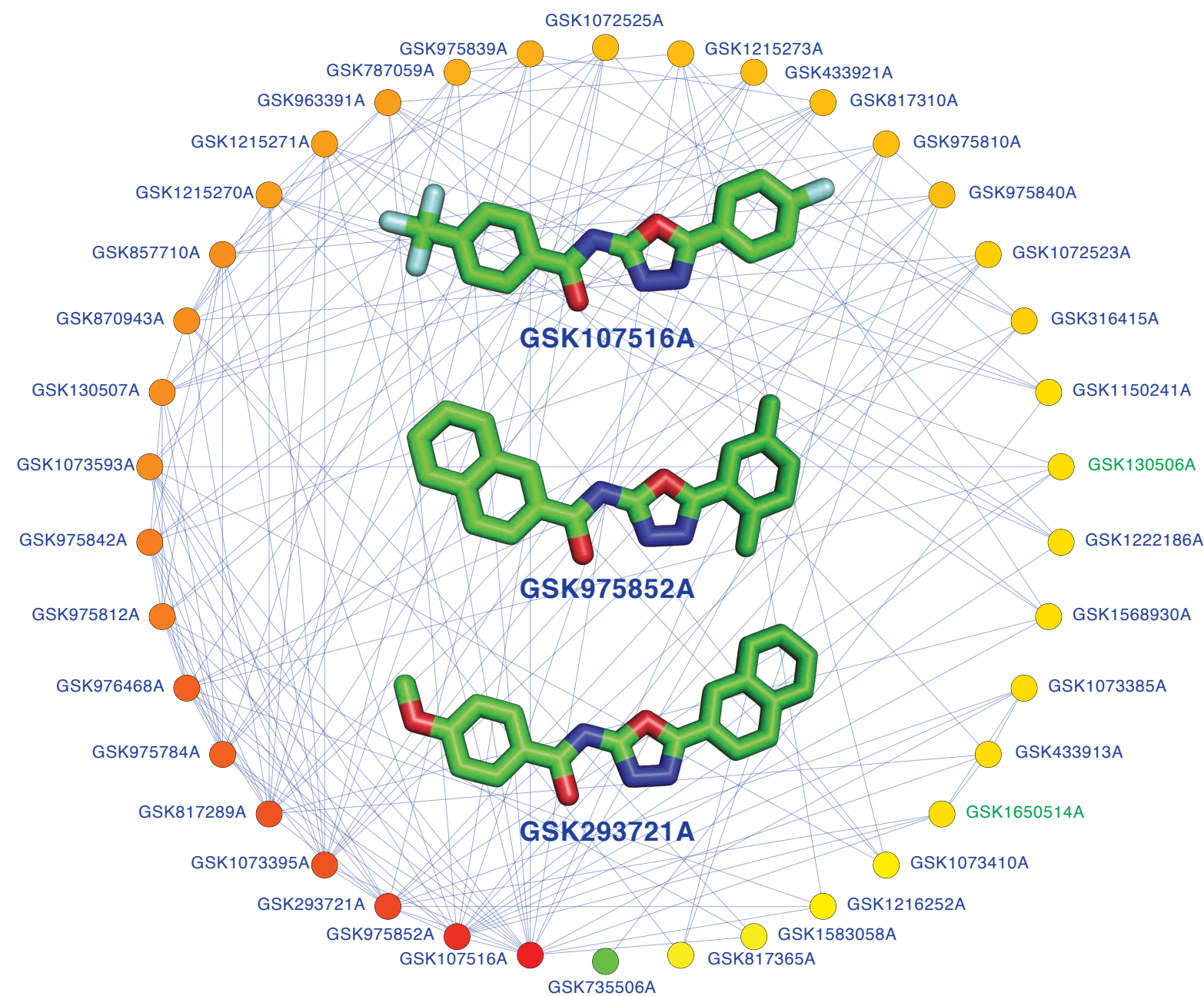
the TCAMS-TB dataset



the TCAMS-TB dataset

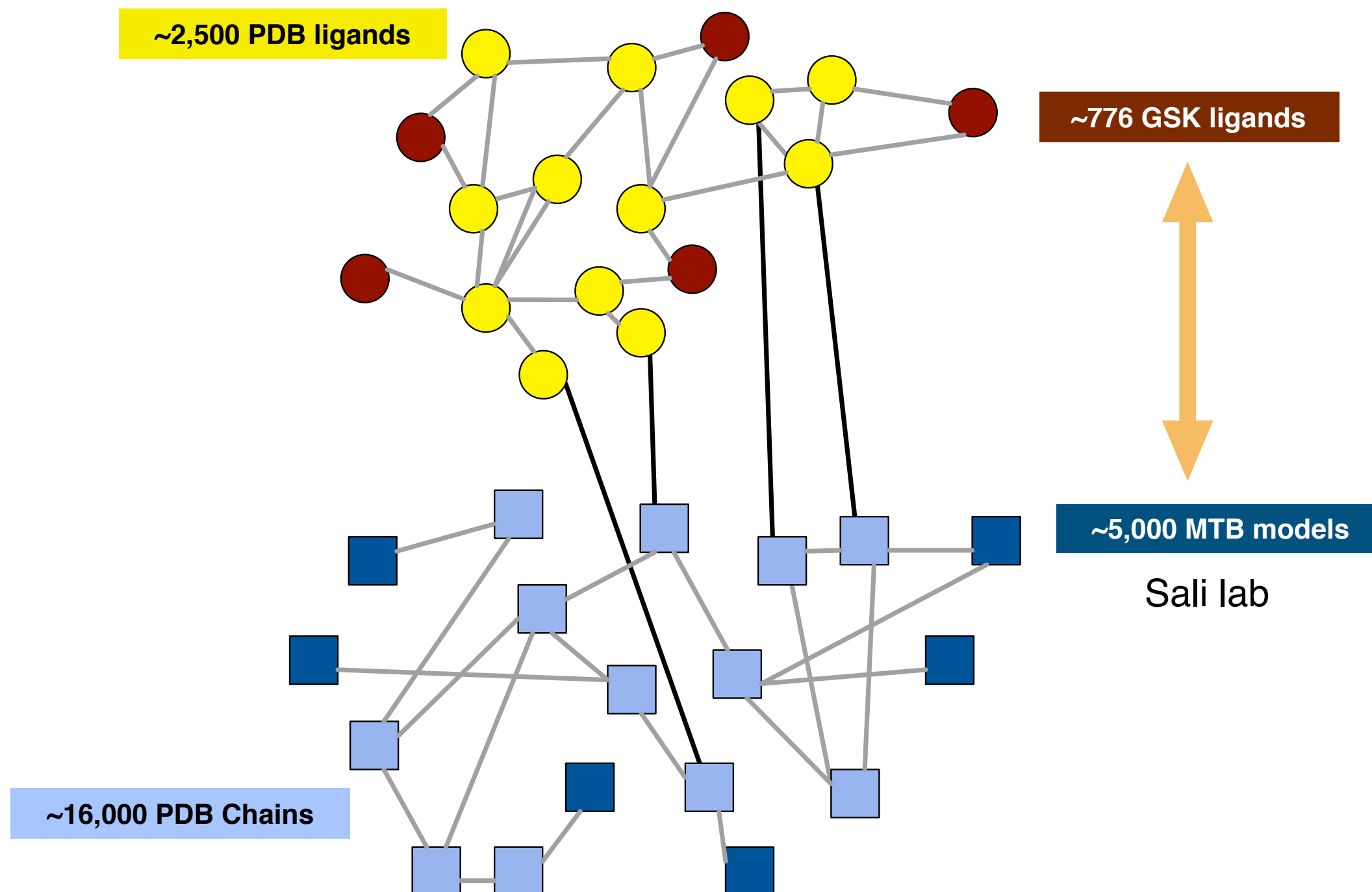


+ 486 singletons



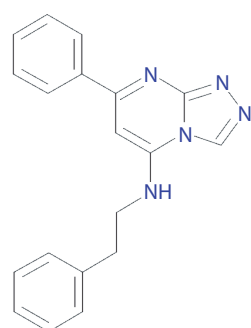
CNAG's nAnnoLyze (STR)

Marti-Renom lab

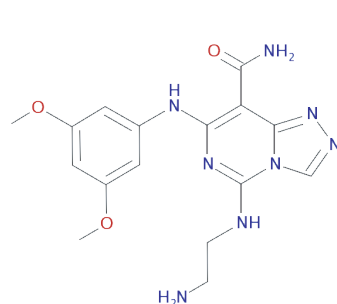


ChEMBL (CHEM)

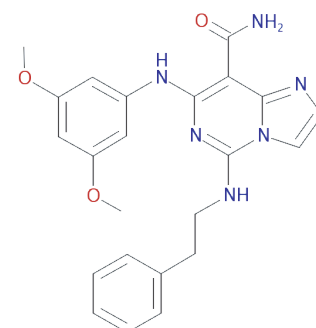
Overington Lab



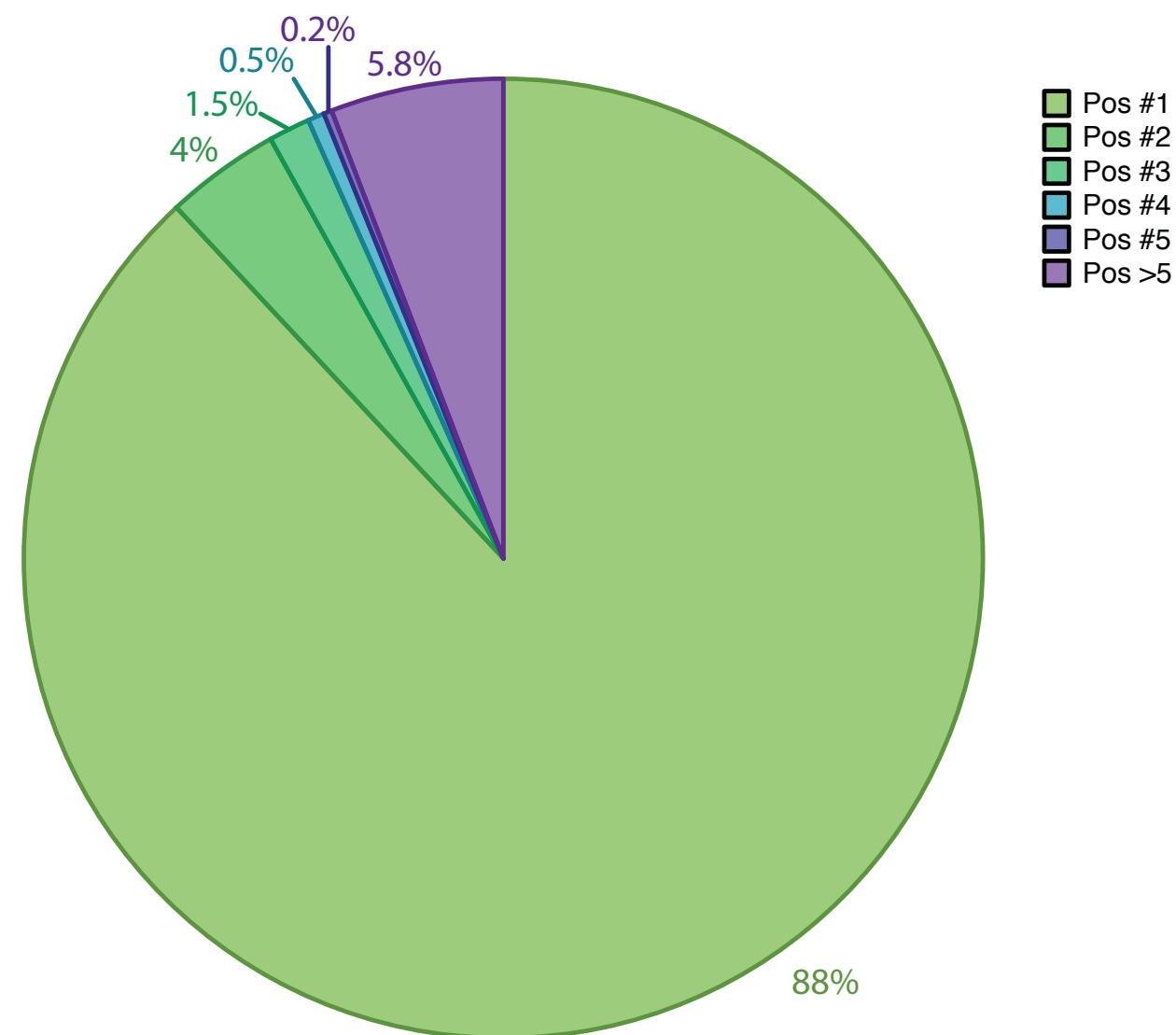
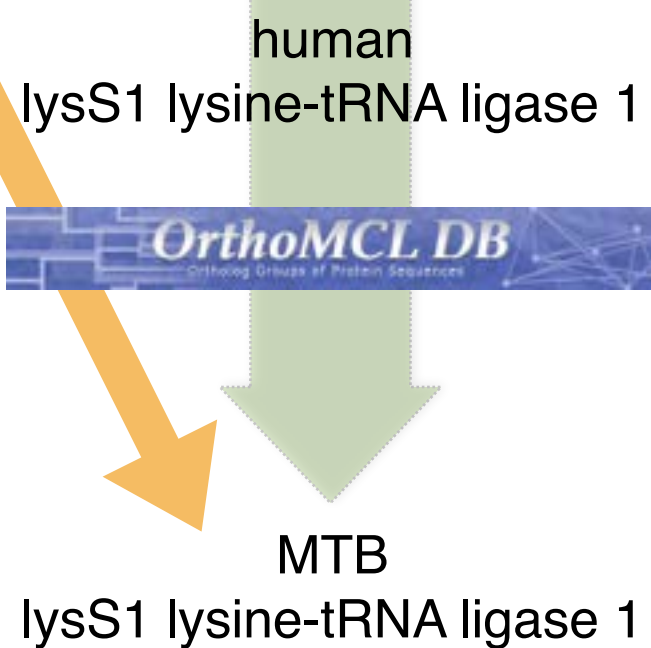
GSK1402290A



CHEMBL474582



CHEMBL508242



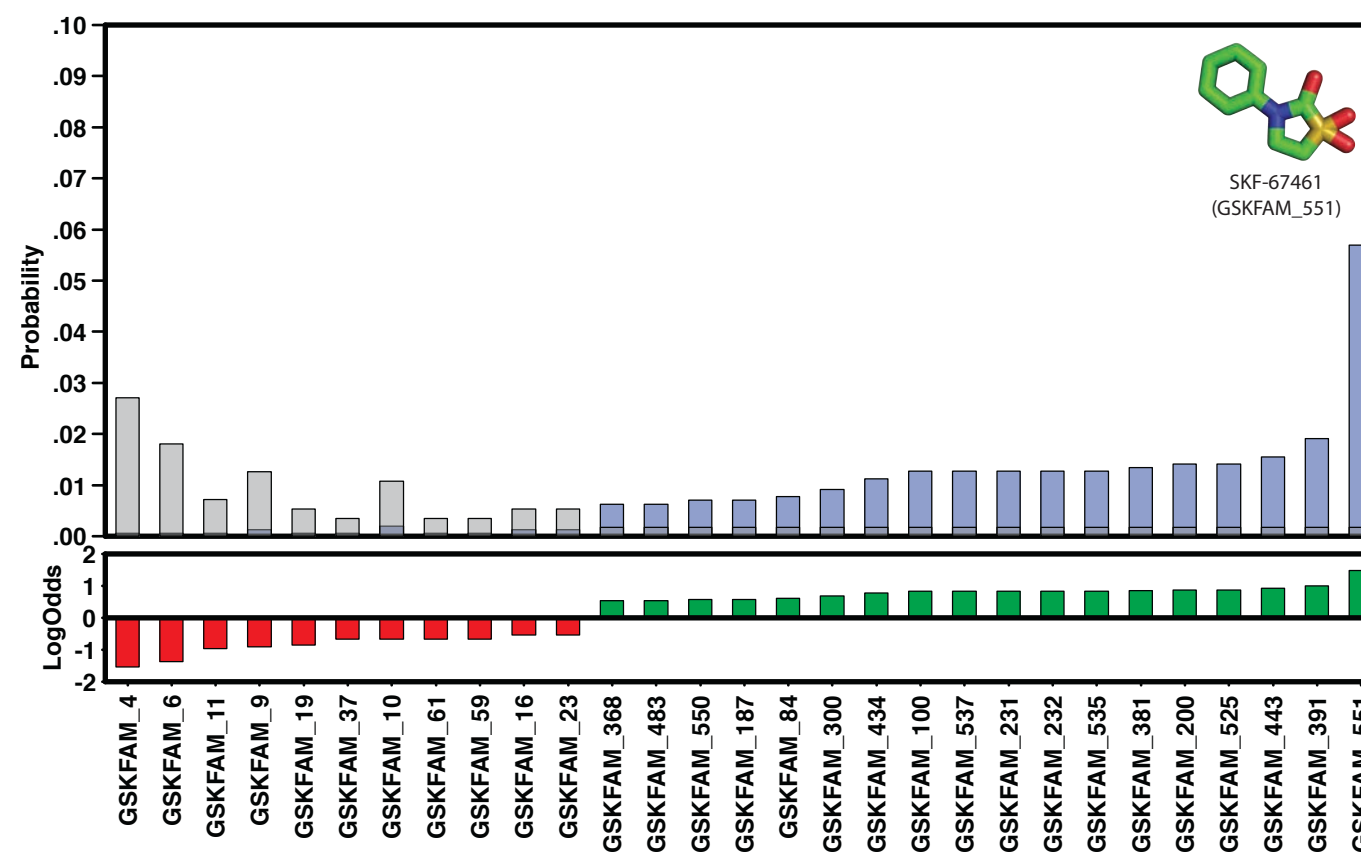
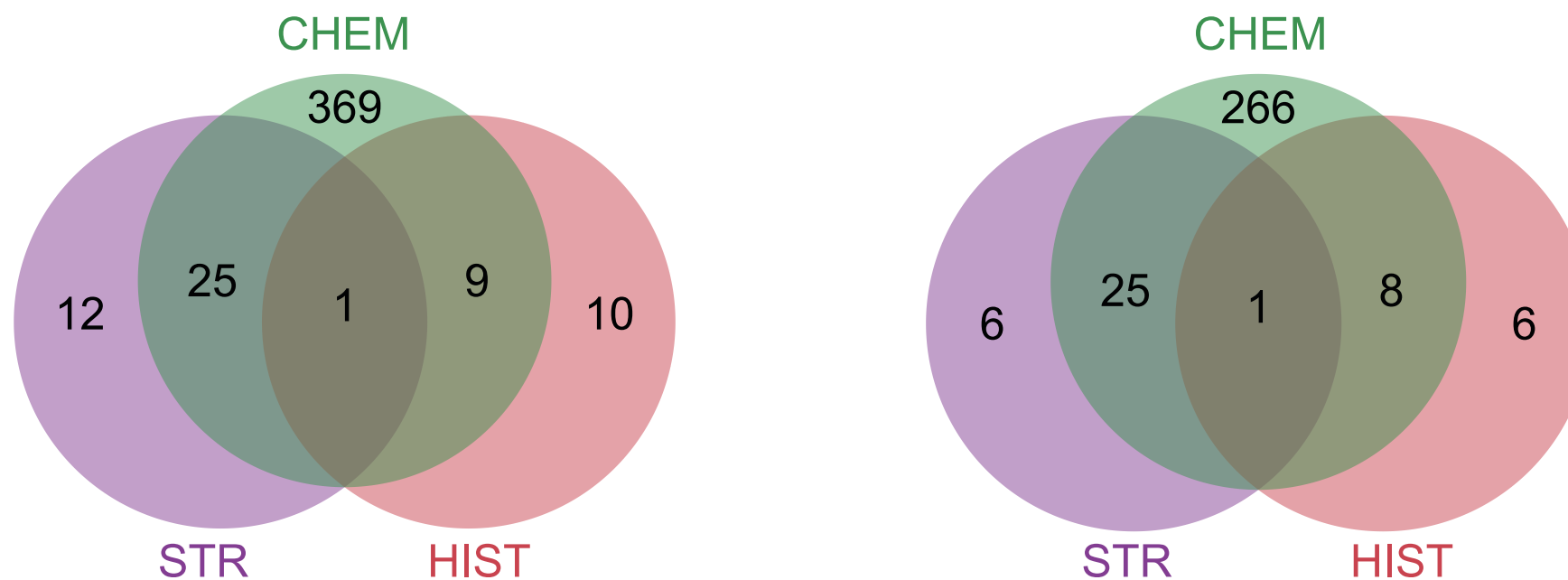
GSK (HIST)

Brown lab

Human Target Class	No. of Compounds		Putative <i>Mycobacterium</i> Target		Accession No.		Essentiality
	BCG	H37Rv	Gene	Product	M. bovis BCG	MTB H37Rv	
Kinase	35	8	pknA	transmembrane serine/threonine- protein kinase A	YP_976148.1	NP_214529.1	Essential
			pknB	transmembrane serine/threonine- protein kinase B	YP_976147.1	NP_214528.1	Essential
			pknD	Ser/Thr protein kinase	YP_977078.1	NP_215446.1	NE
			pknH	putative transmembrane serine/threonine- protein kinase H	YP_977417.1	NP_215782.1	NE
			pknJ	putative transmembrane serine/threonine- protein kinase J	YP_978197.1	NP_216604.1	NE
			pknL	putative transmembrane serine/threonine- protein kinase L	YP_978280.1	NP_216692.1	NE
	None	None	pknF	anchored- membrane serine/threonine- protein kinase F	YP_977877.1	NP_216262.1	NE
			pknK	putative serine/threonine- protein kinase transcriptional regulatory protein K	YP_979189.1	NP_217596.1	NE

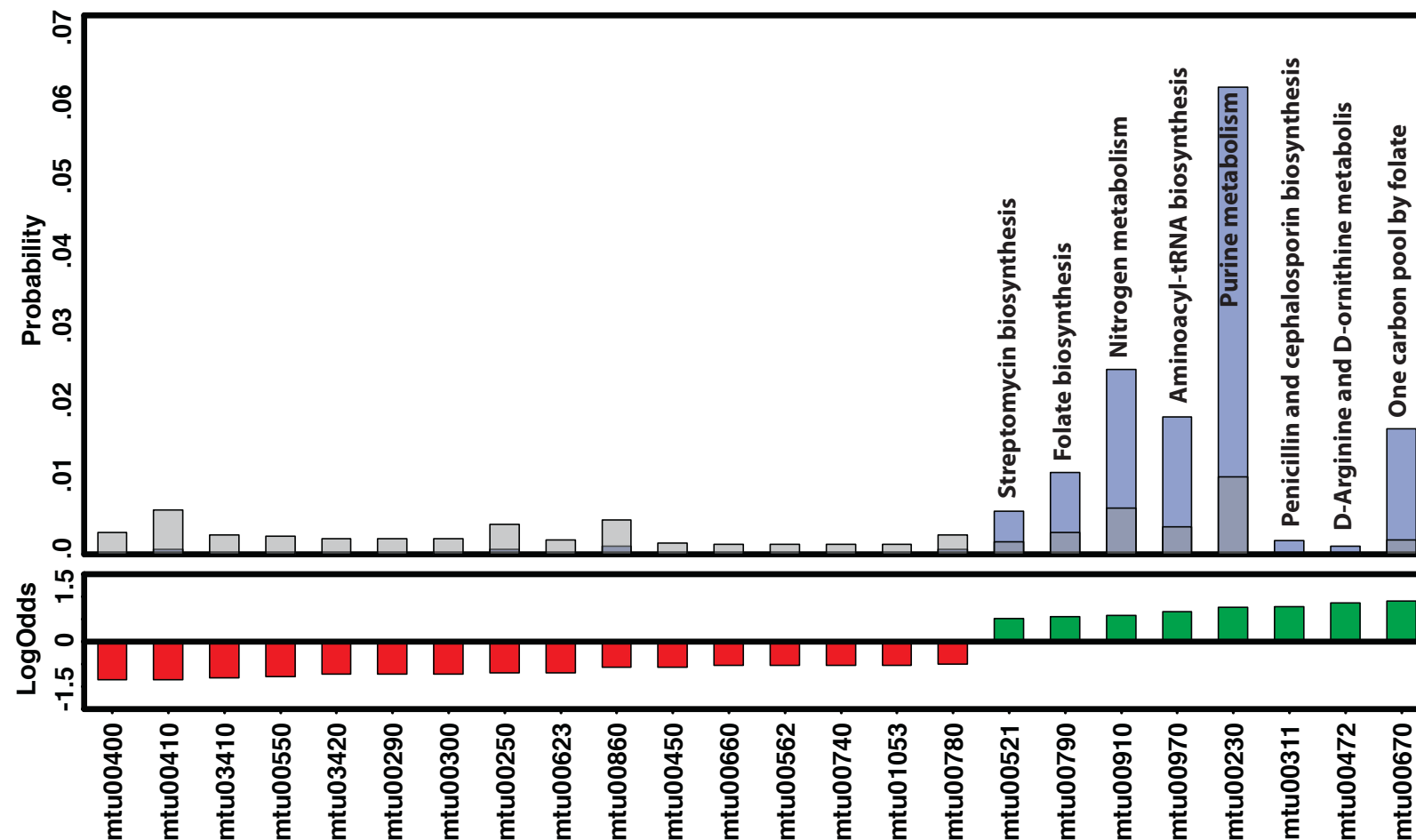
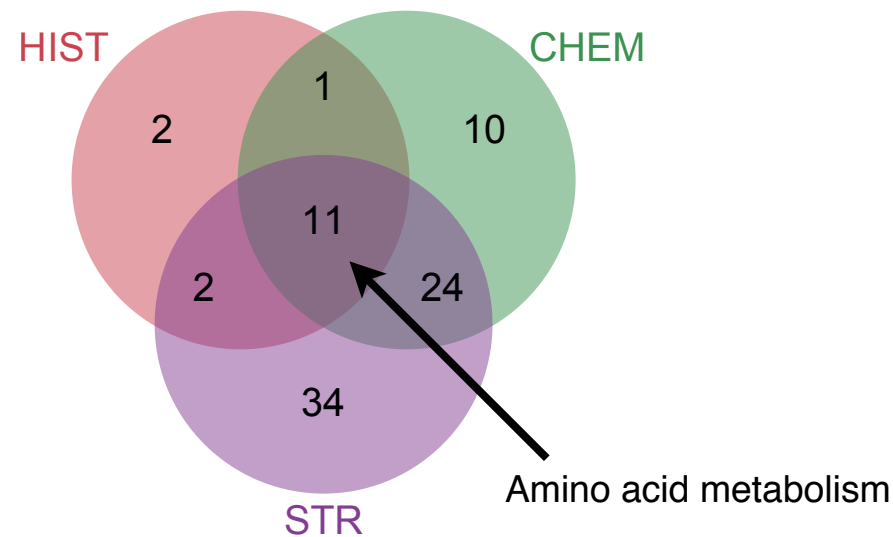
Compounds

776 compounds



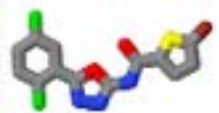
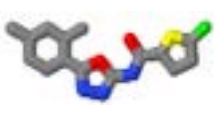
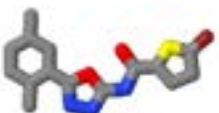

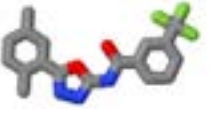
Pathways

1,044 unique targets in 112 (KEGG) pathways



Compounds-pathways

8 compound family “statistically” associated to 13 pathways

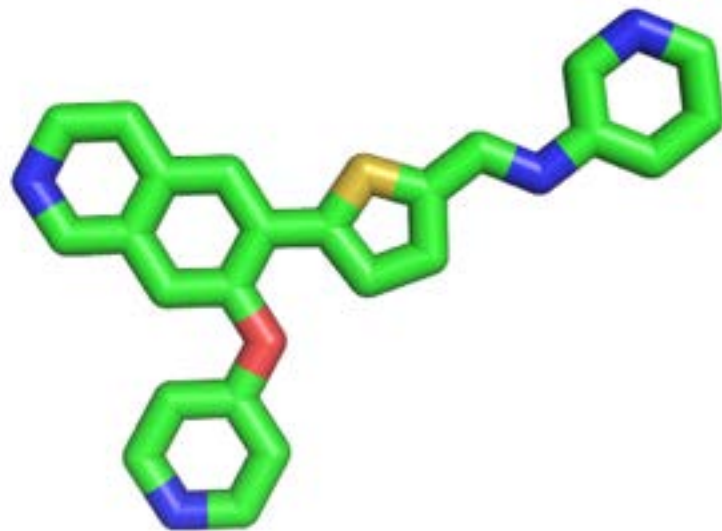
GSK Family	Compound	Target	Pathways
1	GSK975784A 	<i>Rv2182c</i> <i>Rv2483c</i>	Glycerolipid metabolism (mtu00561) Glycerophospholipid metabolism (mtu00564) No Pathway
	GSK975810A 	<i>Rv2182c</i> <i>Rv2483c</i>	Glycerolipid metabolism (mtu00561) Glycerophospholipid metabolism (mtu00564) No Pathway
	GSK975839A 	<i>Rv2182c</i> <i>Rv2483c</i> <i>Rv2299c</i>	Glycerolipid metabolism (mtu00561) Glycerophospholipid metabolism (mtu00564) No Pathway No Pathway
	GSK975840A 	<i>Rv2182c</i> <i>Rv2483c</i>	Glycerolipid metabolism (mtu00561) Glycerophospholipid metabolism (mtu00564) No Pathway
	GSK975842A 	<i>Rv2182c</i> <i>Rv2483c</i> <i>Rv2045c</i> <i>Rv2139</i> <i>Rv2299c</i> <i>Rv2483c</i>	Glycerolipid metabolism (mtu00561) Glycerophospholipid metabolism (mtu00564) No Pathway No Pathway Pyrimidine metabolism (mtu00240) No Pathway No Pathway

8 compound families with significant links to pathways

All available @ <http://www.tropicaldisease.org/TCAMSTB>

pknB target

GSK1598164A



```

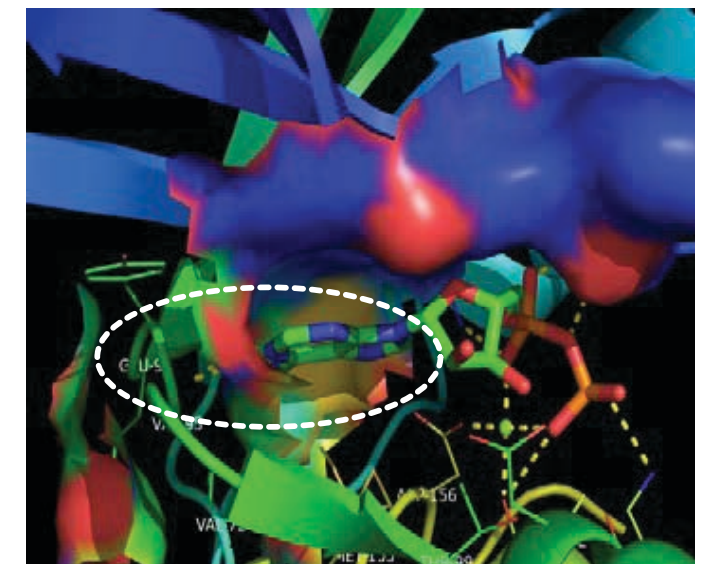
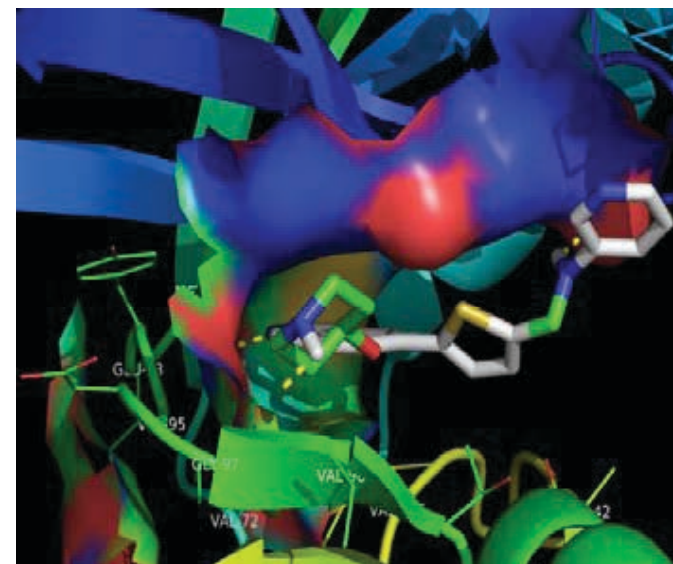
PKNB -----MTTPSHLSDRVLEGEILGFGMSEVHLARDLRDHRDVAVKVLRADLARDPS-FYLRFRREAQNAANHPAIVAVYDTGEAETPAGPLPVVMEYVVGVTI 100
CAMK2D -----SMTDEVOLFEBLGKCA SVVRRCKMKTPTGOEYAKTITNTKK SARD--HQKTEREARCRLKHFNIRHDSISEEG----FHLVLFVLVGGGL 90
MARK3 -----GAMGSDEQPHIGNRLLKTLGKCNFAKVKLARHILTGREVAITITIDKTO NPTS--LOKIFREVRMKIINHPNIMKFEVIETEK----TLVILMEYASGGGV 98
MARK2 -----MADLHIGNRLLKTLGKCNFAKVKLARHILTGREVAITITIDKTO NSSS--LOKIFREVRMKVINHPNIMKFEVIETEK----TLVILMEYASGGGV 93
AKT2 -----KVTMNDLDYLKLLGKCTGKVILVREKATGRYYAMKITIRKEVIAKDEVAHTTES-RVQLNTRHPFLTAQKYAFQTHD----RLCFVMEYANGGGL 92
SGK1 GISQPQEPELMNANPAPPAPSQQINLGSSNPNAKPSDHFLLKVLGKCSGKVLLARHKAEVFYAVKVLQKKA LKKKEEKHI SERNV LKNVKKHFLVGLHFSFOTAD----LVVLDITGGGL 126

PKNB RDI VHTEGPMTPKRAIEVIADACQALNFESHONGTTHRDVKNANIMISATNA---VVMDFGTARAIDSGNSVTQTAAVICTAQVLSPEQARGDSVDARSDVYSLGCVLYEVLITGEPPTGDSPDSDVAY 226
CAMK2D FEDIVAREYYSADASHCQITLESVNHCNLNGIVHRDLKPNLLASKSKGAAYKLTADFGLAIVQGDQQAWFG---ACTEGVLSPEVLRKDPGKPVDMMACVILYILLYGYPPWDEQHRLYQ 216
MARK3 FDYLVAGRMKKEARSKFRIVSAVOYCHQKRIVHRDLKPNLLDADMN---KITADFGFSNFTVG-GKLDT---CGSPYAAPFQGGKDGPEVDVSLGVILYTLVSGSLPDGQNLKELRE 221
MARK2 FDYLVAGWMKKEARAKFRIVSAVOYCHQKFIHRDLKPNLLDADMN---KITADFGFSNFTFG-NKLDT---CGSPYAAPFQGGKDGPEVDVSLGVILYTLVSGSLPDGQNLKELRE 216
AKT2 FHL SRERVFTERRARFYGAIVSALFYLHSDVVRDILKPNLLDKDGH---KITDFGLCKGISDGATMKT---CGTEYLAPVLEDNDGRAVDWGLSVVMYEMMCRLPEYNDHERLPE 215
SGK1 FHLQRERCFLPRARFYAAIASALGYLSLNIVYRDLKPNLLDSQGH---IVLTDFGLCKNIEHNSTTST---CGTEYLAPVLEHQPDRTDWVCLSAVLYEMLYGLPPYSRTAEMVD 249

PKNB QHVREDPIP SARHEGLSADLDVVLKATANKENRYQ-----TAAERADLVRVHNGEPPEAPKVLTDARTSLSSAAGNLSGPR----- 308
CAMK2D QKAGAYDF SPEWDTVPEAKDLINKMTINAKRIT-----ASTAKHPICORSTVASMHRQETVDCCLKFNARRKLKGAILTTLATRNFSAAKSLKKPDGVKESTESSN----- 327
MARK3 RVLRGKYRI---FYMSTDCENLLKRFILNLIKRGTL---LEIILKDRINAGHEEDEKPFVEPELDISDQKRIDIMVGMGYSQEEIQESLSMKYDEITATYLLGRKSSE----- 328
MARK2 RVLRGKYRI---FYMSTDCENLLKRFILNLIKRGTL---LEIILKDRMNVGHEDDEKPYVEPLPDYKDPRTETLMVSMGYTREEIQDSLVLGQRYNEVMATYLLGLY----- 319
AKT2 LLMEEIRFPR---TSSPEAKSLAGLKKDKORLGGGSPDAKLVVPHRFLSINWQDVQKLLPFPKQVTSVDTRYFDDEFTAQSSITITPPDRYDSLGL----- 315
SGK1 NLLNKPLQLKP---NUN SARHDEGLLQKDRTKRLG-AKDDFMILKSHVFSLINWDDLNNKITPPFPNPVSGPNDLRHFDPEFTTEVPVNAIGKAPDSVLVTASVKEAAEAFGLFSYAPPTDSFL 373

```

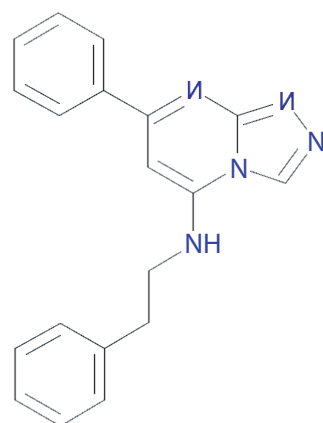
GSK number	pIC50	Gene	Score _{2PZ1} *	Score _{3F69} *
GW623128X	7.3	CNR2	-8.83	-8.63
GSK1519001A	5.7	NPSR1	-9.09	-8.65
GSK547481A	6.1	HTR4	-8.93	-9.27
GSK2043267A	7.5	CYP2C19	-8.8	-8.56
GSK381407A	5.6	P2RY14	-9.02	-8.71
GSK547543A	6.2	GPR55	-9.03	-10.37
GSK547511A	6.2	GPR55	-9.01	-8.58
GSK1598164A	8	IKBKB	-9.19	-8.96
GSK1635139A	5.6	CHRNA7	-8.6	-9.59



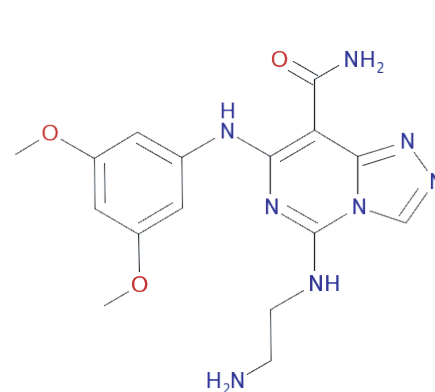
PknB kinase docking to GSK1598164A. **A)** Multiple sequence alignment of *Mycobacterium* PknB kinase with selected human kinases. Human kinases were selected on the criteria of having available PDB structures and top Psi-BLAST scores to *M. bovis* transmembrane serine/threonine-protein kinase B (pknB). First sequence in the alignment (gene name; PDB identifier) is *M. tuberculosis* transmembrane serine/threonine-protein kinase B (PknB; 3F69), which is 99% identical to *M. bovis* PknB and was used in compound docking models. Other sequences are CAMK2D (2EWL), MARK3 (2QNJ), MARK2 (3IEC), AKT2 (1GZK) and SGK1 (2R5T). Residues known to interact with ADP in pknB are highlighted in red. The amino acids aligned with Glu93, which may be essential for the binding of the GSK1132084A, are highlighted in green. **B)** Binding models of the GSK1598164A and ADP within pknB binding site (left and right panels, respectively).

serS target

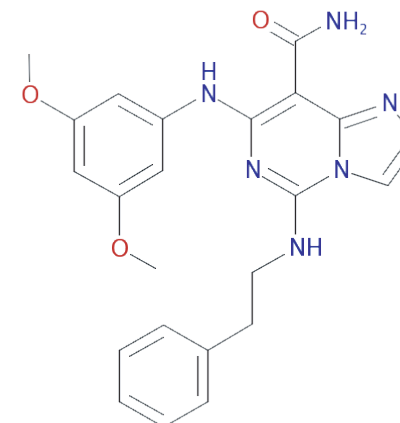
GSK1402290A



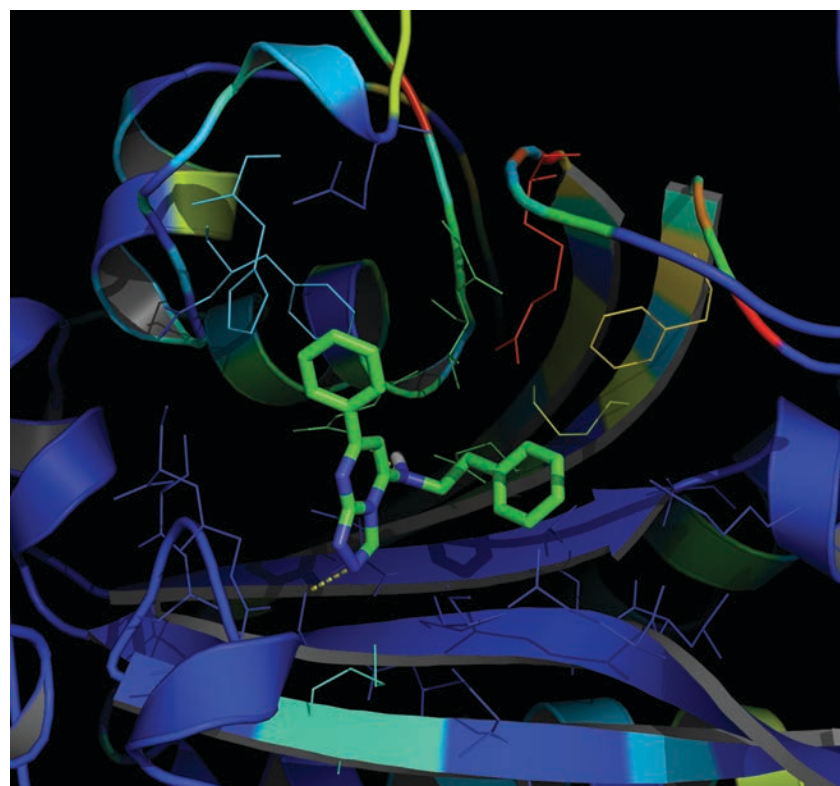
GSK1402290A



CHEMBL474582



CHEMBL508242




Targeting the aminoacyl-tRNA biosynthesis pathway. CHEM results show that GSK1402290A shared several substructural features with compounds reported as potent lysyl-tRNA synthetase inhibitors in the ChEMBL database (e.g., CHEMBL474582 and CHEMBL508242). STR results predicted the serS as a target of GSK1402290A with its binding site including residues F205, H209, G225, T226, E228, R257, F276, K278, and E280, which are conserved in the PFAM family PF00587 (tRNA synthetase class II core domain). The image shows the pose for GSK1402290A predicted by AutoDock and the binding site residues (i.e., within 6Å from the compound) colored from low sequence conservation (blue) to high sequence conservation (red).

<http://www.tropicaldisease.org/TCAMSTB>

TCAMSTB Id: CW335158X
 Polar Surface Area: 56.15
 HB Acceptor: 4
 HB Donor: 2
 ALogP: 1.9166
 Mol. Weight: 240.257
 QED: 0.487
 Compound family: CISKFAM_476

[Download SDF](#)



Label: **TCAMSTB**

TCAM5TB Model id: ttg
Target beginning: 1
Target end: 412
Sequence Identity: 38%
Evalue: 0
ga341: 1
MPQS: 1.470d
edops: -0.64
pdb template: 3pb
pdb chain: A
pdb beginning: 1
pdb end: 417

[Download model](#)

Spin: ☐

jmol: 5

Article

<http://www.tropicaldisease.org/TCAMSTB>

Martínez-Jiménez, et al. (2013). PLoS CB, 9(10), e1003253

OPEN ACCESS Freely available online



Target Prediction for an Open Access Set of Compounds Active against *Mycobacterium tuberculosis*

Francisco Martínez-Jiménez^{1,2}, George Papadatos³, Lun Yang⁴, Iain M. Wallace³, Vinod Kumar⁴, Ursula Pieper⁵, Andrej Sali⁵, James R. Brown^{4*}, John P. Overington^{3*}, Marc A. Marti-Renom^{1,2*}

1 Genome Biology Group, Centre Nacional d'Anàlisi Genòmica (CNAG), Barcelona, Spain, **2** Gene Regulation Stem Cells and Cancer Program, Centre for Genomic Regulation (CRG), Barcelona, Spain, **3** European Molecular Biology Laboratory – European Bioinformatics Institute (EMBL-EBI), Wellcome Trust Genome Campus, Hinxton, Cambridge, United Kingdom, **4** Computational Biology, Quantitative Sciences, GlaxoSmithKline, Collegeville, Pennsylvania, United States of America, **5** Department of Bioengineering and Therapeutic Sciences, University of California, San Francisco, San Francisco, California, United States of America

Abstract

Mycobacterium tuberculosis, the causative agent of tuberculosis (TB), infects an estimated two billion people worldwide and is the leading cause of mortality due to infectious disease. The development of new anti-TB therapeutics is required, because of the emergence of multi-drug resistance strains as well as co-infection with other pathogens, especially HIV. Recently, the pharmaceutical company GlaxoSmithKline published the results of a high-throughput screen (HTS) of their two million compound library for anti-mycobacterial phenotypes. The screen revealed 776 compounds with significant activity against the *M. tuberculosis* H37Rv strain, including a subset of 177 prioritized compounds with high potency and low *in vitro* cytotoxicity. The next major challenge is the identification of the target proteins. Here, we use a computational approach that integrates historical bioassay data, chemical properties and structural comparisons of selected compounds to propose their potential targets in *M. tuberculosis*. We predicted 139 target - compound links, providing a necessary basis for further studies to characterize the mode of action of these compounds. The results from our analysis, including the predicted structural models, are available to the wider scientific community in the open source mode, to encourage further development of novel TB therapeutics.

Citation: Martínez-Jiménez F, Papadatos G, Yang L, Wallace IM, Kumar V, et al. (2013) Target Prediction for an Open Access Set of Compounds Active against *Mycobacterium tuberculosis*. PLoS Comput Biol 9(10): e1003253. doi:10.1371/journal.pcbi.1003253

Editor: Alexander Donald MacKerell, University of Maryland, Baltimore, United States of America

Received: May 2, 2013; **Accepted:** August 11, 2013; **Published:** October 3, 2013

Copyright: © 2013 Martínez-Jiménez 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.

Funding: MAMR acknowledges the support from the Spanish Government (BFU2010-19310) and the Era-Net Pathogenomics (PIM2010EPA-00719). JPO and GP were supported by funding from the EMBL Member States. JRB acknowledges the support of GlaxoSmithKline R&D and AS acknowledges support from NIH (U54 GM094662 and P01 GM71790). GlaxoSmithKline participated in this research through the co-authorship of VK, LY, and JRB. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: VK, LY, and JRB are paid employees of GlaxoSmithKline (GSK). This does not alter our adherence to the PLOS policies on sharing data and materials. All other authors have declared that no competing interests exist.

* E-mail: James.R.Brown@gsk.com (JRB); jpo@ebi.ac.uk (JPO); mmarti@pcb.ub.cat (MAMR)

Introduction

One third of the world's population is infected with *Mycobacterium tuberculosis* (MTB), the causative agent of tuberculosis [1]. Approximately 95% of infected individuals are thought to have persistent, latent MTB infections that remain dormant until activated by specific environmental and host response events. Approximately 10% of latent infections eventually progress to active disease, which, if left untreated, kills more than half of the infected patients [2]. Moreover, there is an increasing clinical occurrence of MTB strains with extensive multi-drug-resistance (eg, MTB MDR and MTB XDR), where mortality rates can approach 100% [3]. In some countries, the MTB MDR and XDR strains may account for up to 22% of infections [1]. In addition, current TB therapeutic regimes involve a combination of antibiotics, administered at regular intervals over a 6-month period, which makes patient compliance an issue, especially in developing countries [1,2].

The discovery and development of new antibiotics is widely recognized as one of the major global health emergencies, yet it is also a major pharmaceutical challenge. Most currently used antibiotics were discovered during the golden era from the 1940s

to 1960s through large scale screening of compound collections for anti-bacterial activity – the so-called whole cell or phenotypic screens [4]. The emergence of bacterial molecular genomics technologies and the availability of whole genome sequences in the 1990s led to dramatic changes in anti-bacterial drug discovery, where the emphasis was placed on screening essential targets for inhibitory compounds. However, despite intensive efforts, target-based screening has been largely unsuccessful in producing clinical candidate molecules [5]. As a result, a return to whole cell screening has been widely advocated, in combination with novel technologies and bioinformatics to rapid identify targets associated with a compound's mechanism of action (MOA) [4,6].

Recently, the pharmaceutical company GlaxoSmithKline (GSK) completed an anti-mycobacterial phenotypic screening campaign against *M. bovis* BCG, a non-virulent, vaccine *Mycobacterium* strain, with a subsequent secondary screening in *M. tuberculosis* H37Rv (MTB H37Rv) for hit confirmation [7]. A total of 776 potent compound hits (including 177 MTB H37RV hits with limited human cell line toxicity) were made openly available to the wider scientific community through the ChEMBL database (<http://dx.doi.org/10.6019/CHEMBL2095176>). The aim of this release was to stimulate mechanism of action analyses using

Fast Sequencing for TB treatment

Köser & Peacock

Whole-Genome Sequencing for Rapid Susceptibility Testing of *M. tuberculosis*

TO THE EDITOR: Efforts to contain drug-resistant tuberculosis depend on the rapid detection and effective treatment of cases, together with public health interventions to prevent and investigate ongoing transmission. The necessary laboratory support for these activities includes the identification of the *Mycobacterium tuberculosis* complex, antimicrobial susceptibility testing, and bacterial genotyping. However, even in well-resourced countries, it typically takes 1 to 2 months to achieve all these goals because of the slow growth rate of the *M. tuberculosis* complex.^{1,2} Moreover, phenotypic susceptibility testing can be unreliable and is not performed for some agents. Molecular techniques have accelerated some of these diagnostic functions, but they only interrogate a small part of the microbial genome and do not provide all the clinically relevant information.¹⁻³ Whole-genome sequencing has not been used as a diagnostic tool for tuberculosis, in part because of the need to culture *M. tubercu-*

losis complex for several weeks, until sufficient DNA can be extracted.^{2,4}

Here we report the use of rapid whole-genome sequencing to investigate the case of a patient with extensively drug-resistant (XDR) tuberculosis (the case history is provided in the Supplementary Appendix, available with the full text of this letter at NEJM.org). His first sputum sample became culture-positive after 3 days in the mycobacterial growth indicator tube (MGIT) culture system. DNA was extracted directly from the MGIT tube and sequenced with the use of the Illumina MiSeq platform. Two distantly related Beijing strains of *M. tuberculosis* were identified (in a ratio of 7:3) (Fig. 1B). Mixed infection was not apparent when standard genotyping was performed on three additional samples from this patient by means of mycobacterial interspersed repetitive-unit-variable-number tandem-repeat assay, which probably identified the majority strain. These findings have important implications for

Claudio U. Köser, Ph.D.
Public Health England
Cambridge, United Kingdom

Josephine M. Bryant, B.Sc.
Wellcome Trust Sanger Institute
Hinxton, United Kingdom

Jennifer Becq, Ph.D.
Illumina (Cambridge)
Little Chesterford, United Kingdom

M. Estée Török, Ph.D.
University of Cambridge
Cambridge, United Kingdom

Matthew J. Ellington, D.Phil.
Public Health England
Cambridge, United Kingdom

Marc A. Marti-Renom, Ph.D.
Centre Nacional d'Anàlisi Genòmica
Barcelona, Spain

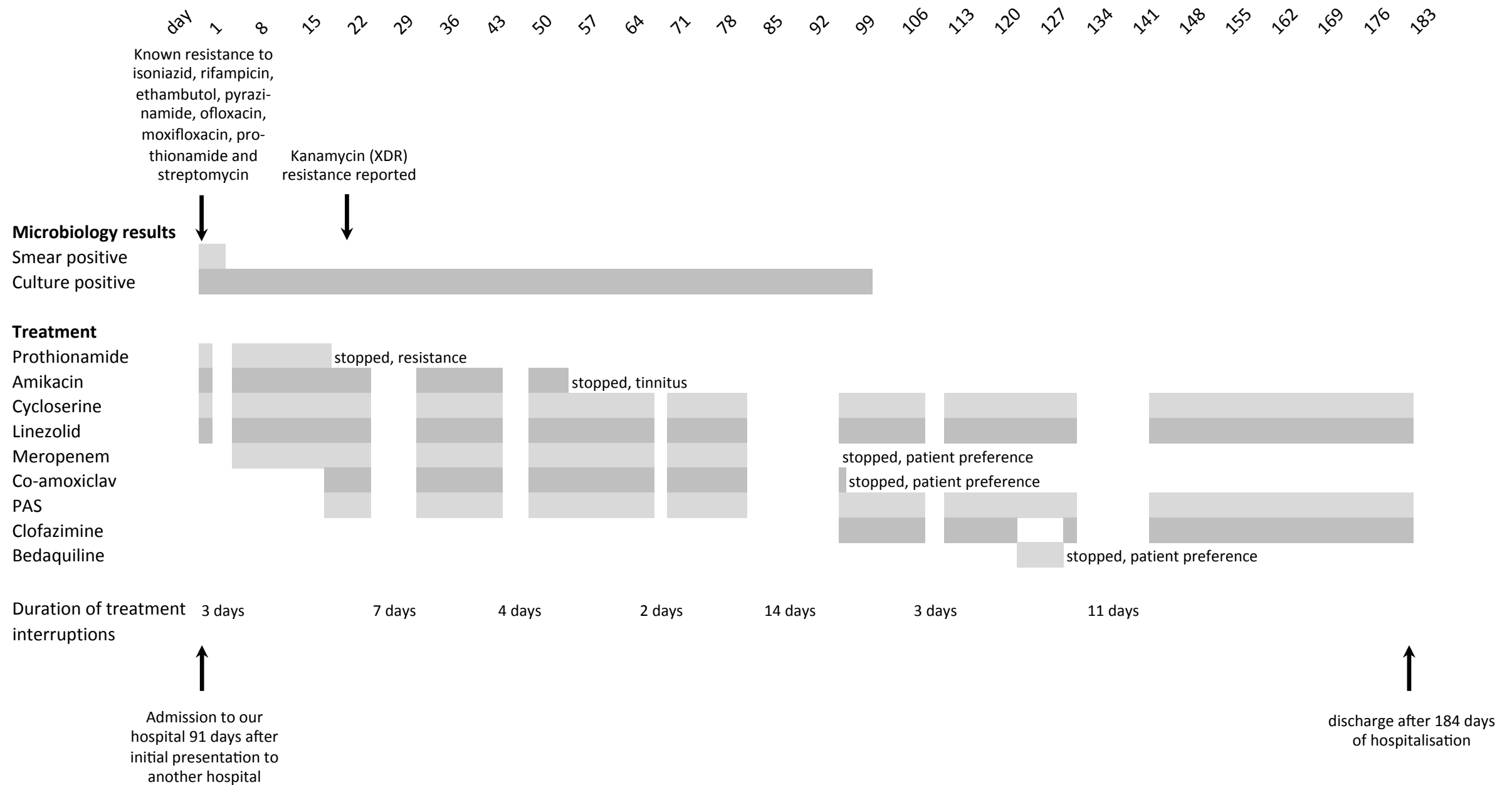
Andrew J. Carmichael, Ph.D.
Cambridge University Hospitals National Health Service
Foundation Trust
Cambridge, United Kingdom

Julian Parkhill, Ph.D.
Wellcome Trust Sanger Institute
Hinxton, United Kingdom

Geoffrey P. Smith, Ph.D.
Illumina (Cambridge)
Little Chesterford, United Kingdom

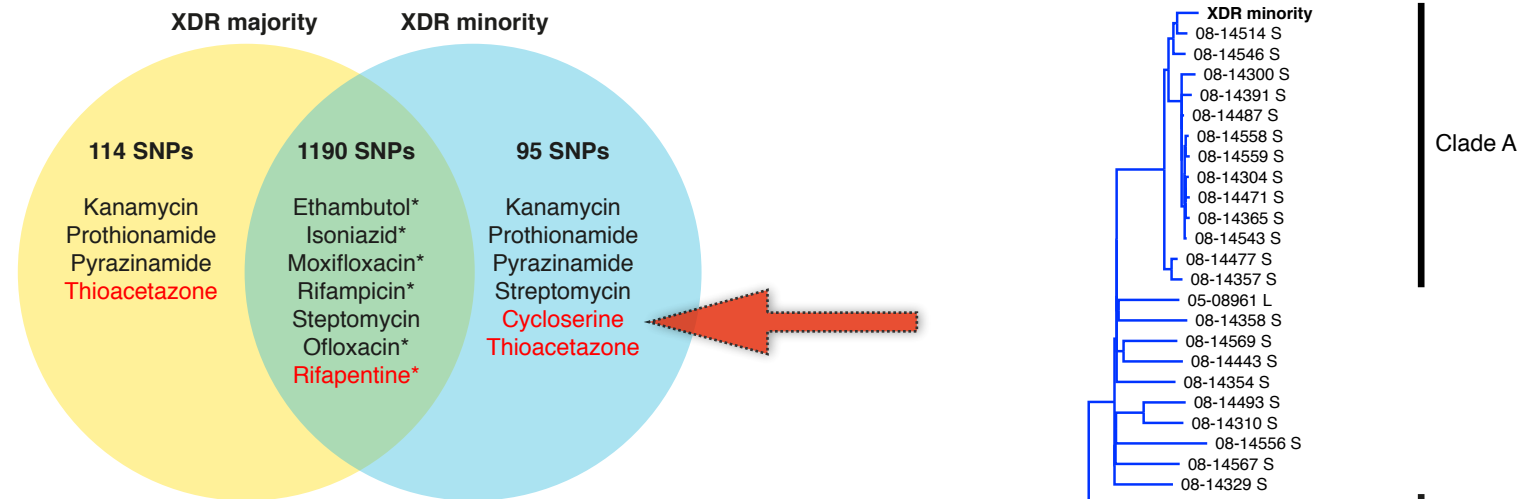
Sharon J. Peacock, Ph.D.
University of Cambridge
Cambridge, United Kingdom
sjp97@medschl.cam.ac.uk

Microbiology Results, Phenotypic Drug Susceptibility Results and Anti-tuberculosis Therapy During Hospitalisation



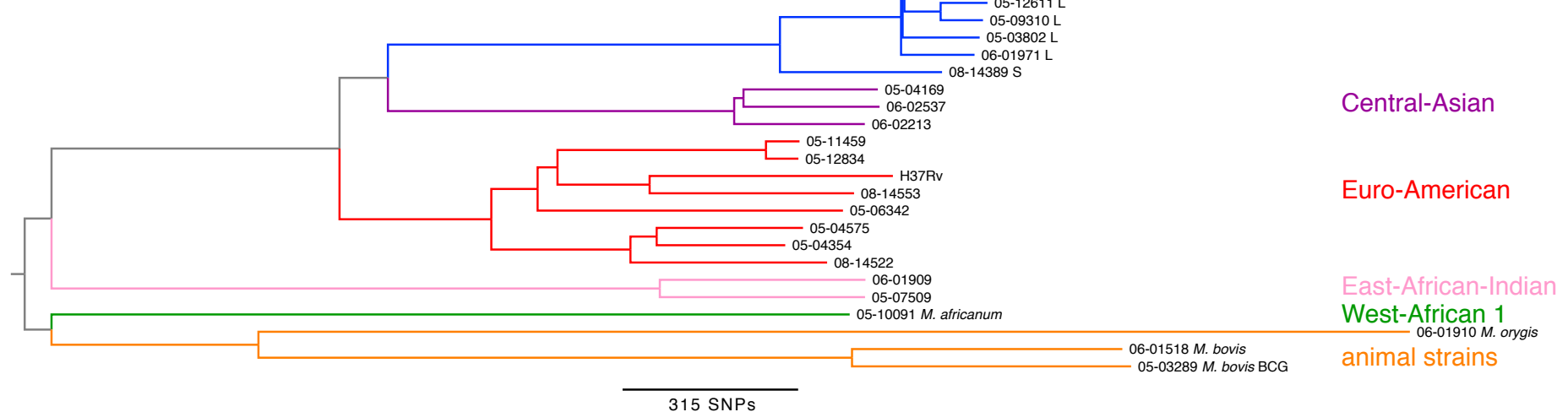
Phylogenetic Analysis and the Distribution of Drug-Resistance Mutations.

a Distribution of SNPs and Resistance Mutations



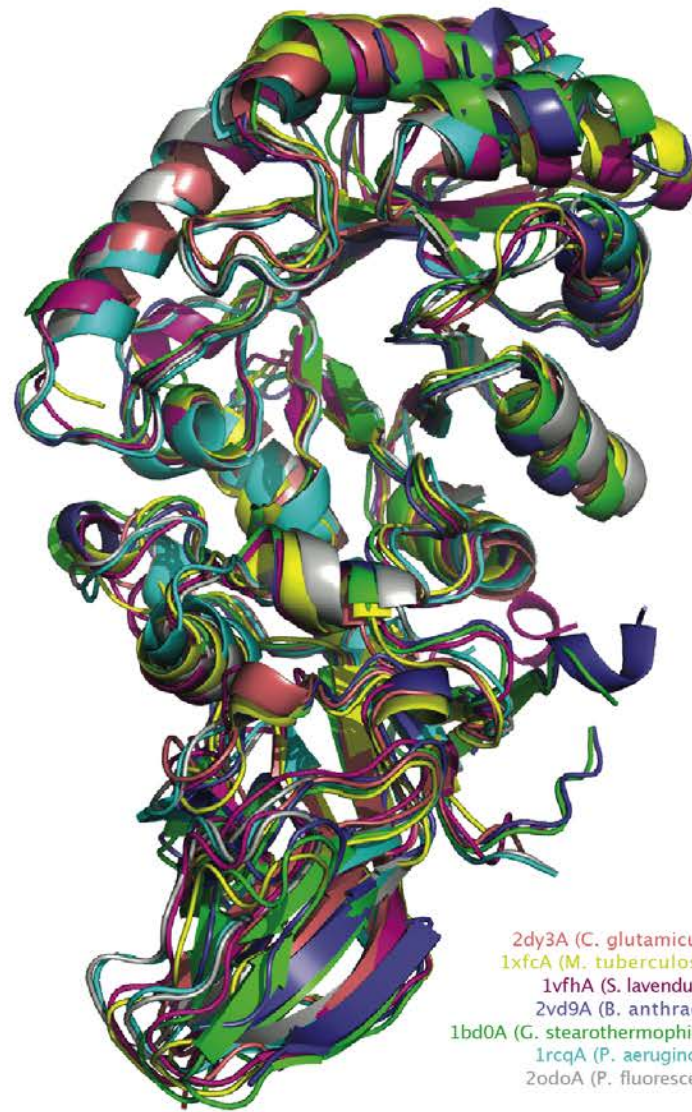
Beijing

b Phylogenetic Analysis



Alanine Racemase Conservation

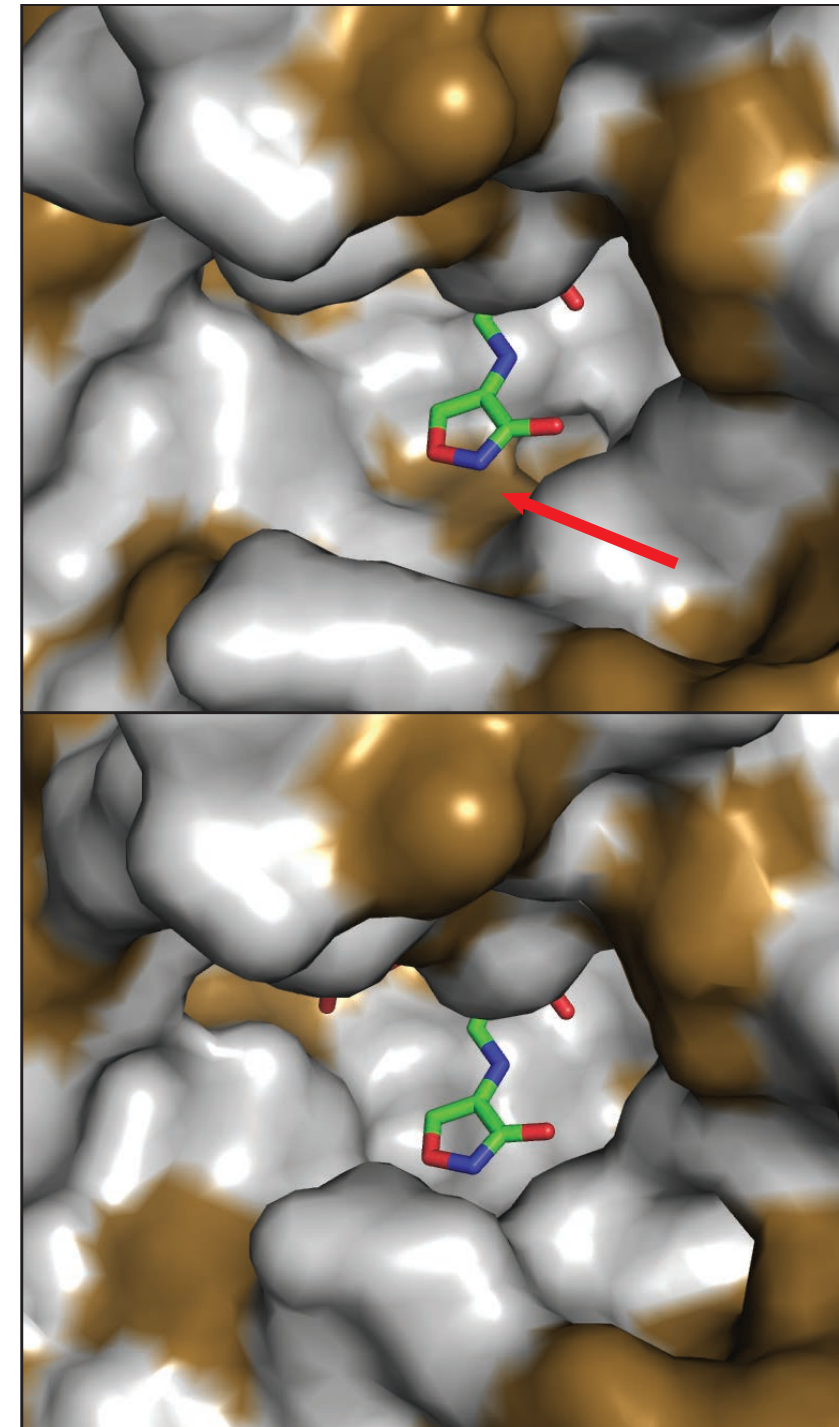
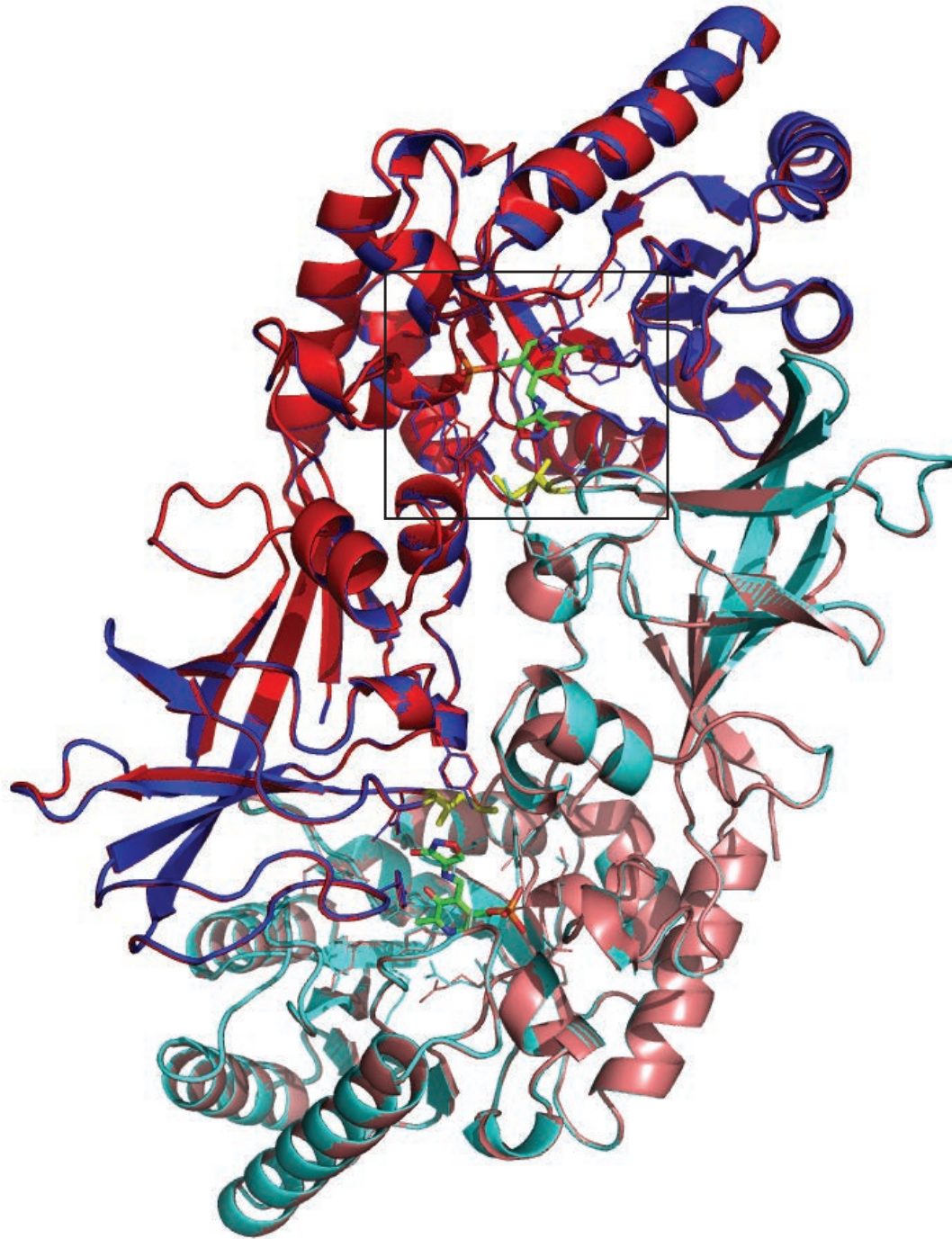
Cycloserine



2dy3A (C. glutamicum)	1	- - - MNL LTTK I D L D A I A H N T R V L K Q M A G P - A K L M A V V K A N A Y N H G V E K V A P V I A A H G D A F G V A T L A	63
1xfcA (M. tuberculosis)	11	- - - - - L A E A M V D L G A I E H N V R V L R E H A G H - A Q L M A V V K A D G Y G H G A T R V A Q T A L G A G A E A L G V A T V D	71
1vfhA (S. lavendulae)	3	- E T P T R V Y A E I D L D A V R A N V R A L R A R A P R - S A L M A V V K S N A Y G H G A V P C A R A A Q E A G A A W L G T A T P E	67
2vd9A (B. anthracis)	4	A P F Y R D T W V E V D L D A I Y N N V T H I - E F I P S D V E I F A V V K G N A Y G H D Y V P V A - I A L E A G A T R L A V A F L D	68
1bd0A (G. stearothermophilus)	2	N D F H R D T W A E V D L D A I Y D N V E N L R R L L P D D T H I M A V V K A N A Y G H G D V Q V A R T A L E A G A S R L A V A F L D	68
1rcqA (P. aeruginosa)	1	- - - M R P A R A L I D L Q A L R H N Y R L A R E A T G - - A R A L A V I K A D A Y G H G A V R C A E A L A A - E A D G F A V A C I E	61
2odoA (P. fluorescens)	1	- - - M R P A R A L I D L Q A L R H N Y Q L A R E V T G - - A K A L A V I - A D A Y G H G A V R C A Q A L E A - E A D G F A V A C I E	60
2dy3A (C. glutamicum)	64	E A M Q L R D I G I S Q E V L C W I - W T P E Q D F R A A I D R N I D L A V I J S P A H A K A L I E T D - - - A E H I R V S I K I D S G	126
1xfcA (M. tuberculosis)	72	E A L A L R A D G I T A P V L A W L - H P P G I D F G P A L L A D V Q V A V S S L R Q L D E L L H A V R R T G R T A T V T K V D T G	137
1vfhA (S. lavendulae)	68	E A L E L R A A G I Q G R I M C W L - W T P G G P W R E A I E T D I D V S V S G M W A L D E V R A A A A A G R T A R I Q L - A D T G	132
2vd9A (B. anthracis)	69	E A L V L R R A G I T A P I L V L G P S P - P R D I N V A A E N D V A L T V F Q - E W V D E A I - L W - D G S S T M - Y H I N F D S G	130
1bd0A (G. stearothermophilus)	69	E A L A L R E K G I E A P I L V L G A S R - P A D A A L A A Q R I A L T V F R S D W L E E A S A L Y - S G P P F I H F L K M D T G	133
1rcqA (P. aeruginosa)	62	E G L E L R E A G I R Q P I L L L E G F F E A S E L E L I V A H D F W C V V H S L W Q L D A I E Q A R L - - A R P L N V W L - M D S G	125
2odoA (P. fluorescens)	61	E A L E L R A A G I R A P I L L L E G F F E A D E L P L I V E H D F W C V V H S L W Q L D A I E Q A R L - - A K P I H V W L - L D S G	124
2dy3A (C. glutamicum)	127	L H R S G V D E Q E W E G V F S A L A A A P - - - H I E V T G M F T H L A - - - - - E T D R D I I A F R R A L A R K H G L E	182
1xfcA (M. tuberculosis)	138	L N R N G V G P A Q F P A M L T A L R Q A M A E D A V R L R G L M S H M - - - - P D D S I N D V Q A Q R T A F L A Q A R E Q G V R	199
1vfhA (S. lavendulae)	133	L G R N G C Q P A D W A E L V G A A V A A Q A E G T V Q V T G V W S H F A C A D E P G H P S I R L Q L D A F R D M L A Y A E K E G V D	199
2vd9A (B. anthracis)	131	M G R I G I R E R E L - - G F L S L - E G A - - P F L E E G V Y T H F A T A D E V E T S Y F D - Q Y N T - L E Q L S W L - - - E F G V	189
1bd0A (G. stearothermophilus)	134	M G R L G V K D E E T K R I V A L I E R H - - P H F V L E G L Y T H F A T A D E V N T D Y F S Y Q Y T R F L H M L E W L - - - P S R	195
1rcqA (P. aeruginosa)	126	M H R V G F F P E D F R A A H E R L R A S G - - - K V A K I V M M S H F S R A D E L D C P R T E E Q L A A F S A A S Q G L E - - - -	184
2odoA (P. fluorescens)	125	M H R V G L H P N D Y K A A Y Q R L Q A S A - - - N V A K I V L M S H F A R A D E L D C T R S V E Q V A V F L G A R A D L T - - - -	183
2dy3A (C. glutamicum)	183	C P V N H V C N S P A F L T R S D L H M E M V R P G L A F Y G L E P V A G L - - - - E H G L K P A M T W E A K V S V V K Q I E - - - -	241
1xfcA (M. tuberculosis)	200	F E V A H L S N S S A T M A R P D L T F D L V R P G I A V Y G L S P V P A L G D - - - M G L V P A M T V K C A V A L V K S I R A G E G	263
1vfhA (S. lavendulae)	200	P E V R H I A N S P A T L T L P E T H F D L V R T G L A V Y G V S P S P E L G T P A Q L G L R P A M T L R A S L A L V K T V P A G H G	266
2vd9A (B. anthracis)	190	D P F V H T A N S A A T L R F Q G I T F A N V R I G I A M Y G L S P S V E I R P L P F - L E P A L S L H T - V A H I K Q V I - G D G	253
1bd0A (G. stearothermophilus)	196	P P L V H C A N S A S L R F P D R T F N M V R F G I A M Y G L A P S P G I K P L L P Y P L K E A F S L H S R L V H V K K L Q P G E K	262
1rcqA (P. aeruginosa)	185	- G E I S L R N S P A V L G W P K V P S D W V R P G I L L Y G A T P F E R A H P L A D - R L R P V M T L E S K V I S V R D L P A G E P	249
2odoA (P. fluorescens)	184	- A E I S L R N S P A V L G W P R V P S D W V R P G L M L Y G S S P F D E P Q A T A S - R L Q P V M T L E S K V I C V R E L P A G E P	248
2dy3A (C. glutamicum)	242	- - - - - R G F V A V V P A G Y A D G M P R H A Q G K F S V T I D G L D Y P Q V G R V C M D Q F V I S L G D N P H G V E A	297
1xfcA (M. tuberculosis)	264	V S Y G H T W I A P R D T N L A L L P I G Y A D G V F R S L G G R L E V L I N G R R C P G V G R I C M D Q F M V D L G P G P L D V A E	330
1vfhA (S. lavendulae)	267	V S Y G H H Y V T E S E T H A L V P A G Y A D G I P R N A S G R G P V L V A G K I R R A A G I A M D Q F V V D L G E D - - L A E A	331
2vd9A (B. anthracis)	254	I S Y N V T Y R T - T E E W I A T V A I G Y A D G W L R R L Q G - F E V L V N G - R V P I G R V T M D Q F M I H L P C - - - E V P L	314
1bd0A (G. stearothermophilus)	263	V S Y G A T Y T A Q T E E W I G T I P I G Y A D G W L R R L Q H - F H V L V D G Q K A P I V G R I C M D Q C M I R L P G - - - P L P V	325
1rcqA (P. aeruginosa)	250	V G Y G A R Y S T E R R Q R I G V V A M G Y A D G Y P R H A A D G T L V F I D G K P G R L V G R V S M D M L T V D L T D H P Q - A G L	315
2odoA (P. fluorescens)	249	V G Y G A R F I T P K P M R I G V V A M G Y A D G Y P R Q A P T G T P V F V D G V R S Q L L G R V S M D M L C I D L T D V P Q - A G L	314
2dy3A (C. glutamicum)	298	G A K A V I F G E - - - N G H D A T D F A E R L D T I N Y E V V C R P T G R T V R A Y V - - - - -	338
1xfcA (M. tuberculosis)	331	G D E A I L F G P G I R G E P T A Q D W A D L V G T I H Y E V V T S P R G R I T R Y R E A - - - - -	376
1vfhA (S. lavendulae)	332	G D E A V I L G D A E R G E P T A E D W A Q A A H T I A Y E I V T R I G G R V P R V Y L G G L E H H H H - - - - -	383
2vd9A (B. anthracis)	315	G T - V T L I G R Q G D E Y I S A T E V A E Y S G T I N Y E I I T T I S F R V P R I F I R N G - V V E V I N Y L N D I	371
1bd0A (G. stearothermophilus)	326	G T K V T L I G R Q G D E V I S I D D V A R H L E T I N Y E V P C T I S Y R V P I F F R H K R I M E V R N A I G - -	382
1rcqA (P. aeruginosa)	316	G S R V E L W G P N - - - - V P V G A L A A Q F G S I P Y Q L L C N L K - R V P R V Y S G A - - - - -	356
2odoA (P. fluorescens)	315	G S T V E L W G K N - - - - I L A S E V A T A A D T I P Y Q I F C N L K - R V P R L Y S G - - - - -	354

Alanine Racemase M343T Modeling

Cycloserine



Open positions in the lab soon
<http://marciuslab.org>

Acknowledgments



<http://marciuslab.org>
<http://integrativemodeling.org>
<http://cnag.cat> · <http://crg.cat>

