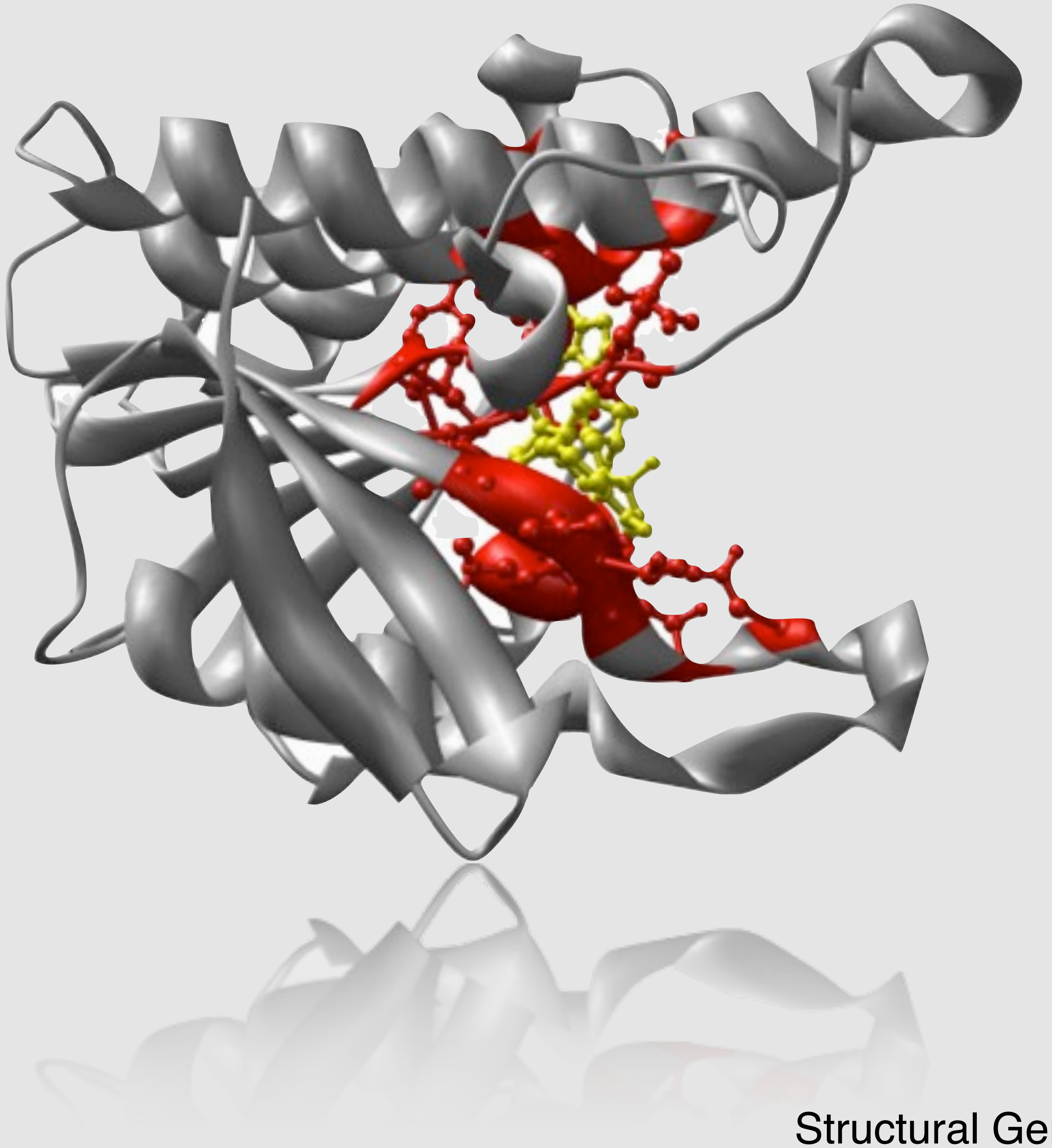
Comparative docking for predicting molecular targets of known drugs

A “kernel” for the Tropical Disease Initiative



Marc A. Marti-Renom

http://sgu.bioinfo.cipf.es

Structural Genomics Unit
Bioinformatics & Genomics Department
Centro de Investigación Príncipe Felipe (CIPF), Valencia, Spain

protein prediction .vs. protein determination

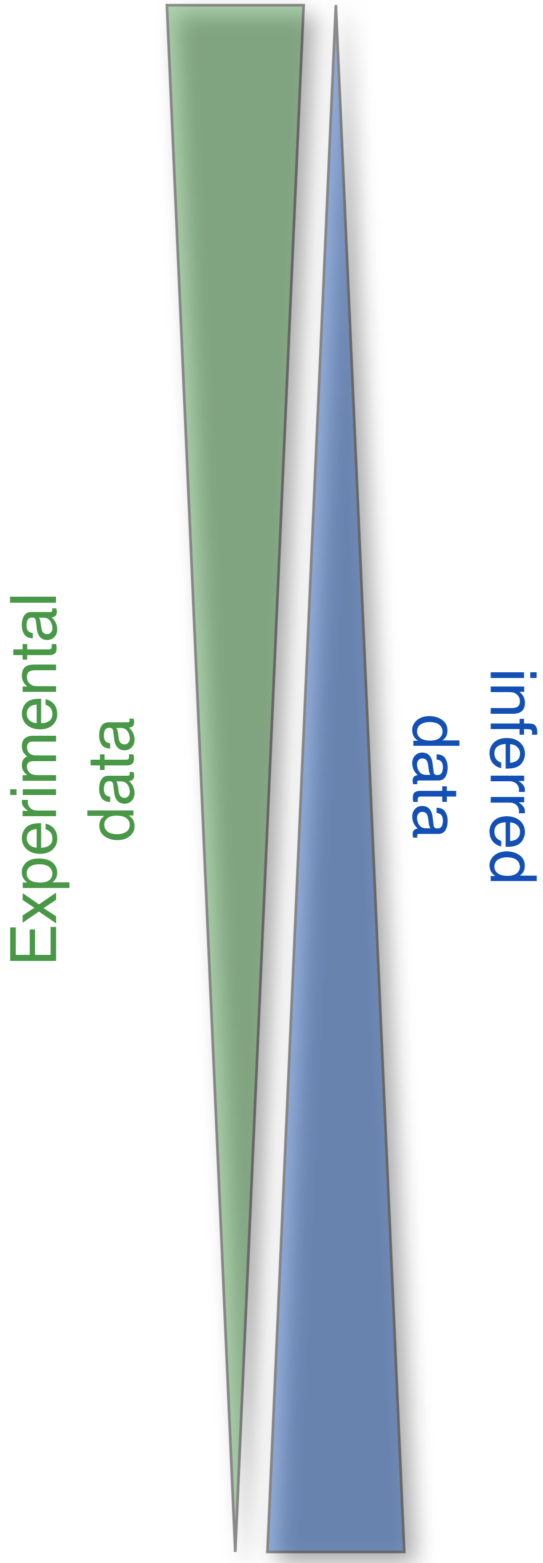
X-Ray

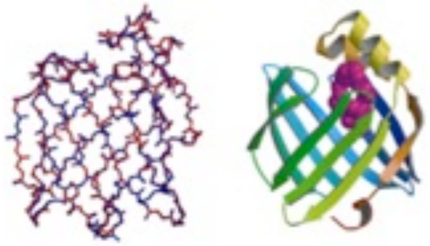
NMR

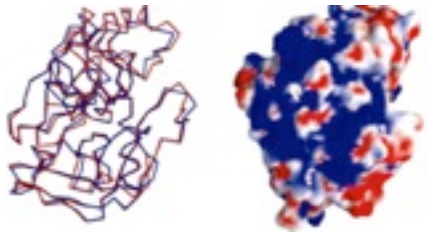
Comparative Modeling

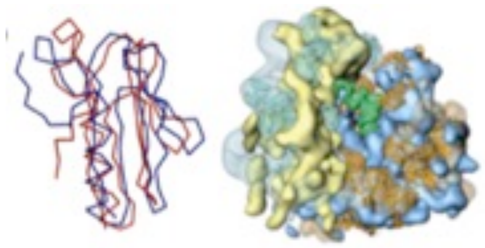
Threading

Ab-initio

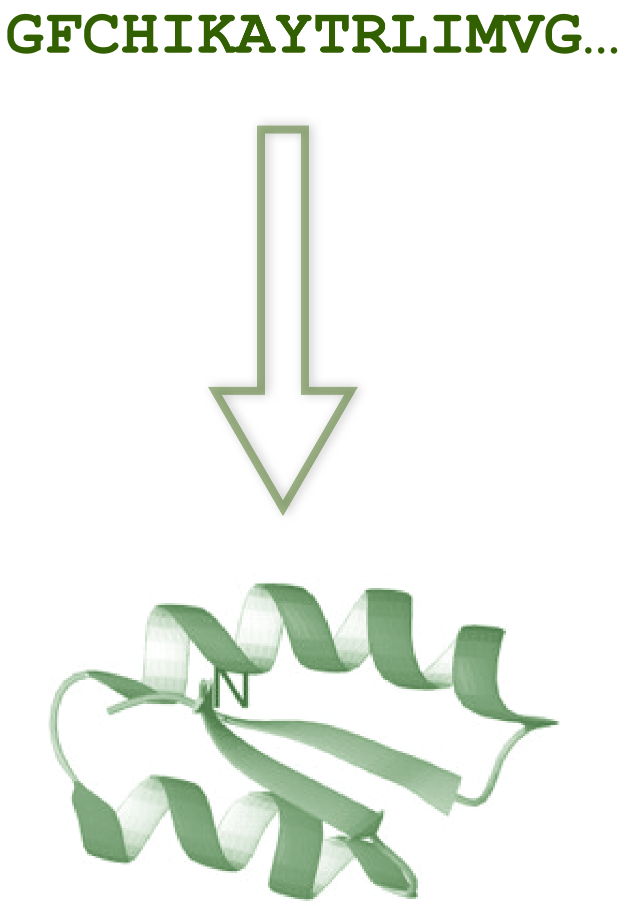


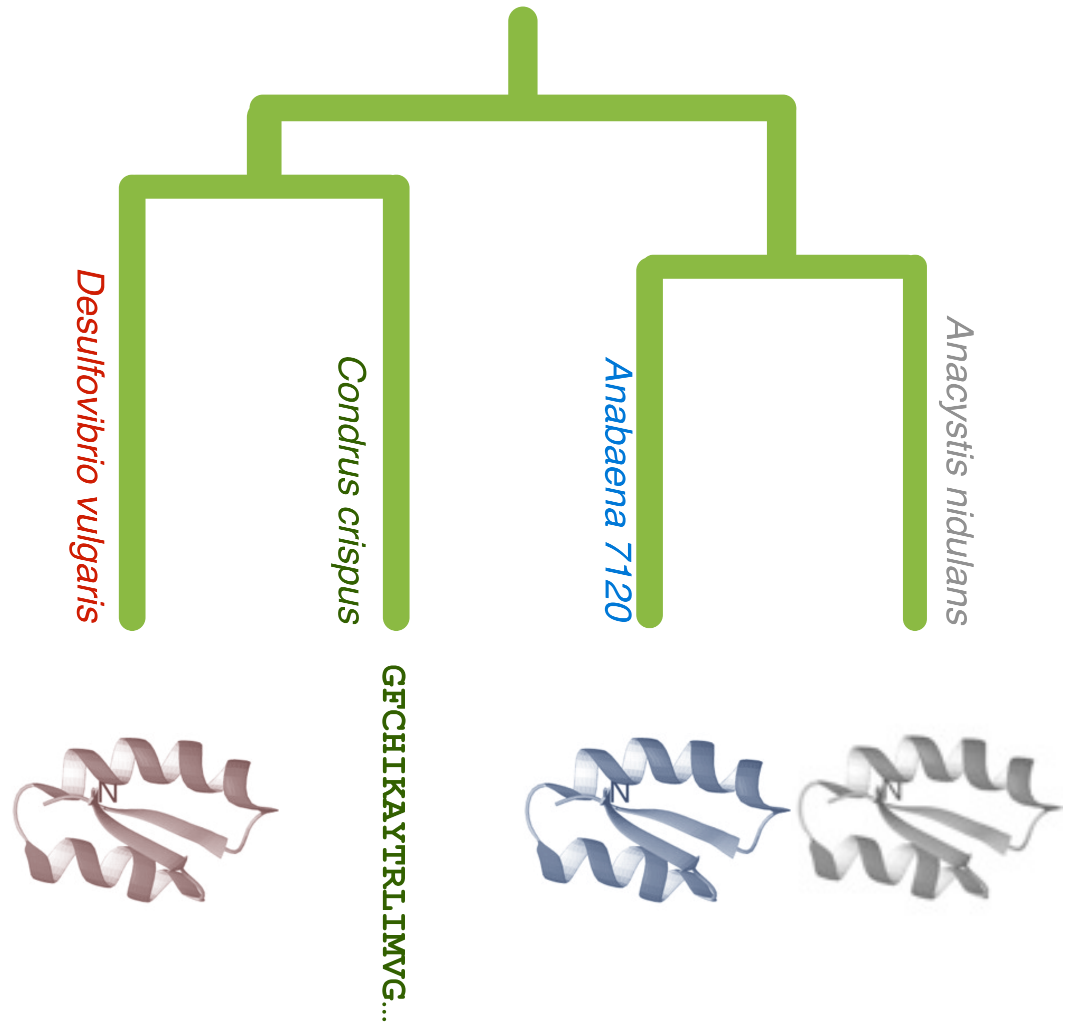






Principles of protein structure





Folding (physics)

Ab initio prediction

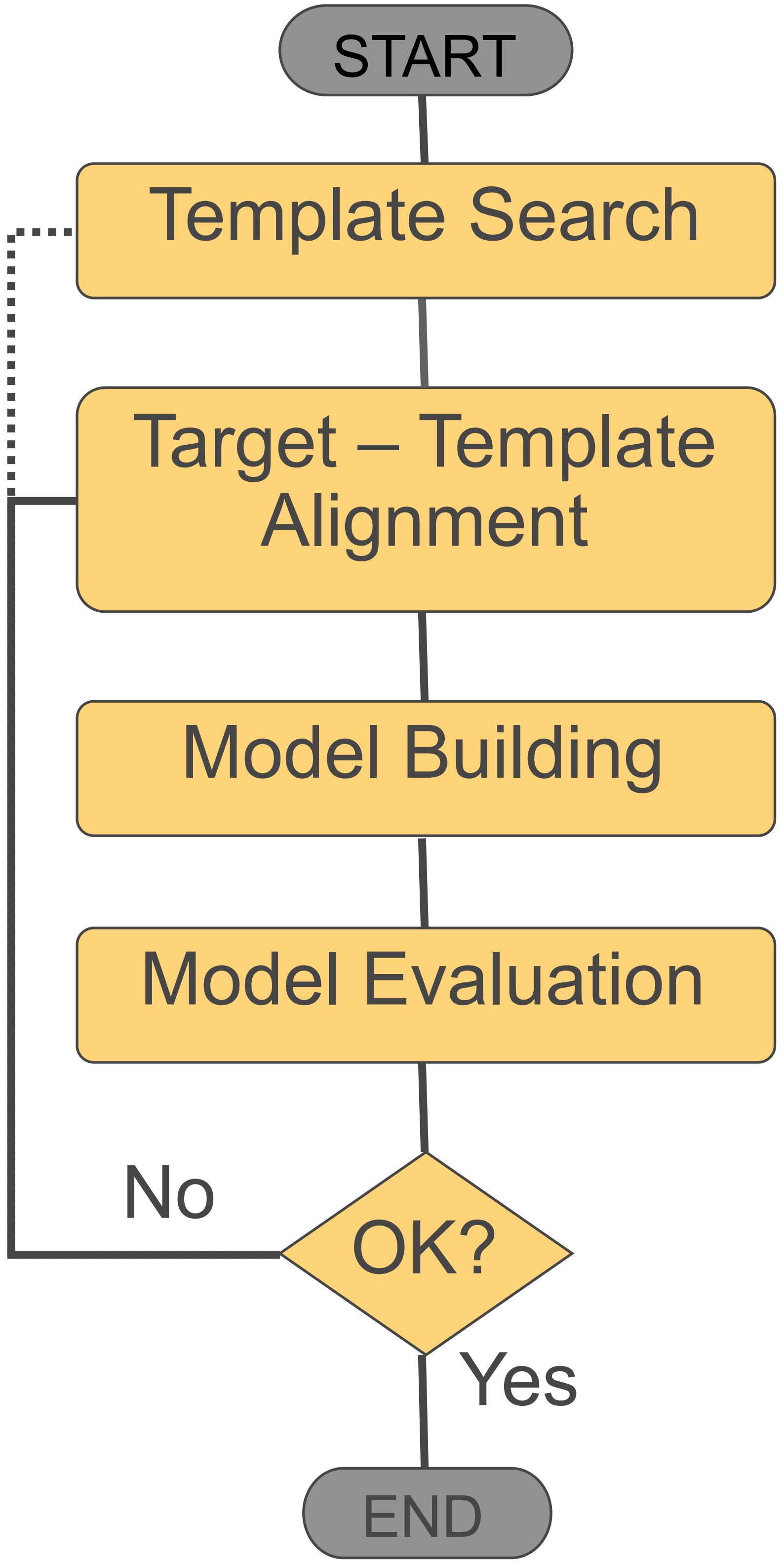
Evolution (rules)

Threading

Comparative Modeling

D. Baker & A. Sali. Science 294, 93, 2001.

Steps in Comparative Protein Structure Modeling


START

Template Search

Target – Template Alignment

Model Building

Model Evaluation

OK?

No

Yes

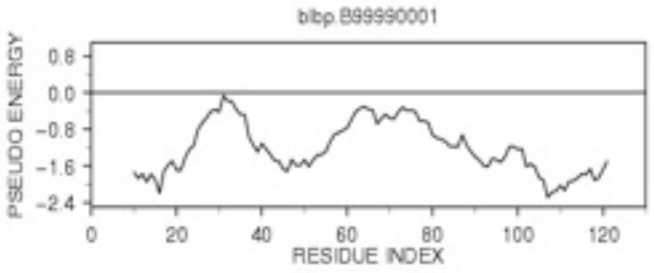
END

TARGET

TEMPLATE

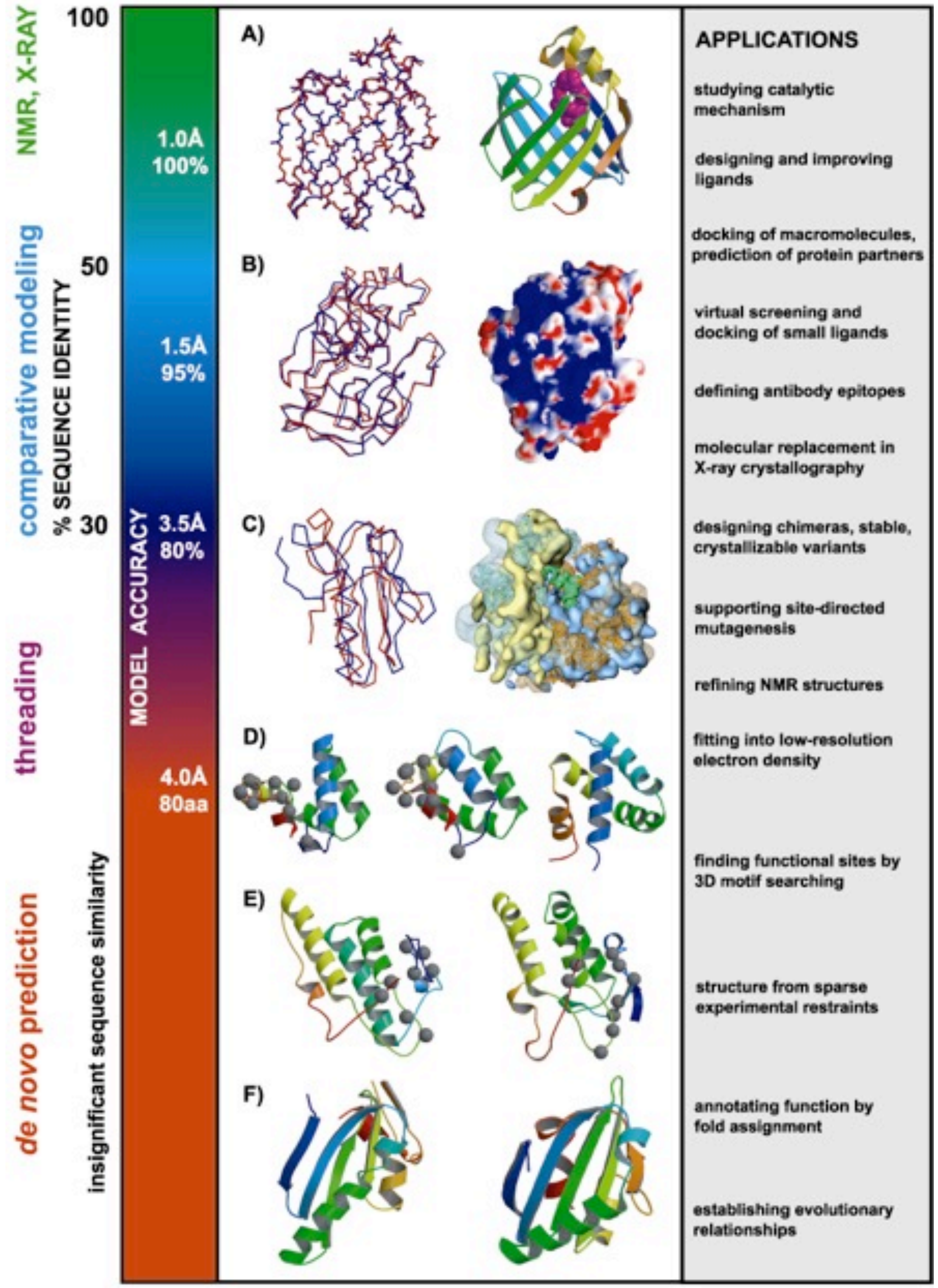
ASILPKRLFGNCEQTSDEGL
KIERTPLVPHISAQNVCLKID
DVPERLIPERASFQWMNDK

```
ASILPKRLFGNCEQTSDEGLKIERTPLVPHISAQNVCLKIDDVPERLIPE
MSVIPKRLYGNCEQTSEEAIRIEDSPIV---TADLVCLKIDEIPERLVGE
```

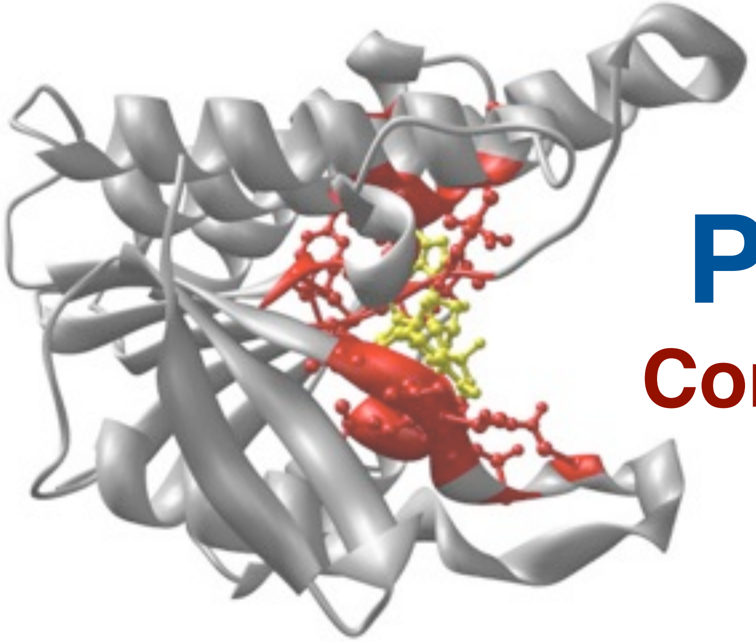


A. Šali, Curr. Opin. Biotech. 6, 437, 1995.
R. Sánchez & A. Šali, Curr. Opin. Str. Biol. 7, 206, 1997.
M.A. Marti-Renom et al. Ann. Rev. Biophys. Biomolec. Struct., 29, 291, 2000.

Utility of protein structure models, despite errors

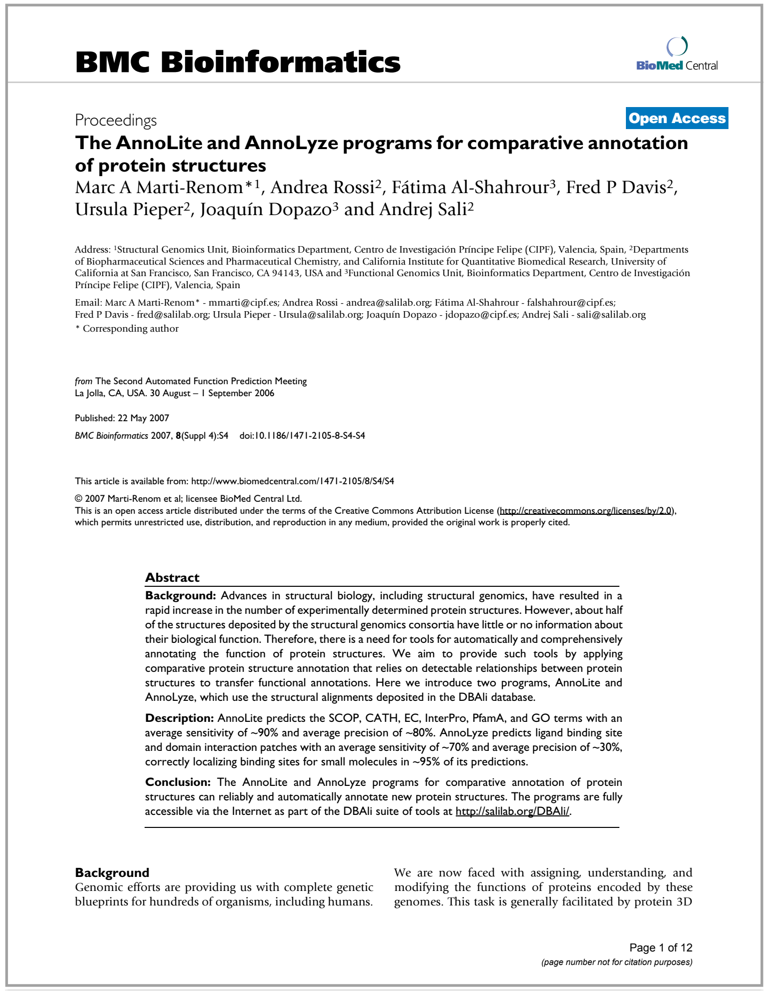

NMR, X-RAY
comparative modeling
threading
de novo prediction
% SEQUENCE IDENTITY
100
50
30
insignificant sequence similarity
MODEL ACCURACY
1.0Å
100%
1.5Å
95%
3.5Å
80%
4.0Å
80aa
A)
B)
C)
D)
E)
F)
APPLICATIONS
studying catalytic mechanism
designing and improving ligands
docking of macromolecules, prediction of protein partners
virtual screening and docking of small ligands
defining antibody epitopes
molecular replacement in X-ray crystallography
designing chimeras, stable, crystallizable variants
supporting site-directed mutagenesis
refining NMR structures
fitting into low-resolution electron density
finding functional sites by 3D motif searching
structure from sparse experimental restraints
annotating function by fold assignment
establishing evolutionary relationships

D. Baker & A. Sali. Science 294, 93, 2001.



Protein function from structure

Comparative binding site prediction by AnnoLyze.



BMC Bioinformatics

BioMed Central

Proceedings

Open Access

The AnnoLite and AnnoLyze programs for comparative annotation of protein structures

Marc A Marti-Renom*[1], Andrea Rossi[2], Fátima Al-Shahrour[3], Fred P Davis[2], Ursula Pieper[2], Joaquín Dopazo[3] and Andrej Sali[2]

Address: [1]Structural Genomics Unit, Bioinformatics Department, Centro de Investigación Príncipe Felipe (CIPF), Valencia, Spain, [2]Departments of Biopharmaceutical Sciences and Pharmaceutical Chemistry, and California Institute for Quantitative Biomedical Research, University of California at San Francisco, San Francisco, CA 94143, USA and [3]Functional Genomics Unit, Bioinformatics Department, Centro de Investigación Príncipe Felipe (CIPF), Valencia, Spain

Email: Marc A Marti-Renom* - mmarti@cipf.es; Andrea Rossi - andrea@salilab.org; Fátima Al-Shahrour - falshahrour@cipf.es; Fred P Davis - fred@salilab.org; Ursula Pieper - Ursula@salilab.org; Joaquín Dopazo - jdopazo@cipf.es; Andrej Sali - sali@salilab.org

* Corresponding author

from The Second Automated Function Prediction Meeting
La Jolla, CA, USA. 30 August – 1 September 2006

Published: 22 May 2007

BMC Bioinformatics 2007, **8**(Suppl 4):S4   doi:10.1186/1471-2105-8-S4-S4

This article is available from: http://www.biomedcentral.com/1471-2105/8/S4/S4

© 2007 Marti-Renom et al; licensee BioMed Central Ltd.
This is an open access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/2.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Abstract

Background: Advances in structural biology, including structural genomics, have resulted in a rapid increase in the number of experimentally determined protein structures. However, about half of the structures deposited by the structural genomics consortia have little or no information about their biological function. Therefore, there is a need for tools for automatically and comprehensively annotating the function of protein structures. We aim to provide such tools by applying comparative protein structure annotation that relies on detectable relationships between protein structures to transfer functional annotations. Here we introduce two programs, AnnoLite and AnnoLyze, which use the structural alignments deposited in the DBAli database.

Description: AnnoLite predicts the SCOP, CATH, EC, InterPro, PfamA, and GO terms with an average sensitivity of ~90% and average precision of ~80%. AnnoLyze predicts ligand binding site and domain interaction patches with an average sensitivity of ~70% and average precision of ~30%, correctly localizing binding sites for small molecules in ~95% of its predictions.

Conclusion: The AnnoLite and AnnoLyze programs for comparative annotation of protein structures can reliably and automatically annotate new protein structures. The programs are fully accessible via the Internet as part of the DBAli suite of tools at http://salilab.org/DBAli/.

Background

Genomic efforts are providing us with complete genetic blueprints for hundreds of organisms, including humans. We are now faced with assigning, understanding, and modifying the functions of proteins encoded by these genomes. This task is generally facilitated by protein 3D

Page 1 of 12
(page number not for citation purposes)

Marti-Renom et al. BMC Bioinformatics (2007)

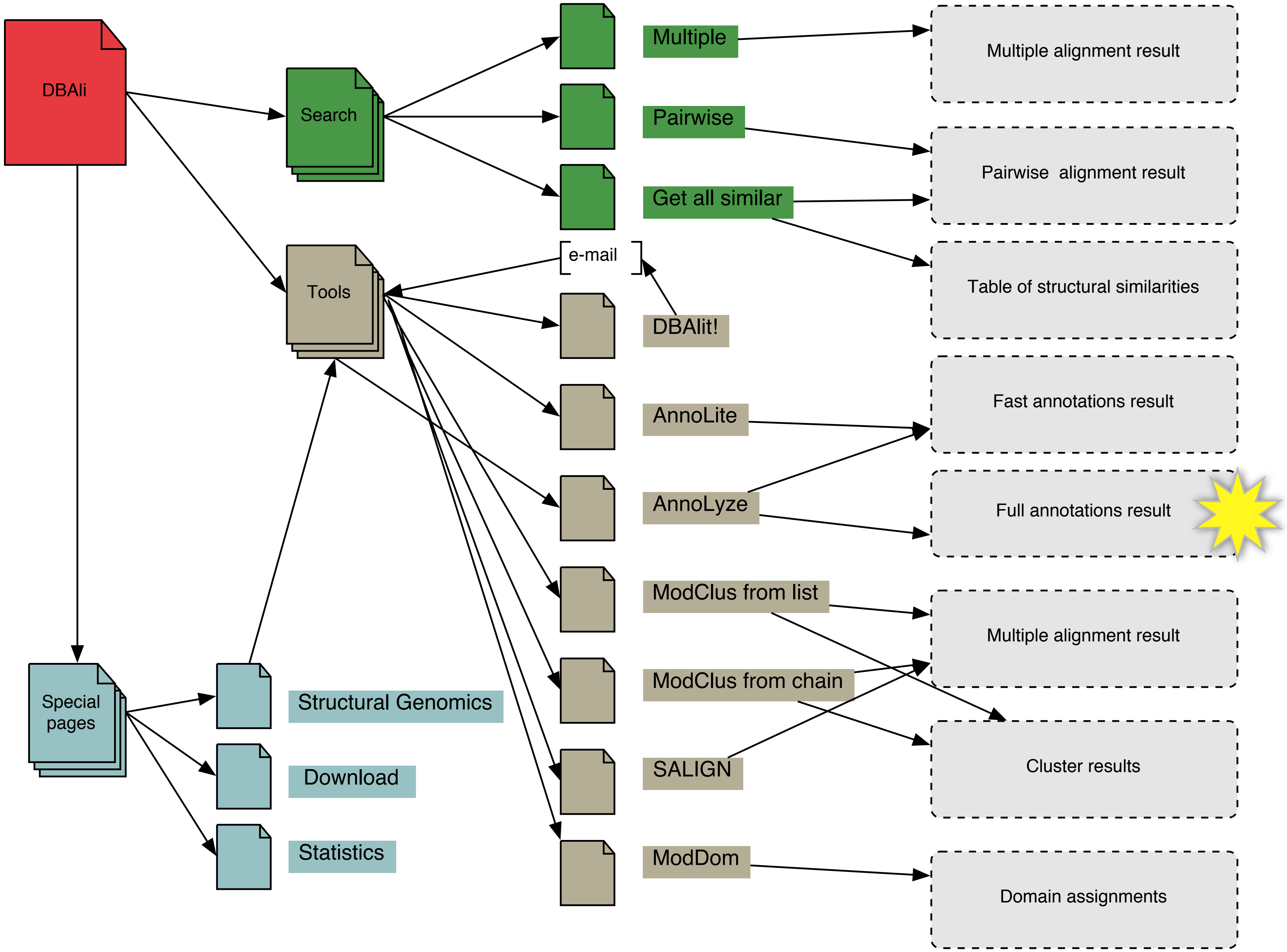
For ~20% protein structures function is *unknown*

	Structural Genomics*	Traditional methods
Annotated**	654	28,342
Not Annotated	506 (43.6%)	6,815 (19,4%)
Total deposited	1,160	35,157

** annotated as STRUCTURAL GENOMICS in the header of the PDB file*
***annotated with either CATH, SCOP, Pfam or GO terms in the MSD database*
36,317 protein structures, as of August 8th, 2006

DBAli$_{v2.0}$ database

http://www.dbali.org



Marti-Renom et al. Nucleic Acids Research (2007)

Nucleic Acids Research, 2007, Vol. 35, Web Server issue ***W393–W397***
doi:10.1093/nar/gkm236

DBAli tools: mining the protein structure space

Marc A. Marti-Renom[1,*], Ursula Pieper[2], M. S. Madhusudhan[2], Andrea Rossi[2], Narayanan Eswar[2], Fred P. Davis[2], Fátima Al-Shahrour[3], Joaquín Dopazo[3] and Andrej Sali[2]

[1]Structural Genomics Unit, [2]Departments of Biopharmaceutical Sciences and Pharmaceutical Chemistry, and California Institute for Quantitative Biomedical Research, University of California at San Francisco, San Francisco, CA 94158-2330, USA and [3]Functional Genomics Unit, Bioinformatics Department, Centro de Investigación Príncipe Felipe (CIPF), Valencia, Spain

Received January 30, 2007; Accepted March 31, 2007

ABSTRACT

The DBAli tools use a comprehensive set of structural alignments in the DBAli database to leverage the structural information deposited in the Protein Data Bank (PDB). These tools include (i) the DBAlit program that allows users to input the 3D coordinates of a protein structure for comparison by MAMMOTH against all chains in the PDB; (ii) the AnnoLite and AnnoLyze programs that annotate a target structure based on its stored relationships to other structures; (iii) the ModClus program that clusters structures by sequence and structure similarities; (iv) the ModDom program that identifies domains as recurrent structural fragments and (v) an implementation of the COMPARER method in the SALIGN command in MODELLER that creates a multiple structure alignment for a set of related protein structures. Thus, the DBAli tools, which are freely accessible *via* the World Wide Web at http://salilab.org/DBAli/, allow users to mine the protein structure space by establishing relationships between protein structures and their functions.

INTRODUCTION

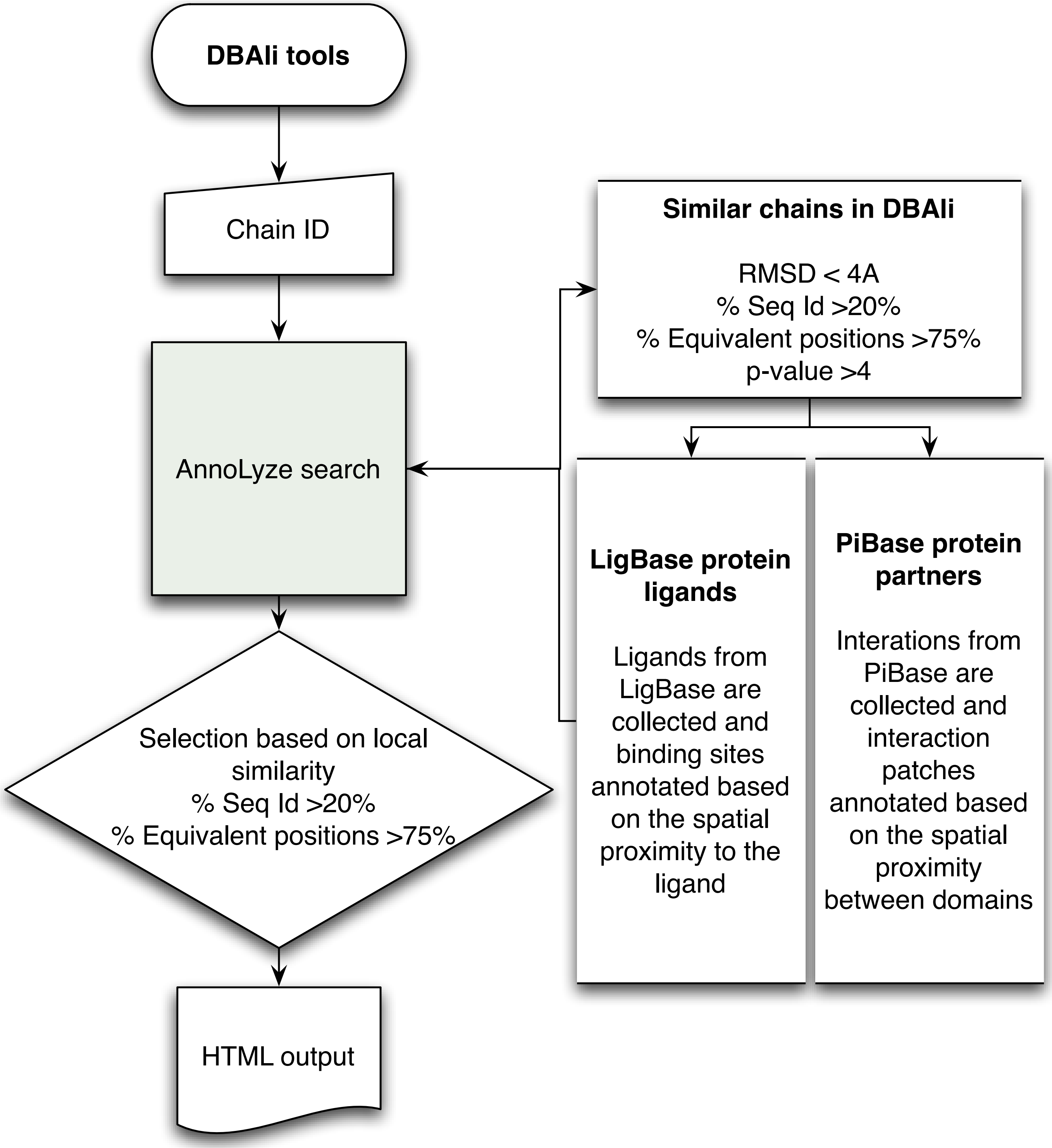
The number of known protein structures deposited in the Protein Data Bank (PDB) has grown exponentially over the years (1). This trend is expected to continue, partly due to the structural genomics efforts (2,3). Currently, there are ~41 000 protein structures deposited in the PDB, containing ~88 000 protein chains. These protein structures constitute a structural space that can be mined to facilitate the understanding, assignment and modification of protein function. Previously developed databases for the classification of protein structure domains, such as SCOP [http://scop.mrc-lmb.cam.ac.uk/scop/ (4)] or CATH [http://www.cathdb.info (5)], and servers for functional annotation of protein structures, such as ProFunc [http://www.ebi.ac.uk/thornton-srv/databases/ProFunc/ (6,7)], ProKnow [http://www.doe-mbi.ucla.edu/Services/ProKnow (8)] and Phunctioner [http://www.sbg.bio.ic.ac.uk (9)], provide an effective way of describing and annotating the protein structure space. However, none of these servers combine a comprehensive database of protein structural alignments with tools for automatically annotating protein structures.

Here, we describe five tools that aid in the analysis of the data stored in DBAli, our comprehensive relational database of pairwise and multiple structural alignments (10). These tools include (i) the DBAlit program that allows users to input their structure for comparison by MAMMOTH (11) against all chains in the PDB; (ii) the AnnoLite and AnnoLyze programs that annotate a target structure based on its stored relationships to other structures; (iii) the ModClus program that clusters structures by sequence and structure similarities; (iv) the ModDom program that identifies recurrent fragments, including domains, from structure; and (v) an implementation of the COMPARER method (12) in the SALIGN command in MODELLER that creates a multiple structure alignment for a set of related protein structures. The DBAli tools allow users to establish relationships between protein structures and their fragments in a flexible and dynamic manner.

The DBAli database is briefly introduced first. Next, we describe each of the five tools that make use of the structural alignments deposited in DBAli. Finally, we discuss the use of the DBAli tools to analyze a structure determined by the New York Structural Genomics Research Consortium (NYSGXRC).

*To whom correspondence should be addressed. Tel: +34 96 3289680; Fax: +34 96 3289701; Email: mmarti@cipf.es

Method



DBAli tools

Chain ID

AnnoLyze search

Similar chains in DBAli

RMSD < 4A
% Seq Id >20%
% Equivalent positions >75%
p-value >4

LigBase protein ligands

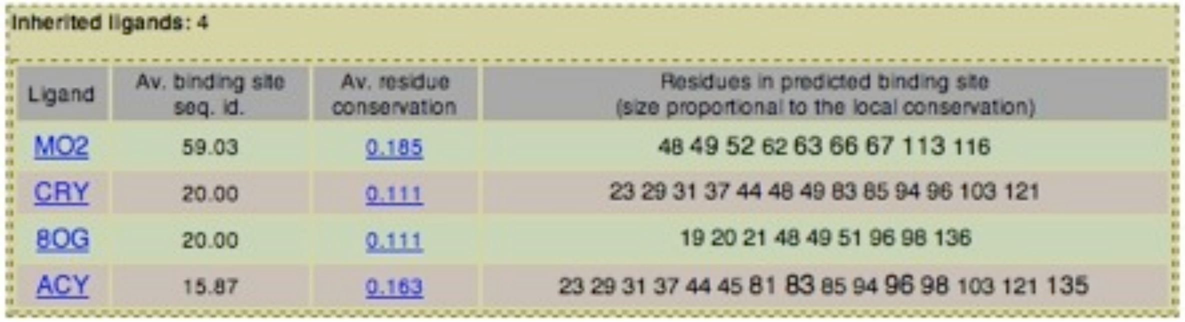
Ligands from LigBase are collected and binding sites annotated based on the spatial proximity to the ligand

PiBase protein partners

Interations from PiBase are collected and interaction patches annotated based on the spatial proximity between domains

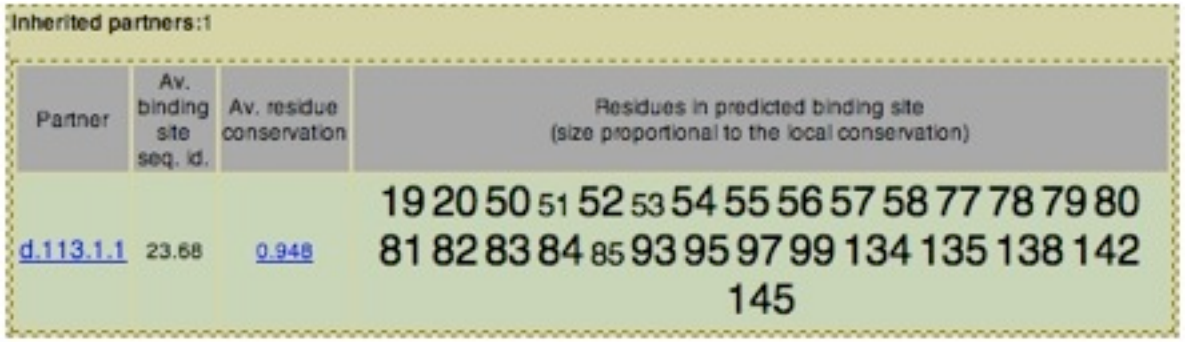
Selection based on local similarity
% Seq Id >20%
% Equivalent positions >75%

HTML output



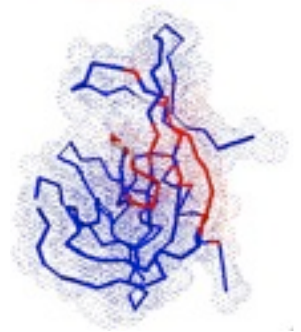
Inherited ligands: 4

Ligand	Av. binding site seq. id.	Av. residue conservation	Residues in predicted binding site (size proportional to the local conservation)
MO2	59.03	0.185	48 49 52 62 63 66 67 113 116
CRY	20.00	0.111	23 29 31 37 44 48 49 83 85 94 96 103 121
8OG	20.00	0.111	19 20 21 48 49 51 96 98 136
ACY	15.87	0.163	23 29 31 37 44 45 81 83 85 94 96 98 103 121 135



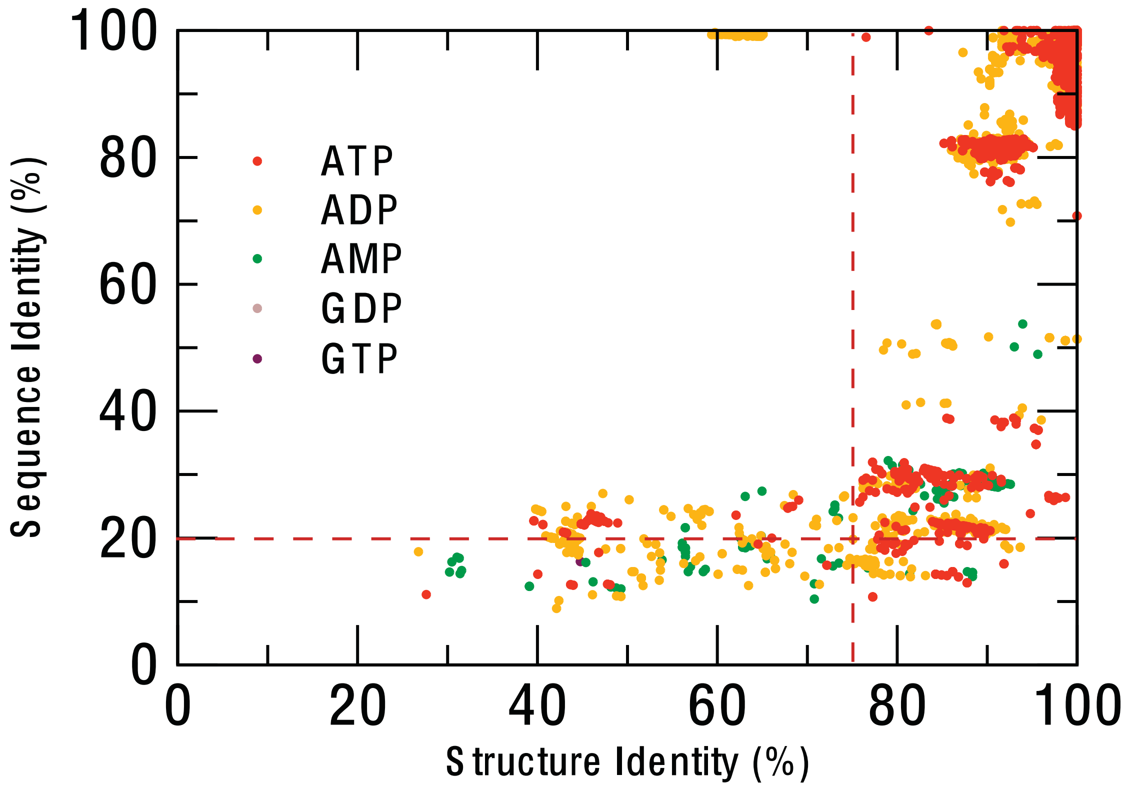
Inherited partners:1

Partner	Av. binding site seq. id.	Av. residue conservation	Residues in predicted binding site (size proportional to the local conservation)
d.113.1.1	23.68	0.948	19 20 50 51 52 53 54 55 56 57 58 77 78 79 80 81 82 83 84 85 93 95 97 99 134 135 138 142 145

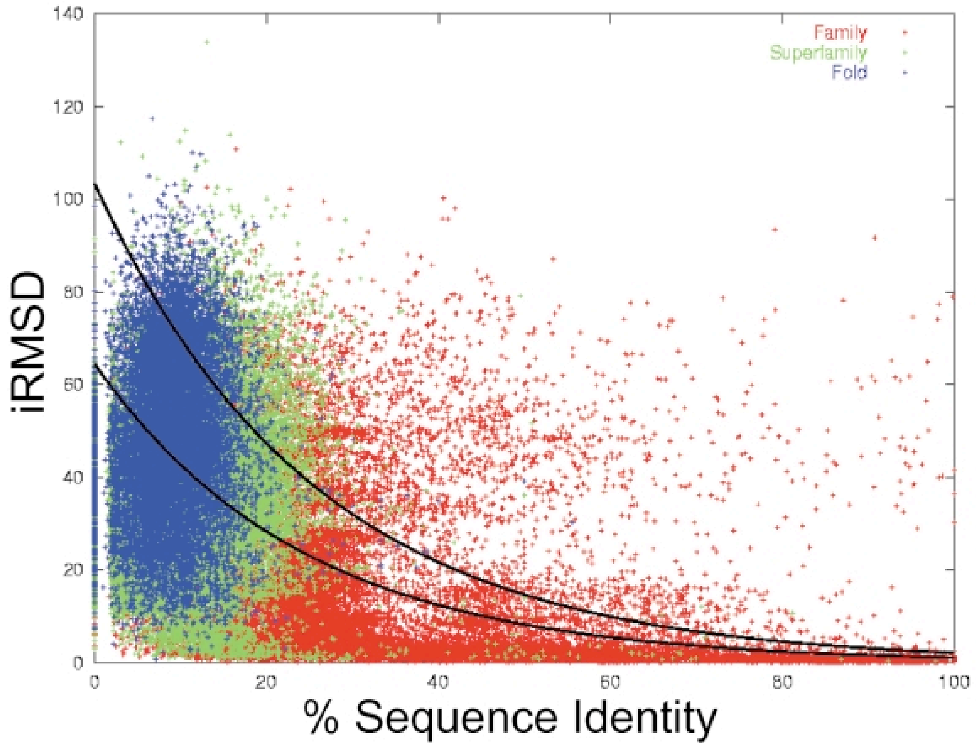


Scoring function

Ligands



Partners



Aloy *et al.* (2003) J.Mol.Biol. 332(5):989-98.

Benchmark

	Number of chains
Initial set*	78,167
LigBase**	30,126
Non-redundant set***	4,948 (8,846 ligands)

*all PDB chains larger than 30 aminoacids in length (8th of August, 2006)

**annotated with at least one ligand in the LigBase database

***not two chains can be structurally aligned within 3A, superimposing more than 75% of their Cα atoms, result in a sequence alignment with more than 30% identity, and have a length difference inferior to 50aa

Sensitivity .vs. Precision

	Optimal cut-off	**Sensitivity (%)** Recall or TPR	Precision (%)
Ligands	30%	71.9	13.7

$$\text{Sensitivity} = \frac{TP}{TP + FN} \qquad \text{Precision} = \frac{TP}{TP + FP}$$

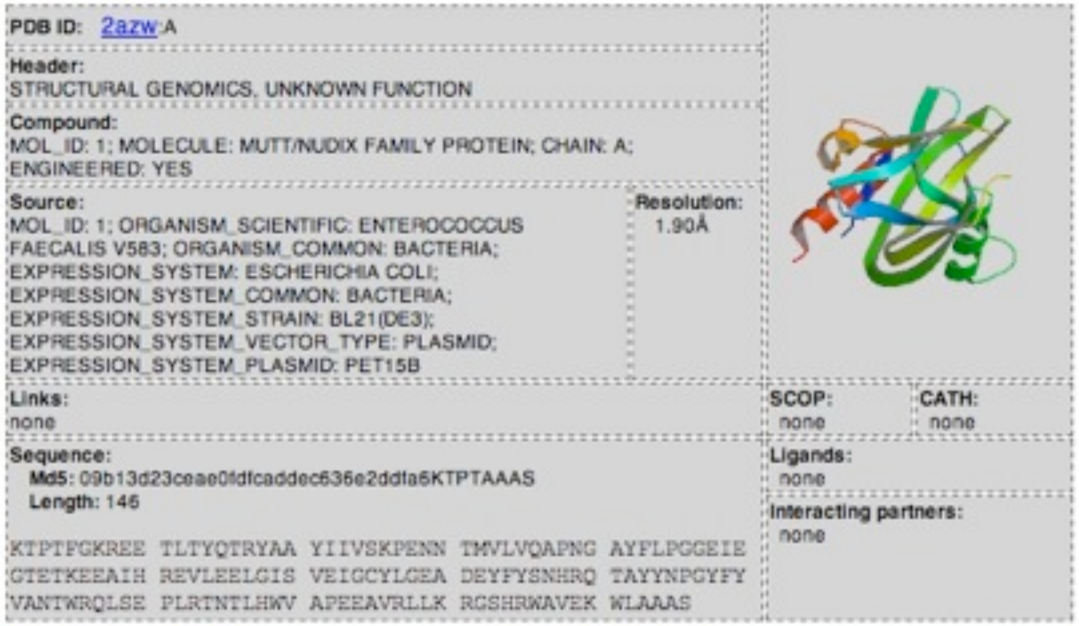
Marti-Renom et al. BMC Bioinformatics (2007) Volume 8. Suppl S4

~90-95% of residues correctly predicted

Example (2azwA)

Structural Genomics Unknown Function

Molecule: MutT/nudix family protein



PDB ID: 2azw:A

Header:
STRUCTURAL GENOMICS, UNKNOWN FUNCTION

Compound:
MOL_ID: 1; MOLECULE: MUTT/NUDIX FAMILY PROTEIN; CHAIN: A;
ENGINEERED: YES

Source:
MOL_ID: 1; ORGANISM_SCIENTIFIC: ENTEROCOCCUS
FAECALIS V583; ORGANISM_COMMON: BACTERIA;
EXPRESSION_SYSTEM: ESCHERICHIA COLI;
EXPRESSION_SYSTEM_COMMON: BACTERIA;
EXPRESSION_SYSTEM_STRAIN: BL21(DE3);
EXPRESSION_SYSTEM_VECTOR_TYPE: PLASMID;
EXPRESSION_SYSTEM_PLASMID: PET15B

Resolution:
1.90Å

Links:
none

SCOP:
none

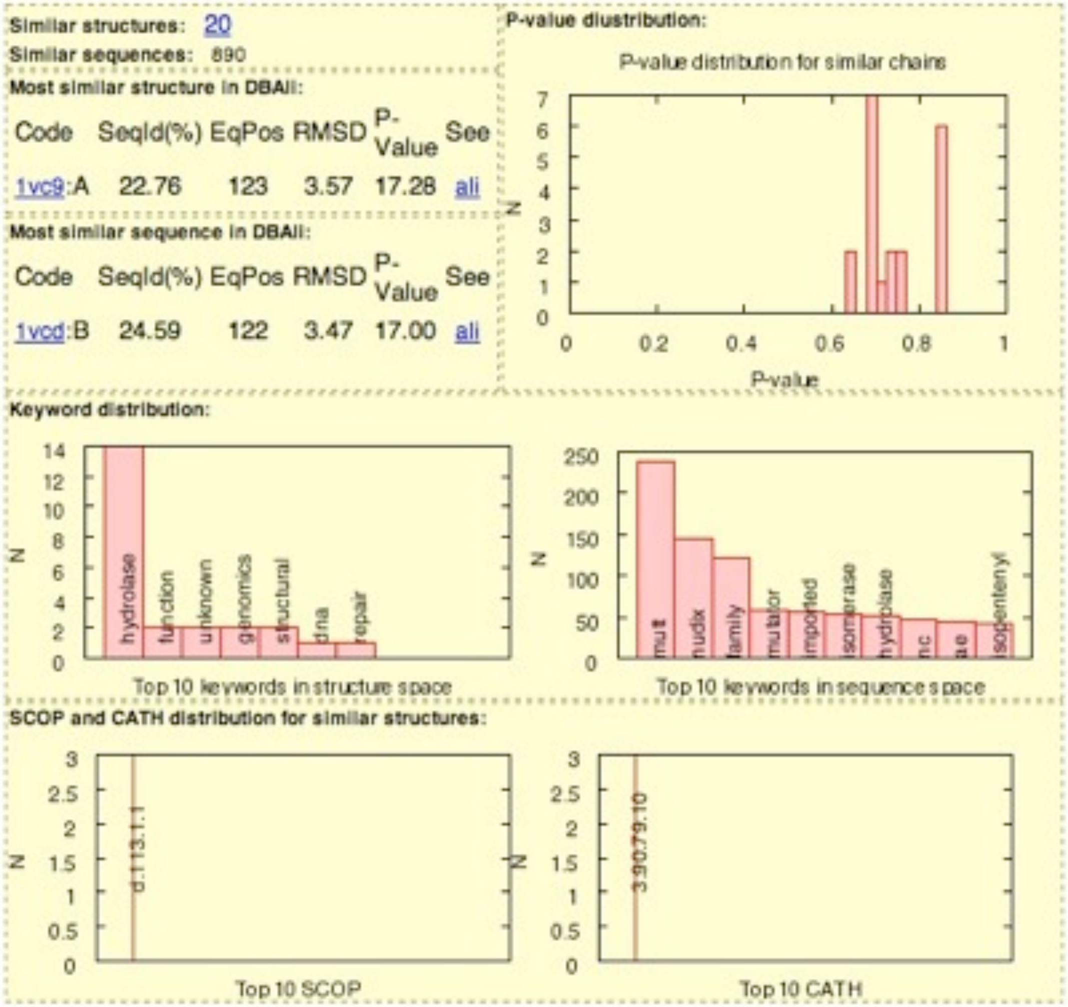
CATH:
none

Sequence:
Md5: 09b13d23ceae0fdfcaddec636e2ddfa6KTPTAAAS
Length: 146

```
KTPTFGKREE TLTYQTRYAA YIIVSKPENN TMVLVQAPNG AYFLPGGEIE
GTETKEEAIH REVLEELGIS VEIGCYLGEA DEYFYSNHRQ TAYYNPGYFY
VANTWRQLSE PLRTNTLHWV APEEAVRLLK RGSHRWAVEK WLAAAS
```

Ligands:
none

Interacting partners:
none



Similar structures: 20
Similar sequences: 890

Most similar structure in DBAli:

Code	SeqId(%)	EqPos	RMSD	P-Value	See
1vc9:A	22.76	123	3.57	17.28	ali

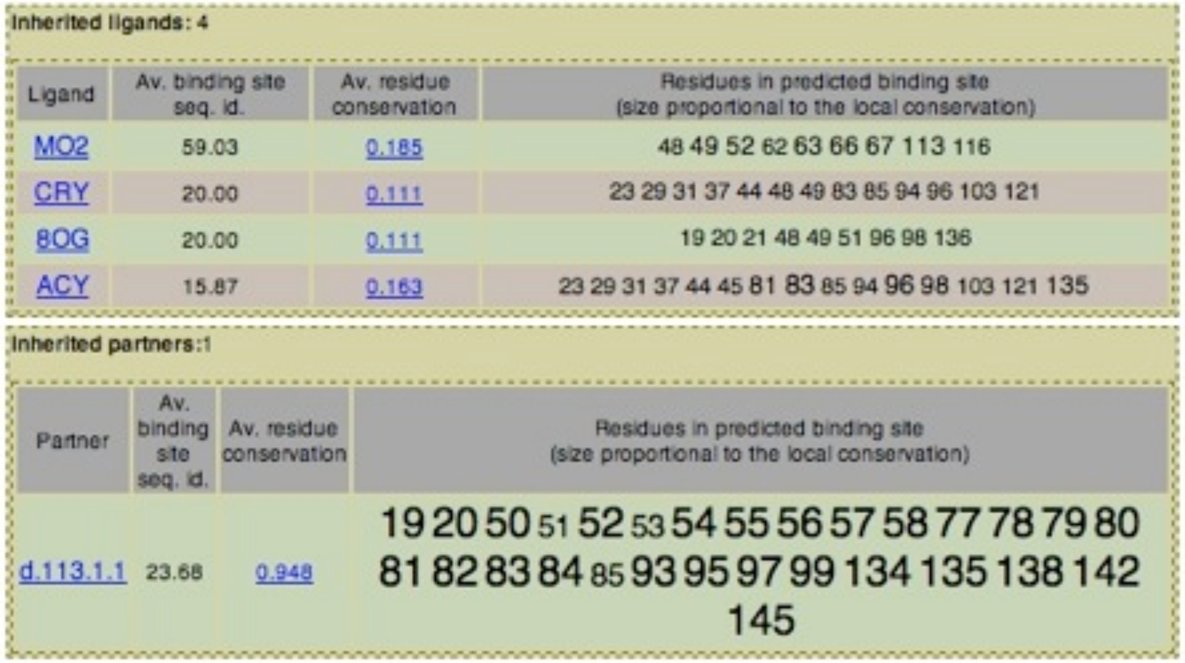
Most similar sequence in DBAli:

Code	SeqId(%)	EqPos	RMSD	P-Value	See
1vcd:B	24.59	122	3.47	17.00	ali

P-value diustribution:

Keyword distribution:

SCOP and CATH distribution for similar structures:

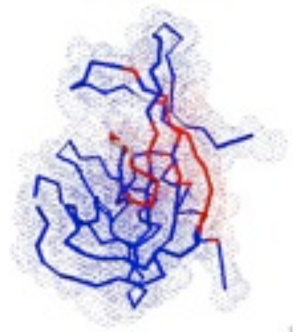


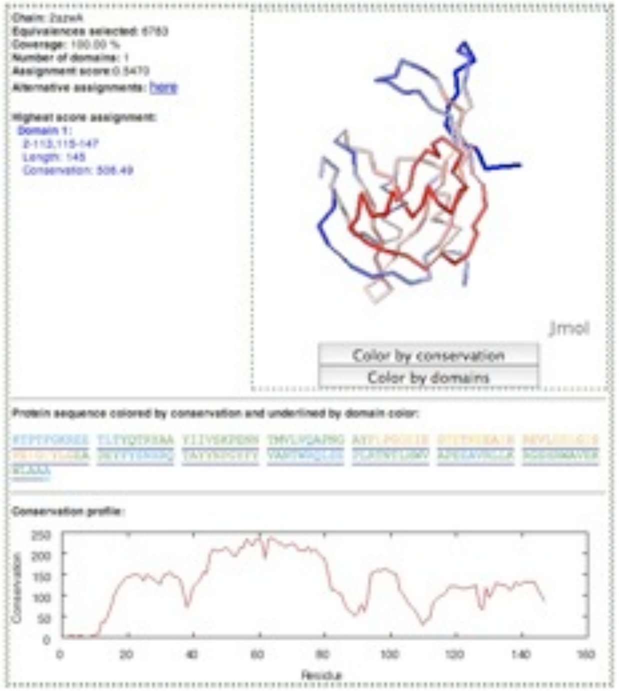
Inherited ligands: 4

Ligand	Av. binding site seq. id.	Av. residue conservation	Residues in predicted binding site (size proportional to the local conservation)
MO2	59.03	0.185	48 49 52 62 63 66 67 113 116
CRY	20.00	0.111	23 29 31 37 44 48 49 83 85 94 96 103 121
8OG	20.00	0.111	19 20 21 48 49 51 96 98 136
ACY	15.87	0.163	23 29 31 37 44 45 81 83 85 94 96 98 103 121 135

Inherited partners:1

Partner	Av. binding site seq. id.	Av. residue conservation	Residues in predicted binding site (size proportional to the local conservation)
d.113.1.1	23.68	0.948	19 20 50 51 52 53 54 55 56 57 58 77 78 79 80 81 82 83 84 85 93 95 97 99 134 135 138 142 145





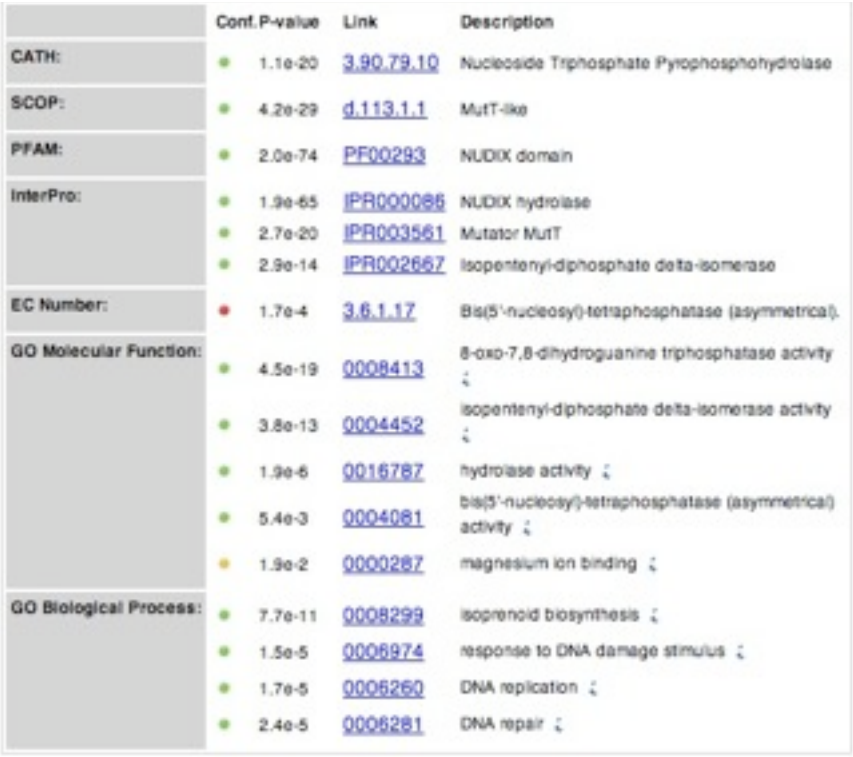
Chain: 2azwA
Equivalences selected: 6763
Coverage: 100.00 %
Number of domains: 1
Assignment score:0.5470
Alternative assignments: here

Highest score assignment:
Domain 1:
2-113,115-147
Length: 145
Conservation: 506.49

Protein sequence colored by conservation and underlined by domain color:

```
KTPTFGKREE TLTYQTRYAA YIIVSKPENN TMVLVQAPNG AYFLPGGEIE GTETKEEAIH REVLEELGIS
VEIGCYLGEA DEYFYSNHRQ TAYYNPGYFY VANTWRQLSE PLRTNTLHWV APEEAVRLLK RGSHRWAVEK
WLAAA
```

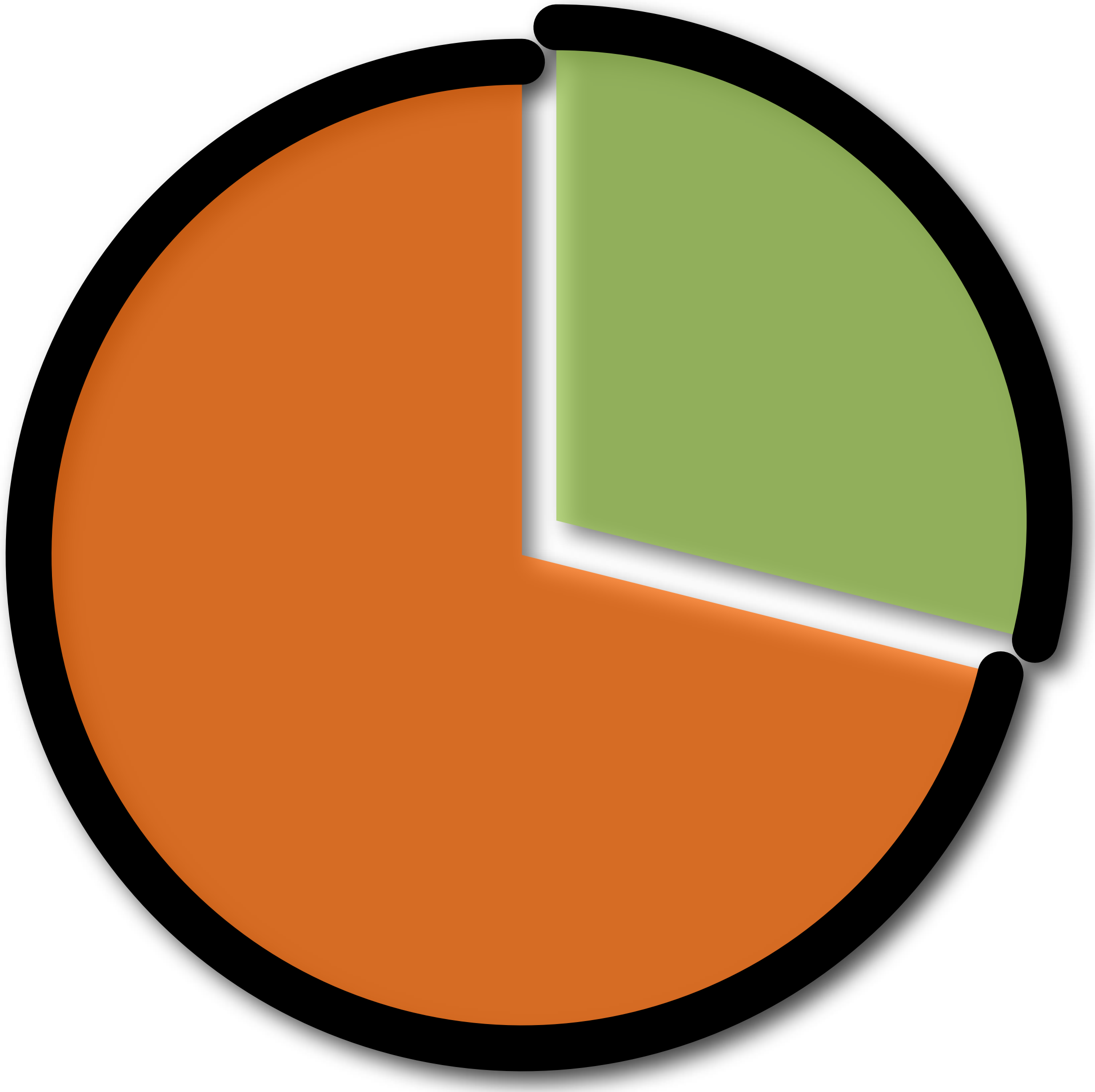
Conservation profile:



	Conf.	P-value	Link	Description
CATH:	●	1.1e-20	3.90.79.10	Nucleoside Triphosphate Pyrophosphohydrolase
SCOP:	●	4.2e-29	d.113.1.1	MutT-like
PFAM:	●	2.0e-74	PF00293	NUDIX domain
InterPro:	●	1.9e-65	IPR000086	NUDIX hydrolase
	●	2.7e-20	IPR003561	Mutator MutT
	●	2.9e-14	IPR002667	Isopentenyl-diphosphate delta-isomerase
EC Number:	●	1.7e-4	3.6.1.17	Bis(5'-nucleosyl)-tetraphosphatase (asymmetrical).
GO Molecular Function:	●	4.5e-19	0008413	8-oxo-7,8-dihydroguanine triphosphatase activity
	●	3.8e-13	0004452	isopentenyl-diphosphate delta-isomerase activity
	●	1.9e-6	0016787	hydrolase activity
	●	5.4e-3	0004081	bis(5'-nucleosyl)-tetraphosphatase (asymmetrical) activity
	●	1.9e-2	0000287	magnesium ion binding
GO Biological Process:	●	7.7e-11	0008299	isoprenoid biosynthesis
	●	1.5e-5	0006974	response to DNA damage stimulus
	●	1.7e-5	0006260	DNA replication
	●	2.4e-5	0006281	DNA repair

Tropical Disease Initiative (TDI)

Predicting binding sites in protein structure models.

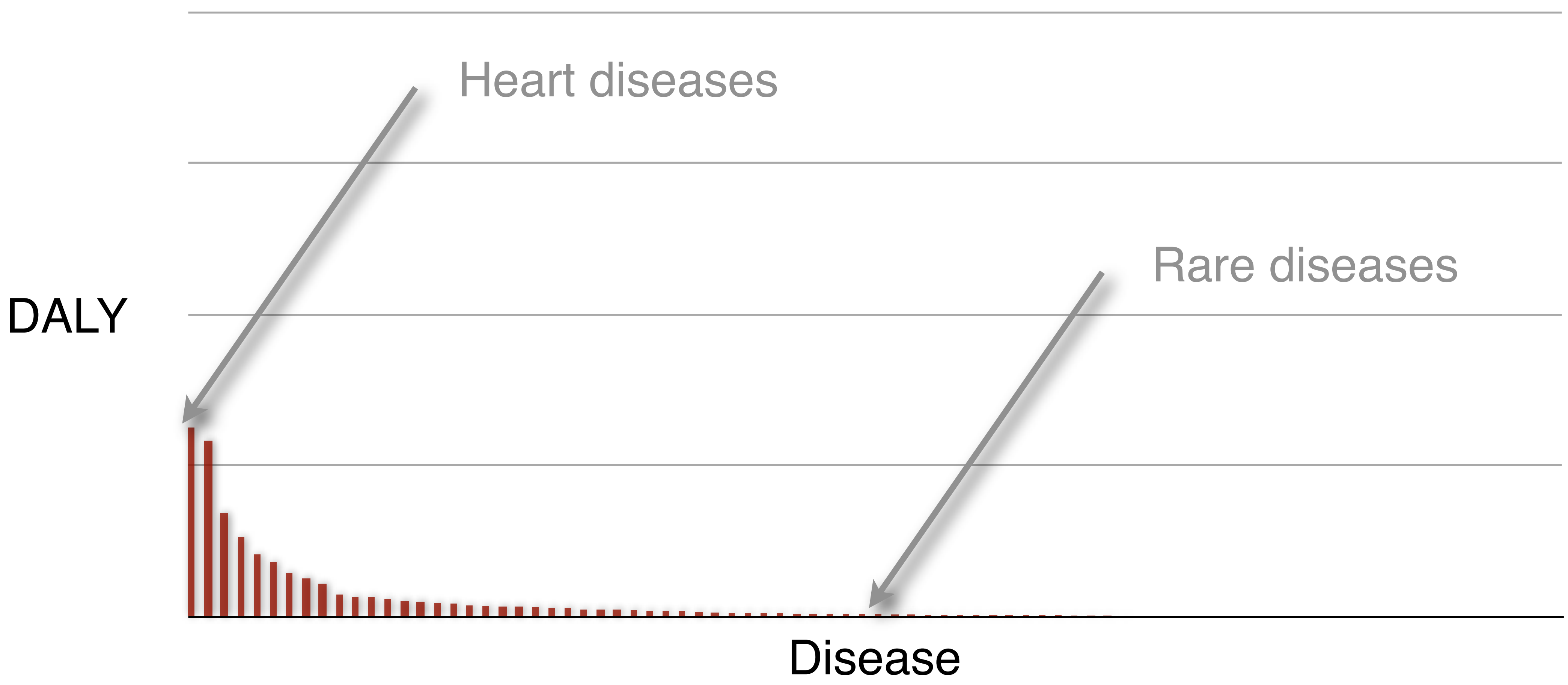


http://www.tropicaldisease.org

Need is High in the Tail

🟥 DALY Burden Per Disease in Developed Countries

🟩 DALY Burden Per Disease in Developing Countries



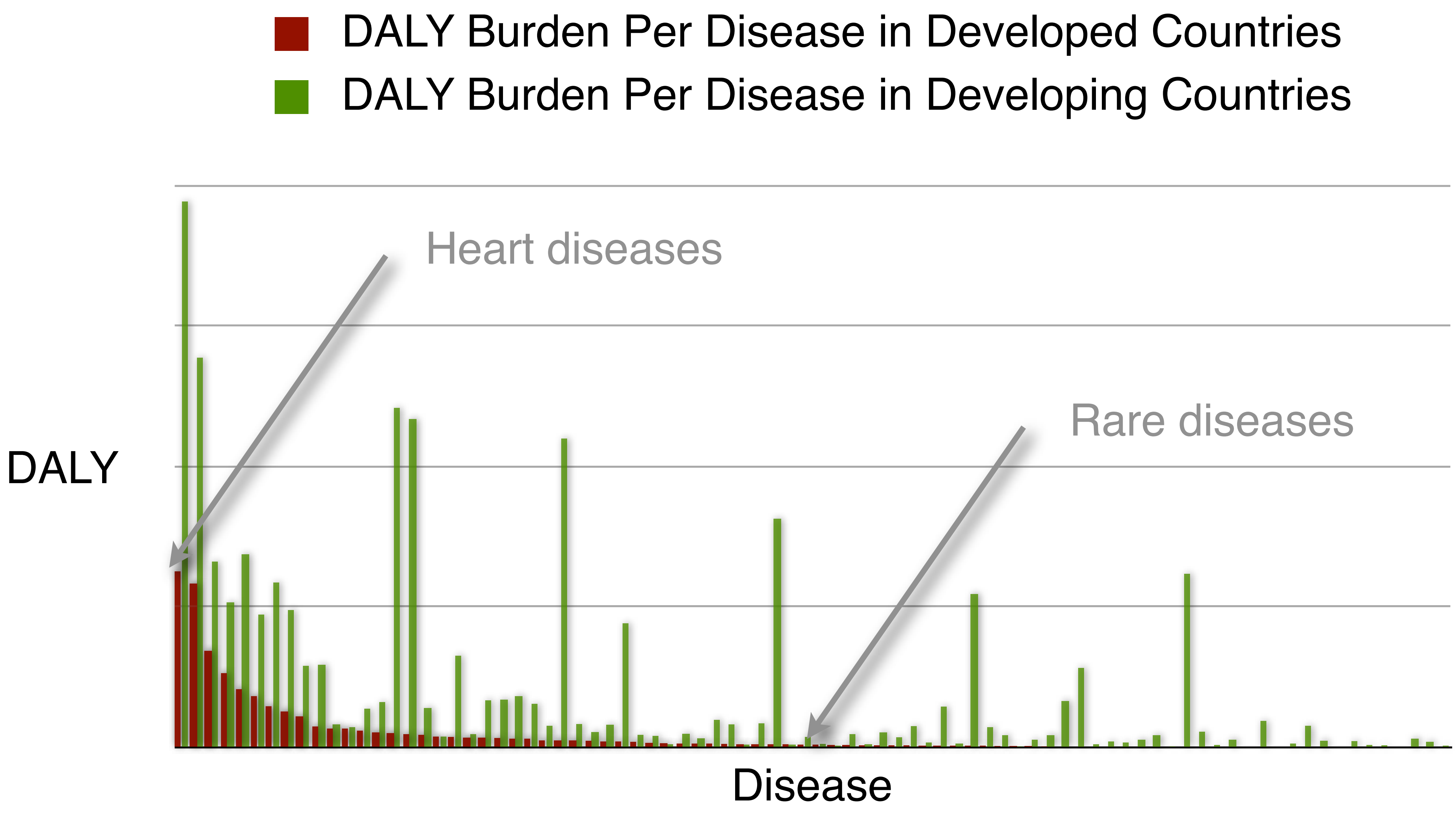
Disease data taken from WHO, *World Health Report 2004*

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.

Need is High in the Tail



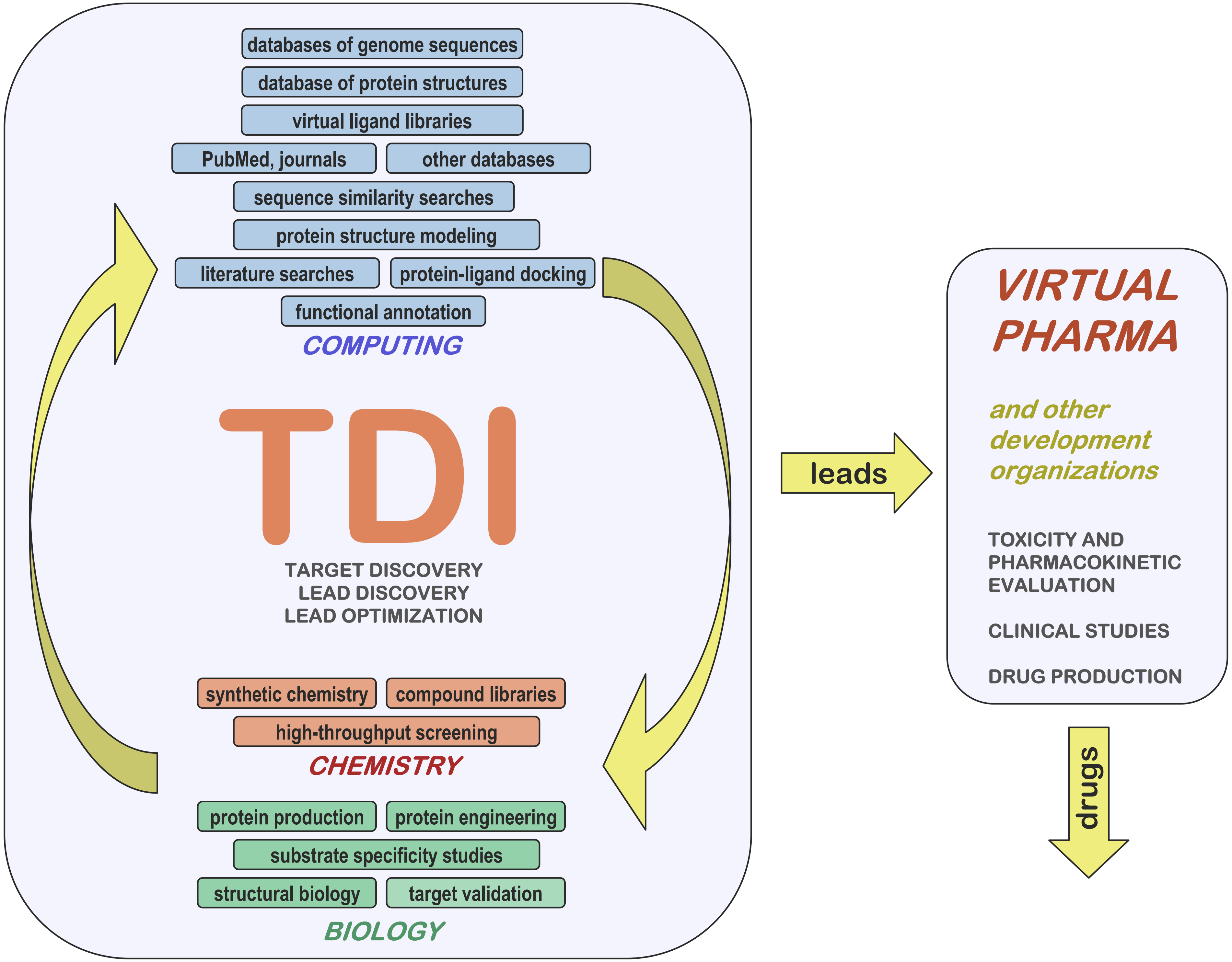
Disease data taken from WHO, *World Health Report 2004*

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.

TDI flowchart



databases of genome sequences

database of protein structures

virtual ligand libraries

PubMed, journals

other databases

sequence similarity searches

protein structure modeling

literature searches

protein-ligand docking

functional annotation

COMPUTING

TDI

TARGET DISCOVERY
LEAD DISCOVERY
LEAD OPTIMIZATION

synthetic chemistry

compound libraries

high-throughput screening

CHEMISTRY

protein production

protein engineering

substrate specificity studies

structural biology

target validation

BIOLOGY

leads

VIRTUAL PHARMA

and other development organizations

TOXICITY AND PHARMACOKINETIC EVALUATION

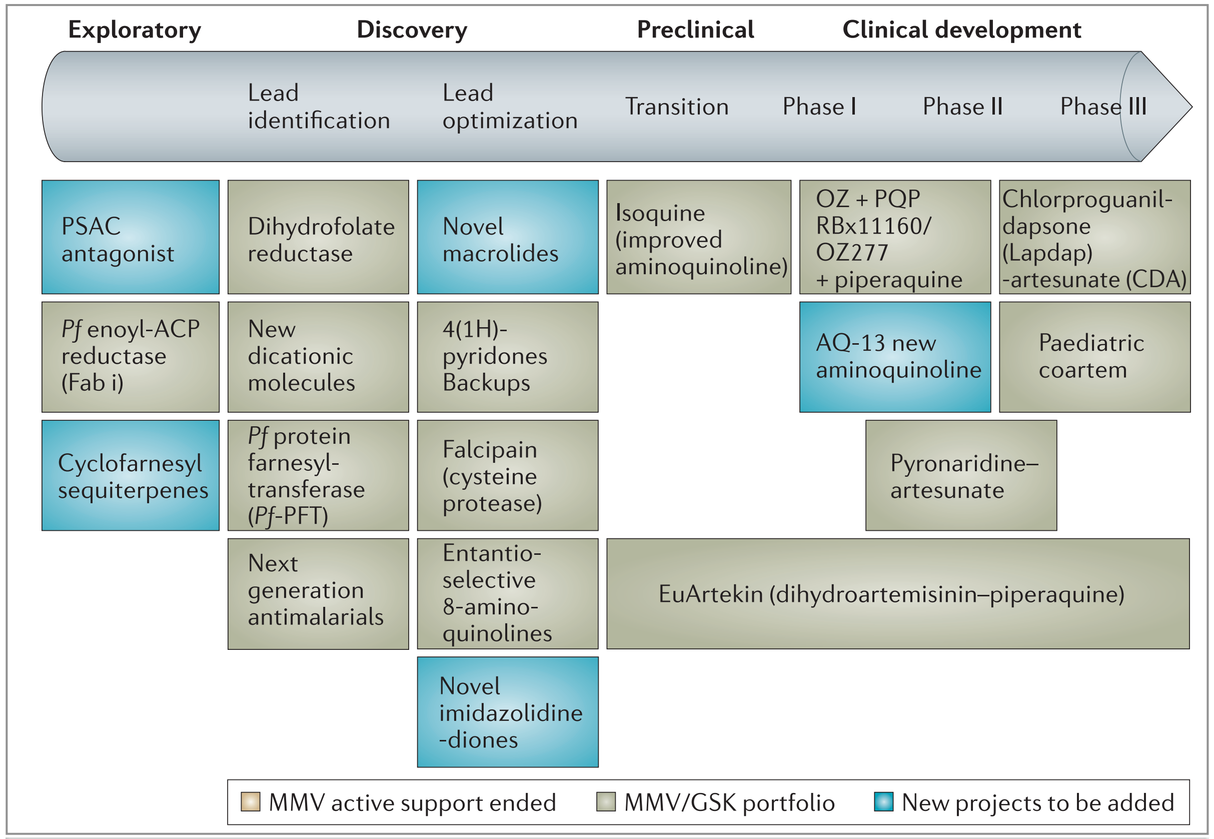
CLINICAL STUDIES

DRUG PRODUCTION

drugs

Non-Profit organizations

Open-Source + Out-Source = low cost business model



Exploratory | Discovery | Preclinical | Clinical development

Lead identification | Lead optimization | Transition | Phase I | Phase II | Phase III

PSAC antagonist

Pf enoyl-ACP reductase (Fab i)

Cyclofarnesyl sequiterpenes

Dihydrofolate reductase

New dicationic molecules

Pf protein farnesyl-transferase (*Pf*-PFT)

Next generation antimalarials

Novel macrolides

4(1H)-pyridones Backups

Falcipain (cysteine protease)

Entantio-selective 8-amino-quinolines

Novel imidazolidine -diones

Isoquine (improved aminoquinoline)

OZ + PQP RBx11160/ OZ277 + piperaquine

AQ-13 new aminoquinoline

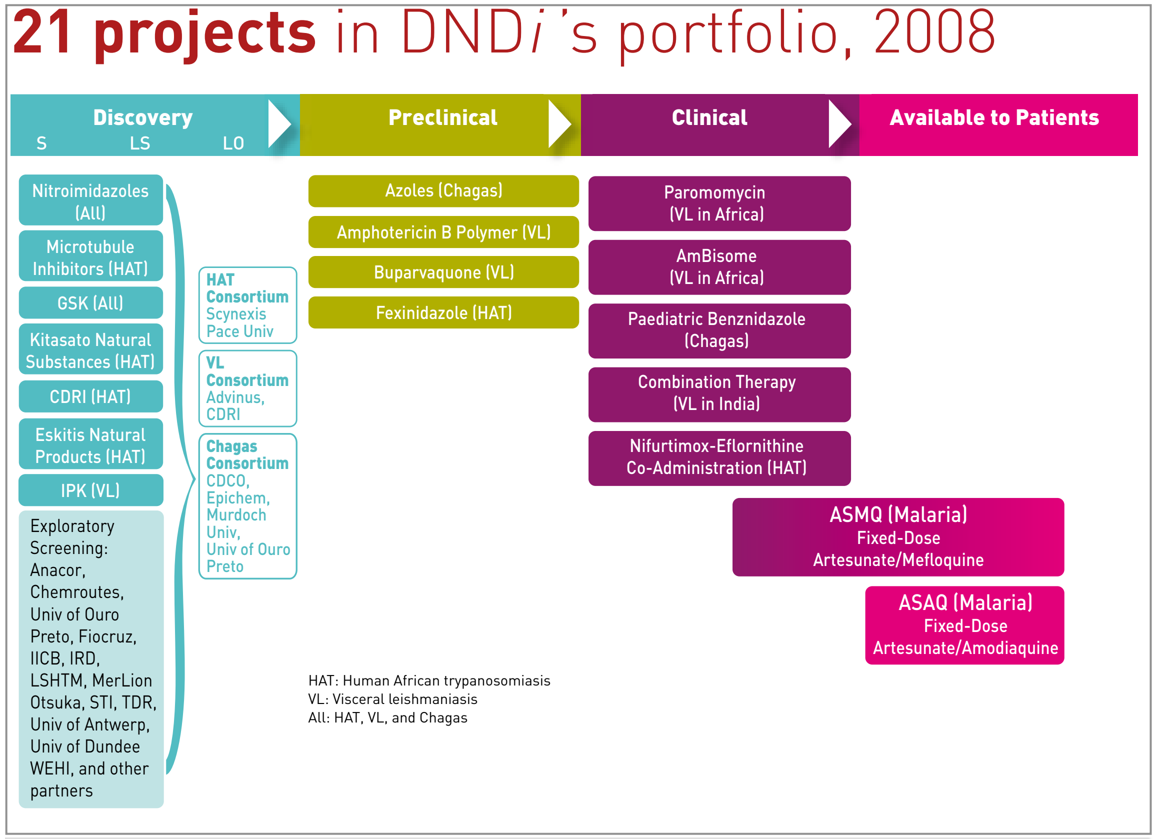
Pyronaridine–artesunate

Chlorproguanil-dapsone (Lapdap) -artesunate (CDA)

Paediatric coartem

EuArtekin (dihydroartemisinin–piperaquine)

▨ MMV active support ended ▨ MMV/GSK portfolio 🟦 New projects to be added



21 projects in DND*i* 's portfolio, 2008

Discovery S LS LO | **Preclinical** | **Clinical** | **Available to Patients**

Nitroimidazoles (All)

Microtubule Inhibitors (HAT)

GSK (All)

Kitasato Natural Substances (HAT)

CDRI (HAT)

Eskitis Natural Products (HAT)

IPK (VL)

Exploratory Screening: Anacor, Chemroutes, Univ of Ouro Preto, Fiocruz, IICB, IRD, LSHTM, MerLion Otsuka, STI, TDR, Univ of Antwerp, Univ of Dundee WEHI, and other partners

HAT Consortium Scynexis Pace Univ

VL Consortium Advinus, CDRI

Chagas Consortium CDCO, Epichem, Murdoch Univ, Univ of Ouro Preto

Azoles (Chagas)

Amphotericin B Polymer (VL)

Buparvaquone (VL)

Fexinidazole (HAT)

Paromomycin (VL in Africa)

AmBisome (VL in Africa)

Paediatric Benznidazole (Chagas)

Combination Therapy (VL in India)

Nifurtimox-Eflornithine Co-Administration (HAT)

ASMQ (Malaria) Fixed-Dose Artesunate/Mefloquine

ASAQ (Malaria) Fixed-Dose Artesunate/Amodiaquine

HAT: Human African trypanosomiasis
VL: Visceral leishmaniasis
All: HAT, VL, and Chagas

Munos (2006) Nature Reviews. Drug Discovery.

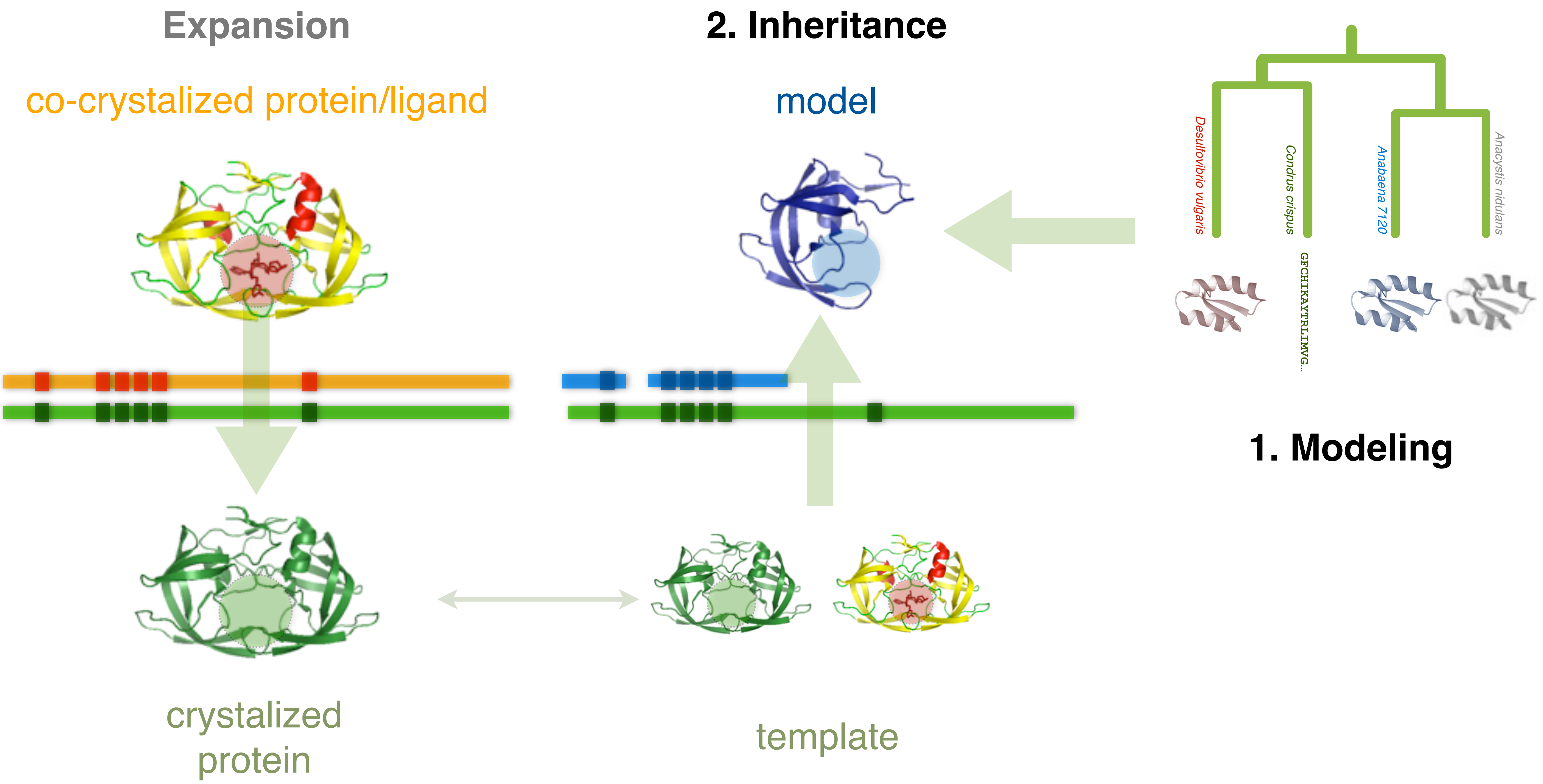
“Unprofitable” Diseases and Global DALY (in 1000’s)

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
Poliomyelitise	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-crystalized protein/ligand

crystalized protein

2. Inheritance

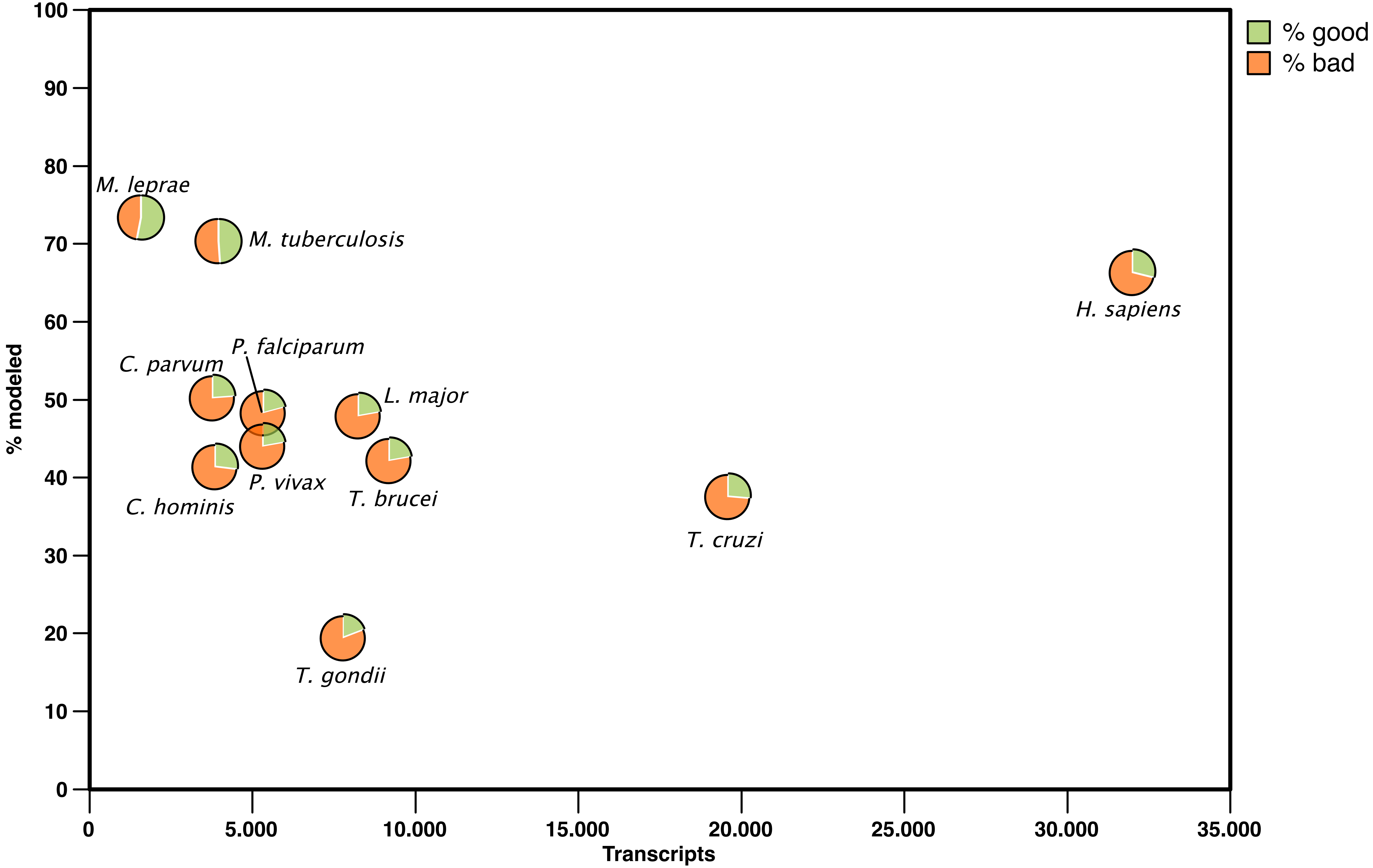
model

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

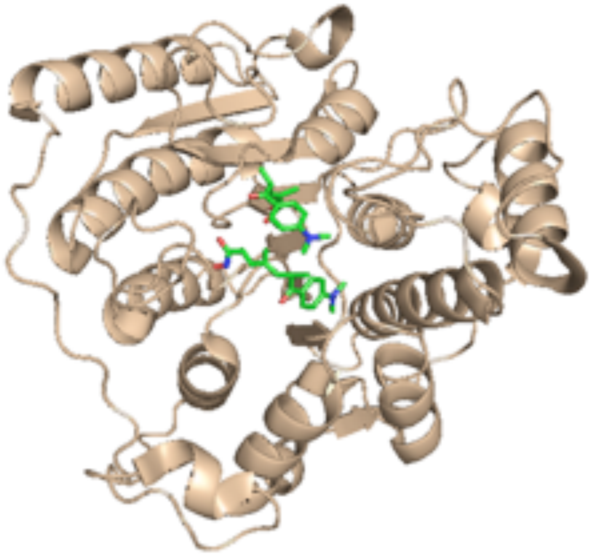
models with inherited ligands

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

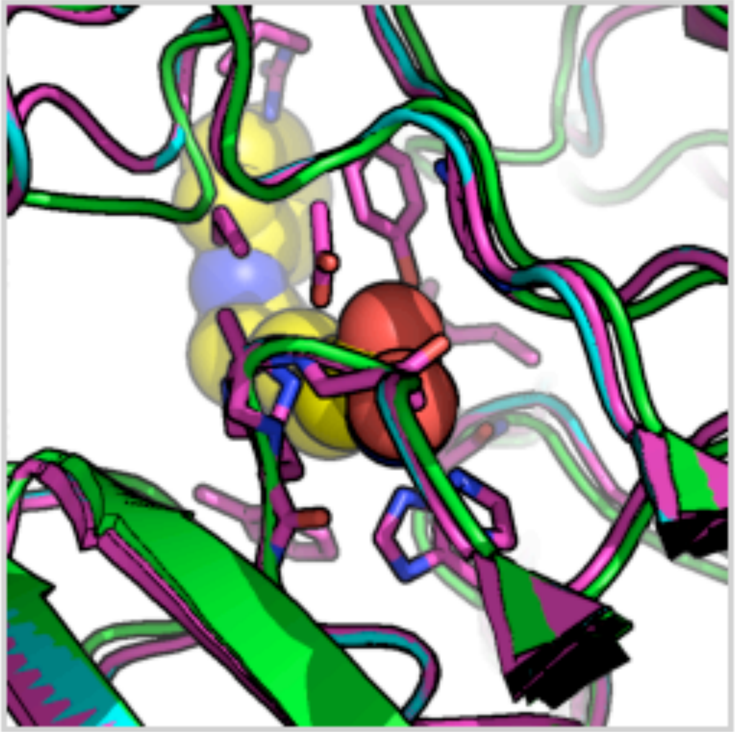
	Transcripts	Modeled targets	Selected models	Inherited ligands	Similar to a drug	Drugs
C. hominis	3,886	1,614	666	197	20	13
C. parvum	3,806	1,918	742	232	24	13
L. major	8,274	3,975	1,409	478	43	20
M. leprae	1,605	1,178	893	310	25	6
M. tuberculosis	3,991	2,808	1,608	365	30	10
P. falciparum	5,363	2,599	818	284	28	13
P. vivax	5,342	2,359	822	268	24	13
T. brucei	7,793	1,530	300	138	13	6
T. cruzi	19,607	7,390	3,070	769	51	28
T. gondii	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		Template		Model		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

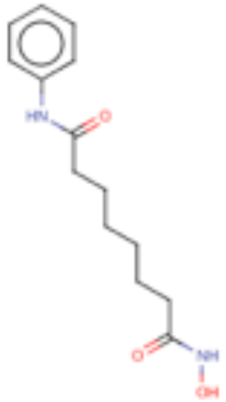
Drug categories:

Anti-Inflammatory Agents, Non-Steroidal

Anticarcinogenic Agents

Antineoplastic Agents

Enzyme Inhibitors

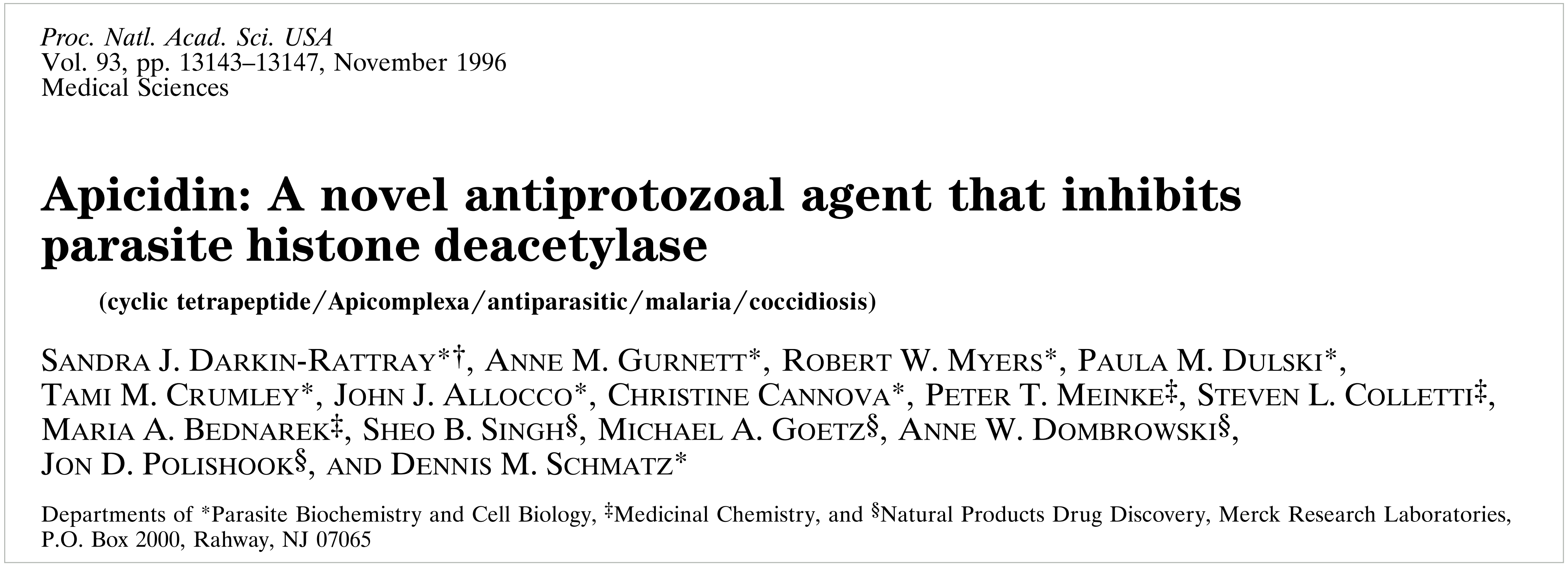


Drug indication:

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



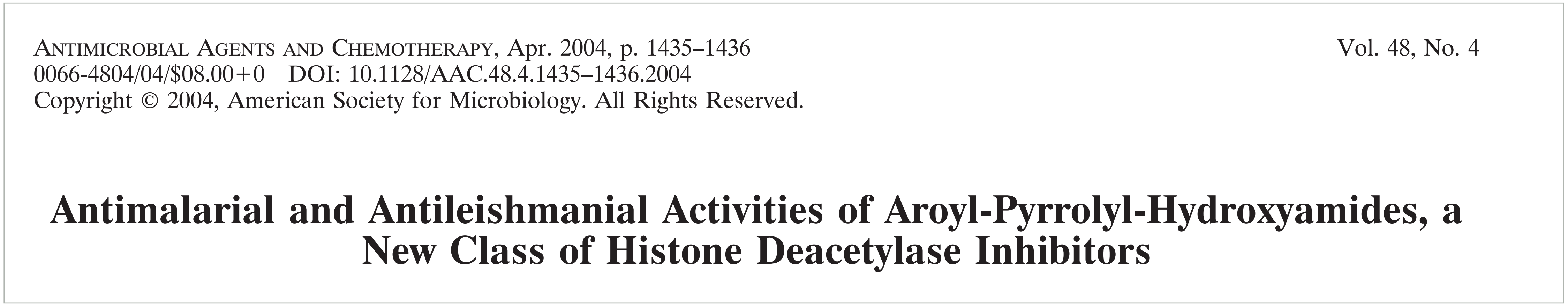
Proc. Natl. Acad. Sci. USA
Vol. 93, pp. 13143–13147, November 1996
Medical Sciences

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

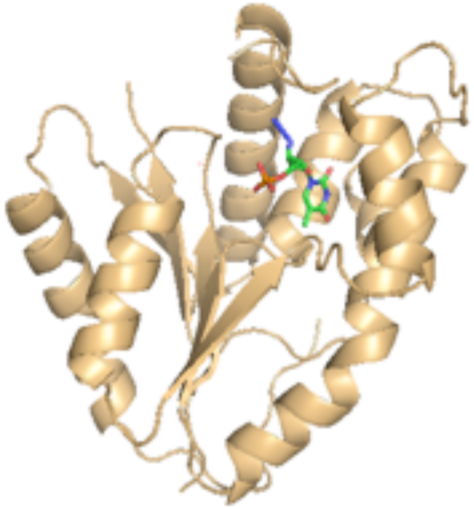


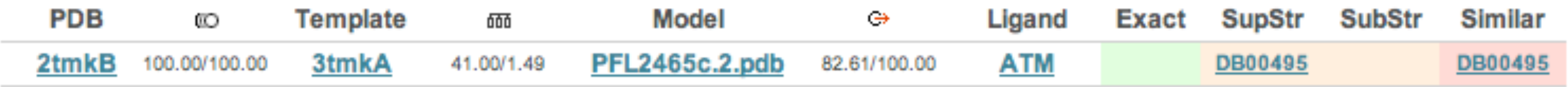
Antimicrobial Agents and Chemotherapy, Apr. 2004, p. 1435–1436 Vol. 48, No. 4
0066-4804/04/$08.00+0 DOI: 10.1128/AAC.48.4.1435–1436.2004
Copyright © 2004, American Society for Microbiology. All Rights Reserved.

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

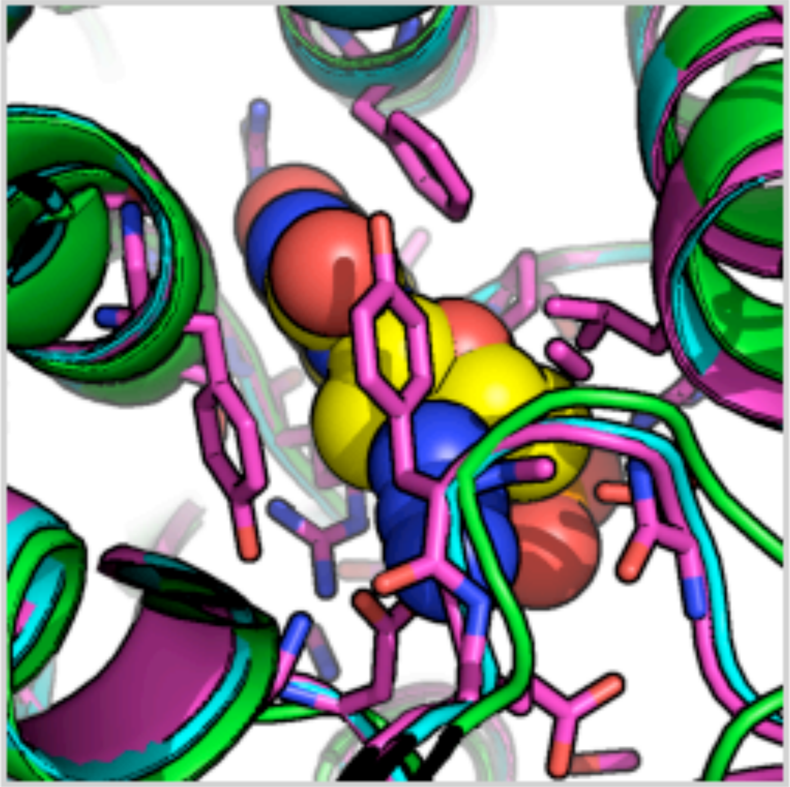
P. falciparum tymidylate kinase + zidovudine

Template 3tmkA a yeast tymidylate 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

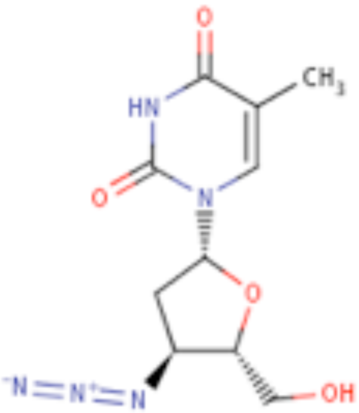
Small Molecule; Approved

Drug categories:

Anti-HIV Agents

Antimetabolites

Nucleoside and Nucleotide Reverse Transcriptase Inhibitors

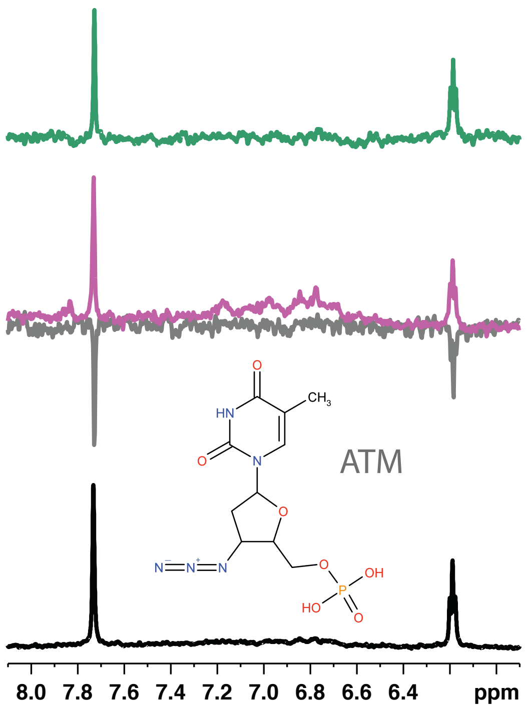


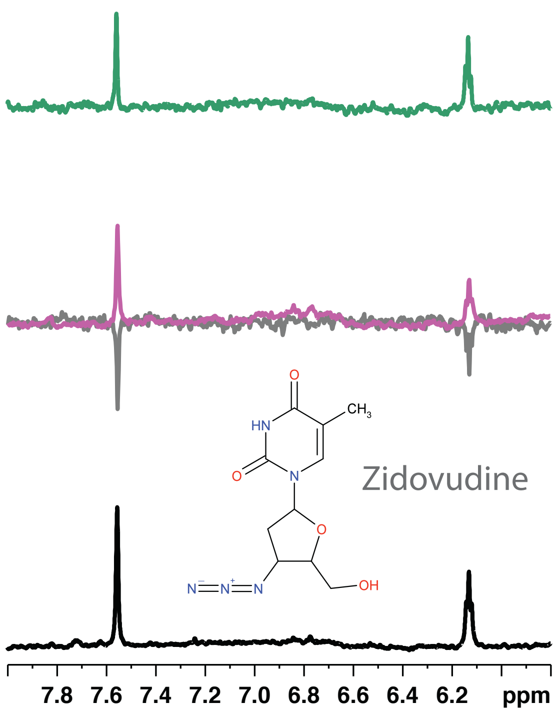
Drug indication:

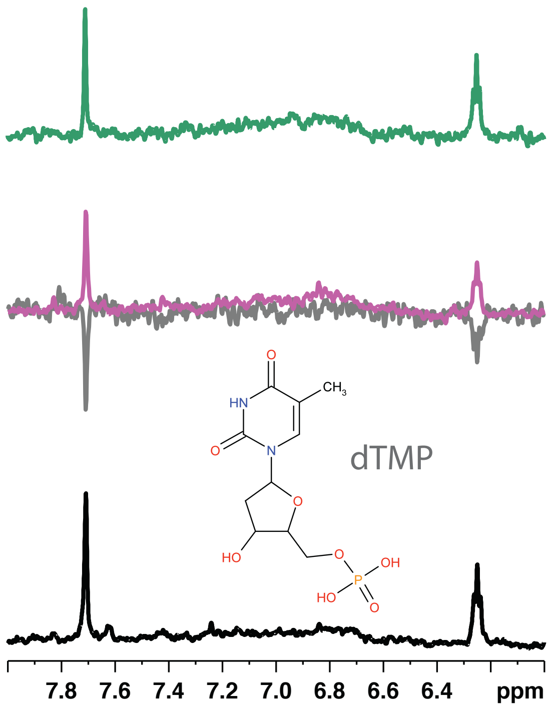
For the treatment of human immunovirus (HIV) infections.

P. falciparum thymidylate kinase + zidovudine

*NMR **Water-LOGSY** and **STD** experiments*



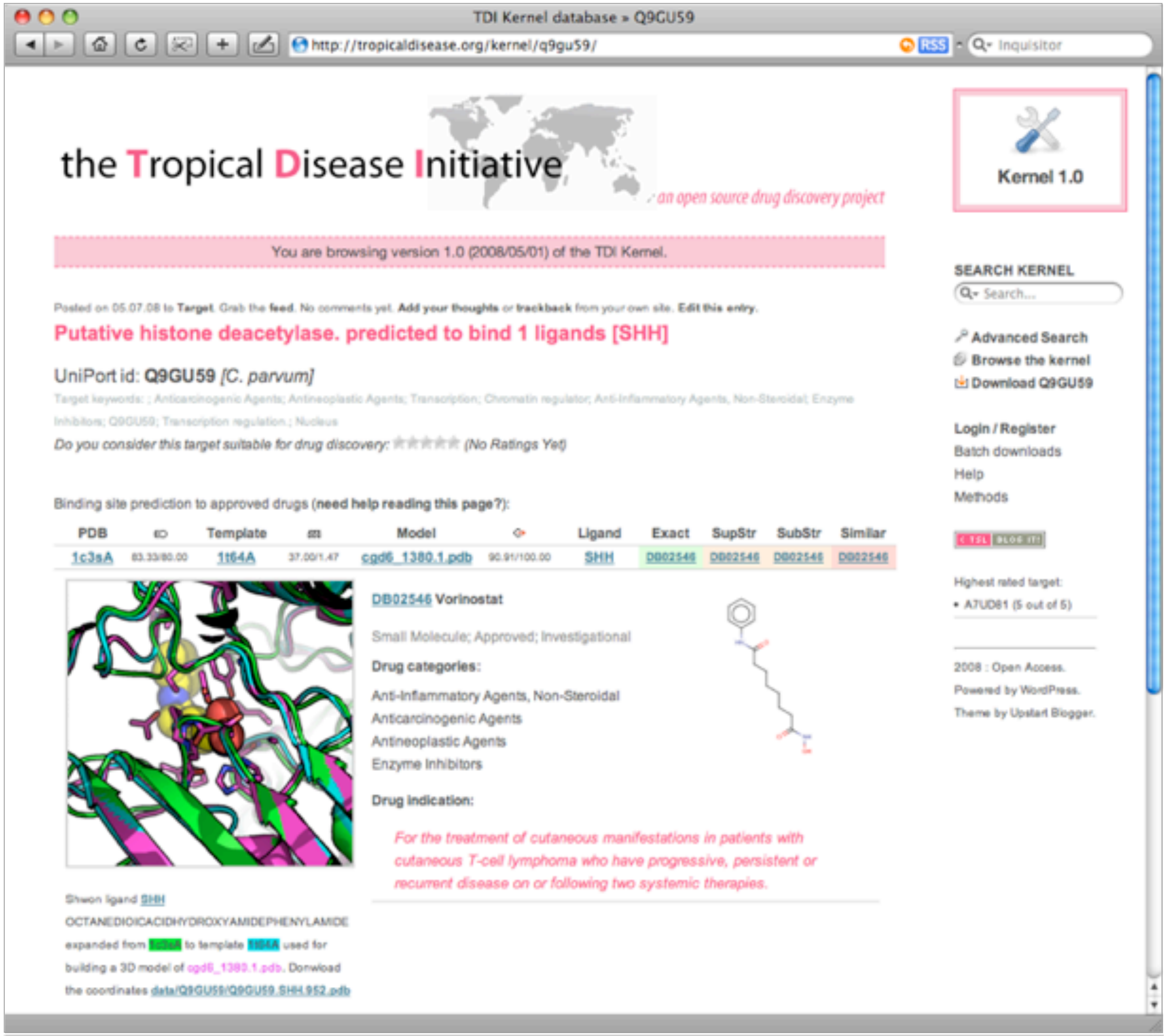




Leticia Ortí, Rodrigo J. Carbajo, and Antonio Pineda-Lucena

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
Kernel 1.0

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		Template		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.

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

SEARCH KERNEL

Advanced Search
Browse the kernel
Download Q9GU59

Login / Register
Batch downloads
Help
Methods

Highest rated target:

- A7UD81 (5 out of 5)

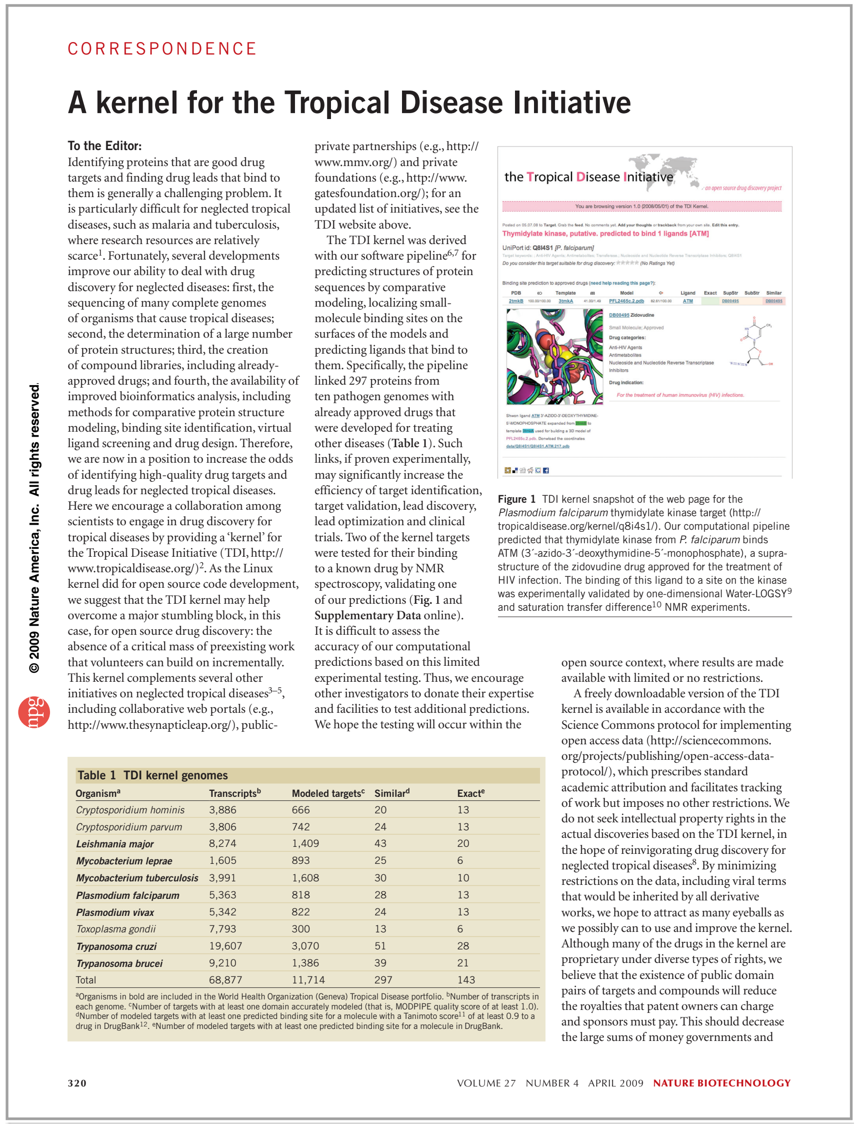
2008 : Open Access.
Powered by WordPress.
Theme by Upstart Blogger.

TDI’s kernel

http://tropicaldisease.org/kernel

L. Orti *et al.*, *Nat Biotechnol* **27**, 320 (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 scarce[1]. 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/)[2]. 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[3–5], 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 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 diseases[8]. 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

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.

Thymidylate kinase, putative. predicted to bind 1 ligands [ATM]

UniPort id: Q8I4S1 *[P. falciparum]*

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

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

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.

Shwon ligand ATM 3'-AZIDO-3'-DEOXYTHYMIDINE-5'-MONOPHOSPHATE expanded from 2tmkB to template 3tmkA used for building a 3D model of PFL2465c.2.pdb. Donwload the coordinates data/Q8I4S1/Q8I4S1.ATM.217.pdb

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 suprastructure 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-LOGSY[9] and saturation transfer difference[10] NMR experiments.

Table 1 TDI kernel genomes

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

[a] Organisms in bold are included in the World Health Organization (Geneva) Tropical Disease portfolio. [b] Number of transcripts in each genome. [c] Number of targets with at least one domain accurately modeled (that is, MODPIPE quality score of at least 1.0). [d] Number of modeled targets with at least one predicted binding site for a molecule with a Tanimoto score[11] of at least 0.9 to a drug in DrugBank[12]. [e] Number of modeled targets with at least one predicted binding site for a molecule in DrugBank.

OPEN ACCESS Freely available online PLoS NEGLECTED TROPICAL DISEASES

A Kernel for Open Source Drug Discovery in Tropical Diseases

Leticia Ortí[1,2], Rodrigo J. Carbajo[2], Ursula Pieper[3], Narayanan Eswar[3¤], Stephen M. Maurer[4], Arti K. Rai[5], Ginger Taylor[6], Matthew H. Todd[7], Antonio Pineda-Lucena[2], Andrej Sali[3*], Marc A. Marti-Renom[1*]

1 Structural Genomics Unit, Bioinformatics and Genomics Department, Centro de Investigación Príncipe Felipe, Valencia, Spain, **2** Structural Biology Laboratory, Medicinal Chemistry Department, Centro de Investigación Príncipe 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 hope that the kernel will facilitate collaborative efforts towards the discovery of new drugs against parasites that cause tropical diseases.

Citation: Ortí 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

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, U54 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@cipf.es (MAM-R)

¤ Current address: DuPont Knowledge Center, Hyderabad, India

Introduction

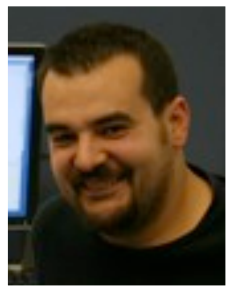
There is a lack of high-quality protein drug targets and drug leads for neglected diseases [1,2]. Fortunately, many genomes of organisms that cause tropical diseases have already been sequenced and published. Therefore, we are now in a position to leverage this information by identifying potential protein 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

Acknowledgments

http://sgu.bioinfo.cipf.es

http://tropicaldisease.org



COMPARATIVE MODELING
Andrej Sali
M. S. Madhusudhan
Narayanan Eswar
Min-Yi Shen
Ursula Pieper
Ben Webb
Maya Topf (Birbeck College)

MODEL ASSESSMENT
David Eramian
Min-Yi Shen
Damien Devos

FUNCTIONAL ANNOTATION
Andrea Rossi (Rinat-Pfizer)
Fred Davis (Janelia Fram)

FUNDING
Prince Felipe Research Center
Ministerio de Educación y Ciencia
STREP UE Grant
Marie Curie Reintegration Grant

MODEL ASSESSMENT
Francisco Melo (CU)
Alejandro Panjkovich (CU)

NMR
Antonio Pineda-Lucena
Leticia Ortí
Rodrigo J. Carbajo

MAMMOTH
Angel R. Ortiz

FUNCTIONAL ANNOTATION
Fatima Al-Shahrour
Joaquin Dopazo

BIOLOGY
Jeff Friedman (RU)
James Hudsped (RU)
Partho Ghosh (UCSD)
Alvaro Monteiro (Cornell U)
Stephen Krilis (St.George H)

Tropical Disease Initiative
Stephen Maurer (UC Berkeley)
Arti Rai (Duke U)
Andrej Sali (UCSF)
Ginger Taylor (TSL)
Matthew Todd (U Sydney)

CCPR Functional Proteomics
Patsy Babbitt (UCSF)
Fred Cohen (UCSF)
Ken Dill (UCSF)
Tom Ferrin (UCSF)
John Irwin (UCSF)
Matt Jacobson (UCSF)
Tack Kuntz (UCSF)
Andrej Sali (UCSF)
Brian Shoichet (UCSF)
Chris Voigt (UCSF)

EVA
Burkhard Rost (Columbia U)
Alfonso Valencia (CNB/UAM)

CAMP
Xavier Aviles (UAB)
Hans-Peter Nester (SANOFI)
Ernst Meinjohanns (ARPIDA)
Boris Turk (IJS)
Markus Gruetter (UE)
Matthias Wilmanns (EMBL)
Wolfram Bode (MPG)