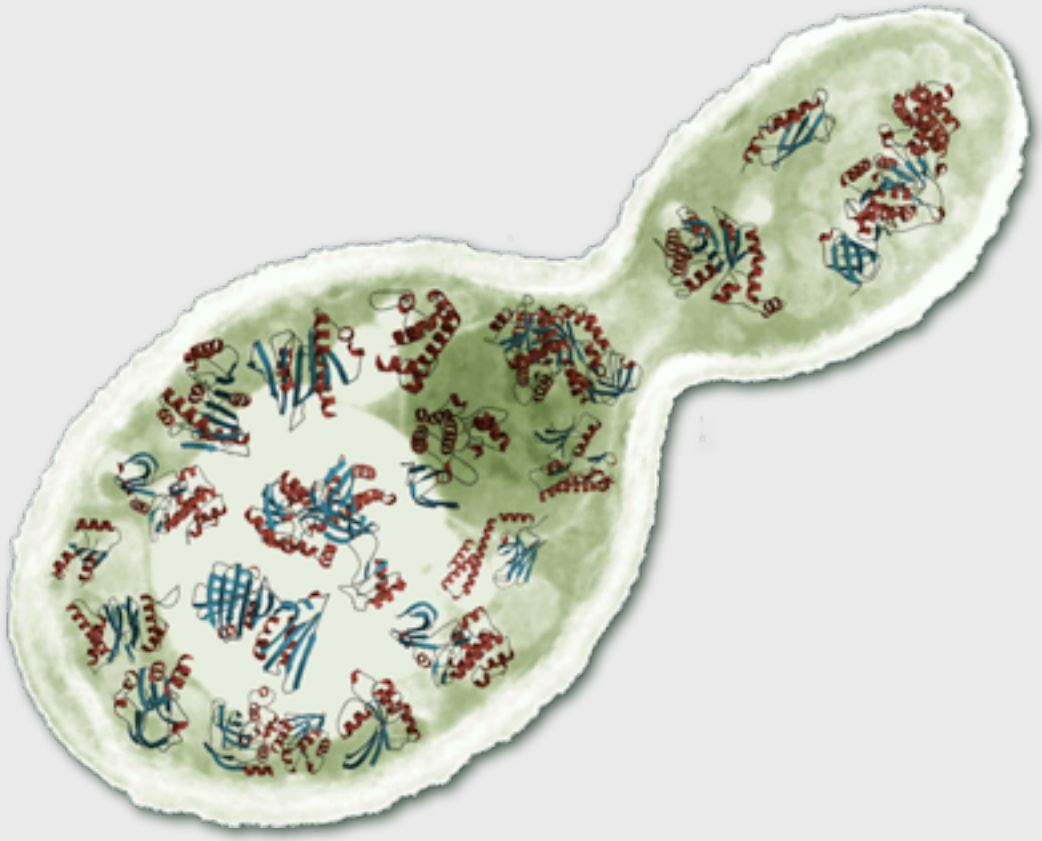
Comparative Protein Structure Prediction



Marc A. Marti-Renom

http://bioinfo.cipf.es/sgu/

Structural Genomics Unit
Bioinformatics Department
Prince Felipe Resarch Center (CIPF), Valencia, Spain

PRINCIPE FELIPE
CENTRO DE INVESTIGACION

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

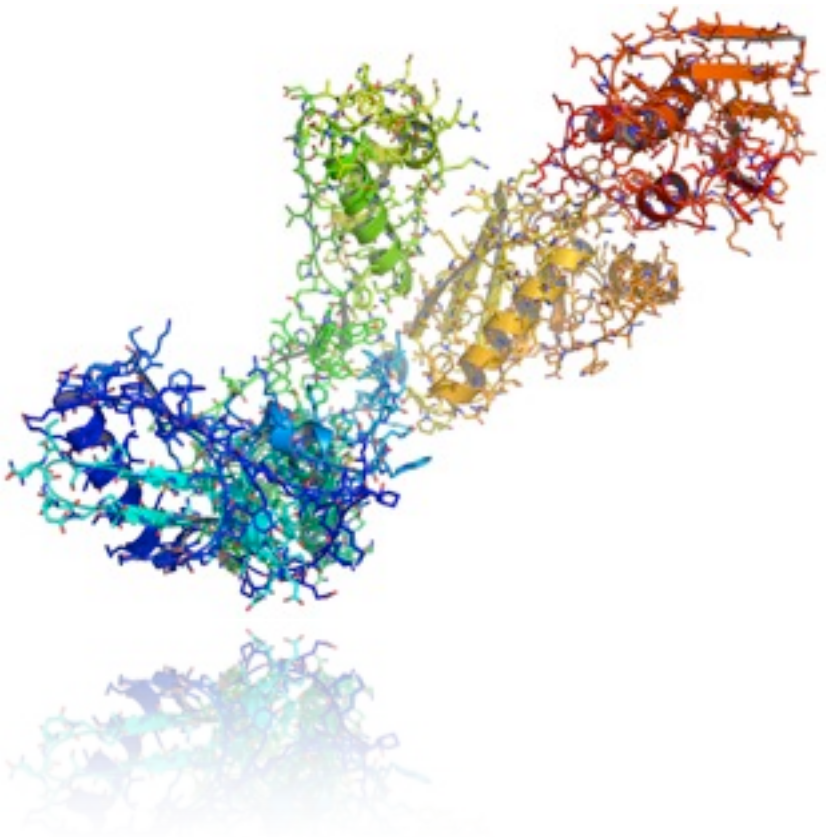
Fold: Three dimensional conformation of a protein sequence (usually at domain level).

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).

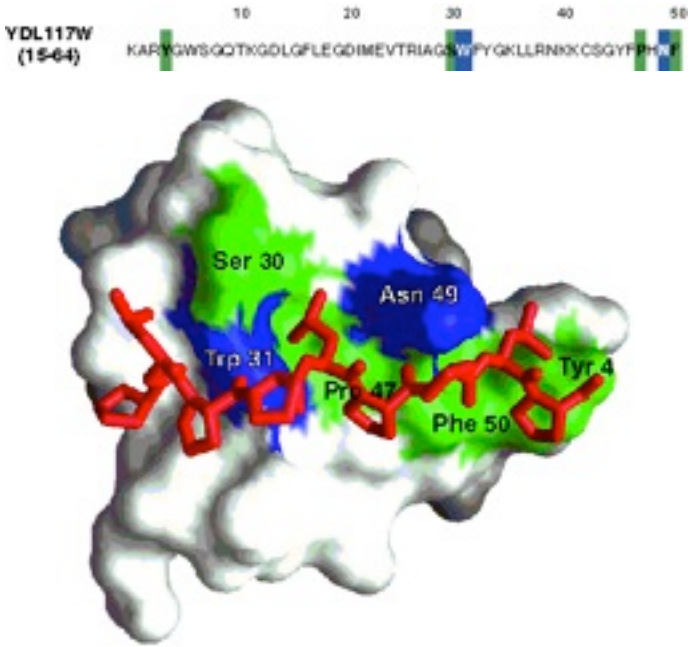
Backbone: Protein structure skeleton composed by the carbon, nitrogen and oxygen atoms.

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



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.

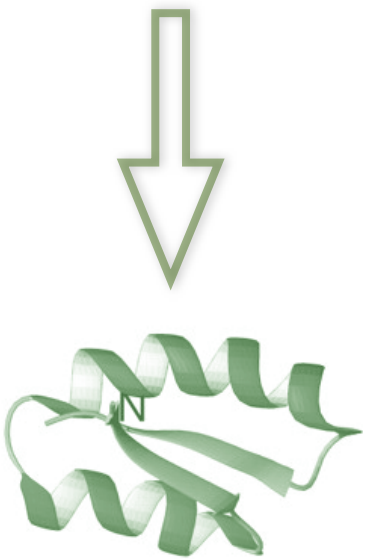


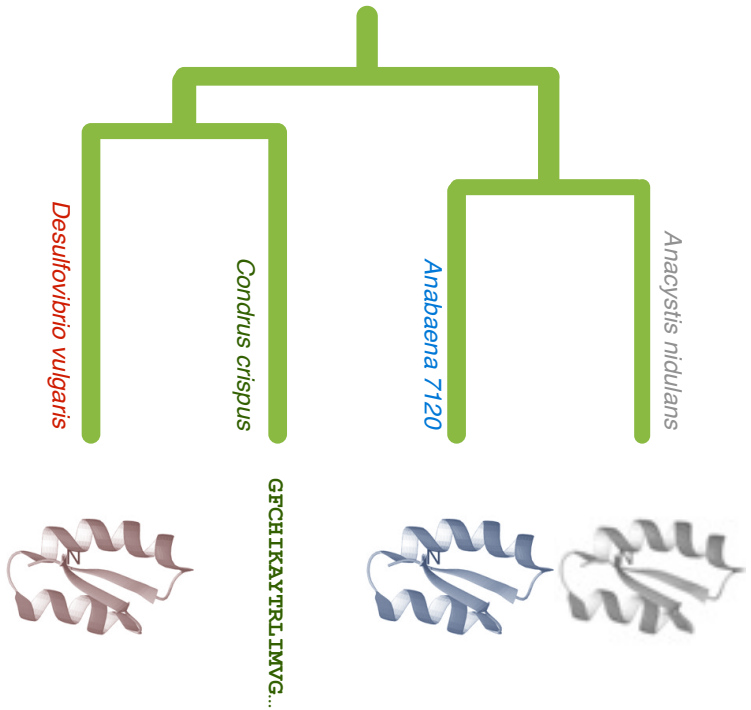
In addition, since evolution tends to conserve function and function depends more directly on structure than on sequence, **structure is more conserved in evolution than sequence**.

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

Principles of protein structure

GFCHIKAYTRLIMVG...





Folding (physics)

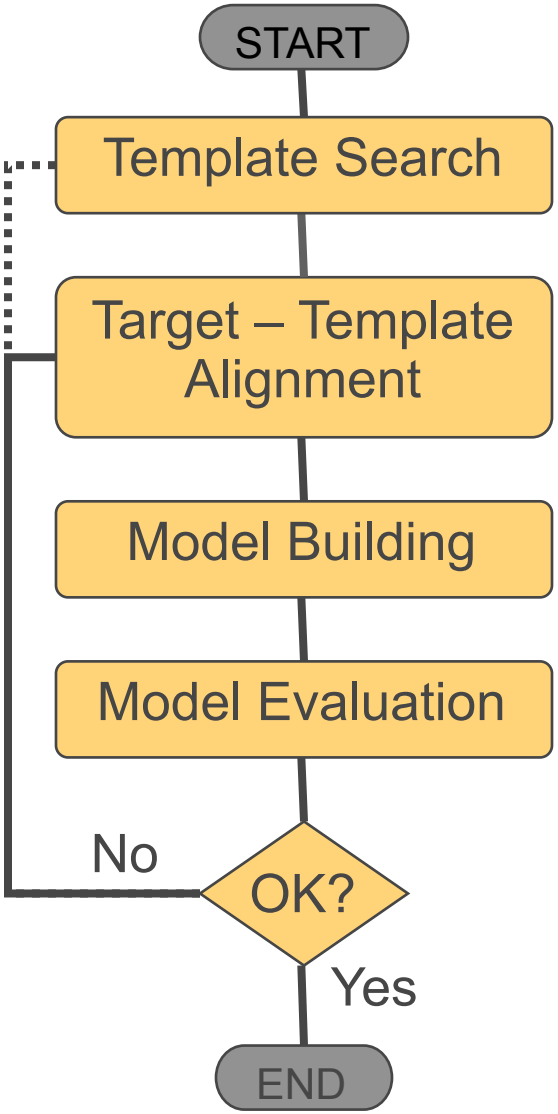
Ab initio prediction

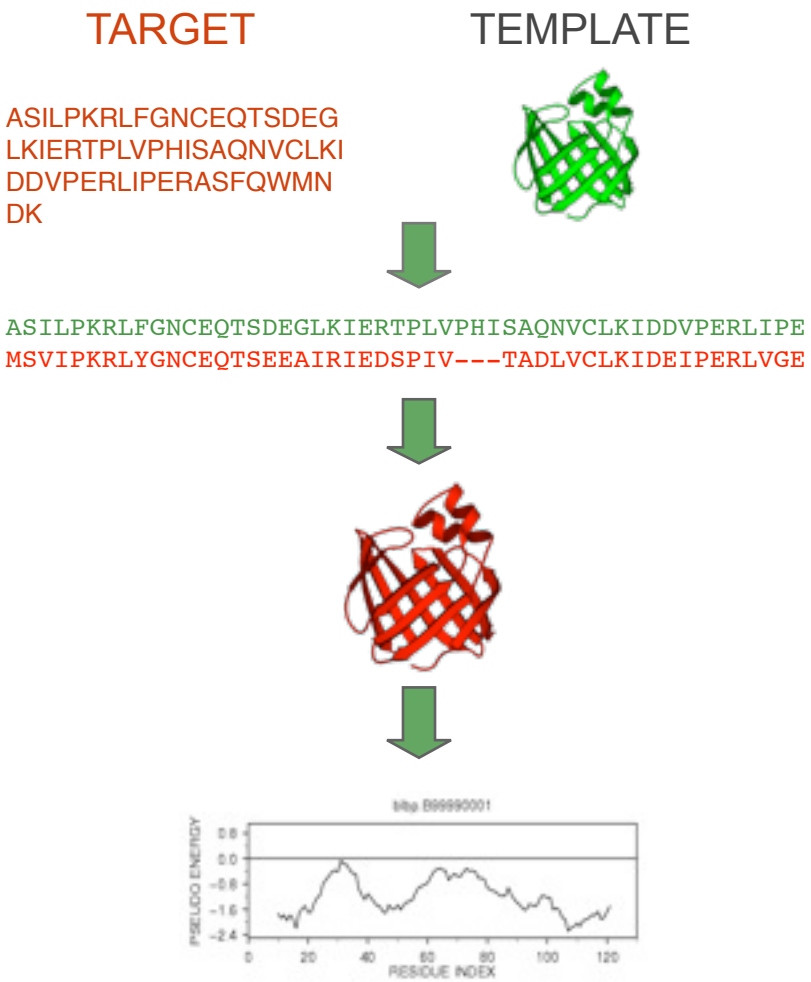
Evolution (rules)

Threading
Comparative Modeling

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

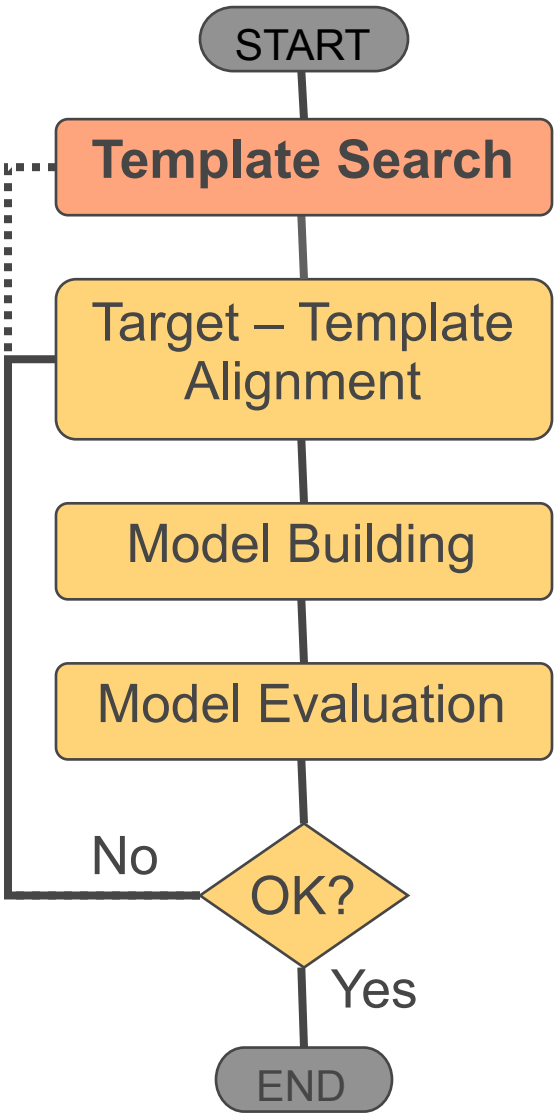
Steps in Comparative Protein Structure Modeling





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

Template Search



Sequence-Sequence search

BLAST

http://www.ncbi.nlm.nih.gov/blast/

Profile-Sequence search

PSI-BLAST

http://www.ncbi.nlm.nih.gov/blast/

Profile-Profile search

pp_scan

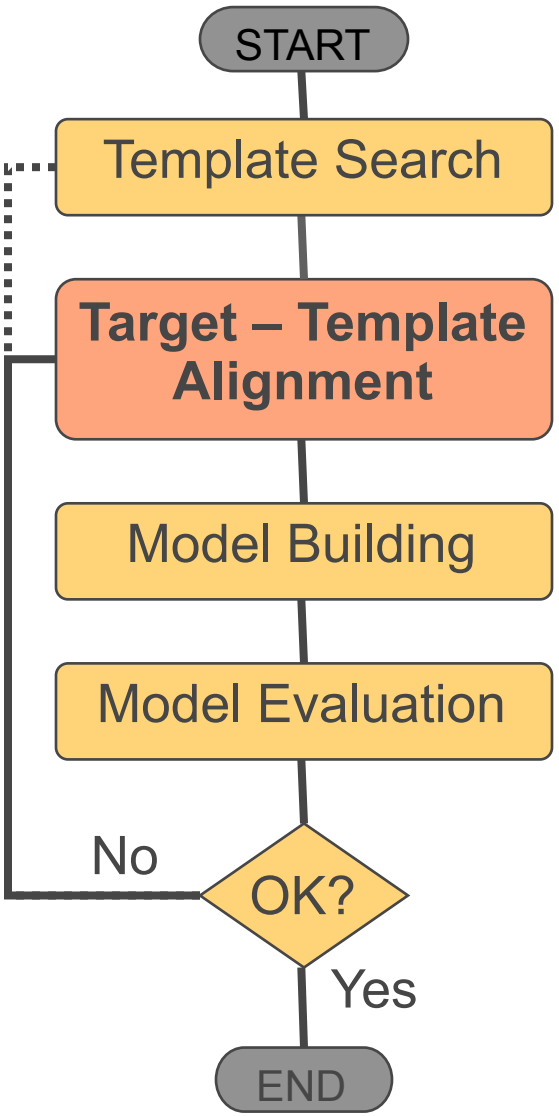
mod9v8

Sequence-Structure search

GenThreader

http://bioinf.cs.ucl.ac.uk/psipred/

Target-Template Alignment



Sequence-Sequence search

BLAST

http://www.ncbi.nlm.nih.gov/blast/

Profile-Sequence search

PSI-BLAST

http://www.ncbi.nlm.nih.gov/blast/

Profile-Profile search

pp_scan

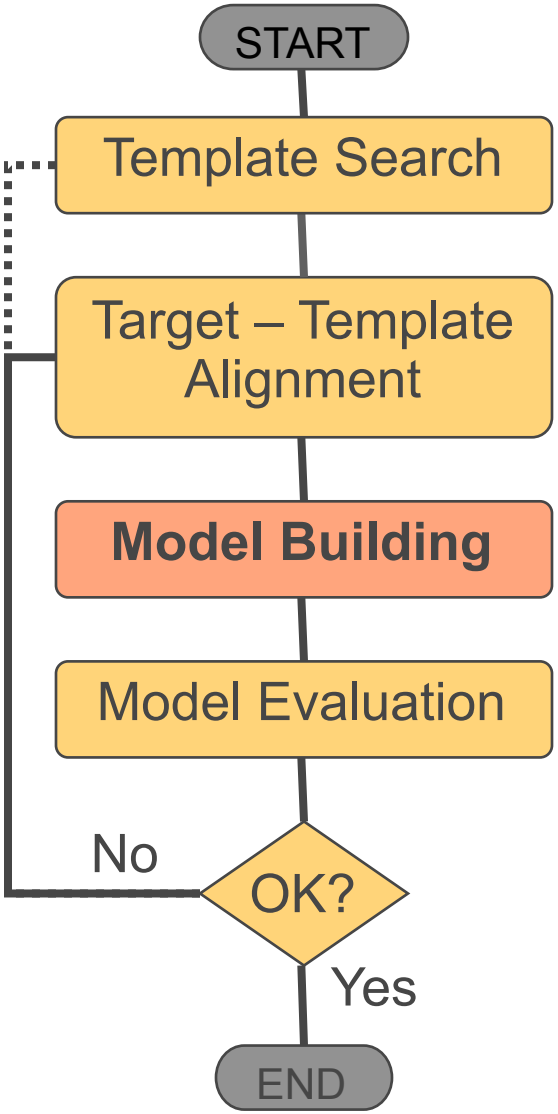
mod9v8

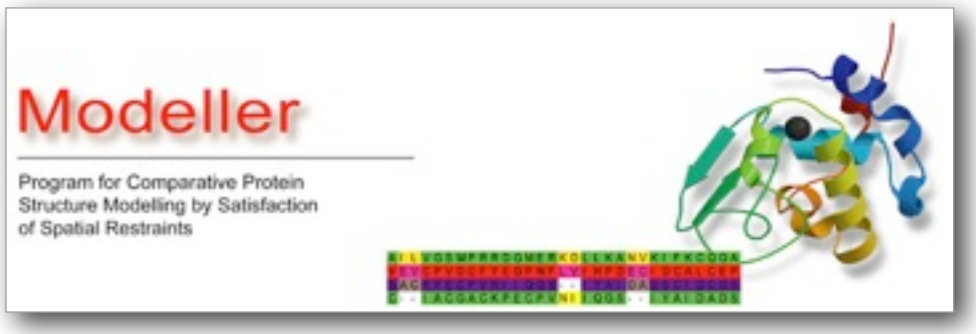
Sequence-Structure search

GenThreader

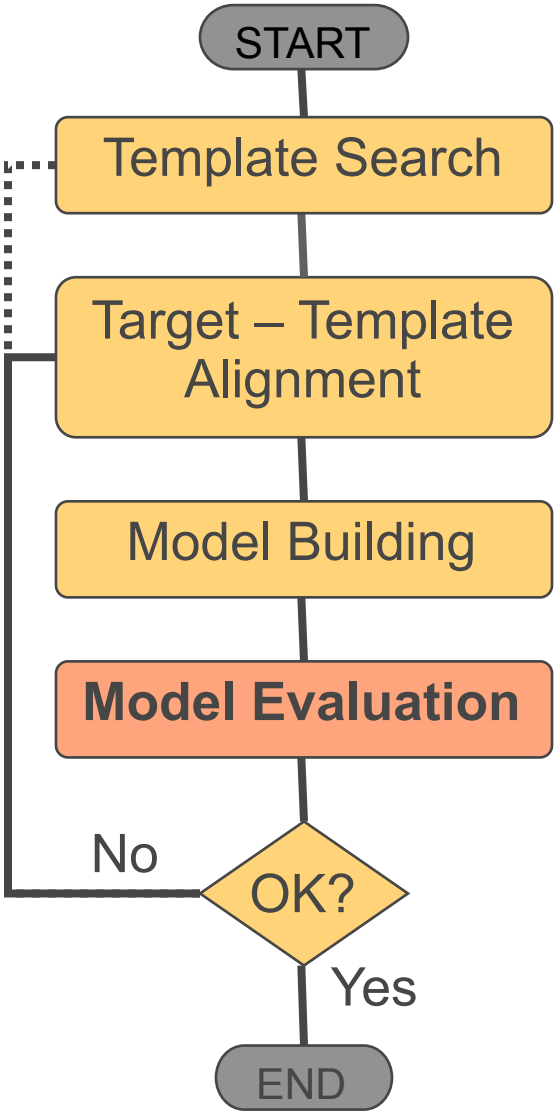
http://bioinf.cs.ucl.ac.uk/psipred/

Model Building





Model Evaluation



Classical potential of mean force

PROSA-WEB

https://prosa.services.came.sbg.ac.at

Protein conformation free energies

DFIRE

http://sparks.informatics.iupui.edu/yueyang/DFIRE/dDFIRE-service

Discrete optimized protein energy

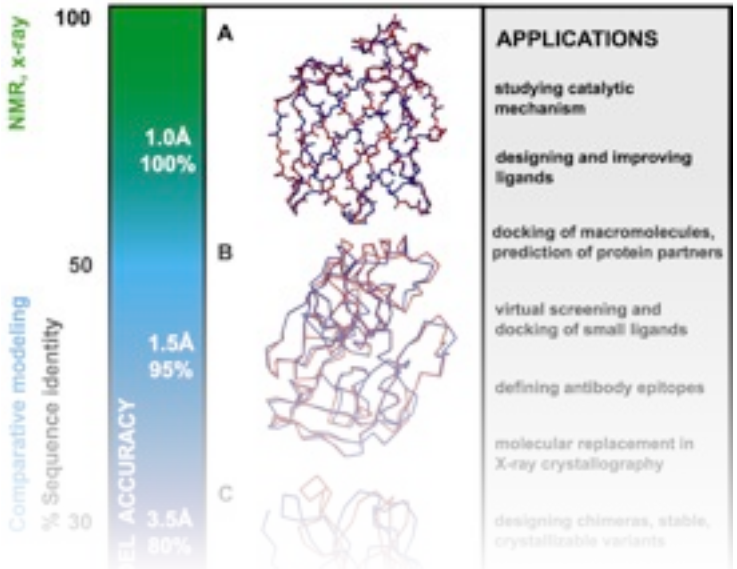
assess_dope

mod9v8

Stereochemistry (not necessary but useful)

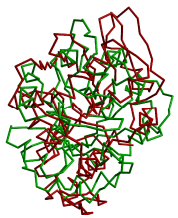
PROCHECK

http://www.ebi.ac.uk/thornton-srv/software/PROCHECK/

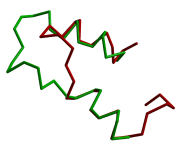


Accuracy and applicability of comparative models

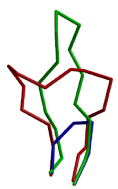
Comparative modeling by satisfaction of spatial restraints Types of errors and their impact



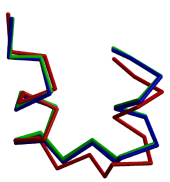
Wrong fold



Miss alignments

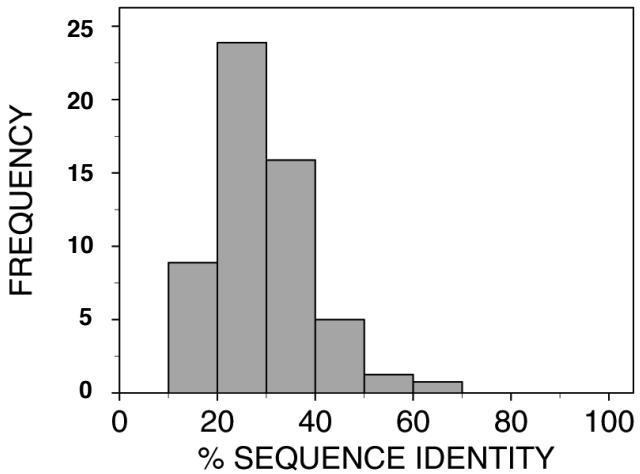


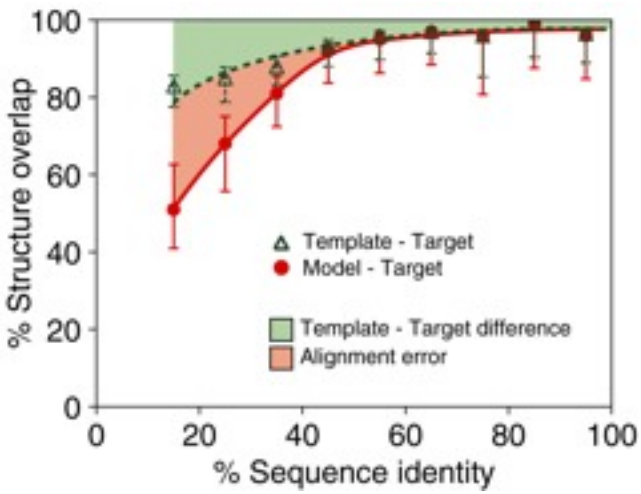
Loop regions



Rigid body distortions

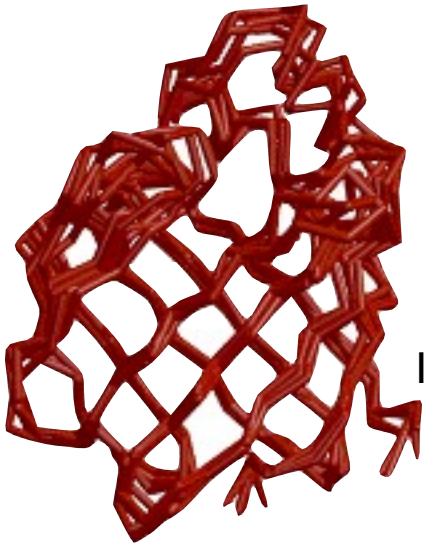
Side-chain packing





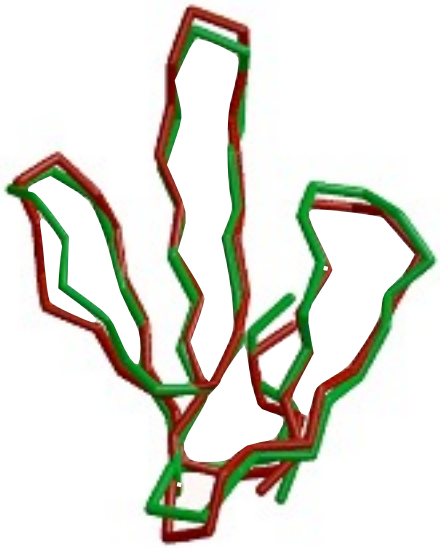
Marti-Renom etal. Ann Rev Biophys Biomol Struct (2000) 29, 291

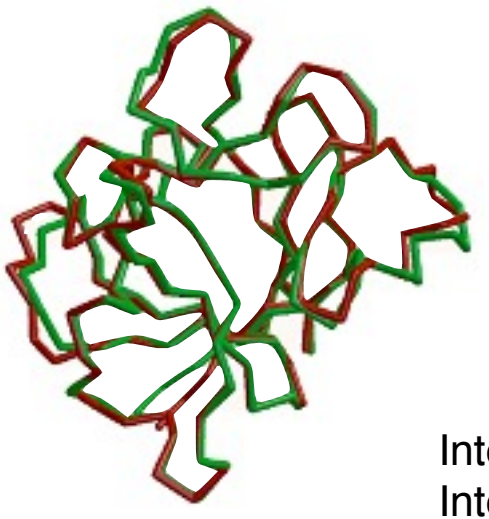
“Biological” significance of modeling errors



NMR
Ileal lipid-binding protein
1eal

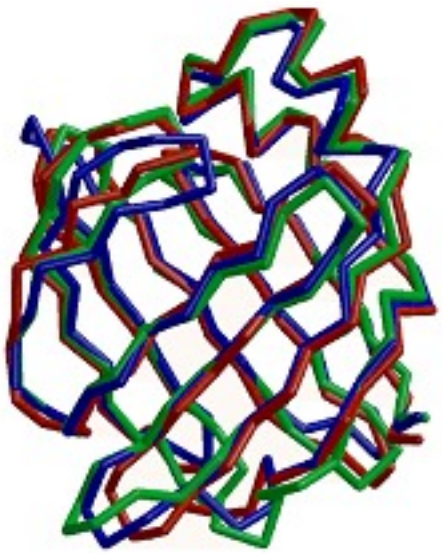
NMR – X-RAY
Erabutoxin 3ebx
Erabutoxin 1era



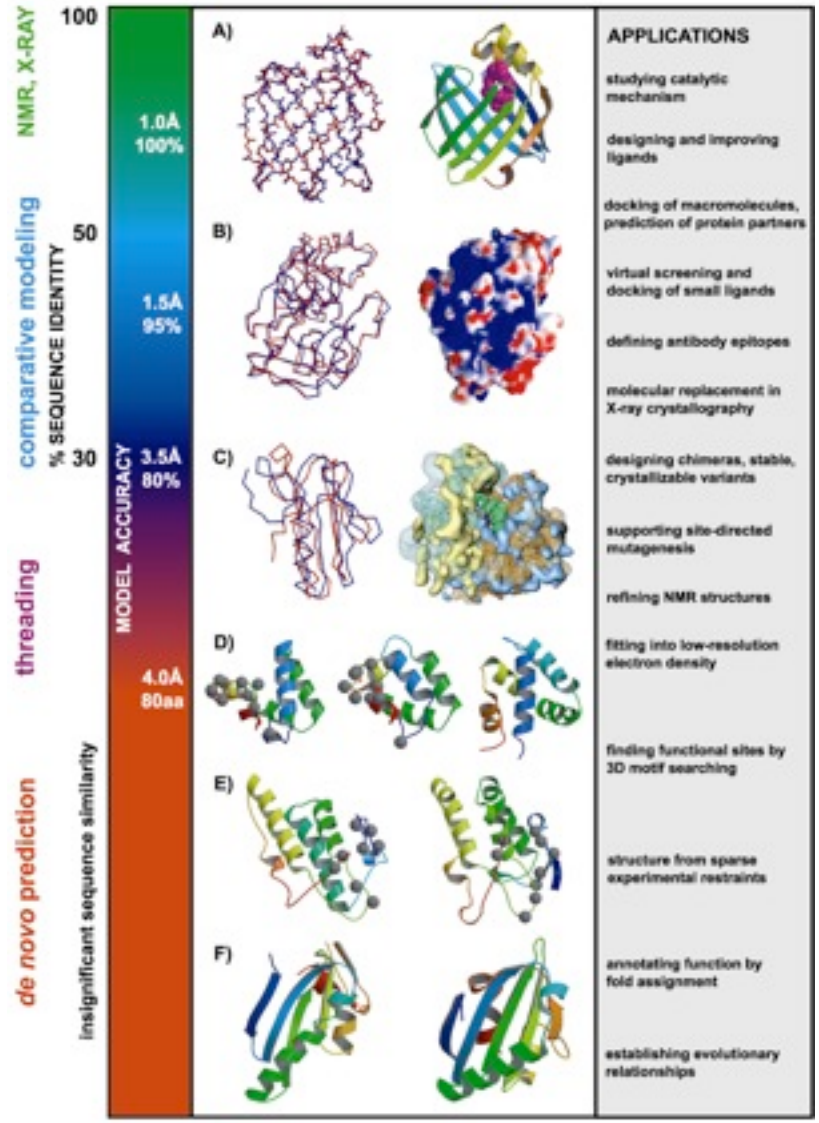


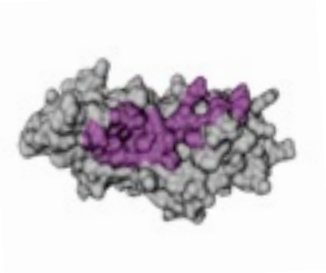
X-RAY
Interleukin 1β 41bi (2.9Å)
Interleukin 1β 2mib (2.8Å)

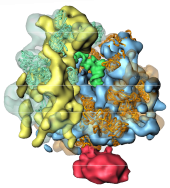
CRABPII 1opbB
FABP 1ftpA
ALBP 1lib
40% seq. id.



Utility of protein structure models, despite errors

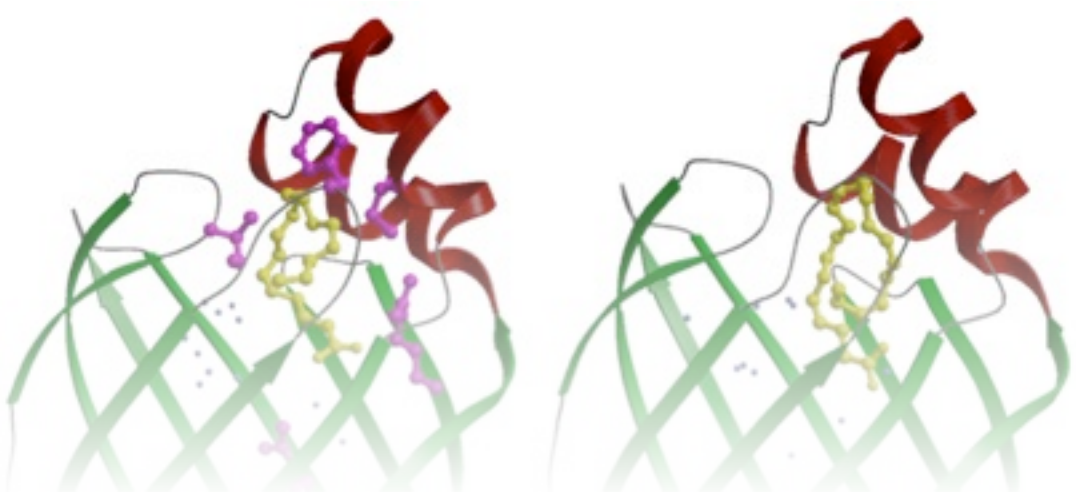






T. cruzi

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



Modeling genes

What is the physiological ligand of Brain Lipid-Binding Protein?

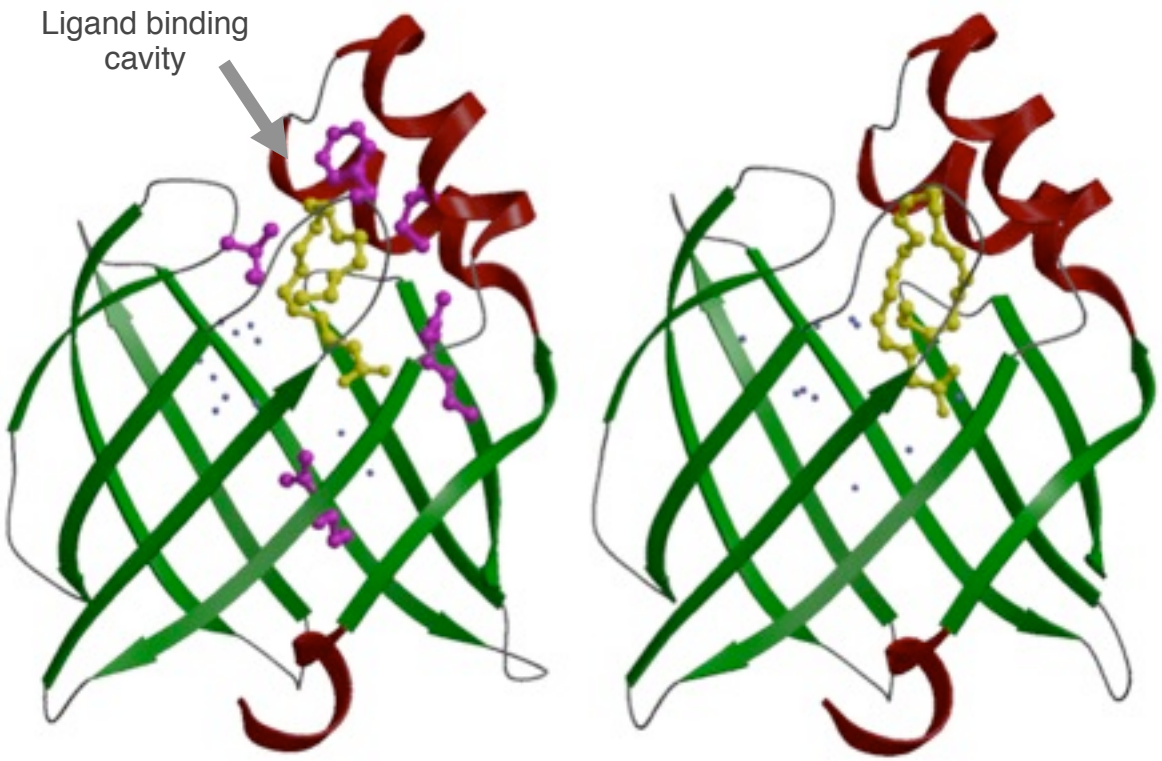
Predicting features of a model that are not present in the template

BLBP/oleic acid

Cavity is not filled

BLBP/docosahexaenoic acid

Cavity is filled



L. Xu, R. Sánchez, A. Šali, N. Heintz, J. Biol. Chem. 271, 24711, 1996.

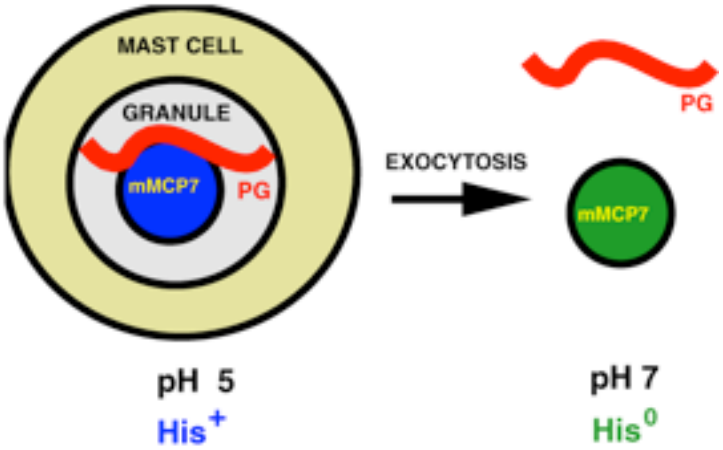
1. BLBP binds fatty acids.

2. Build a 3D model.

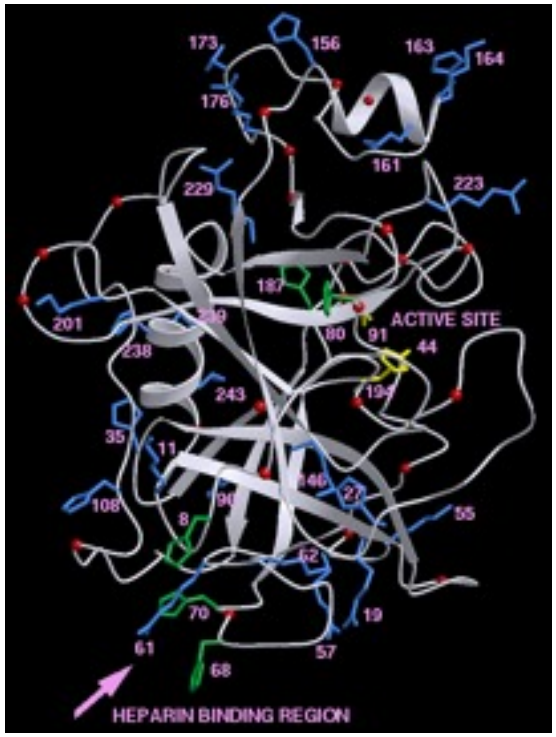
3. Find the fatty acid that fits most snuggly into the ligand binding cavity.

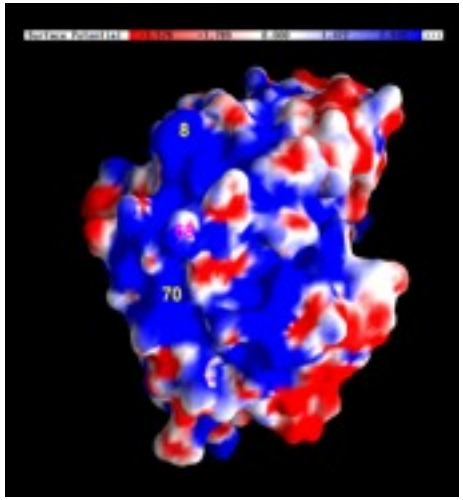
Predicting features of a model that are not present in the template

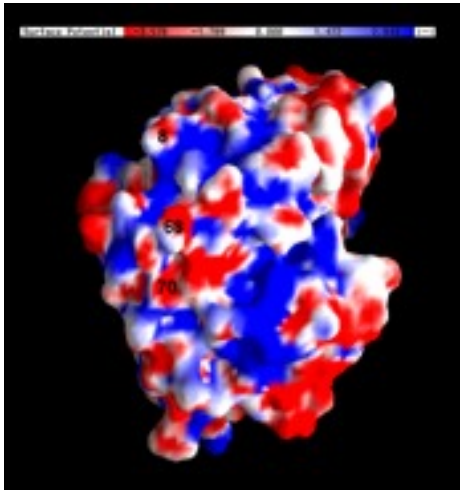
1. mMCPs bind negatively charged proteoglycans through electrostatic interactions
2. Comparative models used to find clusters of positively charged surface residues.
3. Tested by site-directed mutagenesis.



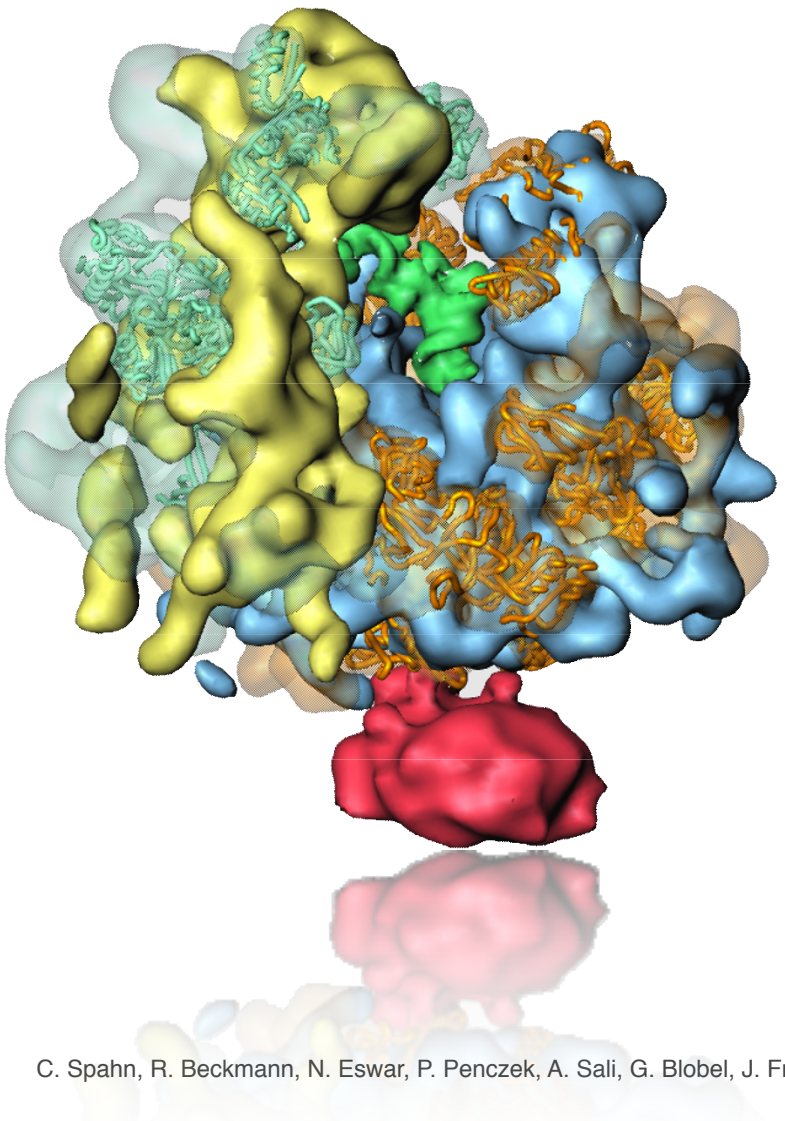
Huang *et al*. *J. Clin. Immunol*. **18**,169,1998.
Matsumoto *et al*. *J.Biol.Chem*. **270**,19524,1995.
Šali *et al*. *J. Biol. Chem*. **268**, 9023, 1993.







S. cerevisiae ribosome



Fitting of comparative models into 15Å cryo-electron density map.

43 proteins could be modeled on 20-56% seq.id. to a known structure.

The modeled fraction of the proteins ranges from 34-99%.

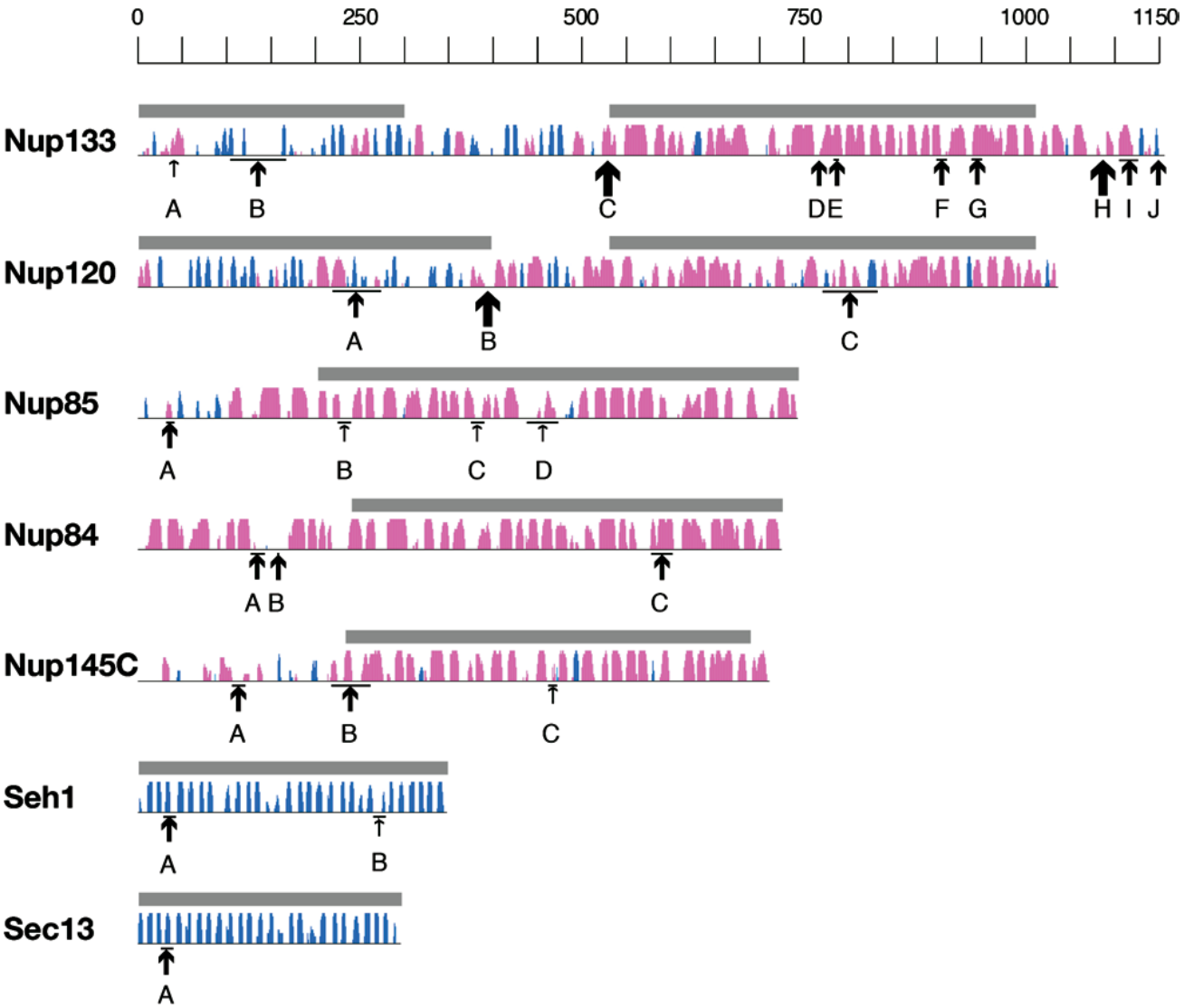
C. Spahn, R. Beckmann, N. Eswar, P. Penczek, A. Sali, G. Blobel, J. Frank. Cell 107, 361-372, 2001.

Common Evolutionary Origin of Coated Vesicles and Nuclear Pore Complexes

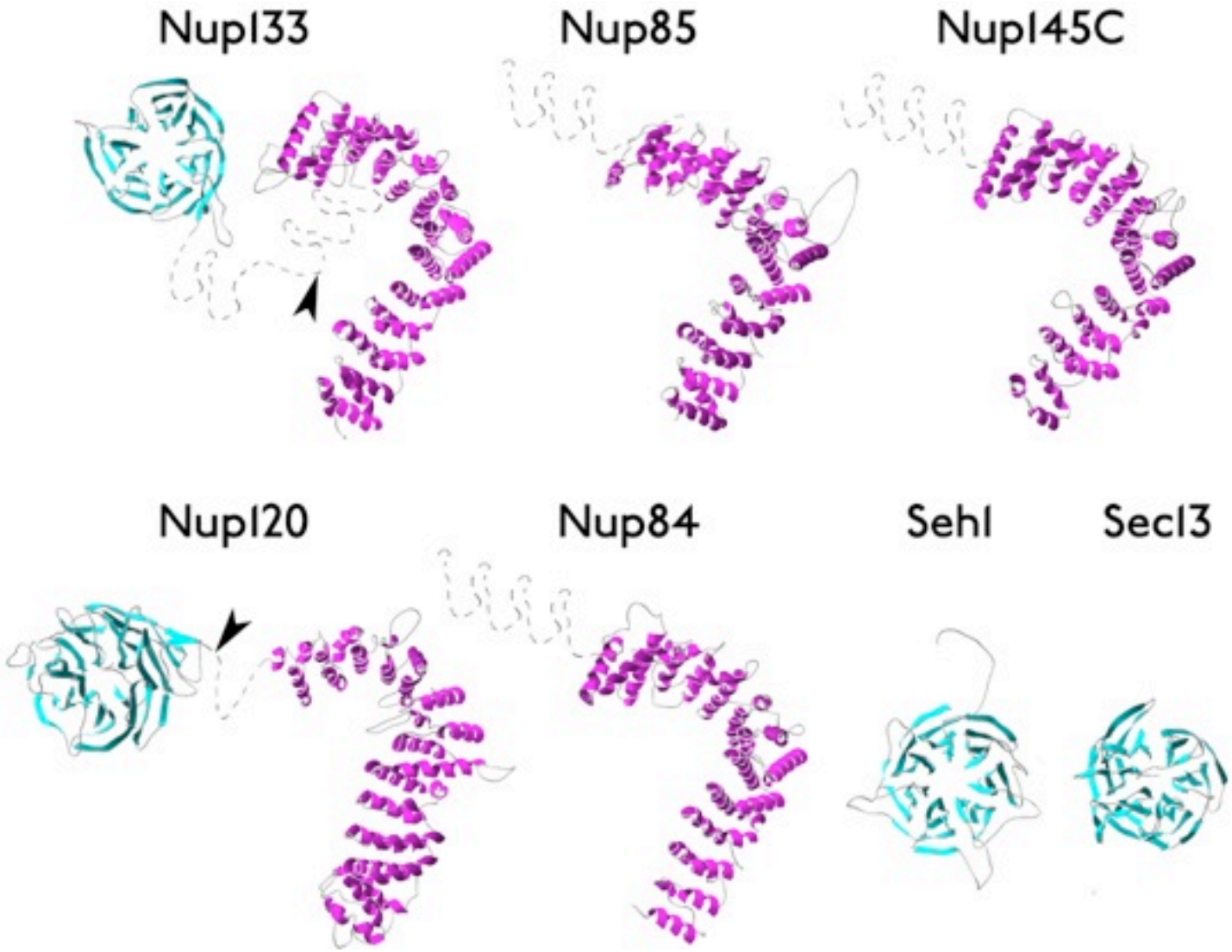
mGenThreader + SALIGN + MOULDER

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

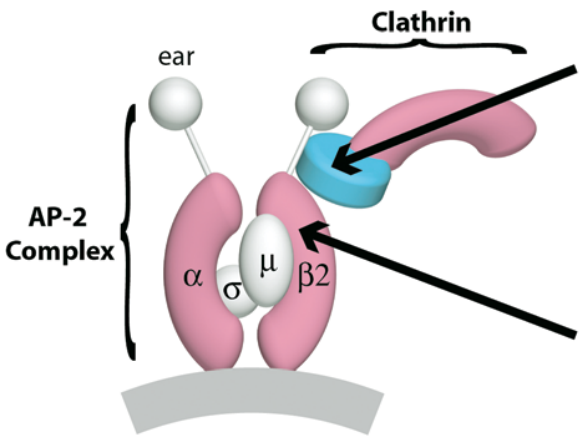
yNup84 complex proteins

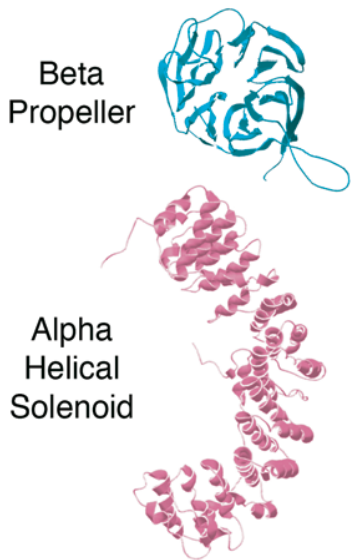


All Nucleoporins in the Nup84 Complex are Predicted to Contain β-Propeller and/or α-Solenoid Folds



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



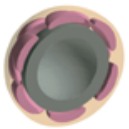


VARIANTS

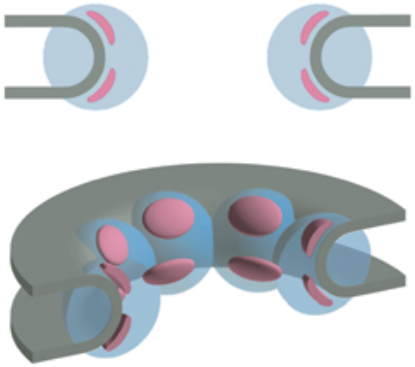
CCVs	*COPI*	*COPII*	*NPC*
Clathrin-N	α-subunit-N β'-subunit-N	Sec13 (Seh1) Sec31-N	Sec13 Seh1 Nup133-N Nup120-N
Clathrin-C α/γ/δ/ε-subunit β-subunits	α-subunit-C β'-subunit-C β-subunit γ-subunit	Sec31-C	Nup133-C Nup120-C Nup85 Nup84 Nup145C

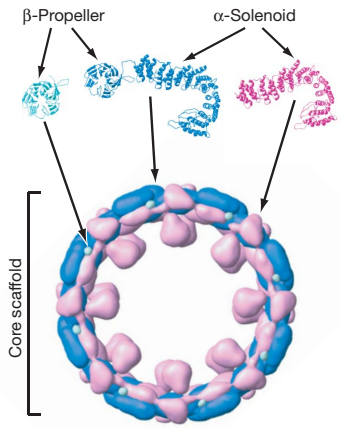
NPC and Coated Vesicles Both Associate with Membranes

Coated Vesicle



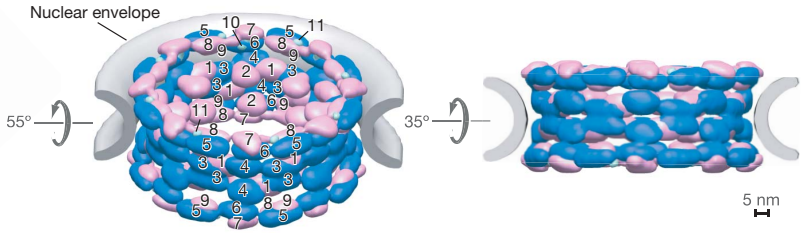
NPC model





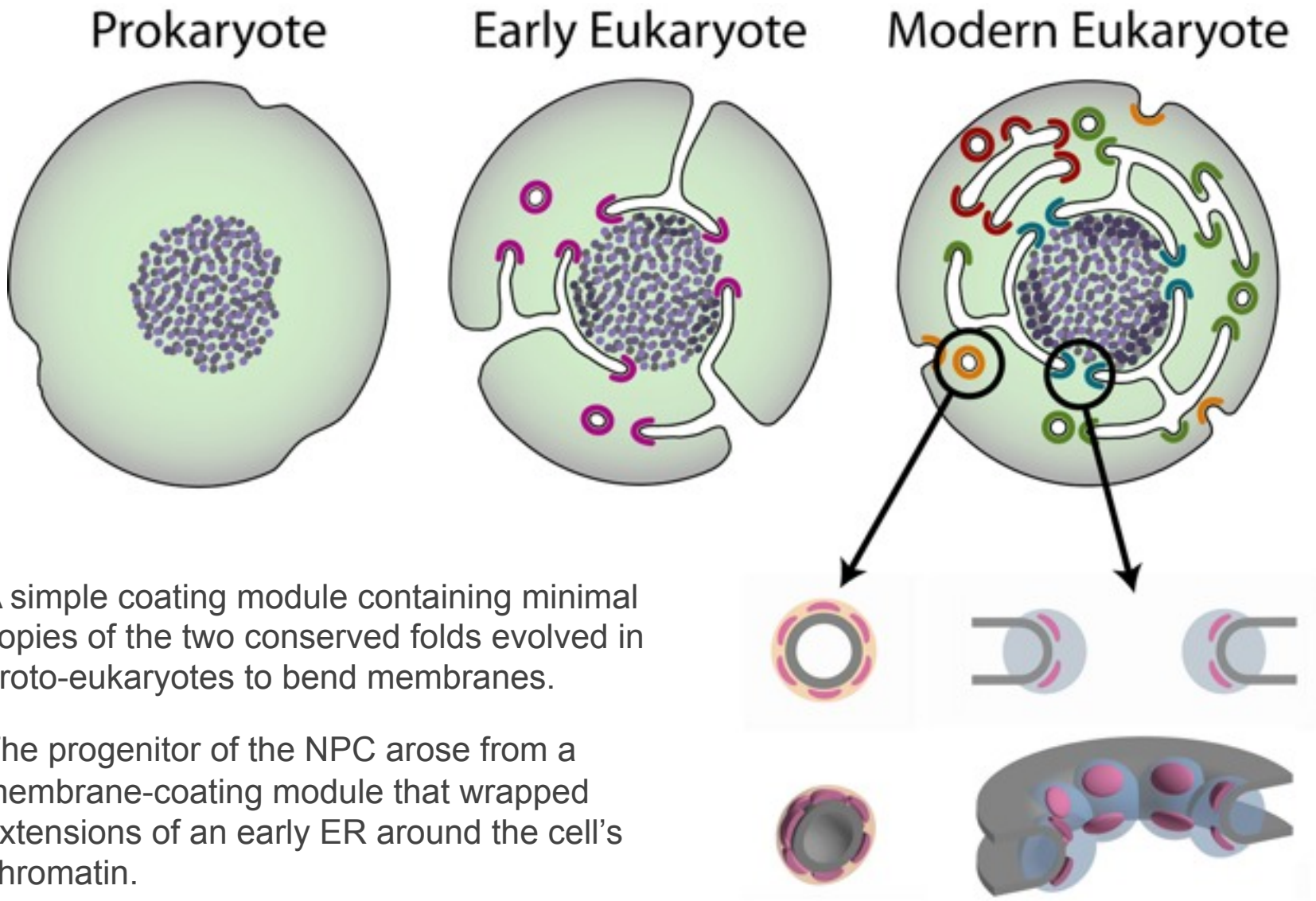
Nup 84 complex

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



Alber et al. The molecular architecture of the nuclear pore complex. Nature (2007) vol. 450 (7170) pp. 695-701

A Common Evolutionary Origin for Nuclear Pore Complexes and Coated Vesicles?
The proto-coatomer hypothesis

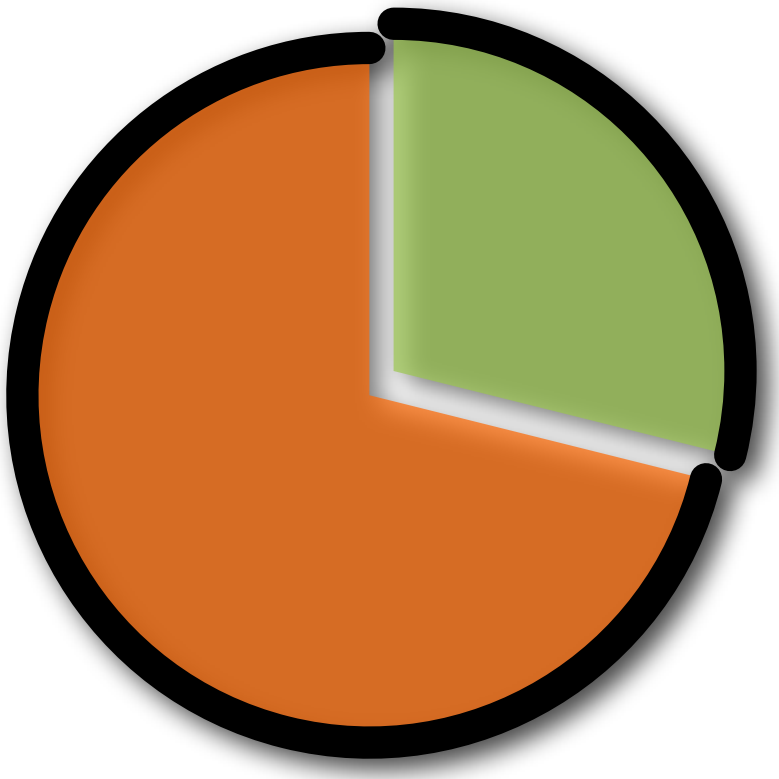


A simple coating module containing minimal copies of the two conserved folds evolved in proto-eukaryotes to bend membranes.

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

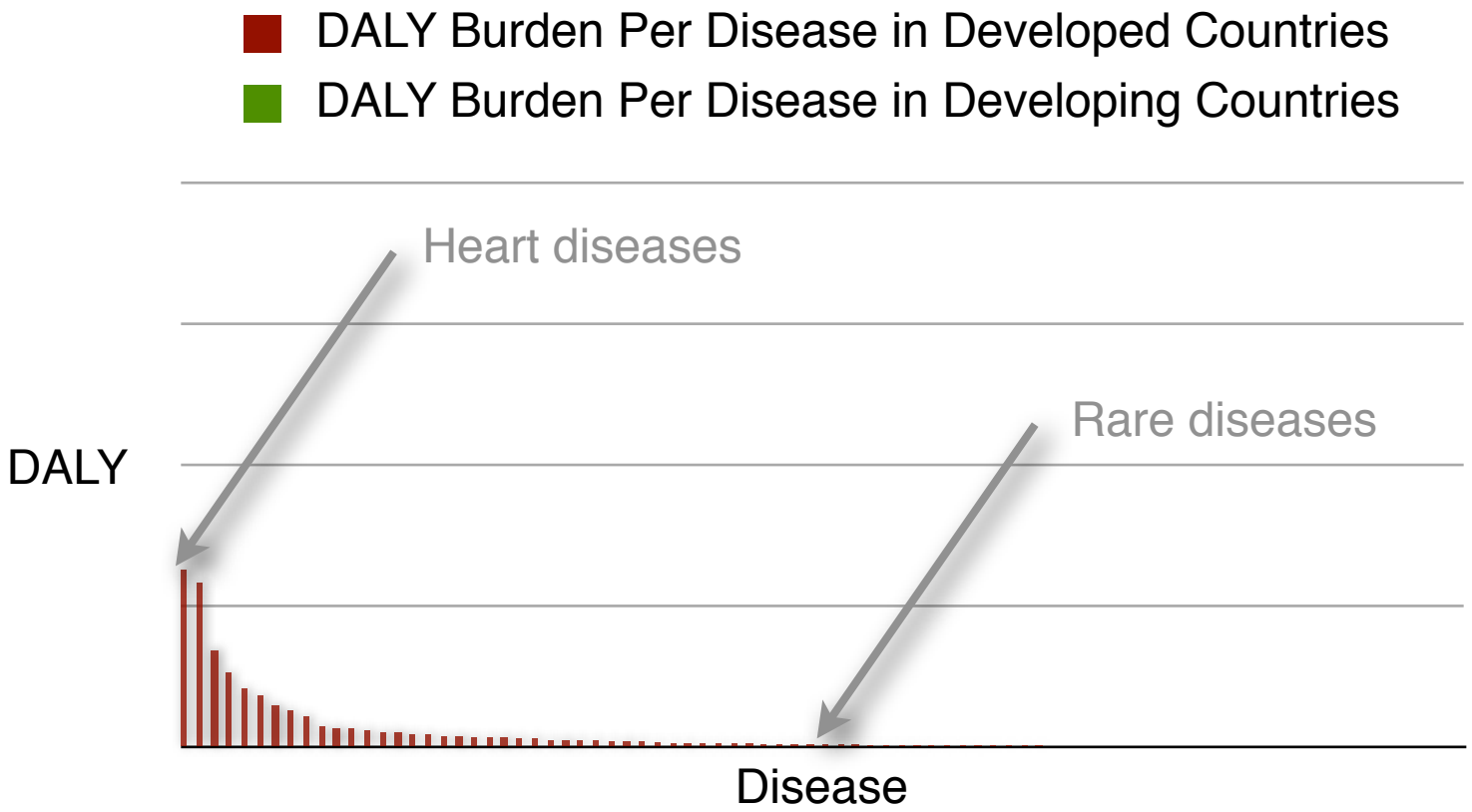
Tropical Disease Initiative (TDI)

Predicting binding sites in protein structure models.



http://www.tropicaldisease.org

Need is High in the Tail



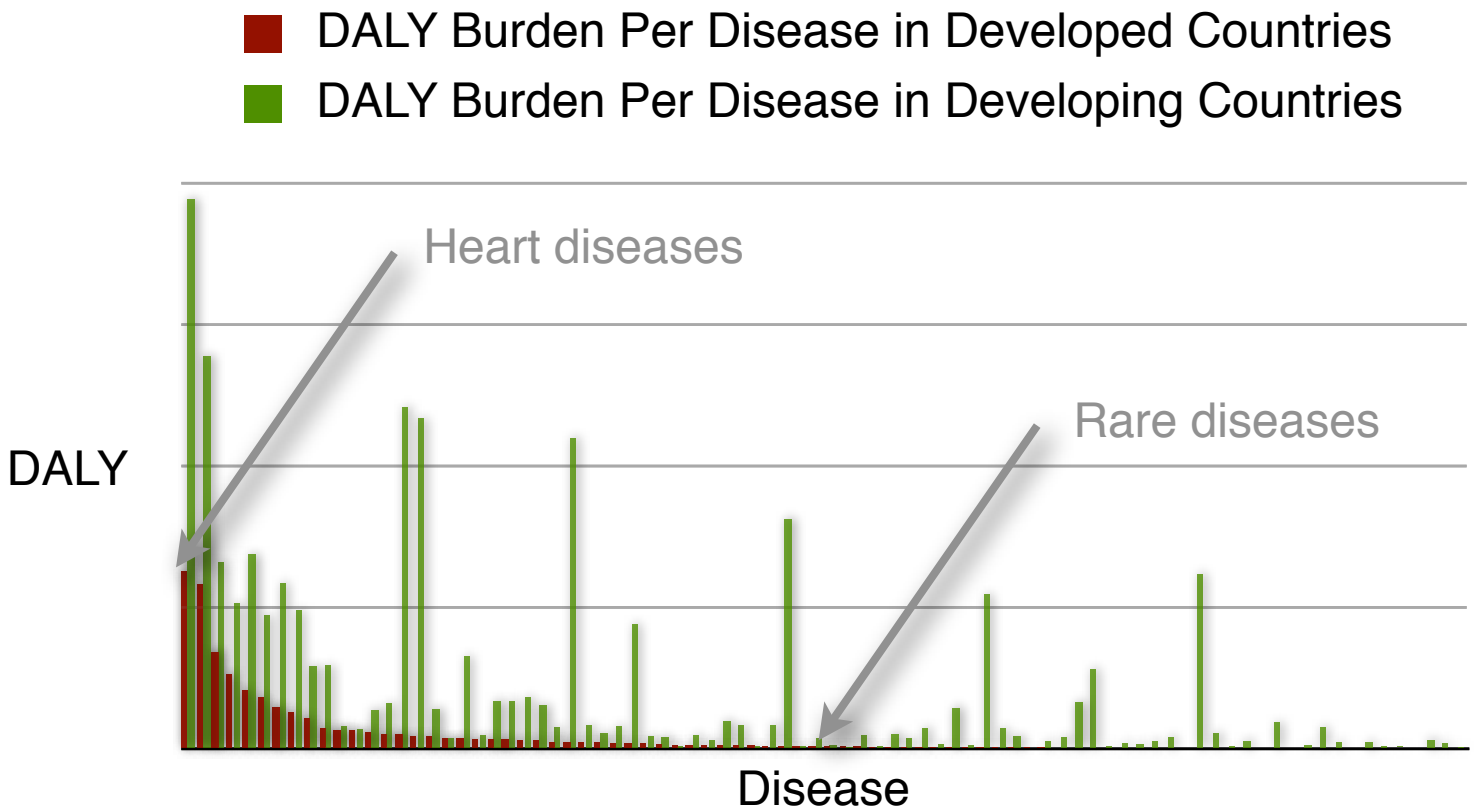
Disease data taken from WHO, *World Health Report 2004*

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.

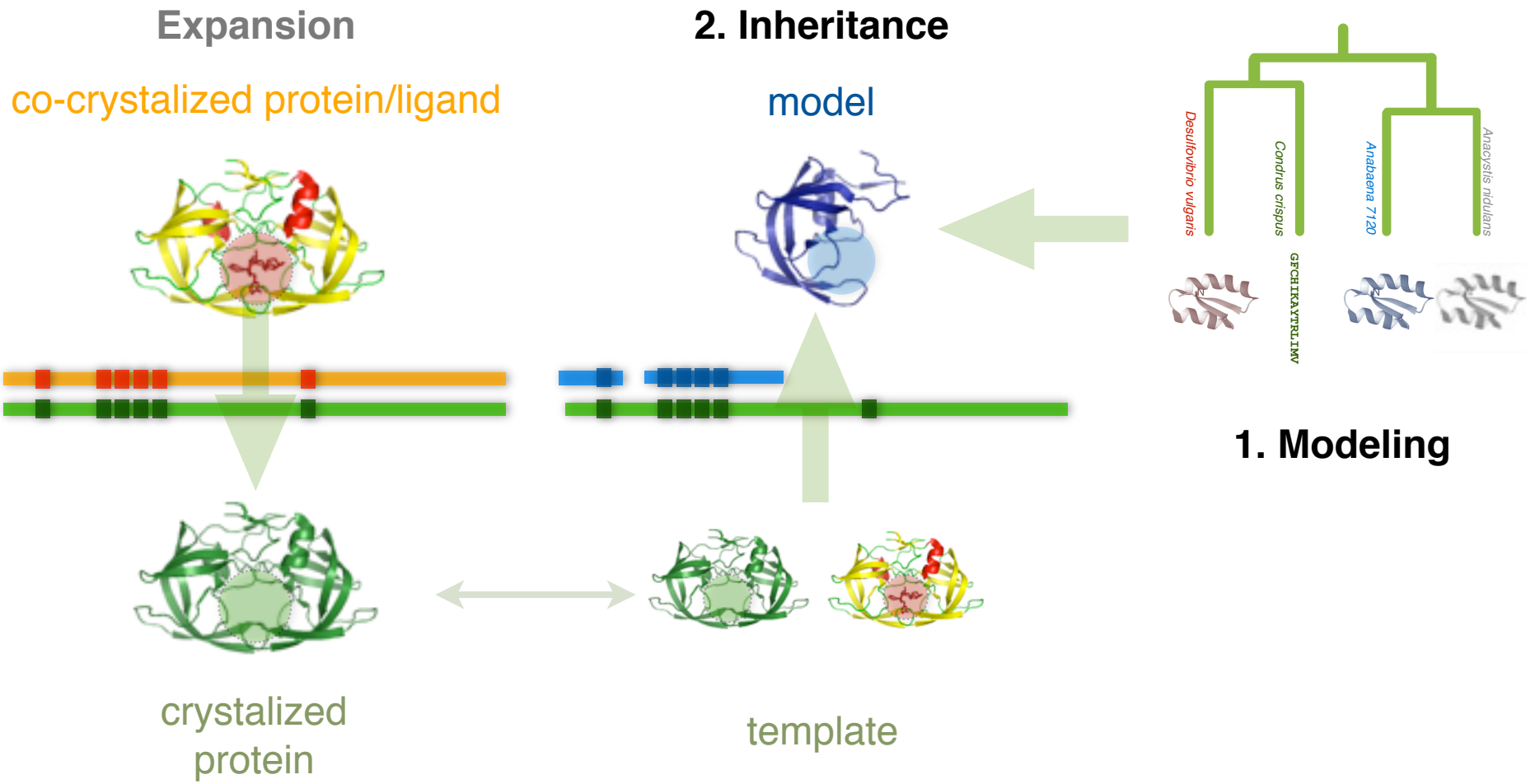
"Unprofitable" Diseases and Global DALY (in 1000's)

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

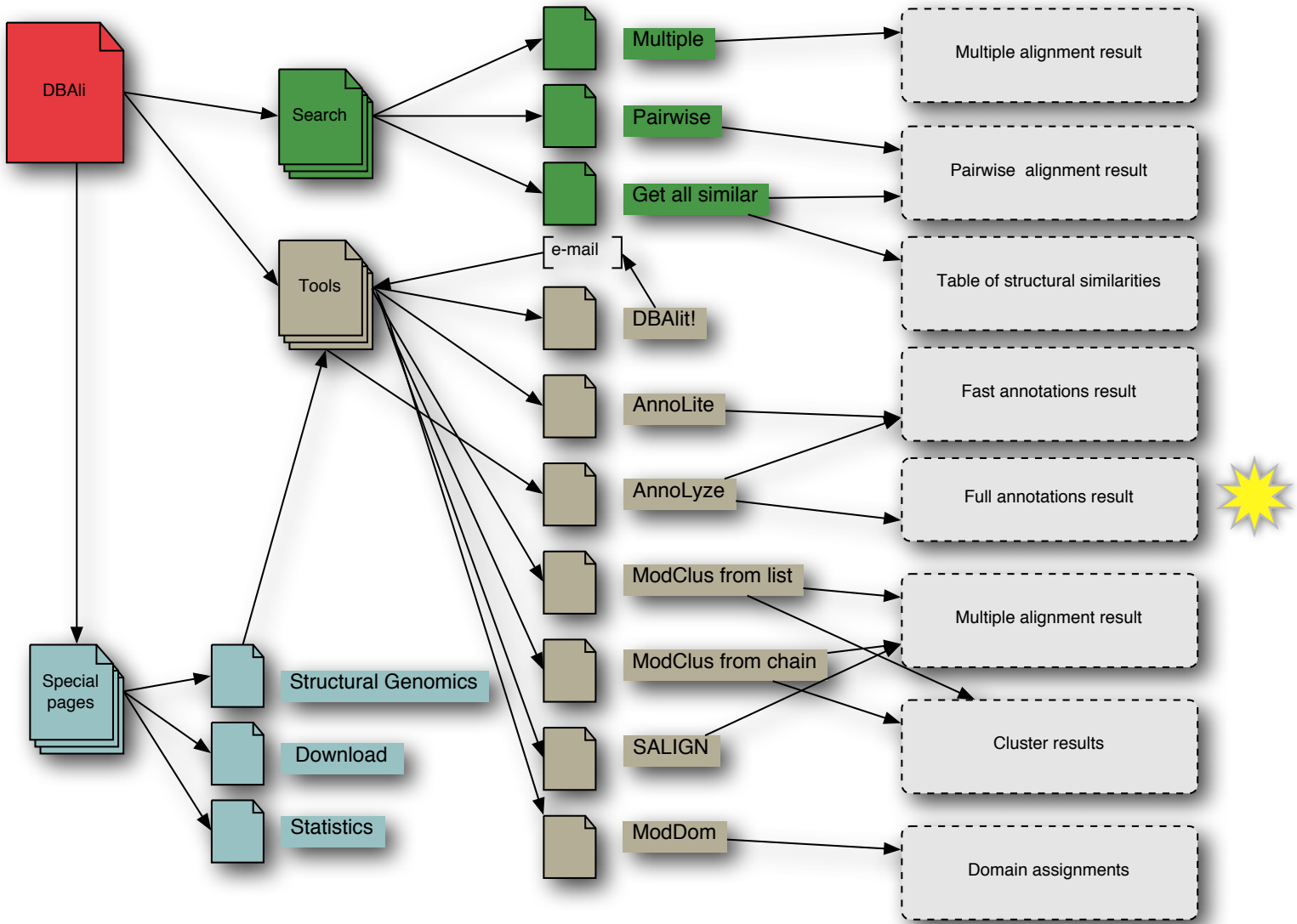
Disease data taken from WHO, *World Health Report 2004*
DALY - Disability adjusted life year in 1000's.
* Officially listed in the WHO Tropical Disease Research disease portfolio.

Comparative docking



DBAli$_{v2.0}$ database

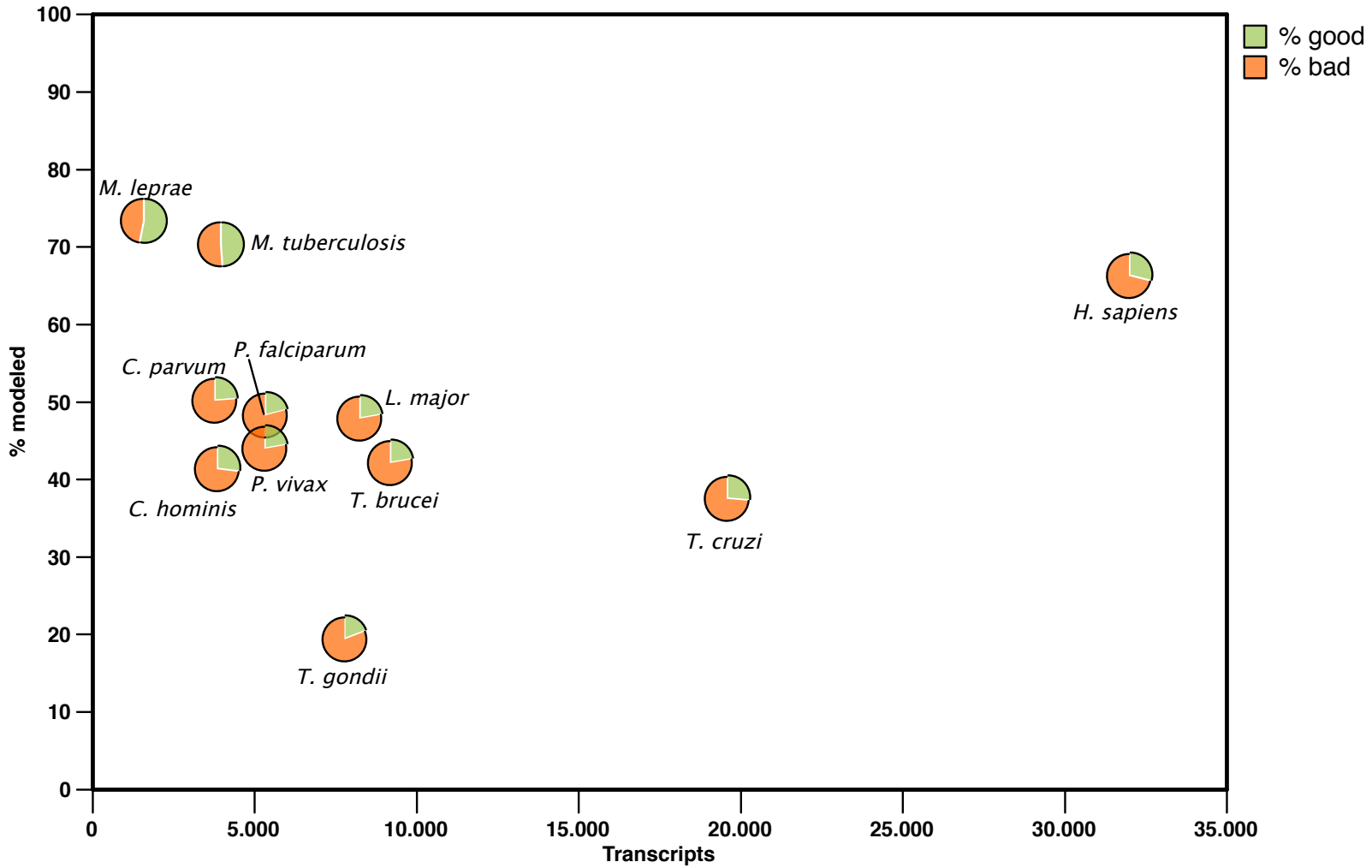
http://www.dbali.org



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

Modeling Genomes

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



A good model has MPQS of 1.0 or higher

Summary table

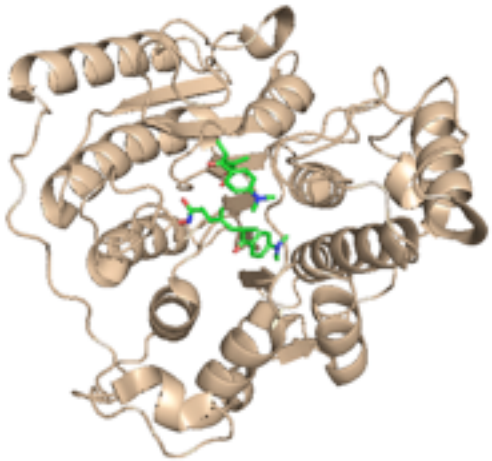
models with inherited ligands

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

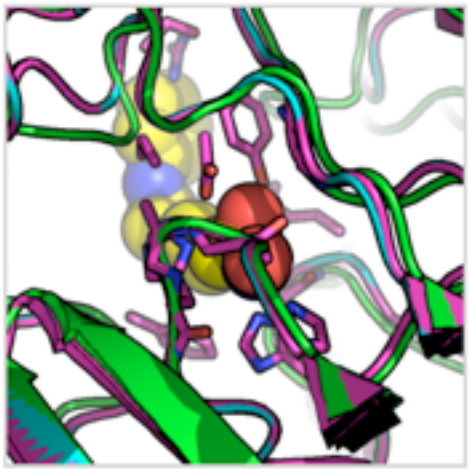
	Transcripts	Modeled targets	Selected models	Inherited ligands	Similar to a drug	Drugs
C. hominis	3,886	1,614	666	197	20	13
C. parvum	3,806	1,918	742	232	24	13
L. major	8,274	3,975	1,409	478	43	20
M. leprae	1,605	1,178	893	310	25	6
M. tuberculosis	3,991	2,808	1,608	365	30	10
P. falciparum	5,363	2,599	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.

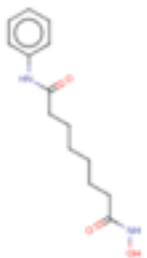


PDB	ID	Template	SR	Model	G	Ligand	Exact	SupStr	SubStr	Similar
1c3sA	83.33/80.00	1t64A	36.00/1.47	LmjF21.0680.1.pdb	90.91/100.00	SHH	DB02546	DB02546	DB02546	DB02546



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*

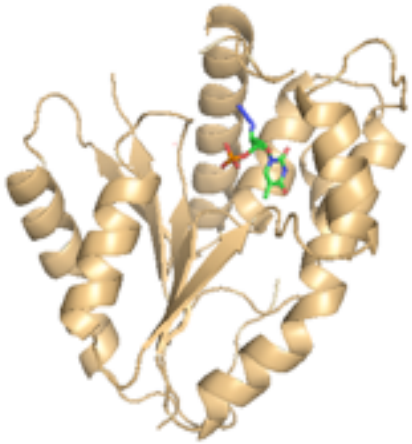
Departments of *Parasite Biochemistry and Cell Biology, ‡Medicinal Chemistry, and §Natural Products Drug Discovery, Merck Research Laboratories, P.O. Box 2000, Rahway, NJ 07065

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

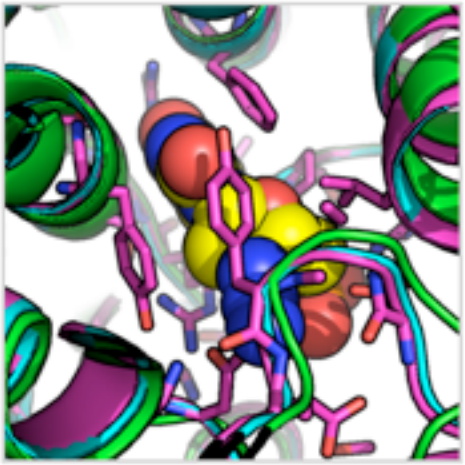
Antimalarial and Antileishmanial Activities of Aroyl-Pyrrolyl-Hydroxyamides, a New Class of Histone Deacetylase Inhibitors

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**

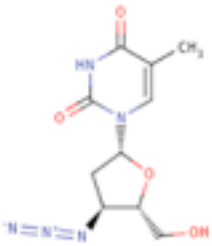
Small Molecule; Approved

Drug categories:

Anti-HIV Agents
Antimetabolites
Nucleoside and Nucleotide Reverse Transcriptase Inhibitors

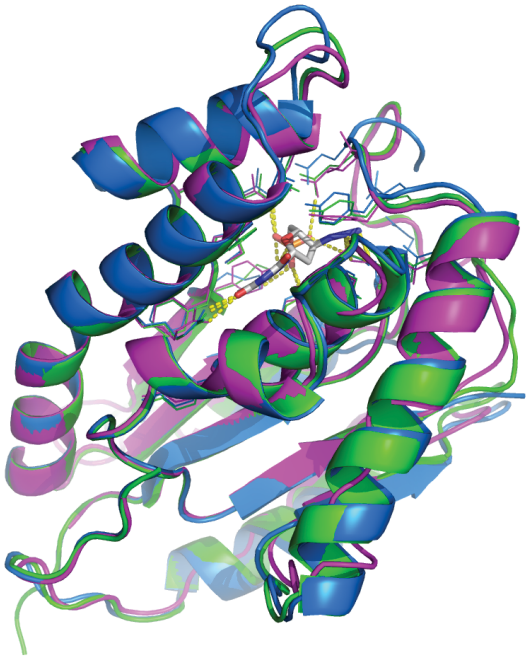
Drug Indication:

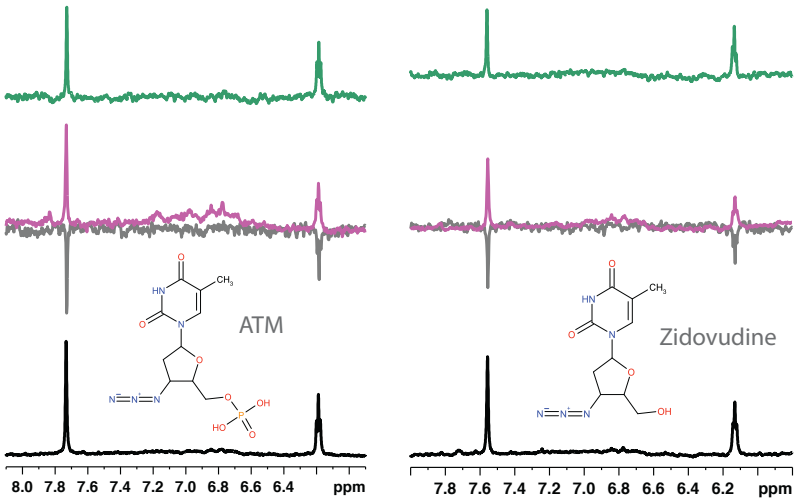
For the treatment of human immunovirus (HIV) infections.

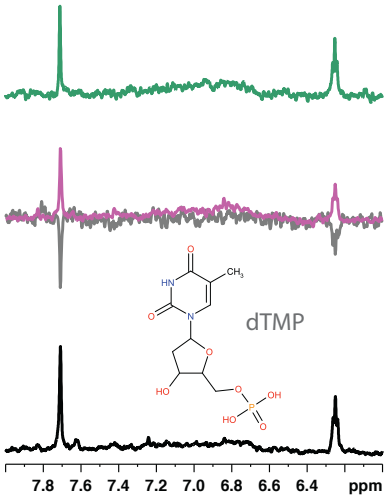


P. falciparum tymydilate kinase + zidovudine

NMR Water-LOGSY and STD experiments



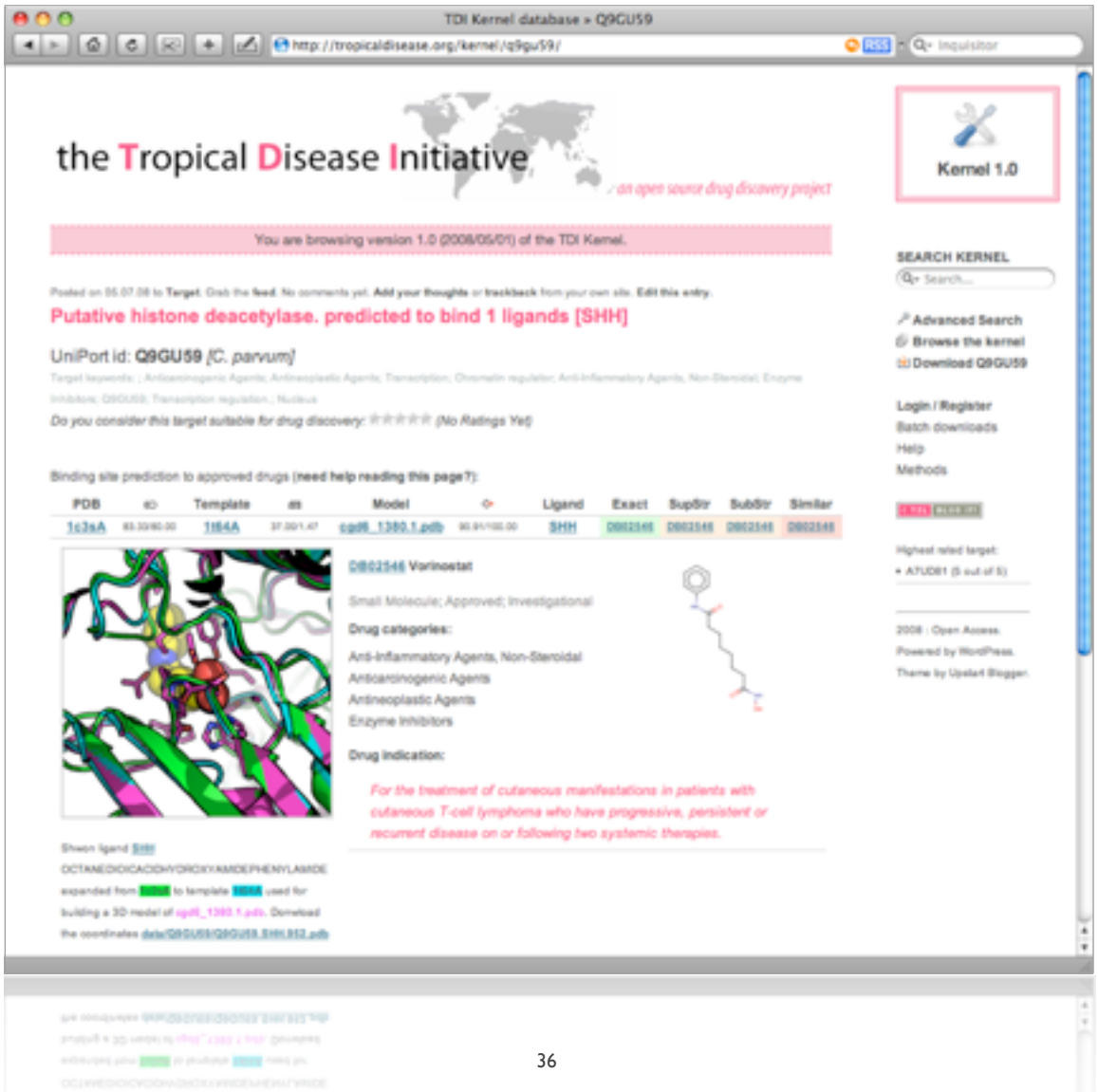




Leticia Ortí, Rodrigo J. Carbajo, and Antonio Pineda-Lucena

TDI's kernel

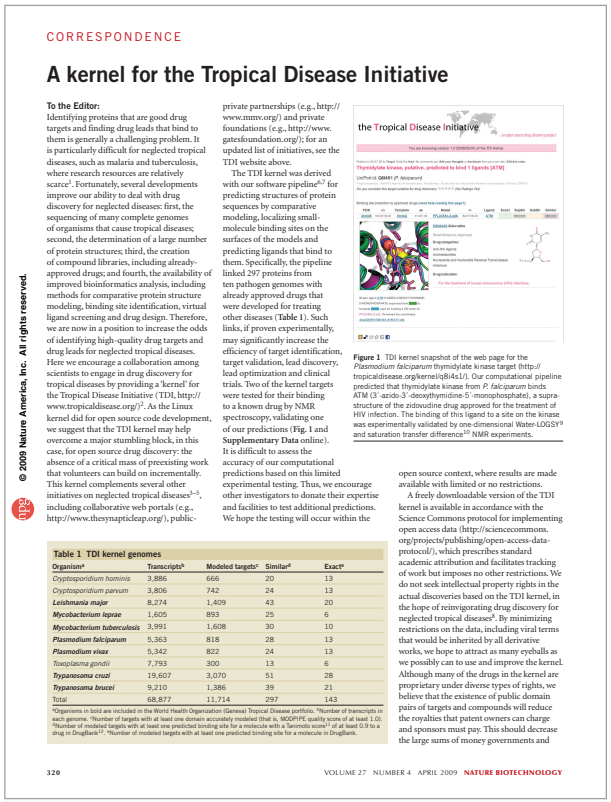
http://tropicaldisease.org/kernel



TDI's kernel

http://tropicaldisease.org/kernel

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

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-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.

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

OPEN ACCESS Freely available online PLoS NEGLECTED TROPICAL DISEASES

A Kernel for Open Source Drug Discovery in Tropical Diseases

Leticia Orti[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: 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

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: salilab@salilab.org (AS); mmarti@cipf.es (MAM-R)

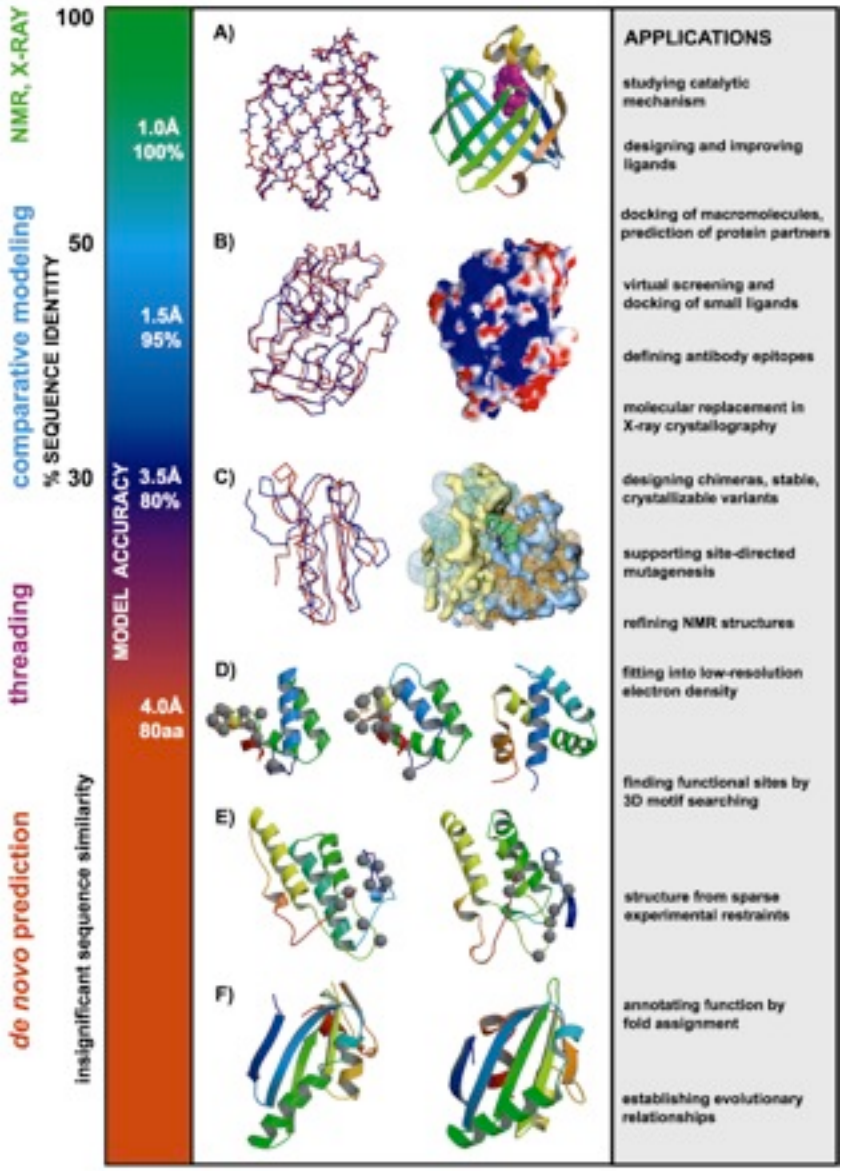
¤ Current address: DuPont Knowledge Center, Hyderabad, India

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

“take home” message



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