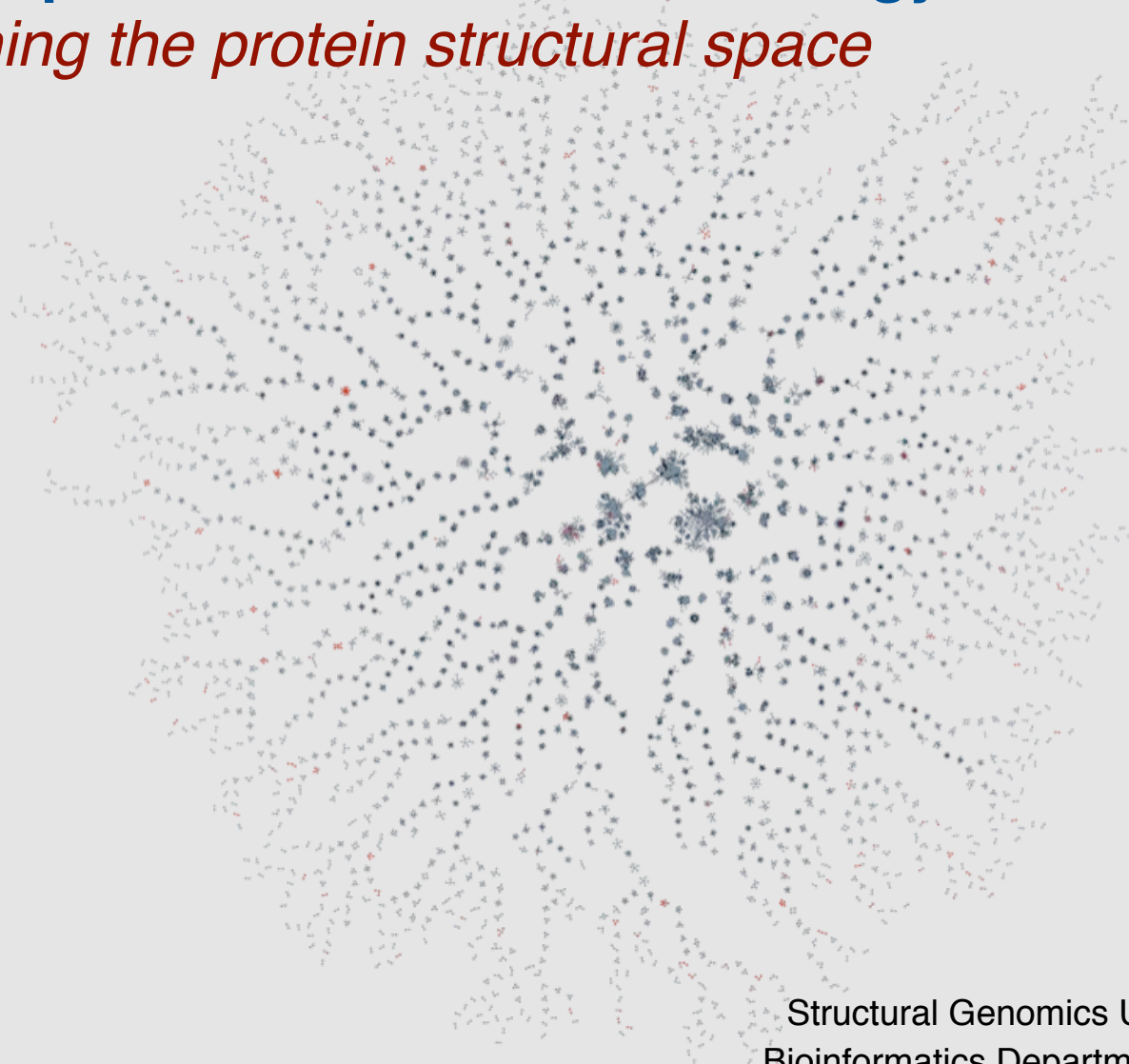


# Comparative STRUCTURE Biology

## *Mining the protein structural space*



**Marc A. Marti-Renom**

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

Structural Genomics Unit  
Bioinformatics Department

Prince Felipe Research Center (CIPF), Valencia, Spain



# Bioinformatics Department <http://bioinfo.cipf.es>



**Functional Genomics**  
Dr. Joaquín Dopazo  
[jdopazo@cipf.es](mailto:jdopazo@cipf.es)

## GEPAS



## BABELOMICS

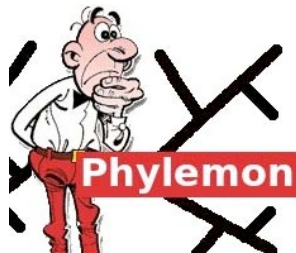
<http://gepas.bioinfo.cipf.es>  
Gene Expression Pattern Analysis Suite

<http://pupasuite.bioinfo.cipf.es>  
SNP Analysis Suite

<http://babelomics.bioinfo.cipf.es>  
Functional Profiling Analysis Suite



**Comparative Genomics**  
**Pharmacogenomics**  
Dr. Hernán Dopazo  
[hdopazo@cipf.es](mailto:hdopazo@cipf.es)



<http://phylemon.bioinfo.cipf.es>  
Molecular Evolution Analysis Suite



**Structural Genomics**  
Dr. Marc A. Martí-Renom  
[mmarti@cipf.es](mailto:mmarti@cipf.es)

**DBAli**<sup>v2.0</sup>  
Home

<http://www.dbali.org>  
Structural Biology Analysis Suite

# Bioinformatics Department <http://bioinfo.cipf.es>

Bioinformatics at CIPF

http://bioinfo.cipf.es/

Google

---

### The Department

People  
Functional Genomics Unit  
Pharmacogenomics & Comparative Genomics Unit  
Structural Genomics Unit

### Tools


DNA array data analysis  
SNPs data analysis  
Functional profiling  
Downloads

### Documents & Publications


Papers  
Communications  
Supplementary material


### Meetings & Courses


Meetings & workshops  
Courses  
On line courses  
Accommodation


**CAMDA 2007, Dec 13-14**  
Seventh international conference for the  
Critical Assessment of Microarray Data  
Analysis


### Coming events...


  
PRINCIPE FELIPE  
CENTRO DE INVESTIGACION


  
INB INSTITUTO NACIONAL  
DE BIOINFORMATICA

  
CEGEN  
Centro Nacional de Genotipo

  
ciberer  
Centro de Investigación Biomédica En Red  
de Enfermedades Raras

  
GECOBIO  
GEOGENOMICS AND  
COMPARATIVE BIOLOGY

  
Genómica  
Genómica y Bioinformática

  
INDIGO  
Integrated High-Density Fluorescence-based  
Assessment for Diagnostic applications

### Bioinformatics

DNA array analysis SNPs analysis Functional Profiling

Today's research in biomedicine cannot be understood without the concourse of bioinformatics. Our department aims to tackle biomedical problems from a system's biology perspective. Following this, the general objective we seek through the main lines of research is to relate the mutations (Pharmacogenomics and Comparative Genomics) to their effect at cellular and phenotypic level (Functional Genomics) trying to understand the mechanism of action (Structural Genomics).

### News

**May 2007** - Positions available: one biostatistician and one bioinformatician  
**April 2007** - Blast2GO, a universal Gene Ontology annotation, visualization and analysis suite for functional genomics research.  
**January 2007** - Phylemon, a suite of tools for molecular evolution, phylogenetics and phylogenomics.  
**September 2006** - Prophet, a tool for building a class predictor.  
**March 2006** - Computing journal award to the best R&D project.  
**February 2006** - New releases: GEPAS v3.0 and Babelomics v2.0. Also PupaSuite, interactive selection of optimal sets of SNPs for large-scale genotyping.  
**December 2005** - CIPF receives autelsi award: "The First European Cluster for Scientific Computing with Free software on GNU/Linux."

#### More news...

Google

☐ Search WWW ☐ Search bioinfo.cipf.es

---

**Postal address**  
Bioinformatics Department (CIPF)  
Av. Autopista del Saler, 16  
(Camino de las moreras)  
46013 Valencia, Spain  
Tel.: + 34 96 328 96 80  
Fax: + 34 96 328 97 01  
(see map)

Send comments to the webmaster. Last updated: January 15, 2007

### Spotlight Tools

**Access**  
Tool usage around the World.

**Prophet**  
A tool for building a class predictor

**PupaSuite**  
Interactive selection of optimal sets of SNPs for large-scale genotyping.

**MARMITE**  
Functional profiling with PubMed words.

**FatIGO+**  
Functional interpretation of large-scale experiments using GO, KEGG, Interpro, Transfac, CisRed...

**CAAT**  
Draw, browse, analyze and validate interactively your hierarchical clustering results.

**FatIScan**  
Detect blocks of functionally related genes (GO, KEGG) with significant coordinate (although modest) over- or under-expression.

**SIDE**  
Interactive design of Small interfering RNA.

**Blast2GO**  
A universal Gene Ontology annotation, visualization and analysis suite for functional genomics research.

### Packages

**GEPAS**  
Gene Expression Pattern Analysis Suite (v3.0).

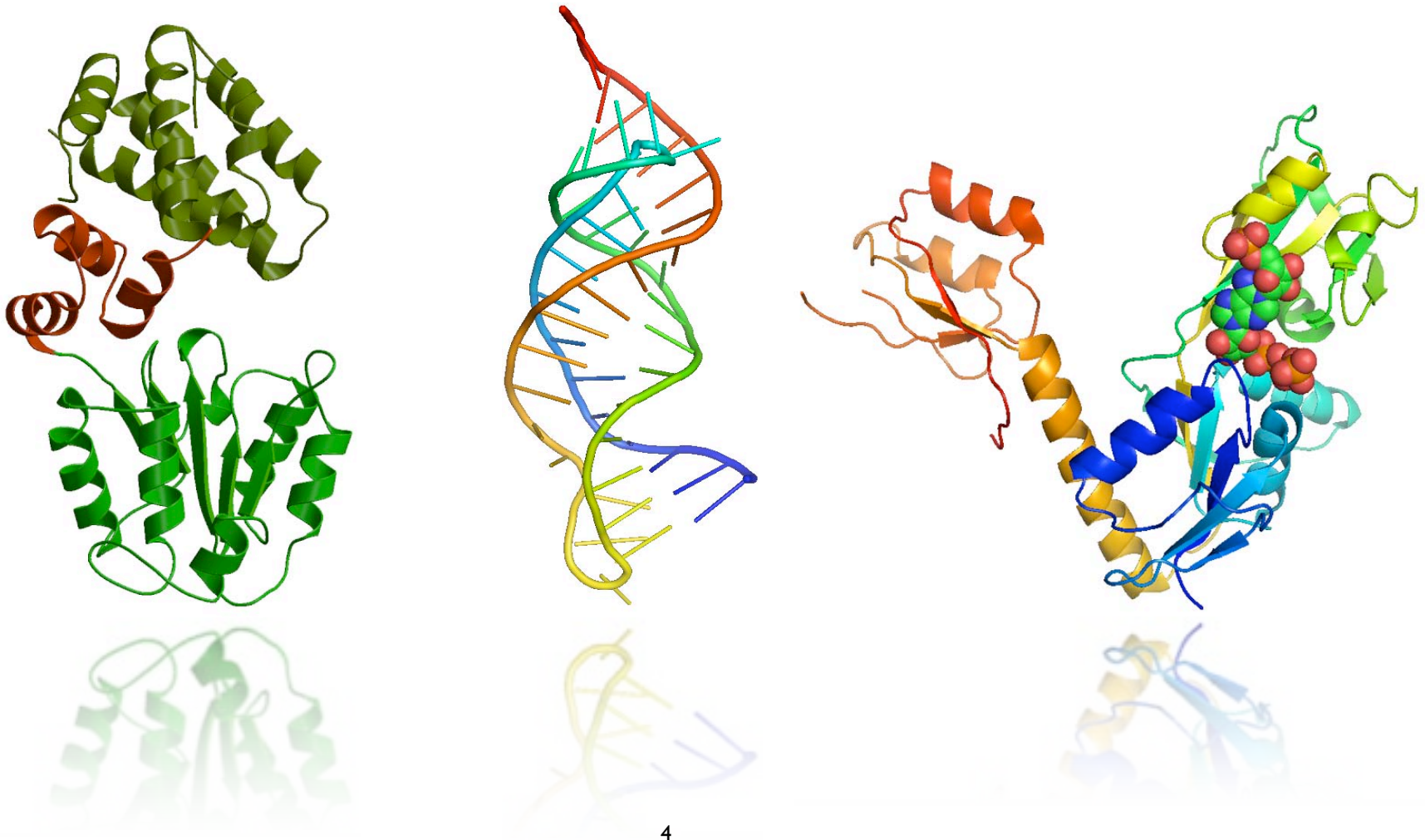
**BABELOMICS**  
Suite of tools for Functional Profiling (v2.0).

**Phylemon**  
Suite of tools for molecular evolution & phylogenetics (v1.0).

**Data Analysis and Visualization in Genomics and Proteomics.**  
Francisco Azuaje, Joaquín Dopazo (Editors)  
ISBN: 0-470-09439-7

# Structural Genomics Unit

*Bioinformatics Department, CIPF*





# **Comparative Modeling** **AnnoLyze - AnnoLite** **Tropical Disease Initiative**

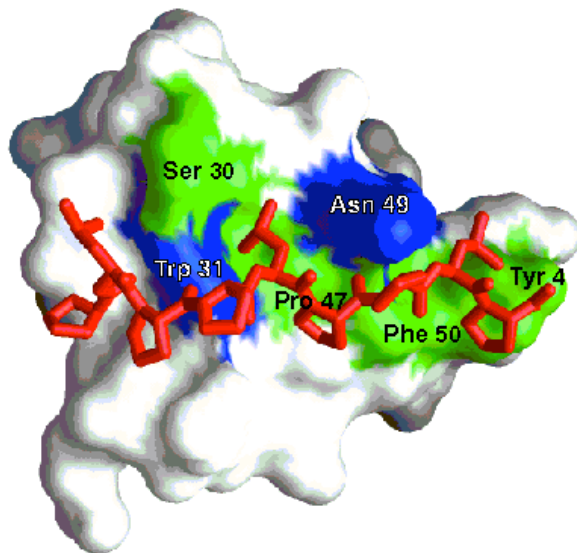
# Why is it useful to know the **structure** of a protein, not only its sequence?

- ◆ 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 N F



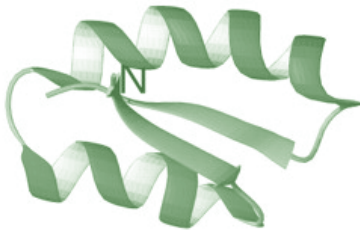
In addition, since evolution tends to conserve function and function depends more directly on structure than on sequence, **structure is more conserved in evolution than sequence.**

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



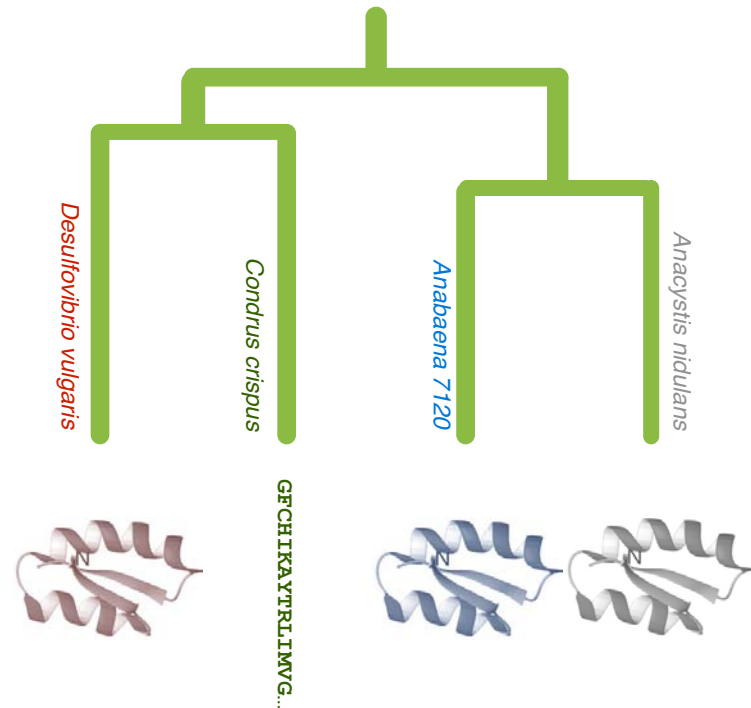
# Principles of protein structure

GFCHIKAYTRLIMVG...



Folding (physics)

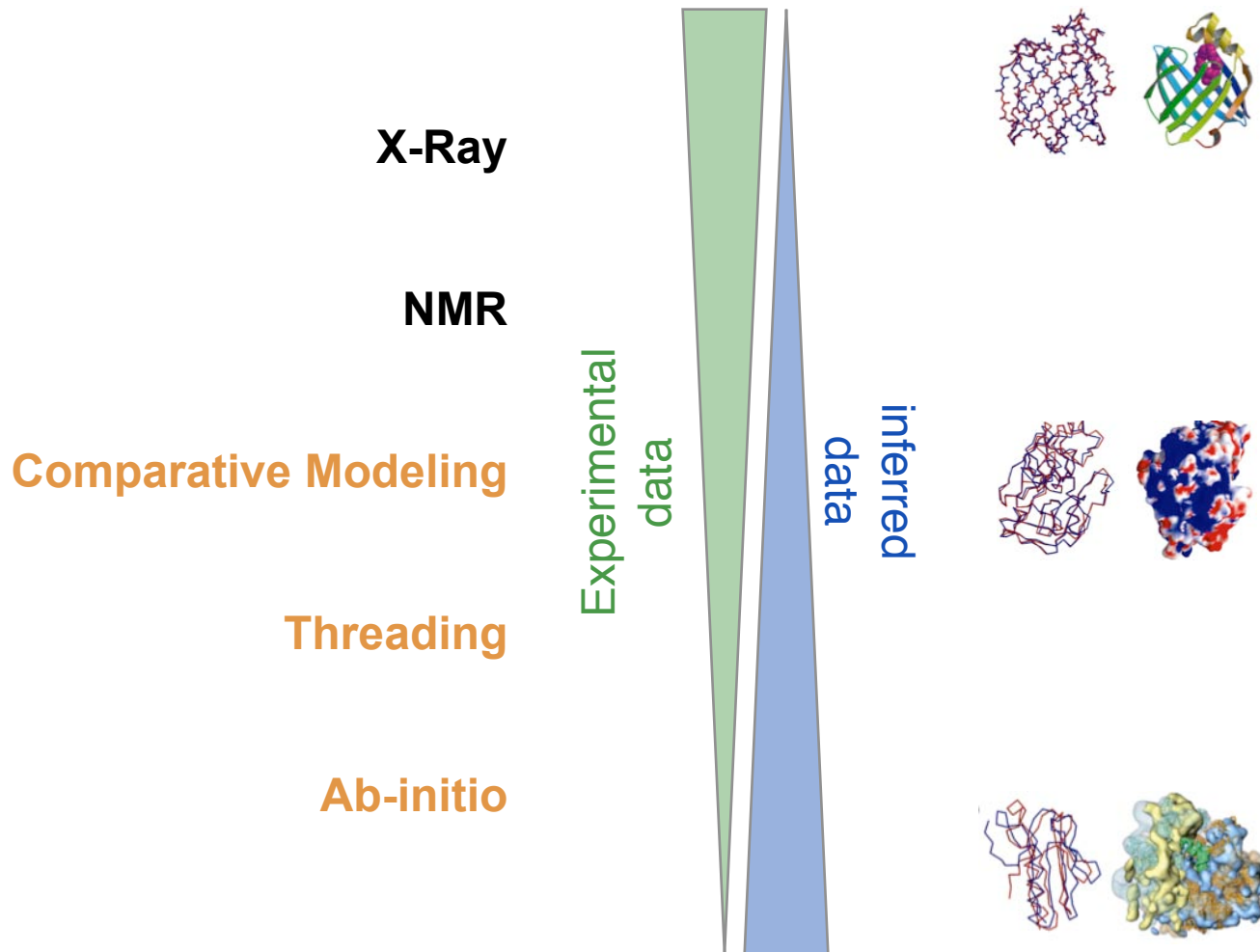
*Ab initio* prediction



Evolution (rules)

Threading  
Comparative Modeling

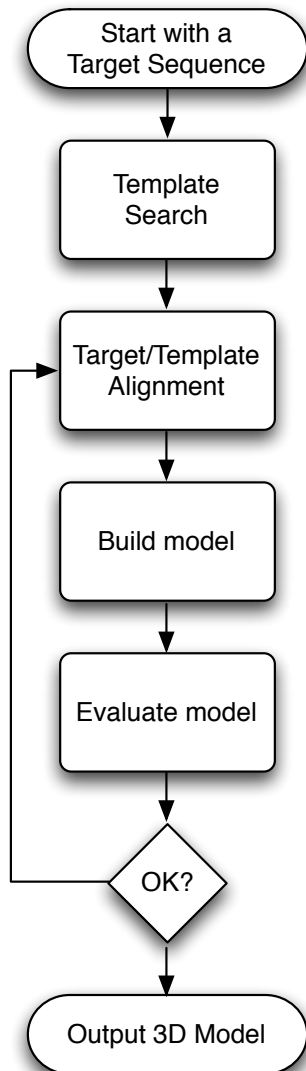
# protein prediction **vs** protein determination





# Comparative modeling by satisfaction of spatial restraints

## MODELLER



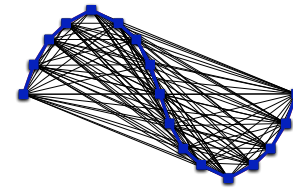
Given an alignment...

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

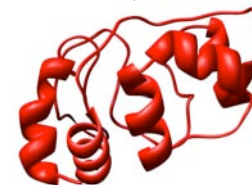
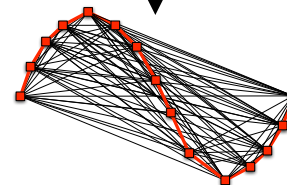
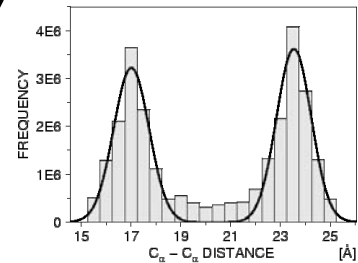
apply these features as restraints on your target sequence

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

MSVIPKR--GNCEQTSE  
ASILPKRLFGNCEQTSD



+

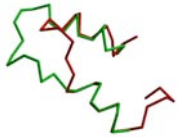


A. Šali & T. Blundell, *J. Mol. Biol.* 234, 779, 1993.  
J.P. Overington & A. Šali, *Prot. Sci.* 3, 1582, 1994.  
A. Fiser, R. Do & A. Šali, *Prot. Sci.*, 9, 1753, 2000.

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



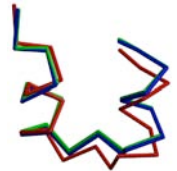
**Wrong fold**



**Miss alignments**



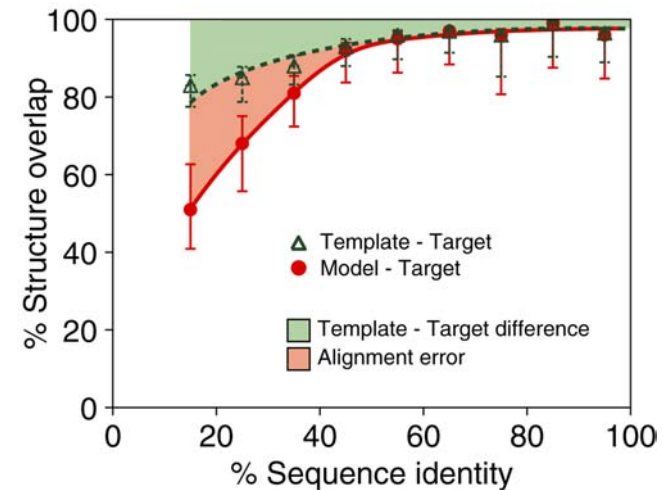
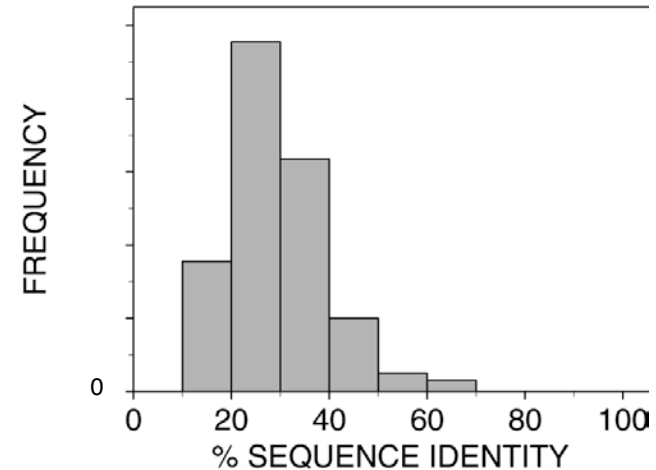
**Loop regions**



**Rigid body distortions**

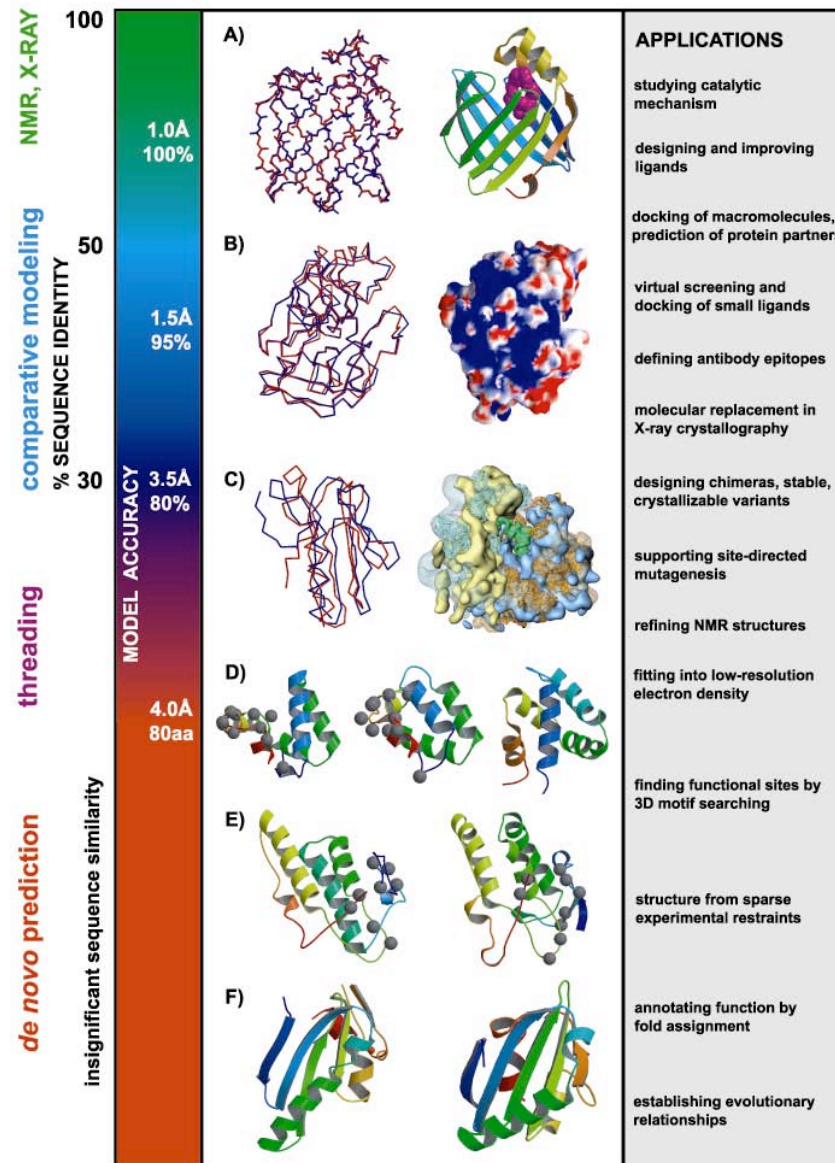


**Side-chain packing**



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

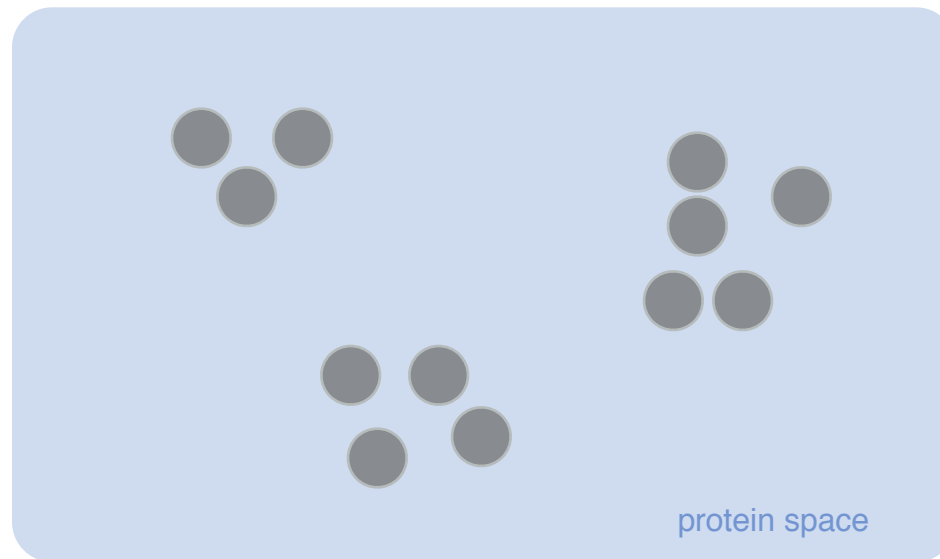
# Utility of protein structure models, despite errors



# Structural Genomics

Characterize most protein **sequences** based on related known **structures**

1. The number of “**families**” is much **smaller** than the number of proteins.
2. **Any one** of the members of a family is **fine**.



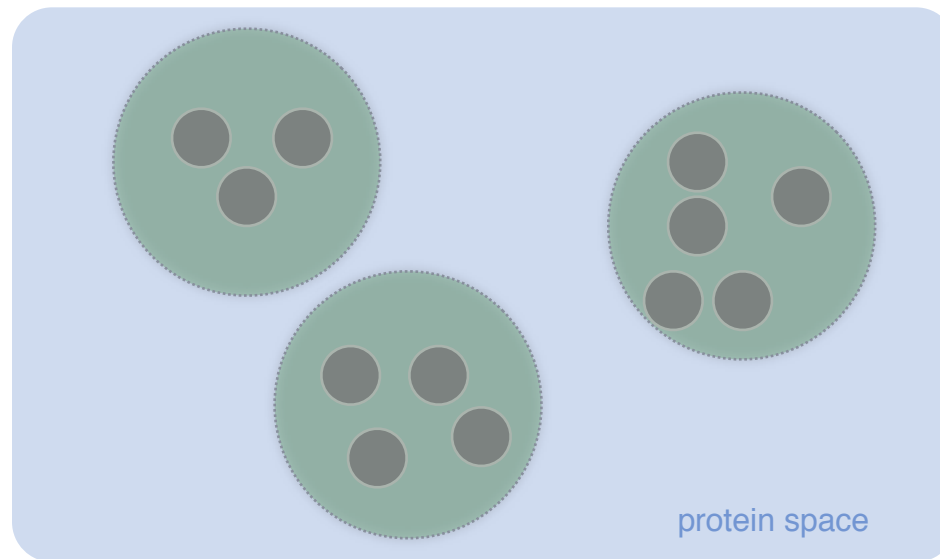
There are **~16,000** families (90%)  
@ 30% sequence identity cutoff

*Sali. Nat. Struct. Biol.* **5**, 1029, 1998.  
*Sali et al. Nat. Struct. Biol.*, **7**, 986, 2000.  
*Sali. Nat. Struct. Biol.* **7**, 484, 2001.  
*Baker & Sali. Science* **294**, 93, 2001.  
*Vitkup et al. Nat. Struct. Biol.* **8**, 559, 2001

# Structural Genomics

Characterize most protein **sequences** based on related known **structures**

1. The number of “**families**” is much **smaller** than the number of proteins.
2. **Any one** of the members of a family is **fine**.



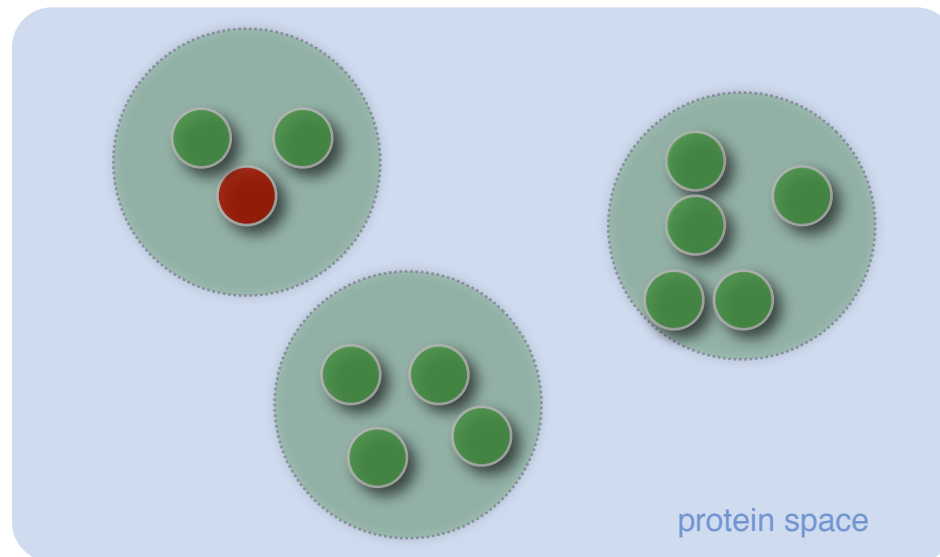
There are **~16,000** families (90%)  
@ 30% sequence identity cutoff

*Sali. Nat. Struct. Biol.* **5**, 1029, 1998.  
*Sali et al. Nat. Struct. Biol.*, **7**, 986, 2000.  
*Sali. Nat. Struct. Biol.* **7**, 484, 2001.  
*Baker & Sali. Science* **294**, 93, 2001.  
*Vitkup et al. Nat. Struct. Biol.* **8**, 559, 2001

# Structural Genomics

Characterize most protein **sequences** based on related known **structures**

1. The number of “**families**” is much **smaller** than the number of proteins.
2. **Any one** of the members of a family is **fine**.



There are **~16,000** families (90%)  
@ 30% sequence identity cutoff

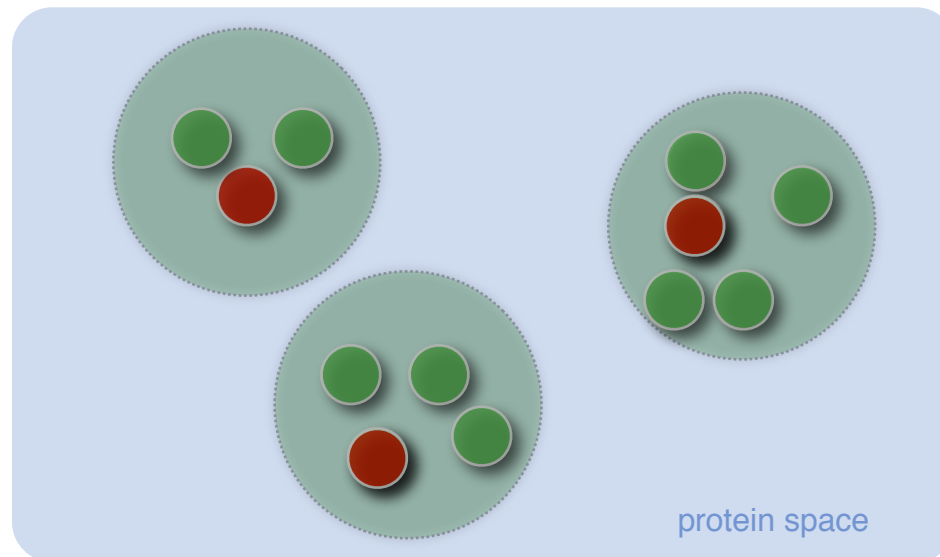
*Sali. Nat. Struct. Biol.* **5**, 1029, 1998.  
*Sali et al. Nat. Struct. Biol.*, **7**, 986, 2000.  
*Sali. Nat. Struct. Biol.* **7**, 484, 2001.  
*Baker & Sali. Science* **294**, 93, 2001.  
*Vitkup et al. Nat. Struct. Biol.* **8**, 559, 2001



# Structural Genomics

Characterize most protein **sequences** based on related known **structures**

1. The number of “**families**” is much **smaller** than the number of proteins.
2. **Any one** of the members of a family is **fine**.

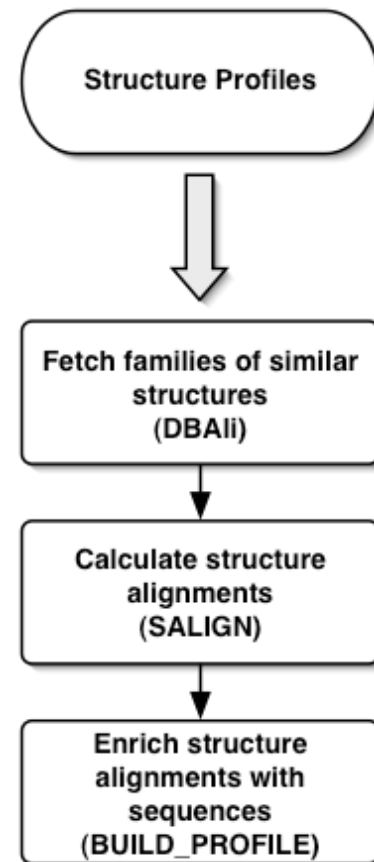
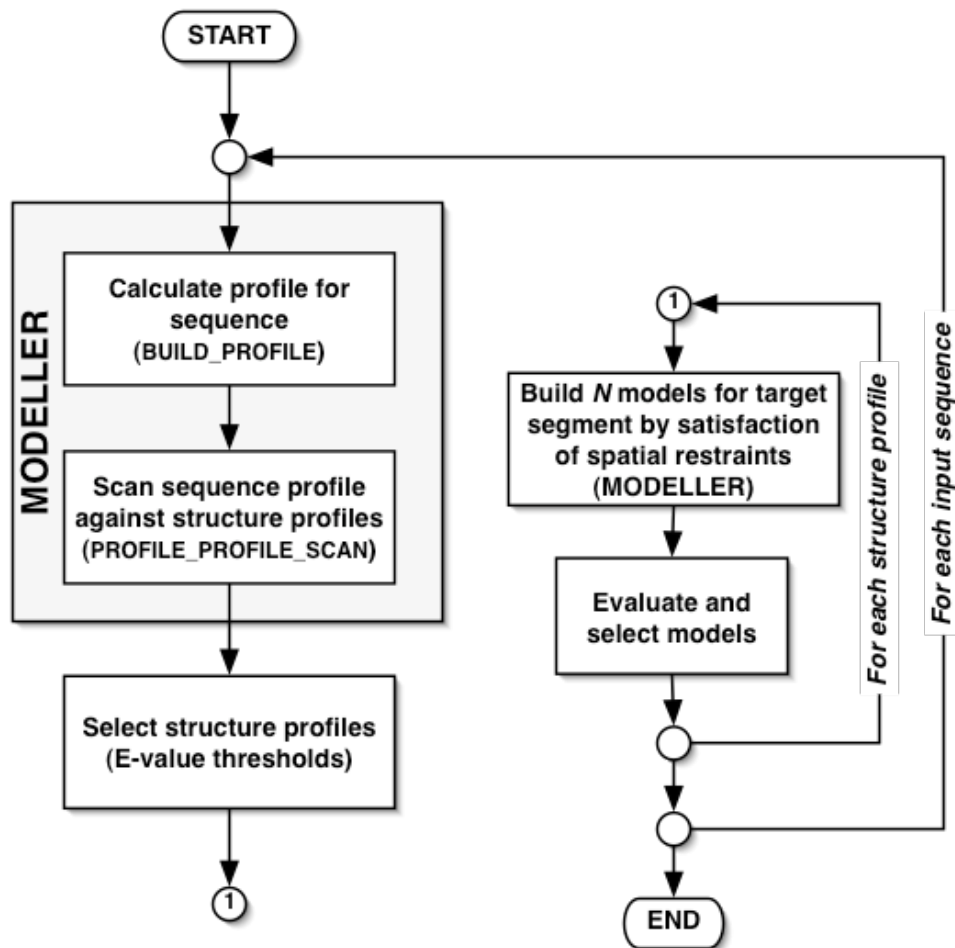


There are **~16,000** families (90%)  
@ 30% sequence identity cutoff

*Sali. Nat. Struct. Biol.* **5**, 1029, 1998.  
*Sali et al. Nat. Struct. Biol.*, **7**, 986, 2000.  
*Sali. Nat. Struct. Biol.* **7**, 484, 2001.  
*Baker & Sali. Science* **294**, 93, 2001.  
*Vitkup et al. Nat. Struct. Biol.* **8**, 559, 2001

# MODPIPE2.0

## Large-Scale Protein Structure Modeling



# ModBase Statistics

Large-scale modeling of the TrEMBL-SWISSPROT databases

<http://www.salilab.org/modbase/>

|                            |           |
|----------------------------|-----------|
| <b>Sequences (total)</b>   | 2,186,210 |
| <b>Sequences (modeled)</b> | 1,340,687 |
| <b>Models</b>              | 4,284,570 |

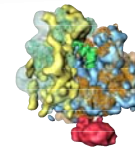
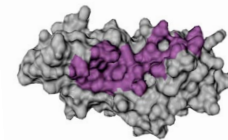
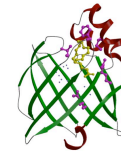
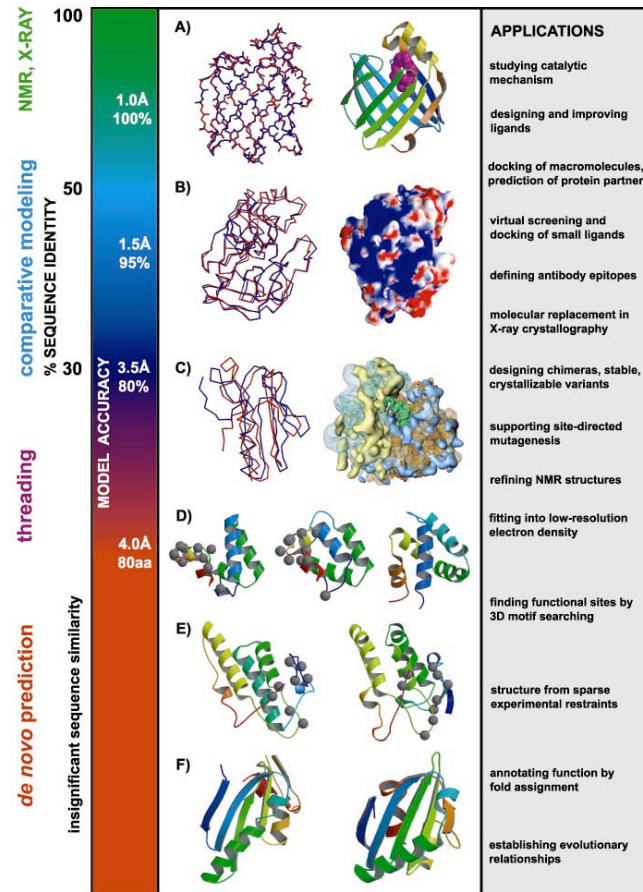
The screenshot shows the ModBase Search Page in a web browser. The page title is "ModBase Search Page" and the URL is "http://modbase.combio.ucsf.edu/modbase-cgi-new/search\_form.cgi". The page features a navigation bar with links: Home, User Login, ModBase Search Page, ModWeb Modelling Server, Help, and Current Logins. The main heading is "Database of Comparative Protein Structure Models" with a subheading "Welcome to ModBase, a database of three-dimensional protein models calculated by comparative modeling." The page includes a "ModBase search form" with a "Search" button. Below the search form, it states "All available datasets are selected" and provides a link "Select specific dataset(s)". There is a "Search by properties" section with dropdown menus for "Property" (set to "Database Accession Number") and "Organism" (set to "ALL"), along with input fields for specific values. A note at the bottom states: "Note: MODBASE contains theoretically calculated models, not experimentally determined structures. The models may contain significant errors." At the very bottom, there is a footer with a list of users requested to cite the article in their publications, including Ursula Pieper, Narayana Eswar, Hannes Bräse, M.S. Madhusudan, Fred Davis, Ashley C. Stuart, Nebojsa Mirkovic, Andrea Rossi, Marc A. Martenson, Andras Fiser, Ben Webb, Daniel Greenblatt, Conrad Huang, Tom Fenn, Andrei Sal, Nucleic Acid Research 32, D291-D292, 2004. The footer also mentions that MODBASE is maintained by Ursula Pieper in the group of Andrej Sali, Departments of Biopharmaceutical Sciences and Pharmaceutical Chemistry, and California Institute for Quantitative Biomedical Research Mission Bay Genentech Hall University of California San Francisco, San Francisco, CA 94143-2240. Please address all inquiries to [modbase@salilab.org](mailto:modbase@salilab.org).



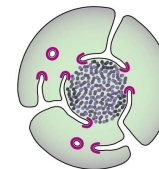
University of California  
San Francisco

*Pieper et al. NAR 34, D291 (2006)*

# Can we use models to infer function?



*T. cruzi*



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

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

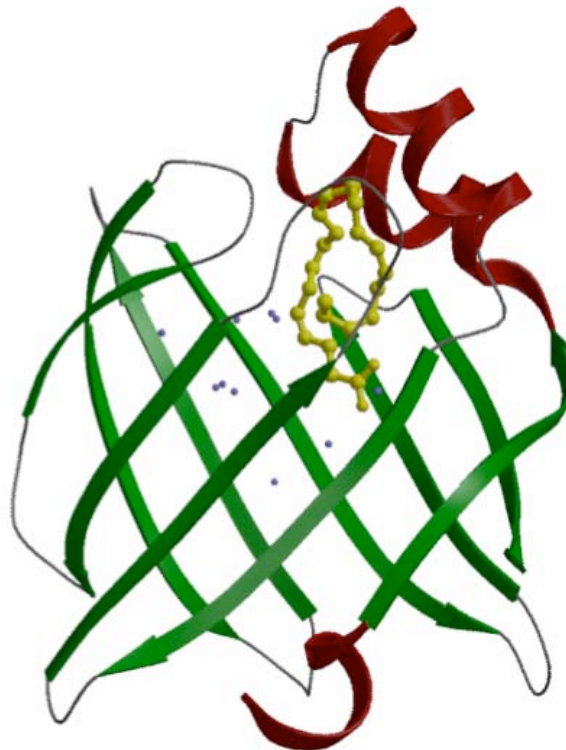
BLBP/oleic acid

Cavity is **not** filled



BLBP/docosahexaenoic acid

Cavity **is** filled



1. BLBP binds fatty acids.
2. Build a 3D model.
3. Find the fatty acid that fits most snugly into the ligand binding cavity.

# Structural analysis of missense mutations in human BRCA1 BRCT domains

Nebojsa Mirkovic, Marc A. Marti-Renom, Barbara L. Weber, Andrej Sali and Alvaro N.A. Monteiro

**Cancer Research (June 2004). 64:3790-97**

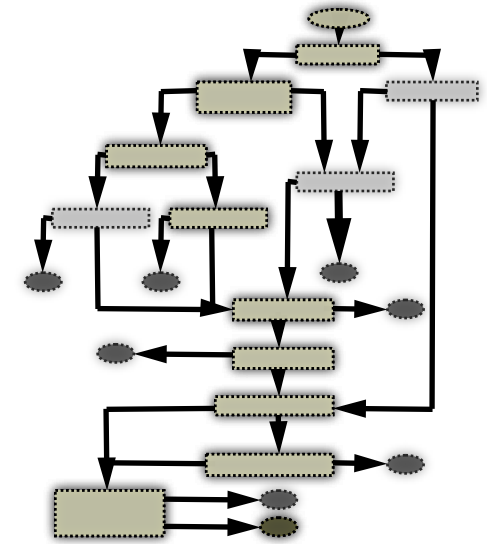
Cannot measure the functional impact of every possible SNP at all positions in each protein!  
Thus, prediction based on general principles of protein structure is needed.



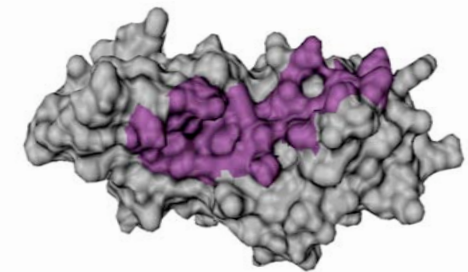
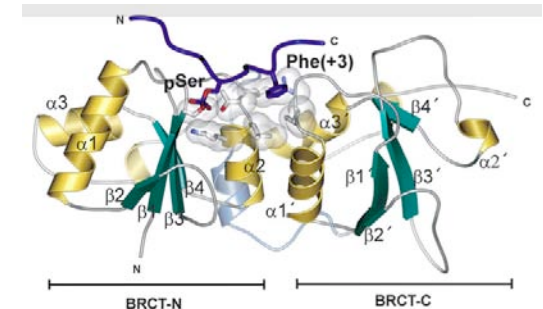
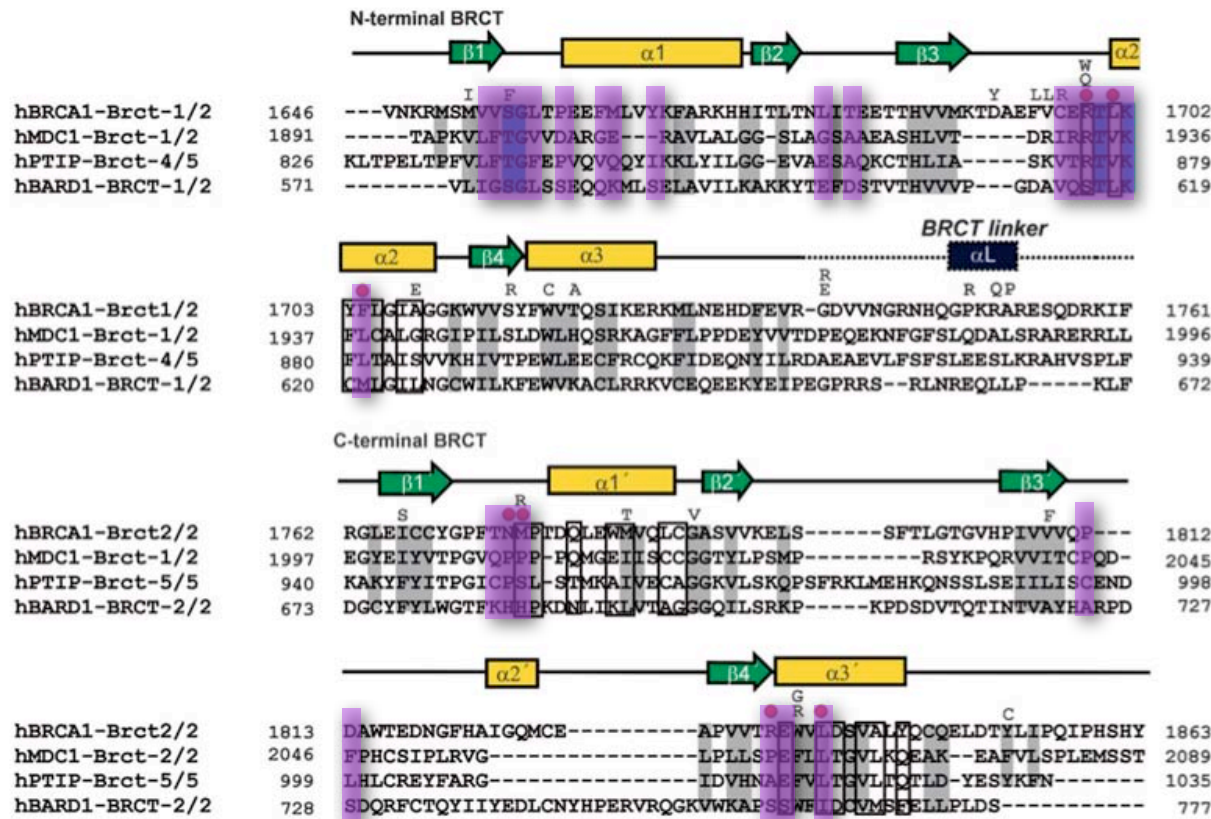


# Missense mutations in BRCT domains by function

| cancer<br>associate                                      | not cancer<br>associated | ?  |  |  |  |  |
|--|--------------------------|--|--|--|--|--|
| C1697R<br>R1699W<br>A1708E<br>S1715R<br>P1749R<br>M1775R |                          | M1652K<br>L1657P<br>E1660G<br>H1686Q<br>R1699Q<br>K1702E<br>Y1703HF<br>1704S   | L1705PS<br>1715NS1<br>722FF17<br>34LG173<br>8EG1743<br>RA1752<br>PF1761I   | F1761S<br>M1775E<br>M1775K<br>L1780P<br>I1807S<br>V1833E<br>A1843T   |  |  |
|  | M1652I<br>A1669S         |  | V1665M<br>D1692N<br>G1706A<br>D1733G<br>M1775V<br>P1806A   |  |  |  |
|  |                          | M1652T<br>V1653M<br>L1664P<br>T1685A<br>T1685I<br>M1689R<br>D1692Y<br>F1695L<br>V1696L<br>R1699L<br>G1706E<br>W1718C | W1718S<br>T1720A<br>W1730S<br>F1734S<br>E1735K<br>V1736A<br>G1738R<br>D1739E<br>D1739G<br>D1739Y<br>V1741G<br>H1746N | R1751P<br>R1751Q<br>R1758G<br>L1764P<br>I1766S<br>P1771L<br>T1773S<br>P1776S<br>D1778N<br>D1778G<br>D1778H<br>M1783T | C1787S<br>G1788D<br>G1788V<br>G1803A<br>V1804D<br>V1808A<br>V1809A<br>V1809F<br>V1810G<br>Q1811R<br>P1812S<br>N1819S | A1823T<br>V1833M<br>W1837R<br>W1837G<br>S1841N<br>A1843P<br>T1852S<br>P1856T<br>P1859R |



# Putative binding site on BRCA1

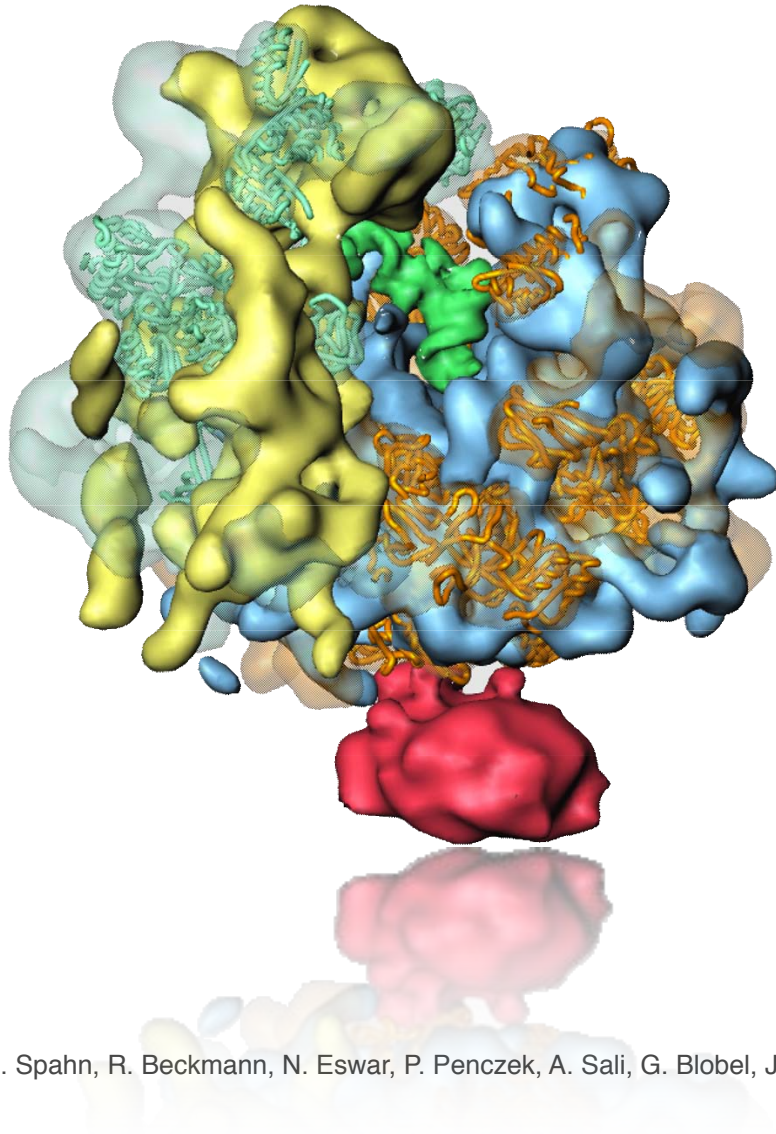


Putative binding site predicted in 2003  
and accepted for publication on March 2004.

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

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

# *S. cerevisiae* ribosome



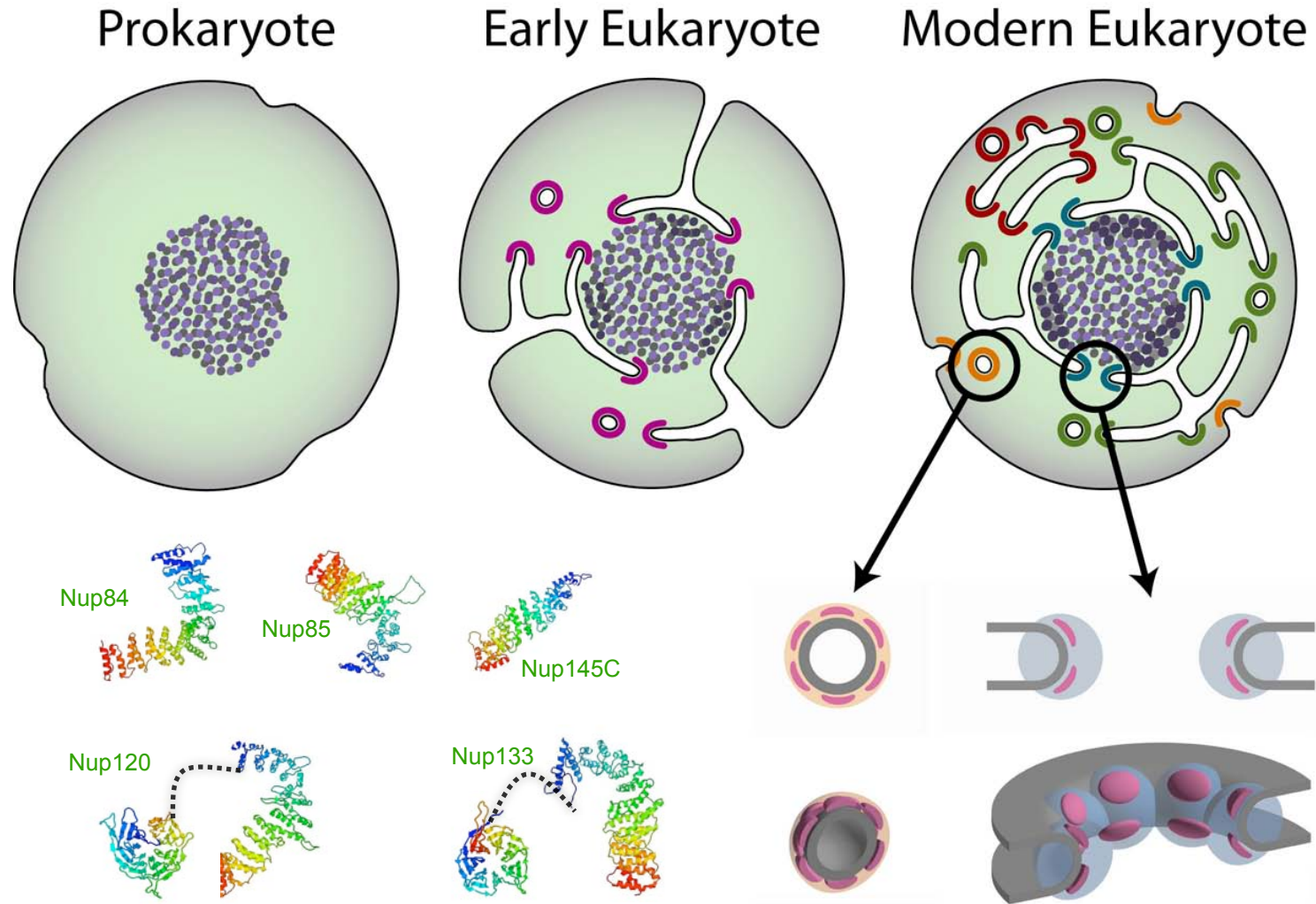
Fitting of comparative models into 15Å cryo-electron density map.

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

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

# The Nucleopore complex

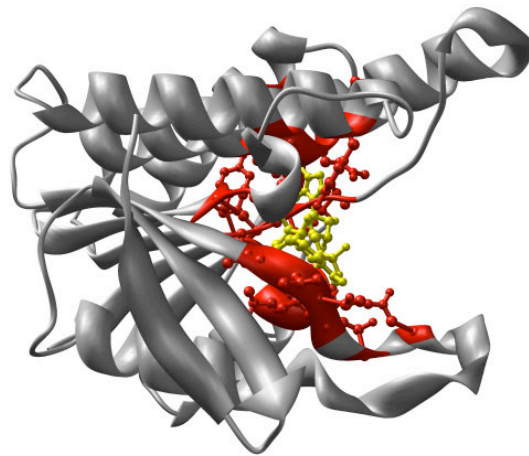
## Cell evolution (?)





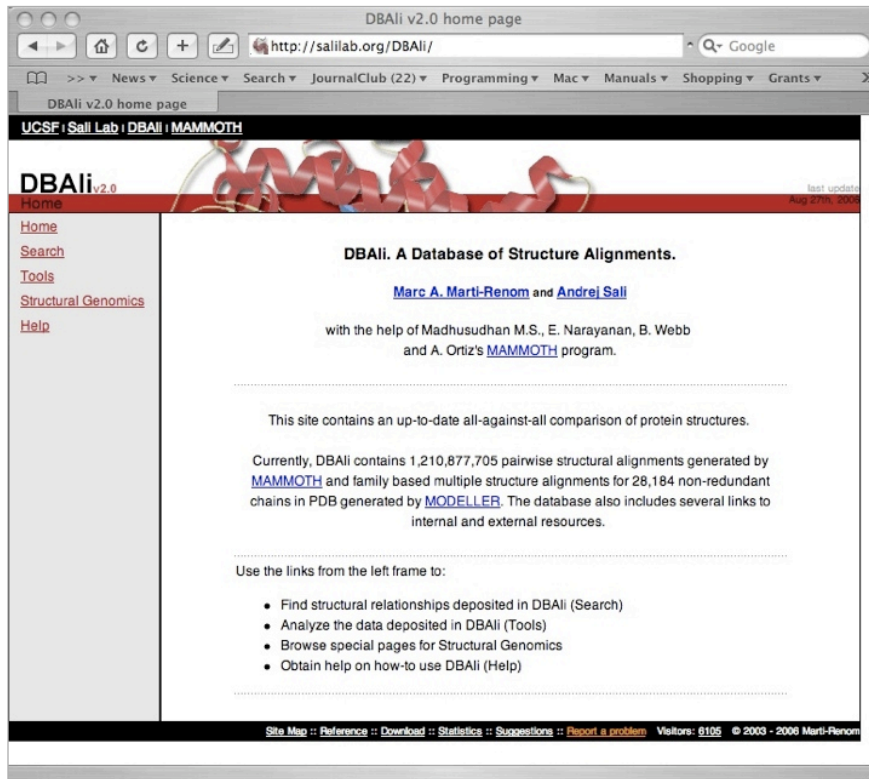
# Protein function from structure

## Comparative annotation. AnnoLite and AnnoLyze.



# DBAli<sub>v2.0</sub> database

<http://www.dbali.org>



- ✓ Fully-automatic
- ✓ Data is kept up-to-date with PDB releases
- ✓ Tools for “on the fly” classification of families.
- ✓ Easy to navigate
- ✓ Provides tools for structure analysis

**Does not provide a stable classification similar to that of CATH or SCOP**

| Pairwise structure alignments               |                   |
|---|-------------------|
| Last update:                                | October 6th, 2007 |
| Number of chains:                           | 96,804            |
| Number of structure-structure comparisons:* | 1,748,371,897     |
| Multiple structure alignments               |                   |
| Last update:                                | August 1st, 2007  |
| Number of representative chains:            | 34,637            |
| Number of families:                         | 12,732            |

Uses MAMMOTH for similarity detection

- ✓ **VERY FAST!!!**
- ✓ **Good scoring system with significance**

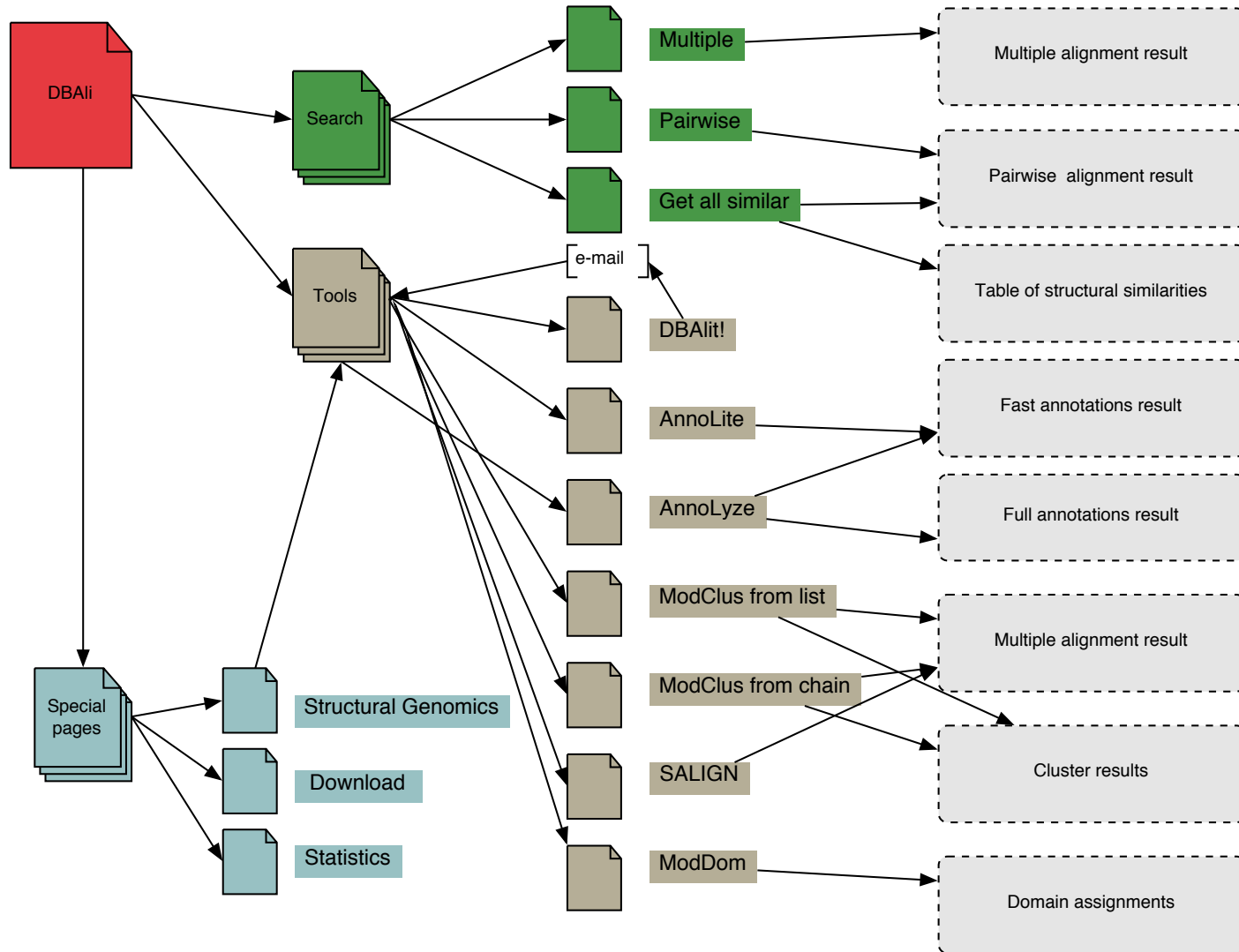
Ortiz AR, (2002) *Protein Sci.* 11 pp2606

Marti-Renom et al. 2001. *Bioinformatics.* 17, 746



# DBAli<sub>v2.0</sub> database

<http://www.dbali.org>



# 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

# AnnoLite

AnnoLite results for chain [1qpl.A](#) based on [45](#) structural similar chains.

|                        | Conf. P-value   | Link                        | Description  |
|------------------------|---|-----------------------------|--|
| CATH:                  |  7.5e-99   | <a href="#">2.70.100.10</a> | 1,4-Beta-D-Glucan Celobiohydrolase I, subunit A  |
| SCOP:                  |  0.00      | <a href="#">b.29.1.10</a>   | Glycosyl hydrolase family 7 catalytic core   |
| PFAM:                  |  0.00      | <a href="#">PF00840</a>     | Glycosyl hydrolase family 7  |
| InterPro:              |  1.3e-99   | <a href="#">IPR001722</a>   | Glycoside hydrolase, family 7  |
|                        |  6.0e-51   | <a href="#">IPR008985</a>   | Concanavalin A-like lectin/glucanase   |
|                        |  1.0e-42   | <a href="#">IPR000254</a>   | Cellulose-binding region, fungal   |
| EC Number:             |  1.2e-44   | <a href="#">3.2.1.91</a>    | Cellulose 1,4-beta-cellobiosidase.   |
|                        |  6.0e-41   | <a href="#">3.2.1.4</a>     | Cellulase.   |
| GO Molecular Function: |  6.0e-36   | <a href="#">0030248</a>     | cellulose binding                                     |
|                        |  8.4e-36   | <a href="#">0016162</a>     | cellulose 1,4-beta-cellobiosidase activity            |
|                        |  1.0e-35   | <a href="#">0004553</a>     | hydrolase activity, hydrolyzing O-glycosyl compounds  |
|                        |  1.4e-30   | <a href="#">0008810</a>     | cellulase activity                                    |
|                        |  3.1e-20 | <a href="#">0016798</a>     | hydrolase activity, acting on glycosyl bonds        |
|                        |  1.0e+0  | <a href="#">0016787</a>     | hydrolase activity                                  |
| GO Biological Process: |  1.1e-63 | <a href="#">0030245</a>     | cellulose catabolism                                |
|                        |  1.2e-54 | <a href="#">0000272</a>     | polysaccharide catabolism                           |
|                        |  3.6e-20 | <a href="#">0005975</a>     | carbohydrate metabolism                             |
| GO Cellular Component: |  1.2e-23 | <a href="#">0005576</a>     | extracellular region                                |

 Information annotated in the MSD database.

 High,  medium and  low confidence annotations not annotated in the MSD database.

 High,  medium and  low confidence annotations already annotated in the MSD database.

# Benchmark set

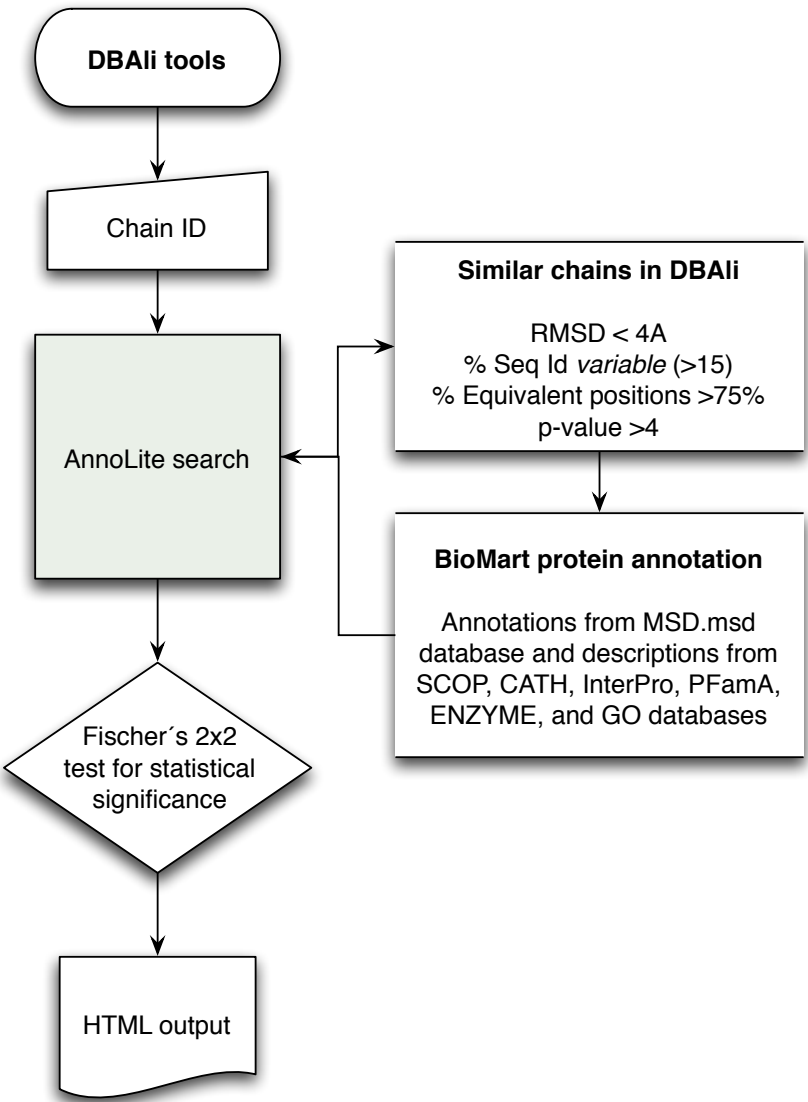
|                             | Number of chains |
|-----------------------------|------------------|
| <b>Initial set*</b>         | 50,223           |
| <b>FULL annotation**</b>    | 10,997           |
| <b>Non-redundant set***</b> | 1,879            |

*\*data from BioMart MSD.3 (release February 2005)*

*\*\*annotated with CATH, SCOP, Pfam, EC, InterPro, and GO terms in the MSD database*

*\*\*\*not two chains can be structurally aligned within 2Å, superimposing more than 60% of their Ca atoms and have a length difference inferior to 30aa*

# Method



AnnoLite results for chain [1gpi:A](#) based on [44](#) structural similar chains.

|                        | Conf. P-value | Link                        | Description  |
|------------------------|---------------|-----------------------------|--|
| CATH:                  | ● 7.5e-99     | <a href="#">2.70.100.10</a> | 1,4-Beta-D-Glucan Cellobiohydrolase I, subunit A       |
| SCOP:                  | ● 0.00        | <a href="#">b.29.1.10</a>   | Glycosyl hydrolase family 7 catalytic core             |
| PFAM:                  | ● 0.00        | <a href="#">PF00840</a>     | Glycosyl hydrolase family 7                            |
| InterPro:              | ● 1.3e-99     | <a href="#">IPR001722</a>   | Glycoside hydrolase, family 7                          |
|                        | ● 6.0e-51     | <a href="#">IPR008985</a>   | Concanavalin A-like lectin/glucanase                   |
|                        | ● 1.0e-42     | <a href="#">IPR000254</a>   | Cellulose-binding region, fungal                       |
| EC Number:             | ● 1.2e-44     | <a href="#">3.2.1.91</a>    | Cellulose 1,4-beta-cellobiosidase.                     |
|                        | ● 6.0e-41     | <a href="#">3.2.1.4</a>     | Cellulase.   |
| GO Molecular Function: | ● 6.0e-36     | <a href="#">0030248</a>     | cellulose binding ↕                                    |
|                        | ● 8.4e-36     | <a href="#">0016162</a>     | cellulose 1,4-beta-cellobiosidase activity ↕           |
|                        | ● 1.0e-35     | <a href="#">0004553</a>     | hydrolase activity, hydrolyzing O-glycosyl compounds ↕ |
|                        | ● 1.4e-30     | <a href="#">0008810</a>     | cellulase activity ↕                                   |
|                        | ● 3.1e-20     | <a href="#">0016798</a>     | hydrolase activity, acting on glycosyl bonds ↕         |
|                        | ● 1.0e+0      | <a href="#">0016787</a>     | hydrolase activity ↕                                   |
| GO Biological Process: | ● 1.1e-63     | <a href="#">0030245</a>     | cellulose catabolism ↕                                 |
|                        | ● 1.2e-54     | <a href="#">0000272</a>     | polysaccharide catabolism ↕                            |
|                        | ● 3.6e-20     | <a href="#">0005975</a>     | carbohydrate metabolism ↕                              |
| GO Cellular Component: | ● 1.2e-23     | <a href="#">0005576</a>     | extracellular region ↕                                 |

● Information annotated in the MSD database.

● High, ● medium and ● low confidence annotations not annotated in the MSD database.

● High, ● medium and ● low confidence annotations already annotated in the MSD database.

# Scoring function

Fisher's 2x2 contingency test

|               | Non-similar | Similar | Total |
|---------------|-------------|---------|-------|
| Annotated     | a           | b       | a+b   |
| Not Annotated | c           | d       | c+d   |
| Total         | a+c         | b+d     | n     |

| 1b78A<br>SCOP<br>c.51.4.1 | Similar | Not similar | Total  |
|---------------------------|---------|-------------|--------|
| Annotated                 | 4       | 2           | 6      |
| Not Annotated             | 0       | 71,096      | 71,096 |
| Total                     | 4       | 71,098      | 71,102 |

$$p = \frac{\binom{a+b}{a} \binom{c+d}{c}}{\binom{n}{a+c}}$$

$$= \frac{(a+b)!(c+d)!(a+c)!(b+d)!}{n!a!b!c!d!}$$

$$p = 1.78e^{-19}$$



# Sensitivity .vs. Precision

|                       | Optimal cut-off | Sensitivity (%)<br>Recall or TPR | Precision (%) |
|-----------------------|-----------------|----------------------------------|---------------|
| SCOP fold             | 1E-06           | 92.7                             | 88.4          |
| CATH fold             | 1E-03           | 95.7                             | 90.1          |
| InterPro              | 1E-03           | 88.4                             | 78.2          |
| PFam family           | 1E-04           | 90.5                             | 82.8          |
| EC number             | 1E-04           | 93.3                             | 79.7          |
| GO Molecular Function | 1E-01           | 84.3                             | 80.9          |
| GO Biological Process | 1E-03           | 85.5                             | 74.8          |
| GO Cellular Component | 1E-02           | 77.6                             | 58.6          |

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

# AnnoLyze

| Inherited ligands: 4      |                           |                          |   |
|---------------------------|---------------------------|--------------------------|---|
| Ligand                    | Av. binding site seq. id. | Av. residue conservation | Residues in predicted binding site (size proportional to the local conservation)                  |
| <a href="#">MO2</a>       | 59.03                     | <a href="#">0.185</a>    | 48 49 52 62 63 66 67 113 116  |
| <a href="#">CRY</a>       | 20.00                     | <a href="#">0.111</a>    | 23 29 31 37 44 48 49 83 85 94 96 103 121  |
| <a href="#">BOG</a>       | 20.00                     | <a href="#">0.111</a>    | 19 20 21 48 49 51 96 98 136   |
| <a href="#">ACY</a>       | 15.87                     | <a href="#">0.163</a>    | 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)                  |
| <a href="#">d.113.1.1</a> | 23.68                     | <a href="#">0.948</a>    | 19 20 50 51 52 53 54 55 56 57 58 77 78 79 80<br>81 82 83 84 85 93 95 97 99 134 135 138 142<br>145 |



# Benchmark

|                             | Number of chains             |
|-----------------------------|------------------------------|
| <b>Initial set*</b>         | 78,167                       |
| <b>LigBase**</b>            | 30,126                       |
| <b>Non-redundant set***</b> | <b>4,948</b> (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 3Å, superimposing more than 75% of their Ca atoms, result in a sequence alignment with more than 30% identity, and have a length difference inferior to 50aa*

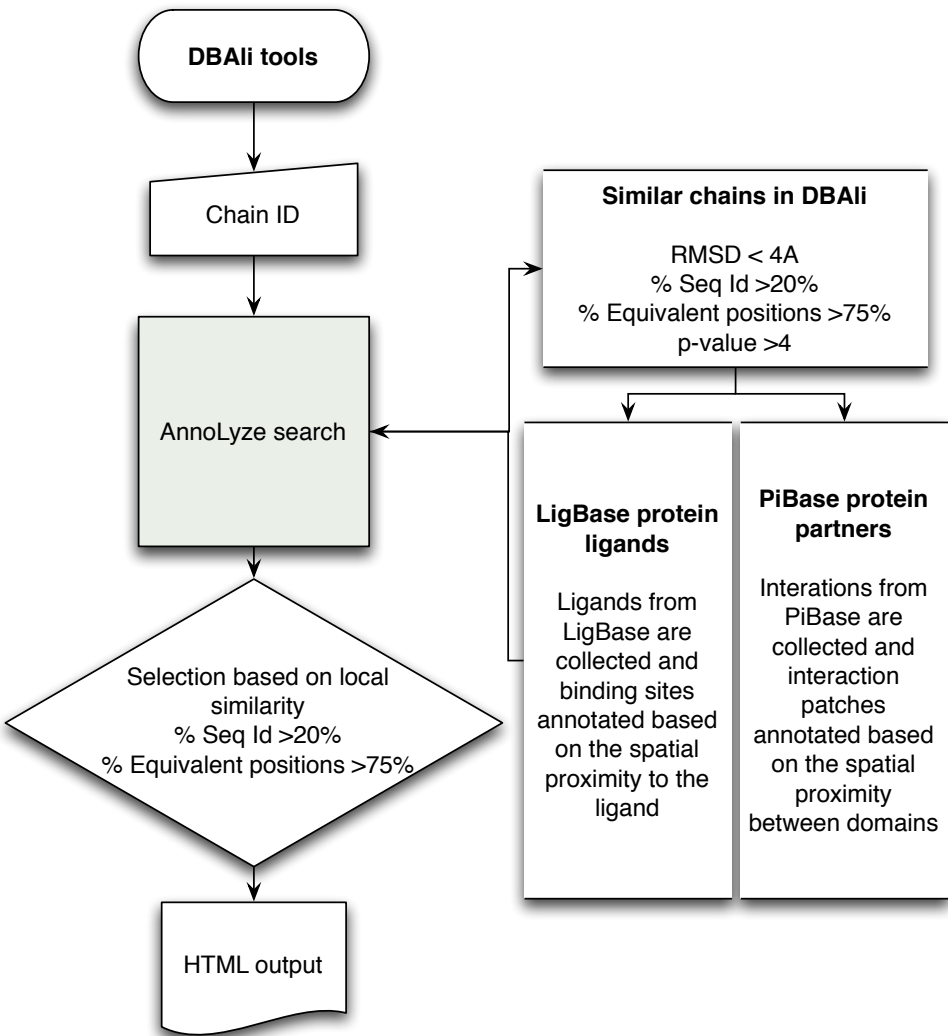
|                               | Number of chains                   |
|-------------------------------|------------------------------------|
| <b>Initial set*</b>           | 78,167                             |
| <b><math>\pi</math>Base**</b> | 30,425                             |
| <b>Non-redundant set***</b>   | <b>4,613</b> (11,641 partnerships) |

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

*\*\*annotated with at least one partner in the  $\pi$ Base database*

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

# Method



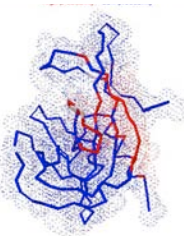
Inherited ligands: 4

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



Inherited partners:1

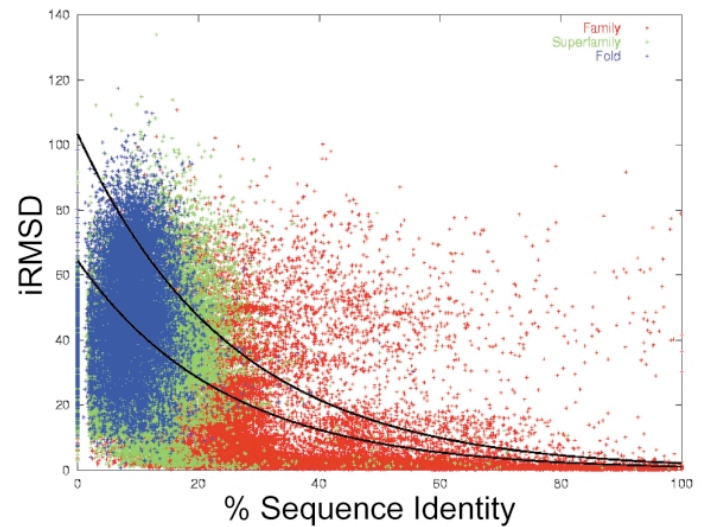
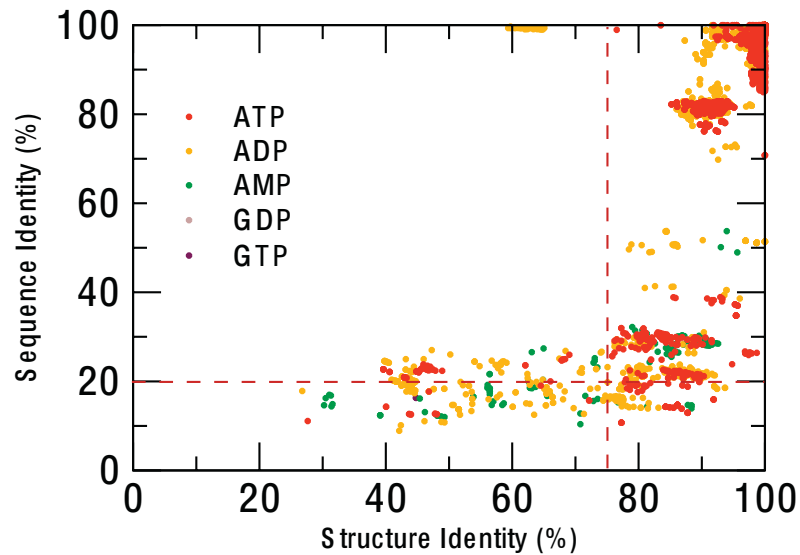
| Partner                   | Av. binding site seq. id. | Av. residue conservation | Residues in predicted binding site (size proportional to the local conservation)            |
|---------------------------|---------------------------|--------------------------|---|
| <a href="#">d.113.1.1</a> | 23.68                     | <a href="#">0.948</a>    | 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.

# Sensitivity .vs. Precision

|          | Optimal cut-off | Sensitivity (%)<br>Recall or TPR | Precision (%) |
|----------|-----------------|----------------------------------|---------------|
| Ligands  | 30%             | 71.9                             | 13.7          |
| Partners | 40%             | 72.9                             | 55.7          |

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



# Example (2azwA)

## Structural Genomics Unknown Function

Molecule: MutT/nudix family protein

PDB ID: [2azwA](#)

**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

**Sequence:**  
Mds: 09b13d23c0ae0dfcaddec636e2ddfa6KTPTAAS  
Length: 146

**SCOP:**  
none

**CATH:**  
none

**Ligands:**  
none

**Interacting partners:**  
none

KIPTFGKREE TLTYQTRYAA YIIIVSKPENN TMVLVQAPNG AYFLPGGSEIE  
GTEKKEAHH REVLLEELGIS VEIGCYLGEA DEYFYSNHRQ TAYYNGYFY  
VANTWRQLSE PLRNTLHWV APEEAVRLK RGSRWAVEK WLAAS

**Similar structures:** [20](#)

**Similar sequences:** 890

**Most similar structure in DBAli:**

| Code                   | SeqId(%) | EqPos | RMSD | P-Value | See                 |
|------------------------|----------|-------|------|---------|---------------------|
| <a href="#">1vc9:A</a> | 22.76    | 123   | 3.57 | 17.28   | <a href="#">ali</a> |

**Most similar sequence in DBAli:**

| Code                   | SeqId(%) | EqPos | RMSD | P-Value | See                 |
|------------------------|----------|-------|------|---------|---------------------|
| <a href="#">1vc9:B</a> | 24.59    | 122   | 3.47 | 17.00   | <a href="#">ali</a> |

**P-value distribution:**

P-value distribution for similar chains

**Keyword distribution:**

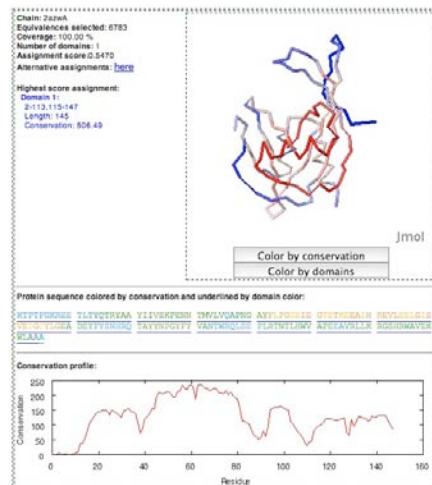
**SCOP and CATH distribution for similar structures:**

Inherited ligands: 4

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

Inherited partners: 1

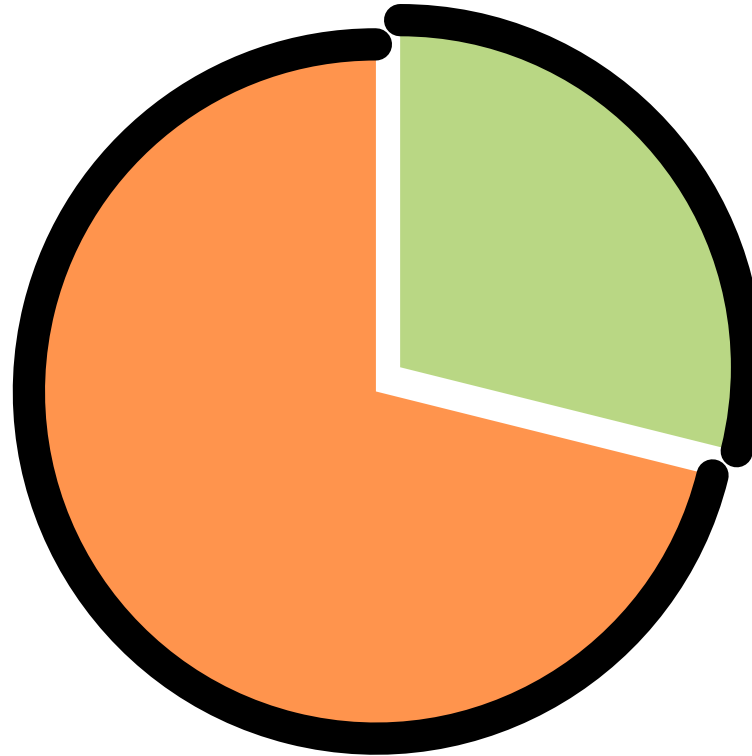
| Partner                   | Av. binding site seq. id. | Av. residue conservation | Residues in predicted binding site (size proportional to the local conservation)            |
|---------------------------|---------------------------|--------------------------|---|
| <a href="#">d.113.1.1</a> | 23.68                     | <a href="#">0.948</a>    | 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 |



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

# Tropical Disease Initiative (TDI)

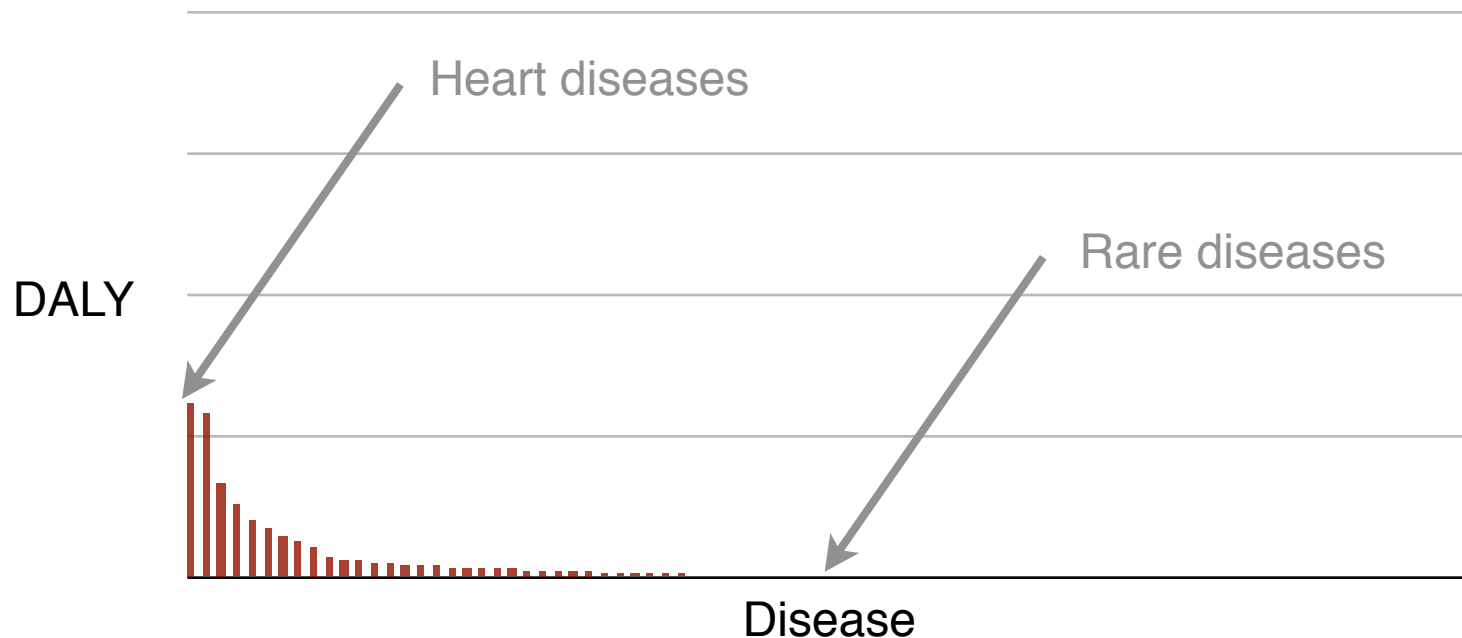
*Predicting binding sites in protein structure models.*



<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*

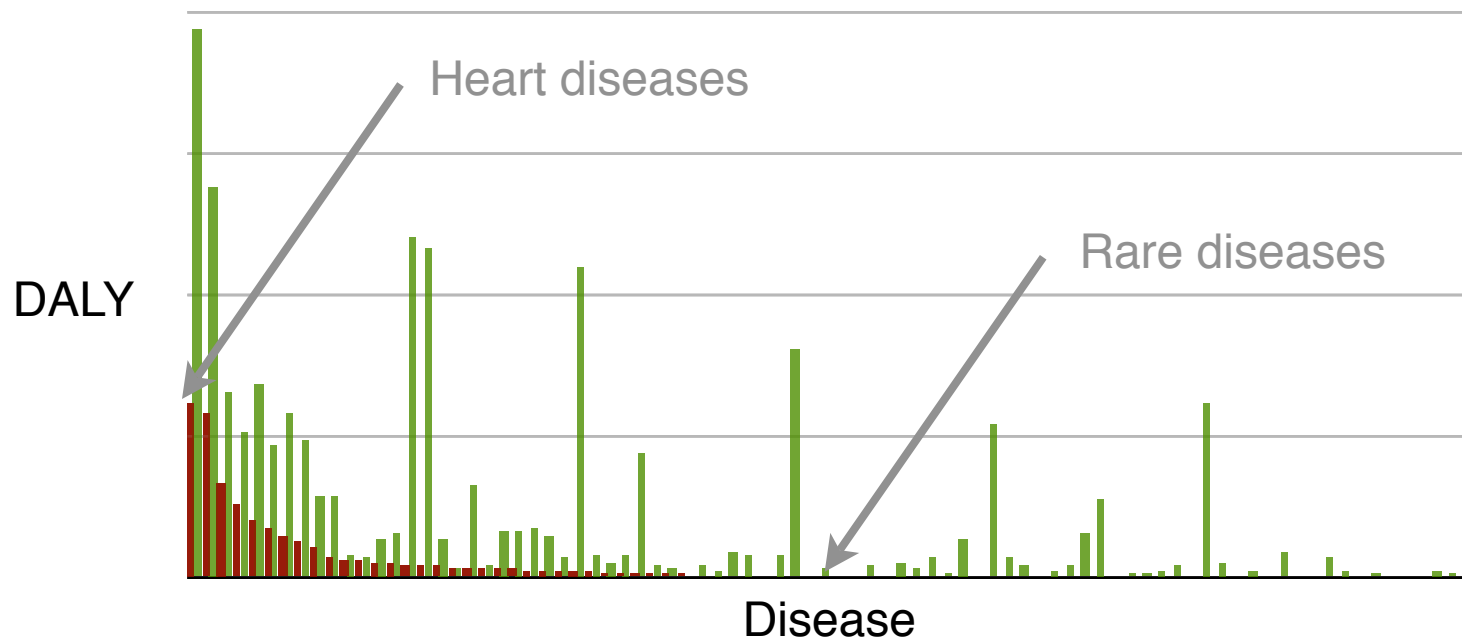
DALY - Disability adjusted life years

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

- 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)

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

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

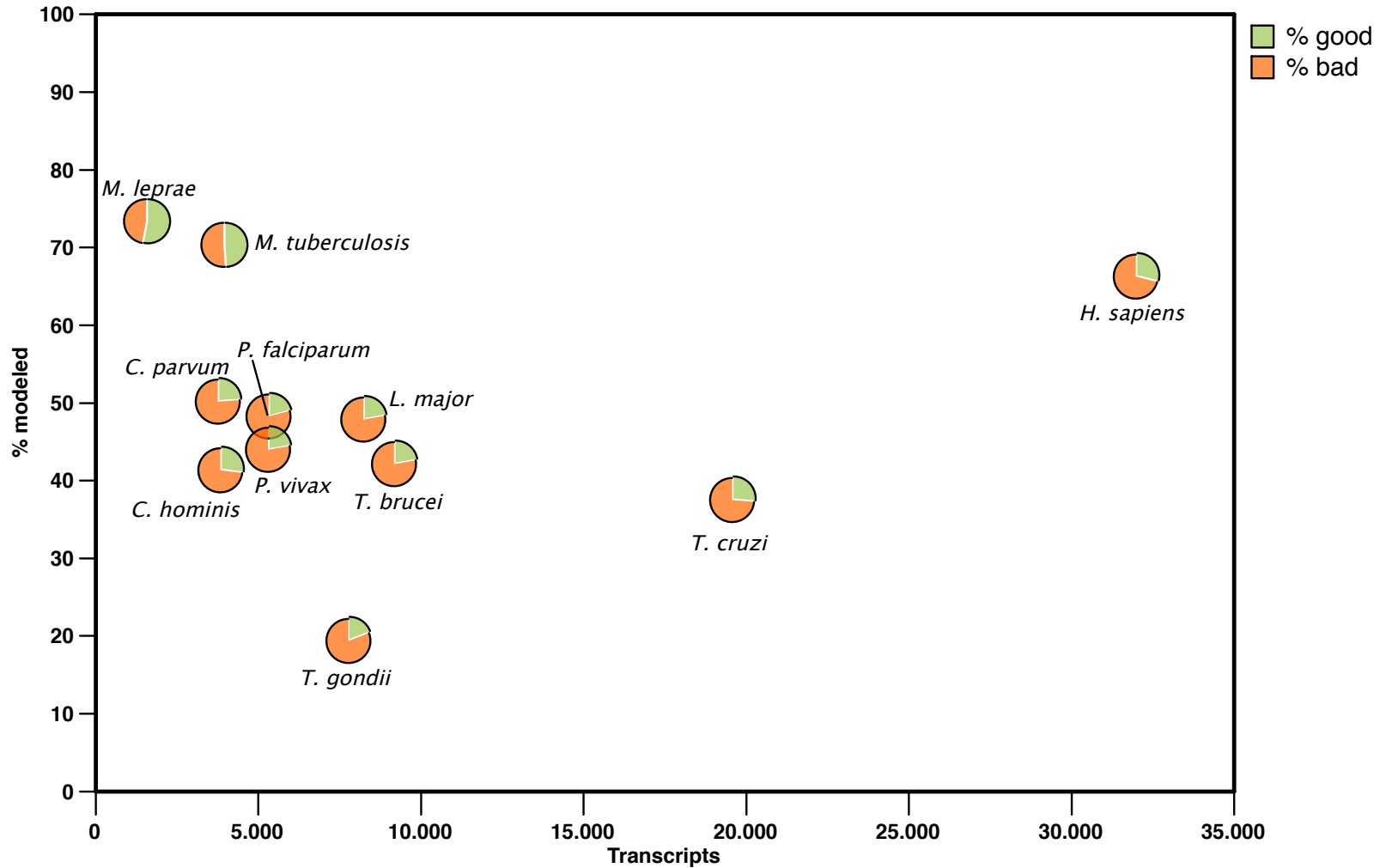
Disease data taken from WHO, *World Health Report 2004*

DALY - Disability adjusted life year in 1000’s.

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

# Modeling Genomes

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



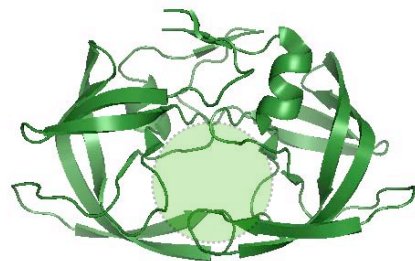
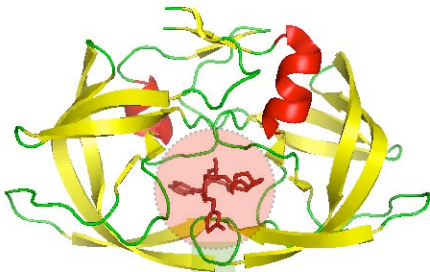
*A good model has MPQS of 1.1 or higher*



# Comparative docking

## 1. Expansion

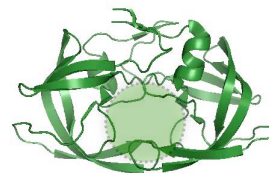
co-crystallized protein/ligand



crystalized protein

## 2. Inheritance

model



template

# Summary table

models with inherited ligands

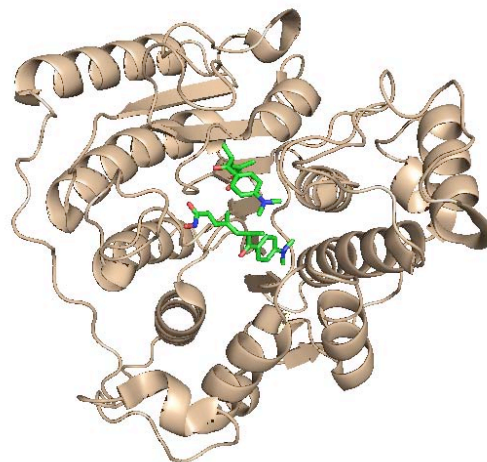
from 16,284 good models, 295 inherited a ligand/substance with at least one compound already approved by FDA and ready to be used from ZINC

|                        | Transcripts   | Good          | Ligands      | Lipinski     | Lipinski+ZINC | FDA+ZINC         |
|------------------------|---------------|---------------|--------------|--------------|---------------|------------------|
| <i>C. hominis</i>      | 3,886         | 886           | 183          | 131          | 28            | 12 (10)          |
| <i>C. parvum</i>       | 3,806         | 949           | 219          | 145          | 30            | 12 (10)          |
| <i>L. major</i>        | 8,274         | 1,845         | 488          | 334          | 84            | 44 (34)          |
| <i>M. leprae</i>       | 1,605         | 1,321         | 286          | 189          | 39            | 29 (25)          |
| <i>M. tuberculosis</i> | 3,991         | 2,887         | 404          | 285          | 71            | 44 (37)          |
| <i>P. falciparum</i>   | 5,363         | 1,057         | 271          | 191          | 48            | 20 (16)          |
| <i>P. vivax</i>        | 5,342         | 1,042         | 267          | 177          | 37            | 18 (15)          |
| <i>T. brucei</i>       | 921           | 1,795         | 440          | 309          | 94            | 46 (36)          |
| <i>T. cruzi</i>        | 19,607        | 3,915         | 730          | 493          | 127           | 62 (52)          |
| <i>T. gondii</i>       | 7,793         | 587           | 174          | 124          | 28            | 8 (7)            |
| <b>TOTAL</b>           | <b>60,588</b> | <b>16,284</b> | <b>3,462</b> | <b>2,378</b> | <b>586</b>    | <b>295 (242)</b> |

# Example of inheritance (expansion)

*LmjF21.0680 from L. major “Histone deacetylase 2” (model 1)*

*Template 1t64A a human HDAC8 protein.*



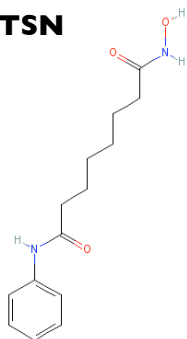
|            | Origen   | Formula   | Name                                      | Cov.   | Seq. Id. (%) |
|------------|----------|---|---|--------|--------------|
| <b>ZN</b>  | X-ray    | Zn <sup>2+</sup>  | Zinc ion                                  | --     | --           |
| <b>NA</b>  | X-ray    | Na <sup>+</sup>   | Sodium ion                                | --     | --           |
| <b>CA</b>  | X-ray    | Ca <sup>2+</sup>  | Calcium ion                               | --     | --           |
| <b>TSN</b> | X-ray    | C <sub>17</sub> H <sub>22</sub> N <sub>2</sub> O <sub>3</sub> | Trichostatin A                            | --     | --           |
| <b>SHH</b> | Expanded | C <sub>14</sub> H <sub>20</sub> N <sub>2</sub> O <sub>3</sub> | Octadenioic acid hydroxyamide phenylamide | 100.00 | 83.8         |

# Example of inheritance (inheritance)

*LmjF21.0680 from L. major "Histone deacetylase 2" (model 1)*

|            | Formula   | Name                                      | Cov.   | Seq. Id. (%) | Residues                                      |
|------------|---|---|--------|--------------|---|
| <b>TSN</b> | C <sub>17</sub> H <sub>22</sub> N <sub>2</sub> O <sub>3</sub> | Trichostatin A                            | 100.00 | 90.9         | 90 131 132 140 141 167<br>169 256 263 293 295 |
| <b>SHH</b> | C <sub>14</sub> H <sub>20</sub> N <sub>2</sub> O <sub>3</sub> | Octadenioic acid hydroxyamide phenylamide | 100.00 | 90.9         |   |

**TSN**



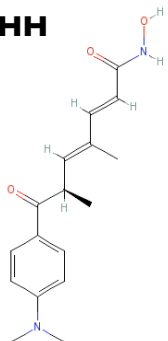
[suberoylanilide hydroxamic acid](#)

**Pharmacological Action:**

[Anti-Inflammatory Agents, Non-Steroidal](#)  
[Antineoplastic Agents](#)  
[Enzyme Inhibitors](#)  
[Anticarcinogenic Agents](#)

Inhibits histone deacetylase 1 and 3

**SHH**



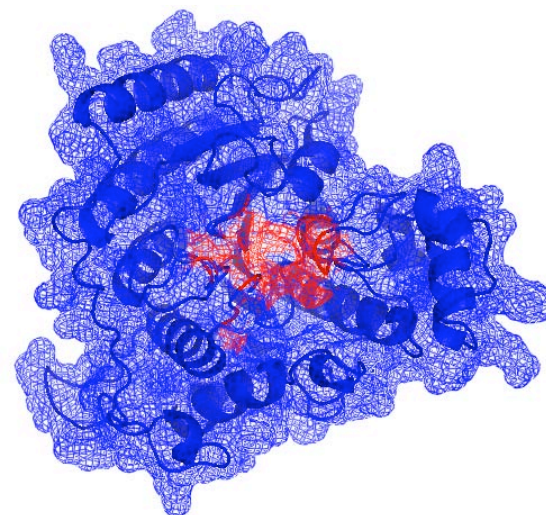
[trichostatin A](#)

**Pharmacological Action:**

[Antibiotics, Antifungal](#)  
[Enzyme Inhibitors](#)  
[Protein Synthesis Inhibitors](#)

chelates zinc ion in the active site of histone deacetylases, resulting in preventing histone unpacking so DNA is less available for transcription

|                    | LmjF21.0680.1.pdb |
|--------------------|-------------------|
| <b>Template</b>    | 1t64A             |
| <b>Seq. Id (%)</b> | <b>38.00</b>      |
| <b>MPQS</b>        | 1.47              |



# Example of inheritance (CDD-Roos-literature)

*LmjF21.0680 from L. major “Histone deacetylase 2” (model 1)*

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

Departments of \*Parasite Biochemistry and Cell Biology, <sup>‡</sup>Medicinal Chemistry, and <sup>§</sup>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**

# Models database

<http://sgu.bioinfo.cipf.es/services/TDIModels/>

The TDIModels server

Results for **O96526** [O96526 Cdc2-related kinase (Cell division related protein)]  
Number of models: 2

**TDIModels** ⓘ

[SGU-HOME]  
DBAli  
Eva-CM  
SeqProfCod  
TDIModels

**Model 1**

JMOL

This model has 1 predicted ligands.

| Lipinski                            | ZINC                     | FDA                      | Coverage | Seq. Id. |
|-------------------------------------|--------------------------|--------------------------|----------|----------|
| <input checked="" type="checkbox"/> | <input type="checkbox"/> | <input type="checkbox"/> | 100.00   | 100.00   |

**NO3**

SEQUENCE IDENTITY: 58.00  
MODPIPE QUALITY SCORE: 1.73  
TEMPLATE PDB: 1gz8  
TEMPLATE CHAIN: A  
TARGET LENGTH: 311  
TARGET BEGIN: 20  
TARGET END: 309  
[Download PDB file](#)

**Model 2**

JMOL

This model has 2 predicted ligands.

| Lipinski                 | ZINC                                | FDA                                 | Coverage | Seq. Id. |
|--------------------------|-------------------------------------|-------------------------------------|----------|----------|
| <input type="checkbox"/> | <input checked="" type="checkbox"/> | <input type="checkbox"/>            | 100.00   | 100.00   |
| <input type="checkbox"/> | <input type="checkbox"/>            | <input checked="" type="checkbox"/> | 100.00   | 93.75    |

**NO3**  
**KCX**

SEQUENCE IDENTITY: 29.00  
MODPIPE QUALITY SCORE: 1.13  
TEMPLATE PDB: 2cn5  
TEMPLATE CHAIN: A  
TARGET LENGTH: 311  
TARGET BEGIN: 1  
TARGET END: 311  
[Download PDB file](#)

<- new search

HELP:

# Acknowledgments

## Structural Genomics Unit (CIPF)

Marc A. Martí-Renom  
Emidio Capriotti  
Peio Ziaresolo Areitioaurtena

## Comparative Genomics Unit (CIPF)

Hernán Dopazo  
Leo Arbiza  
François Serra

## Functional Genomics Unit (CIPF)

Joaquín Dopazo  
Fátima Al-Shahrour  
José Carbonell  
Ignacio Medina  
David Montaner  
Joaquín Tàrraga  
Ana Conesa  
Toni Gabaldón  
Eva Alloza  
Lucía Conde  
Stefan Goetz  
Jaime Huerta Cepas  
Marina Marcet  
Pablo Minguez  
Pablo Escobar  
Rafael Jiménez

## FUNDING

Prince Felipe Research Center  
Marie Curie Reintegration Grant  
STREP EU Grant  
Generalitat Valenciana

## Tropical Disease Initiative

Stephen Maurer (UC Berkeley)  
Arti Rai (Duke U)  
Andrej Sali (UCSF)  
Ginger Taylor (TSL)

## STRUCTURAL GENOMICS

Stephen Burley (SGX)  
John Kuriyan (UCB)  
NY-SGXRC

## MAMMOTH

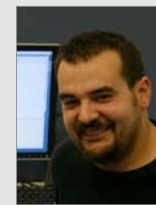
Angel R. Ortiz

## BIOLOGY

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

## FUNCTIONAL ANNOTATION

Fatima Al-Shahrour  
Joaquín Dopazo



## COMPARATIVE MODELING

Andrej Sali  
M. S. Madhusudhan  
Narayanan Eswar  
Min-Yi Shen  
Ursula Pieper  
Bino John  
Maya Topf

## FUNCTIONAL ANNOTATION

Andrea Rossi  
Fred Davis



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