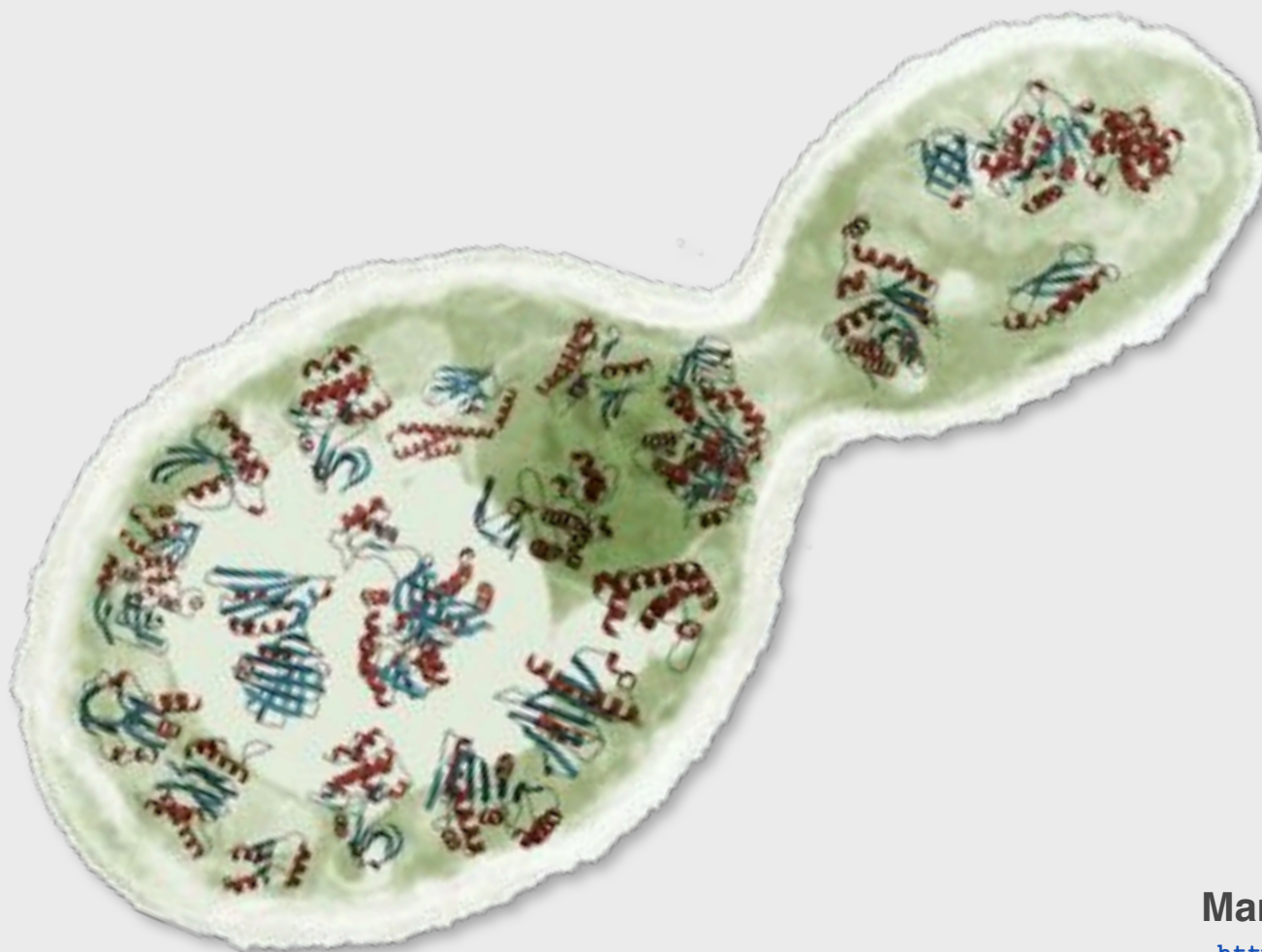


CIPF / Drug Discovery / Bioinformatics



Marc A. Marti-Renom

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

Structural Genomics Unit
Bioinformatics Department

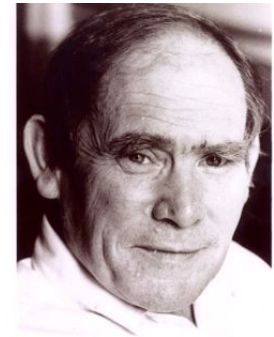
Centro de Investigación Príncipe Felipe (CIPF), Valencia, Spain



Data in the post-genomic era

Progress in science depends on new techniques, new discoveries and new ideas,
probably in that order.

Sydney Brenner, 1980



The introduction and popularization of high-throughput techniques has drastically changed the way in which biological problems **can** be addressed and hypotheses can be tested.

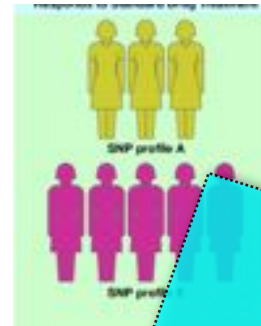
But not necessarily the way in which we really address or test them...

Genes in the DNA...



...code for proteins...

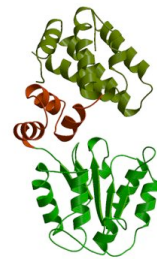
>protein kinase
acctgttgatggcgacagggactgtatgctg
atctatgctgatgcatgcatgctgactactgat



...produces the final phenotype

From genotype to phenotype.

...whose structure accounts



...plus the environment...

...which can be different because of the variability. 10 million SNPs



...whose final effect configures the phenotype...

Genes in the DNA...

Now: 22240 (NCBI build 35 12/04)

50-70% display alternative splicing

25%-60% unknown

Transfrags

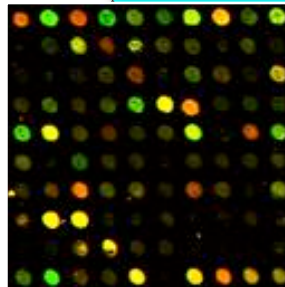
>protein kinase

```
acctgttgatggcgacagggtatgctgatctat
gctgatgcatgctgctgactactgatggtgggcta
ttgactgatgtctatc....
```



...when expressed in the proper moment and place...

A typical tissue is expressing among 5000 and 10000 genes



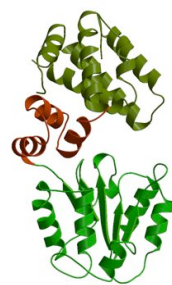
From genotype to phenotype.

(post-genomics scenario)

...code for proteins...

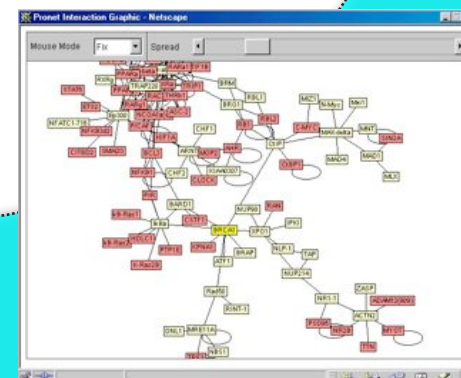
That undergo post-translational modifications, somatic recombination...

100K-500K proteins



...whose structures account for function...

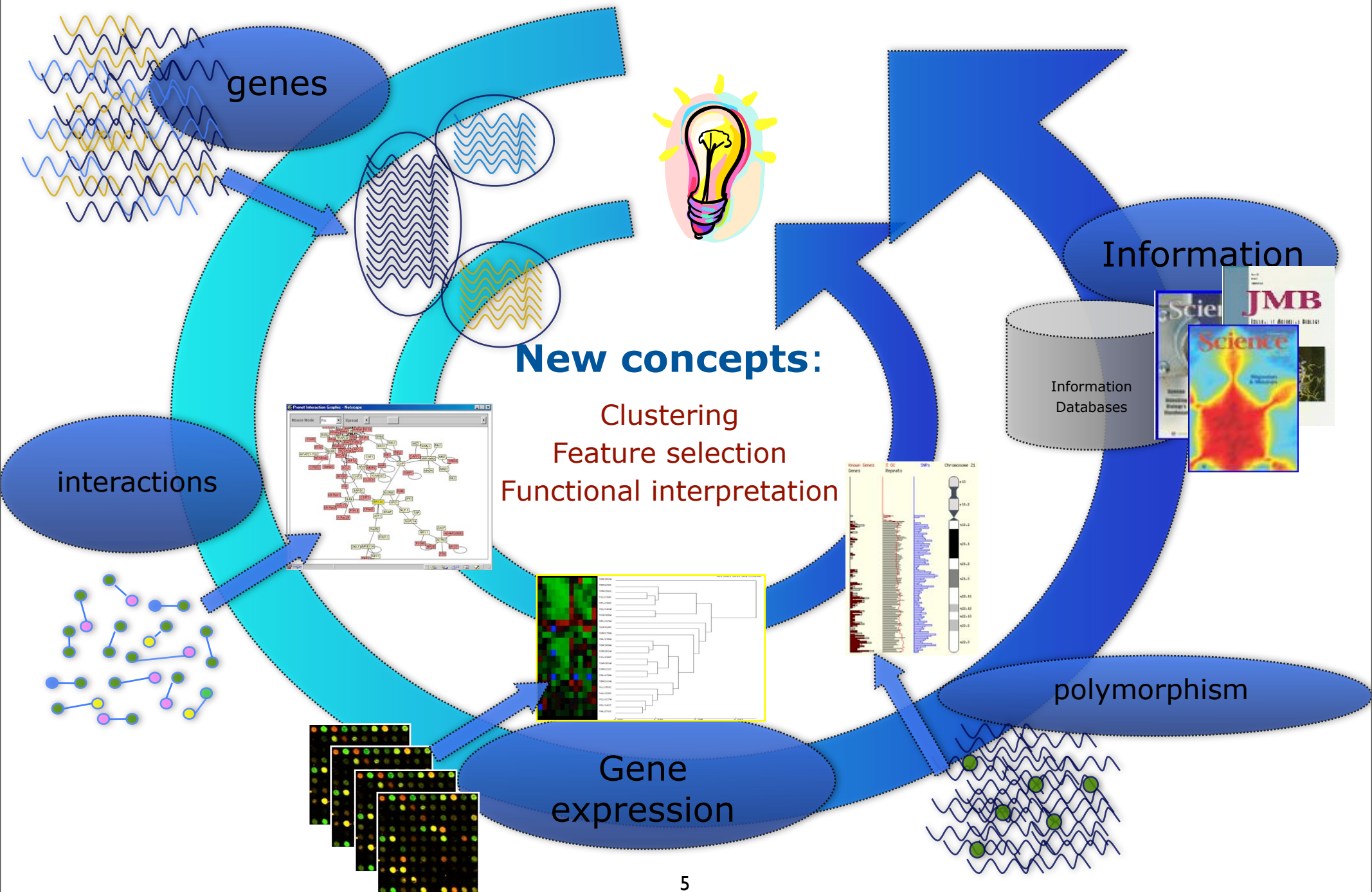
...conforming complex interaction networks...



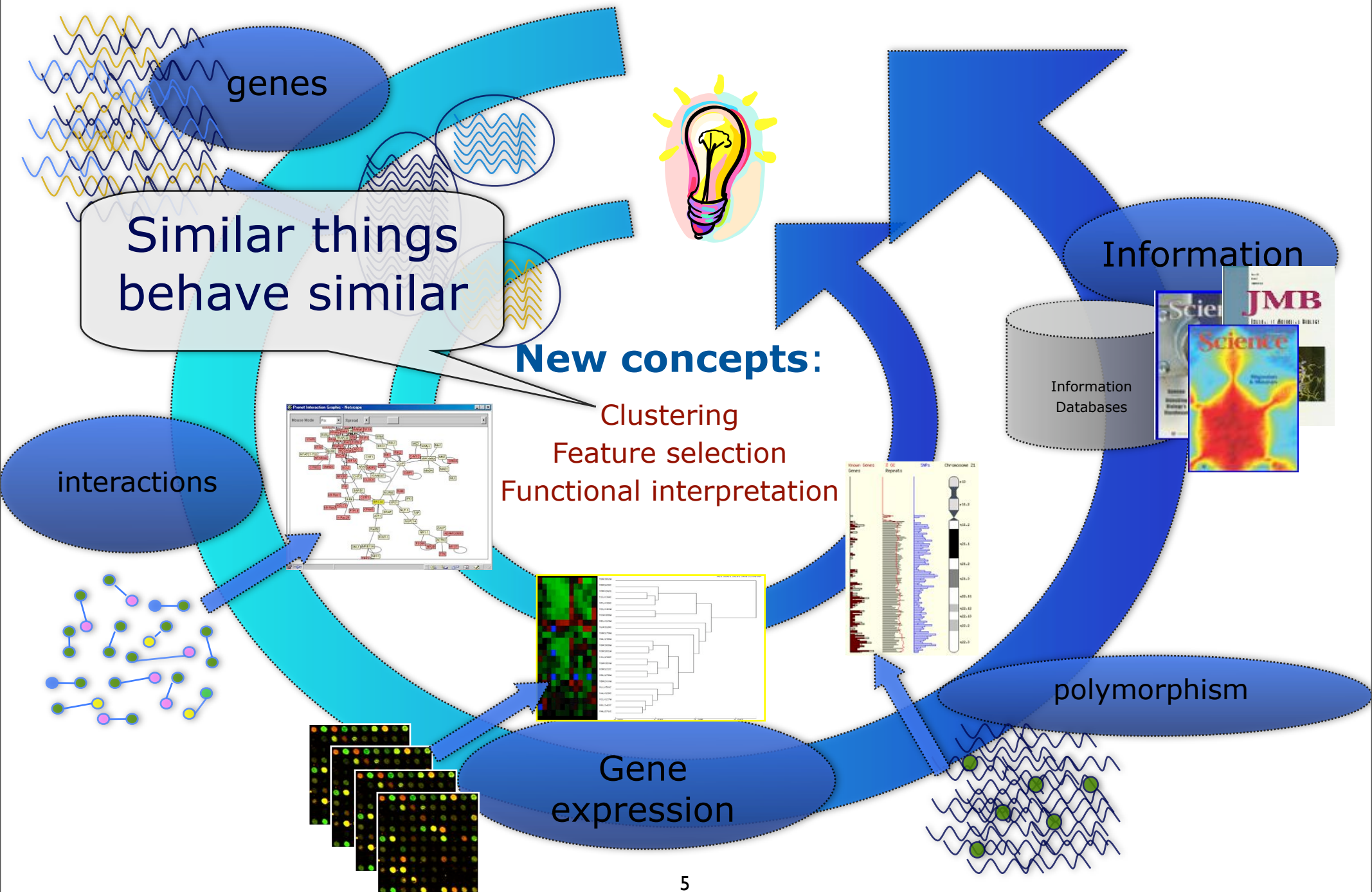
...in cooperation with other proteins...

Each protein has an average of 8 interactions

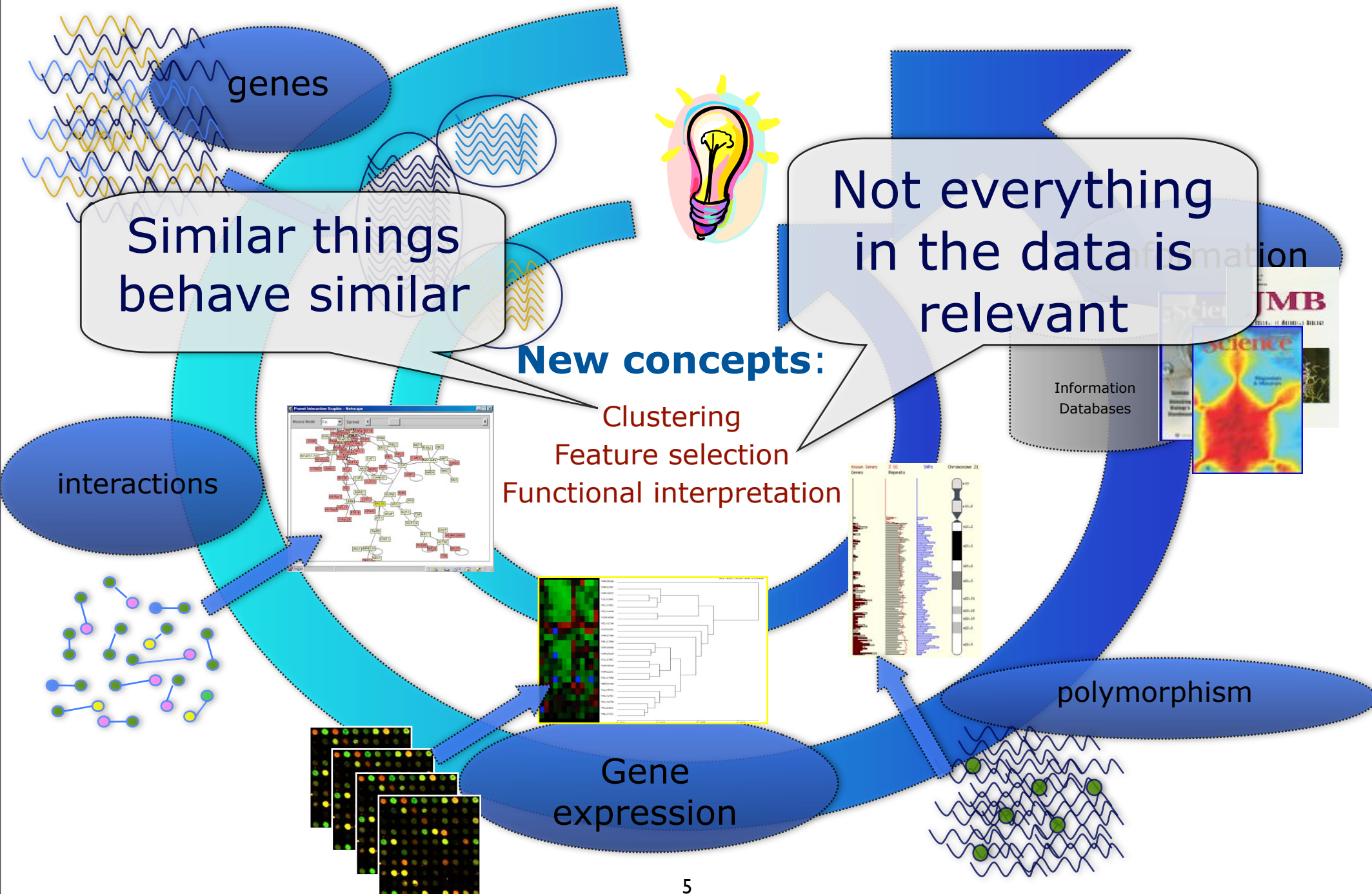
Similar things behave similar...



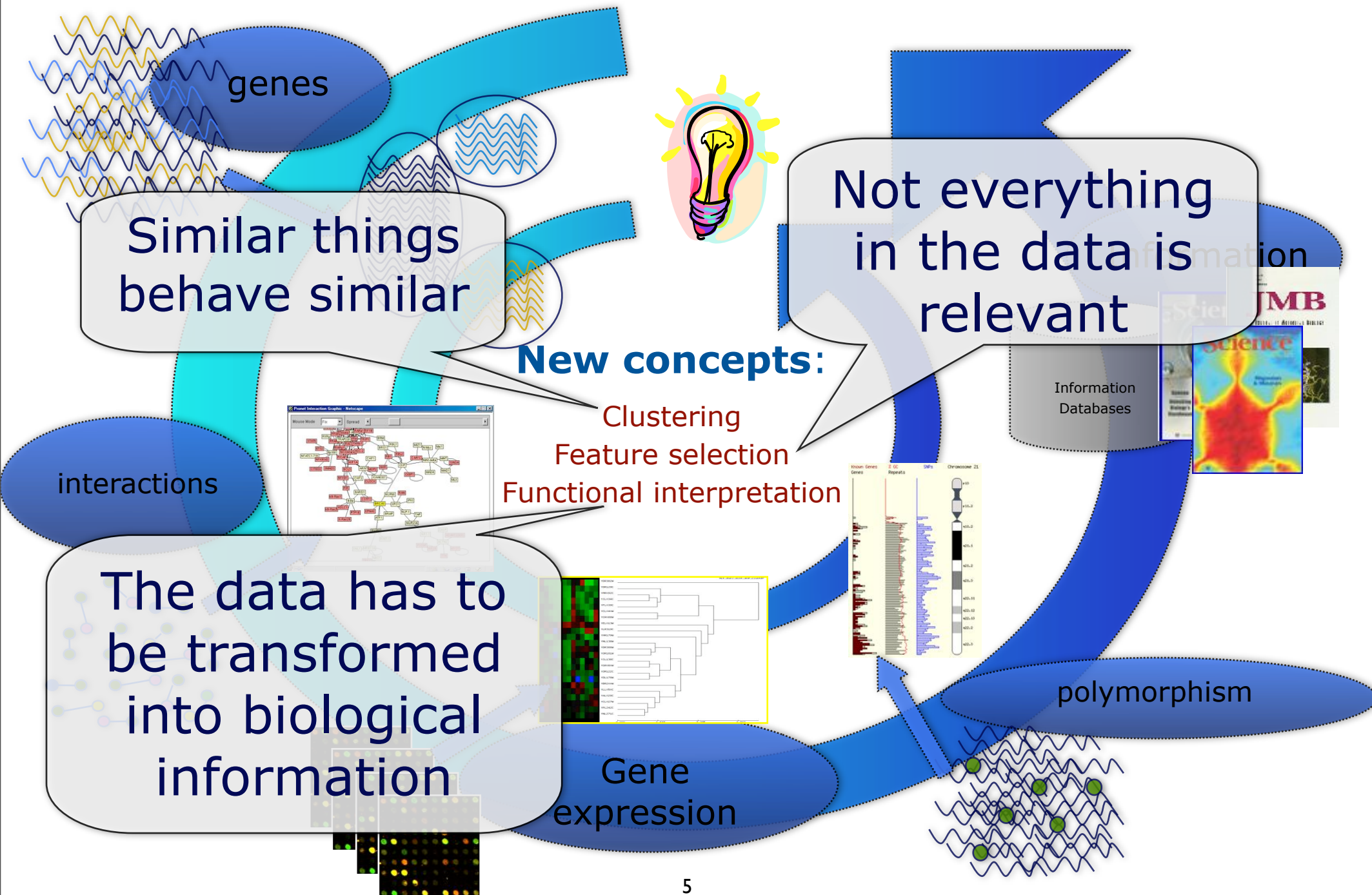
Similar things behave similar...



Similar things behave similar...



Similar things behave similar...



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



The screenshot shows the Bioinformatics Department website at CIPF, with the URL <http://bioinfo.cipf.es/> in the browser address bar. The website is organized into several columns and sections:

- The Department:** Lists the Functional Genomics Unit, Pharmacogenomics & Comparative Genomics Unit, and Structural Genomics Unit.
- Tools:** Lists DNA array data analysis, SNPs data analysis, Functional profiling, and Downloads.
- Documents & Publications:** Lists Papers, Communications, and Supplementary material.
- Meetings & Courses:** Lists Meetings & workshops, Courses, On line courses, and Accommodation.
- Coming events...:** Features logos for PRINCEPE FELITE, INS, C&G, ciber, GECOBIO, and INGENIO.
- Spotlight Tools:** Lists various tools including Access, Prophet, PopScribe, KASITE, FastGO, CAAT, FastScan, SDE, and Blast2GO.
- Packages:** Lists GEPAS, B&B, and a package for Data Analysis and Visualization in Genomics and Proteomics.
- News:** Contains a list of recent news items with dates and brief descriptions.
- More news...:** A link to view more news.
- Google:** A search bar with the text "Search WWW @ Search bioinfo.cipf.es".
- Postal address:** Provides the department's location, including the address, phone numbers, and fax number.

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



Functional Genomics
Dr. Joaquín Dopazo

GEPAS



BABELOMICS

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

Gene Expression Pattern Analysis Suite

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

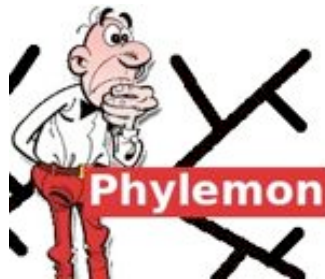
SNP Analysis Suite

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

Functional Profiling Analysis Suite



Comparative Genomics
Pharmacogenomics
Dr. Hernán Dopazo



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

Molecular Evolution Analysis Suite

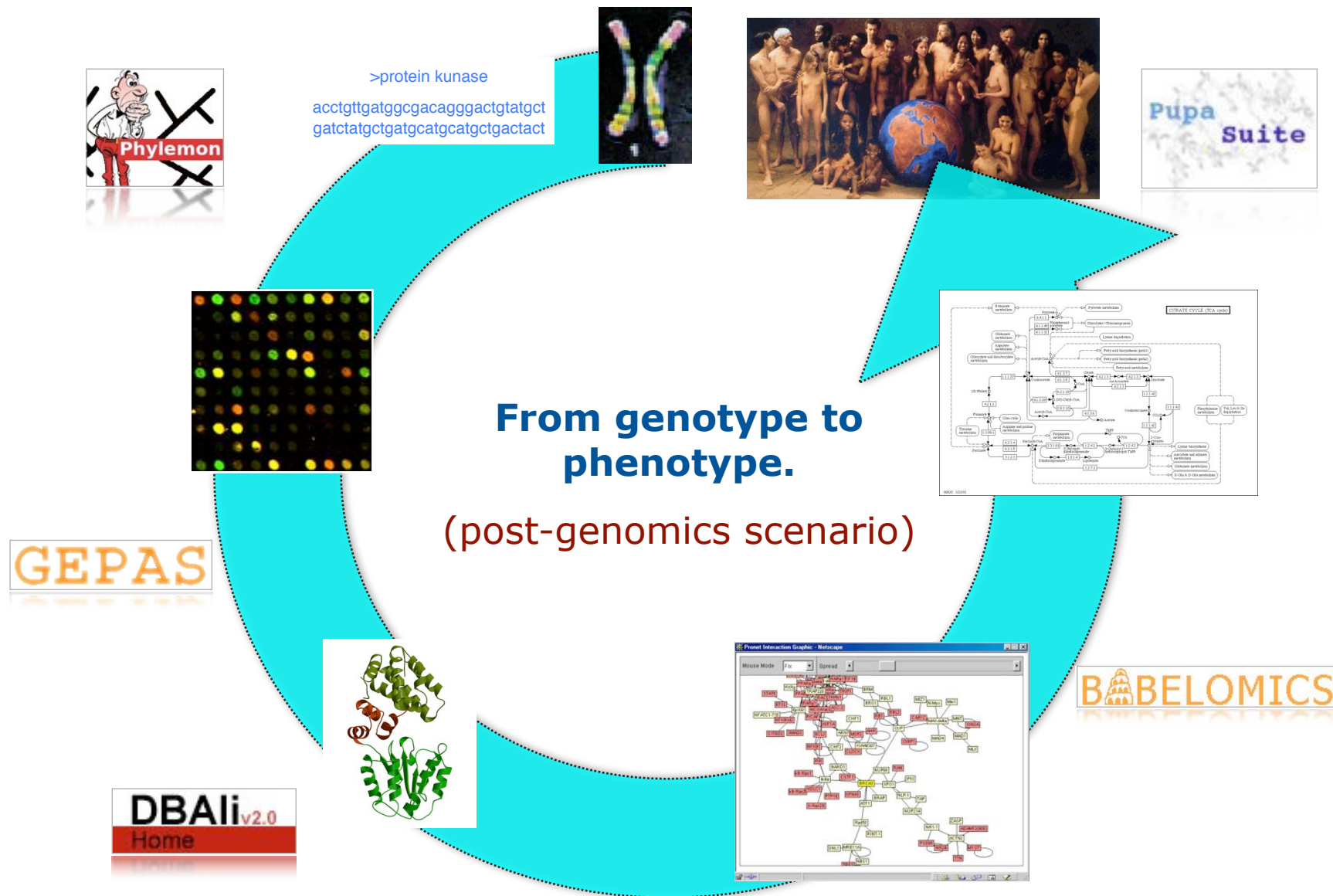


Structural Genomics
Dr. Marc A. Marti-Renom

DBAli_{v2.0}
Home

<http://www.dbali.org>

Structural Biology Analysis Suite



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



The screenshot shows the Bioinformatics Department website at CIPF, with the URL <http://bioinfo.cipf.es/> in the browser address bar. The website is organized into several columns and sections:

- The Department:** Lists the Functional Genomics Unit, Pharmacogenomics & Comparative Genomics Unit, and Structural Genomics Unit.
- Tools:** Lists tools for DNA array data analysis, SNPs data analysis, Functional profiling, and Downloads.
- Documents & Publications:** Lists Papers, Communications, and Supplementary material.
- Meetings & Courses:** Lists Meetings & workshops, Courses, On line courses, and Accommodation.
- Coming events...:** Features logos for PRINCEPE FELITE, INS, C&GEM, ciber, GECOBIO, and INGENIO.
- Spotlight Tools:** Lists tools such as Access, Popsuite, KASITE, FastGO, CAAT, FastScan, SDE, and Blast2GO.
- Packages:** Lists packages such as GEPAS, B&B, and C&GEM.
- News:** Contains a list of recent news items, including the availability of the first European Cluster for Scientific Computing with Free software on GEMMUS.
- More news...:** A link to more news.
- Google:** A search bar with the text "Search WWW @ Search bioinfo.cipf.es".
- Postal address:** Bioinformatics Department (CIPF), Av. Autopista del Sol, 16, (Carretera de los molinos), 46103 Valencia, Spain. Tel.: + 34 96 328 95 90, Fax: + 34 96 328 97 51, (see map).



Introduction to CM

Modeling genomes

Modeling Genes

Tropical Disease Initiative

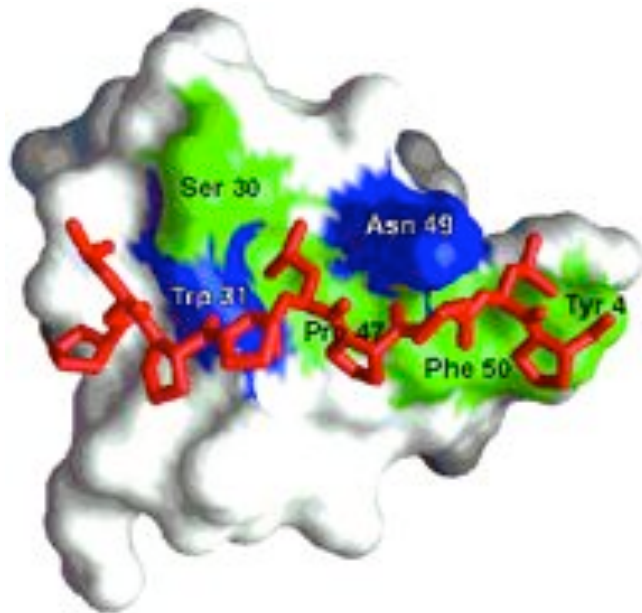
Why is it useful to know the **structure** of a protein, not only its sequence?

- ◆ 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

KARYGWSGQTKGDLGFLEGDIMEVTRIAGSWFYGKLLRNKKCSGYFPHLIF



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.**

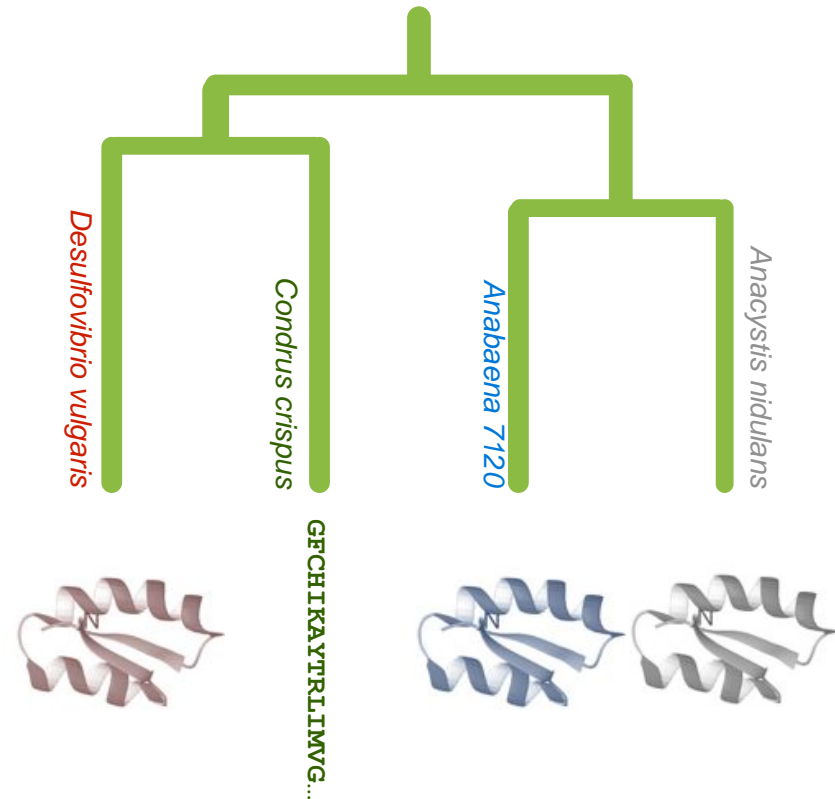
Principles of protein structure

GFCHIKAYTRLIMVG...



Folding (physics)

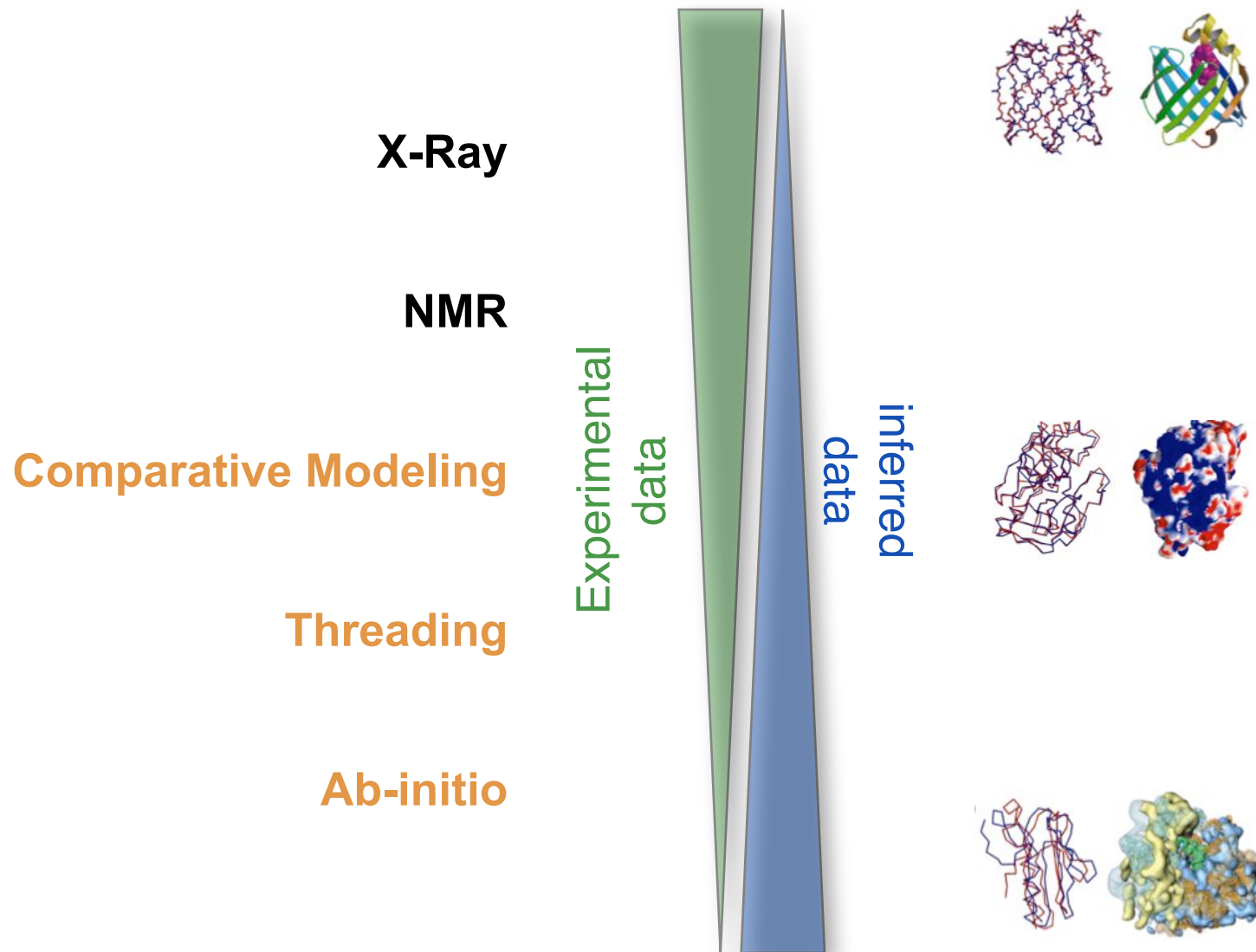
Ab initio prediction



Evolution (rules)

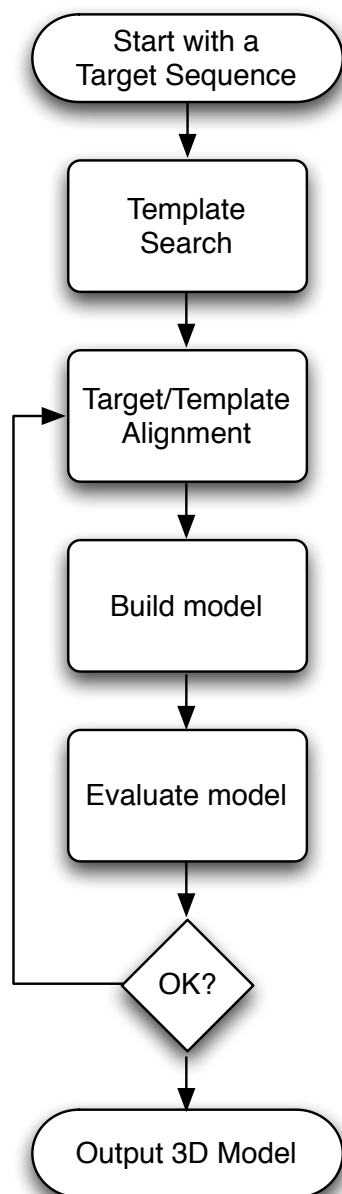
Threading
Comparative Modeling

protein prediction **vs** protein determination



Comparative modeling by satisfaction of spatial restraints

MODELLER



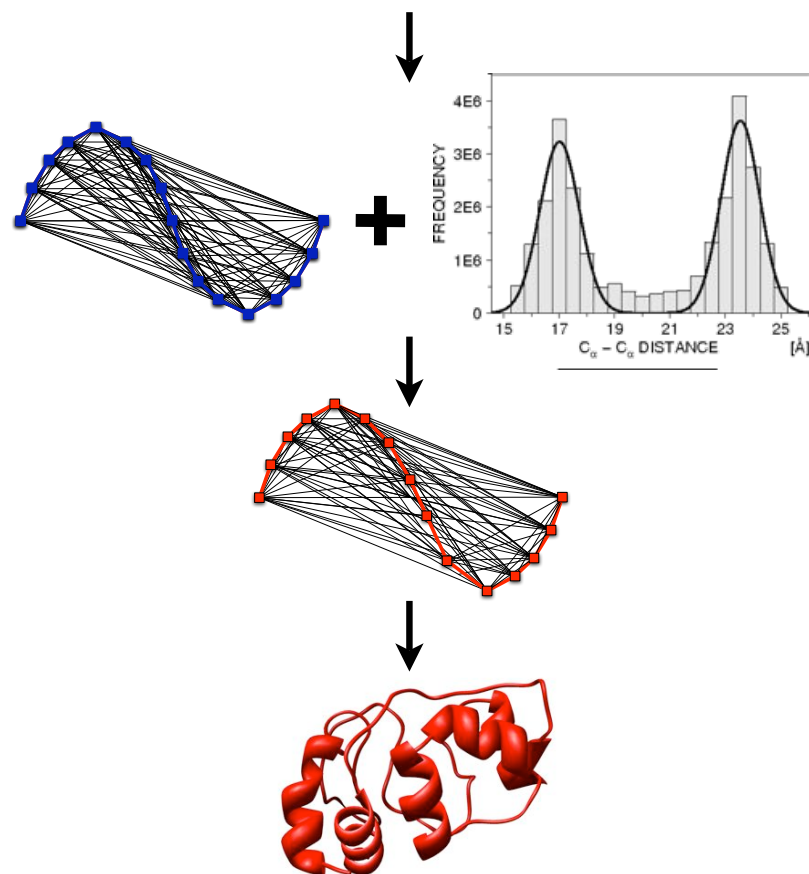
Given an alignment...

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

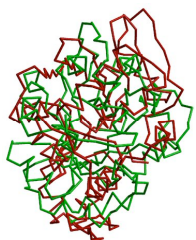
MSVIPKR--GNCEQTSE
ASILPKRLFGNCEQTSD



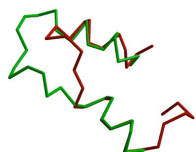
A. Šali & T. Blundell. *J. Mol. Biol.* 234, 779, 1993.
J.P. Overington & A. Šali. *Prot. Sci.* 3, 1582, 1994.
A. Fiser, R. Do & A. Šali, *Prot. Sci.*, 9, 1753, 2000.

Comparative modeling by satisfaction of spatial restraints

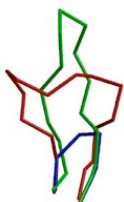
Types of errors and their impact



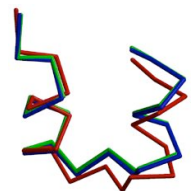
Wrong fold



Miss alignments



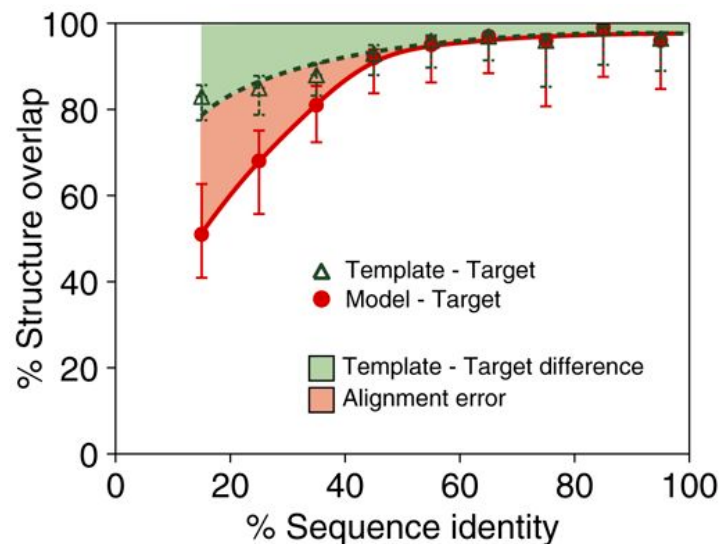
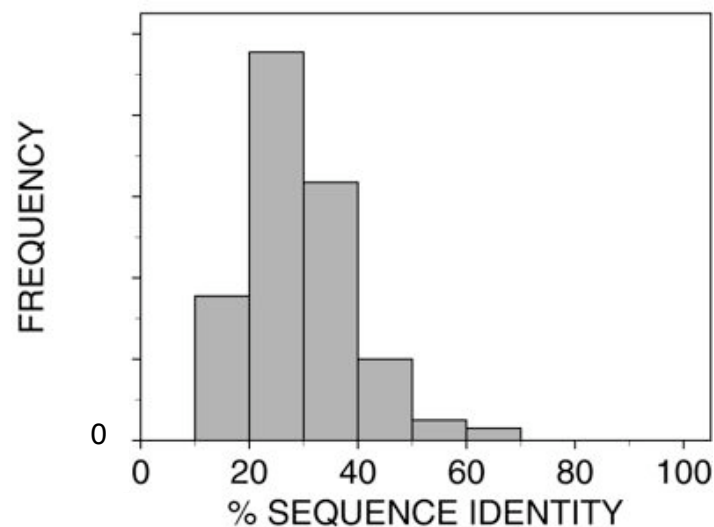
Loop regions



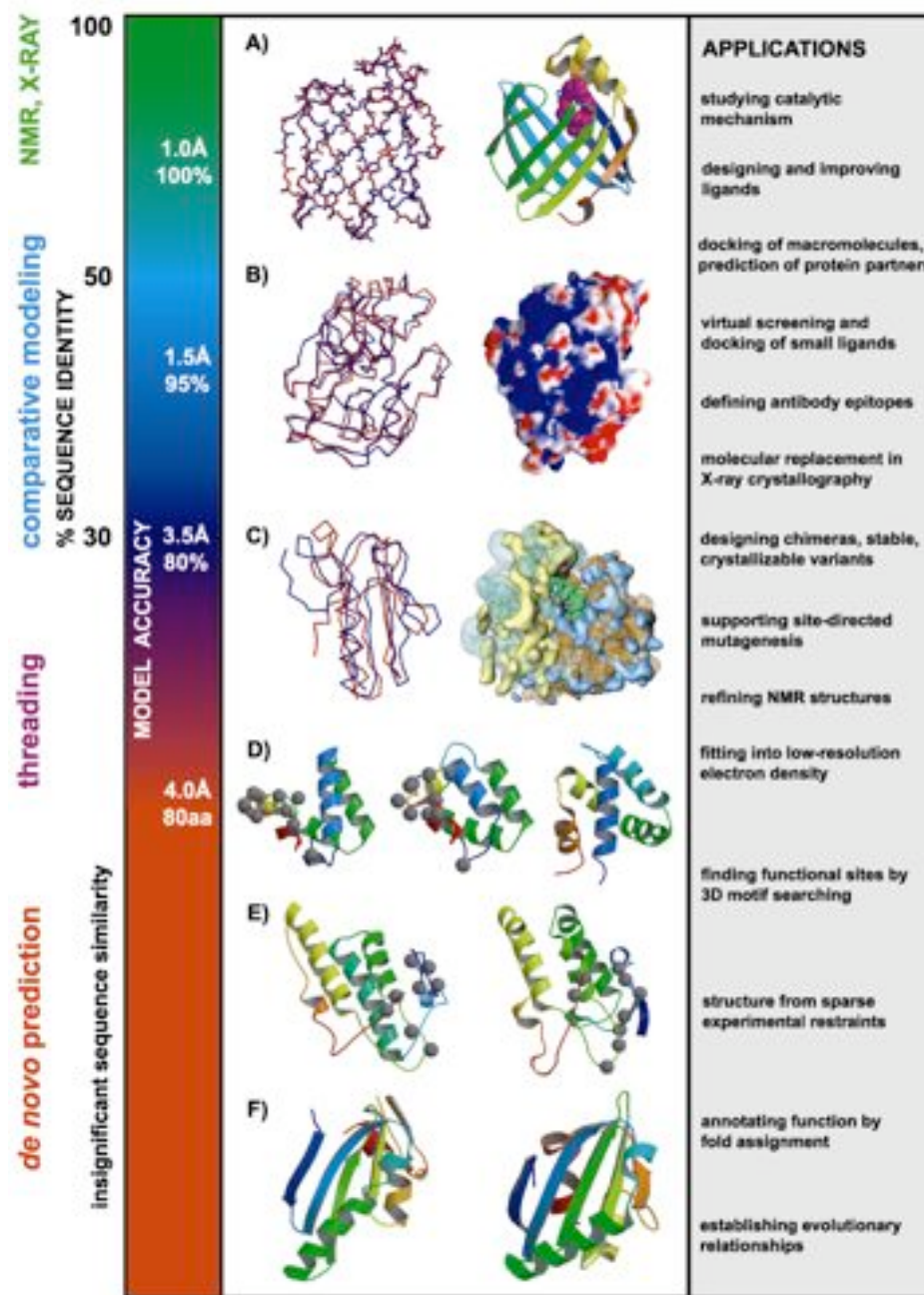
Rigid body distortions



Side-chain packing



Utility of protein structure models, despite errors

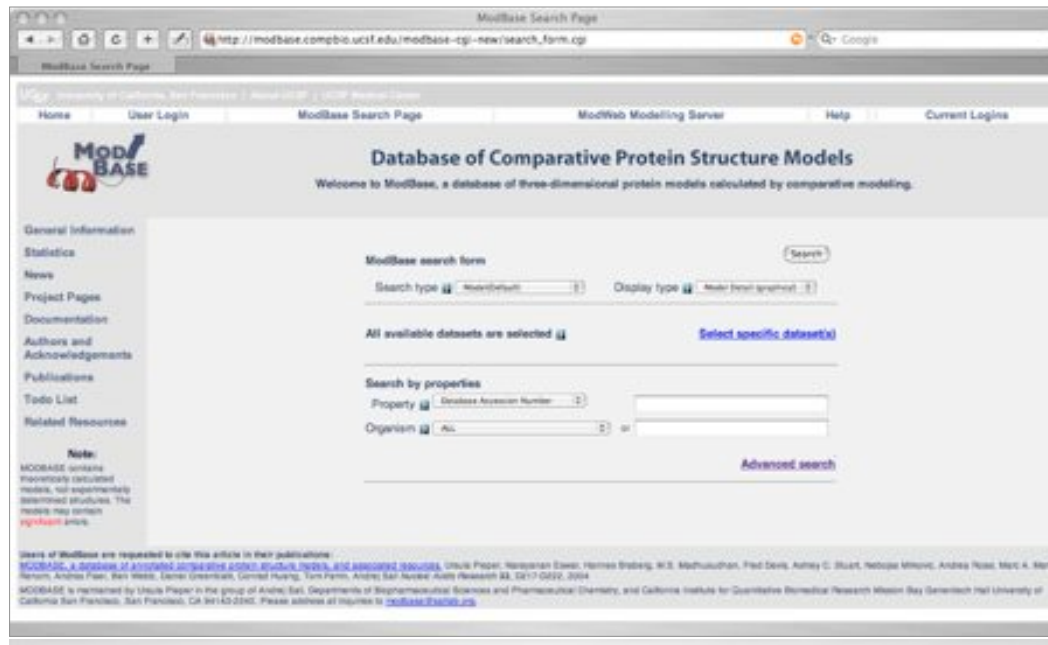


ModBase Statistics

Large-scale modeling of the TrEMBL-SWISSPROT databases

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

Sequences (total)	2,186,210
Sequences (modeled)	1,340,687
Models	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.compbio.ucsf.edu/modbase/cgi-bin/search_form.cgi". The page features a navigation bar with links: Home, User Login, ModBase Search Page, ModWeb Modeling 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 fields for "Search type" (set to "keyword"), "Display type" (set to "text/html"), and "All available datasets are selected" (with a link to "Select specific dataset(s)"). There is also a "Search by properties" section with fields for "Property" (set to "Database Accession Number") and "Organism" (set to "All"). A "Note" section at the bottom left states: "ModBase contains theoretically calculated models, not experimentally determined structures. The models may contain significant errors." The footer contains a list of users of ModBase and a request to cite the article in their publications.



University of California
San Francisco

Pieper et al. NAR 34, D291 (2006)

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

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

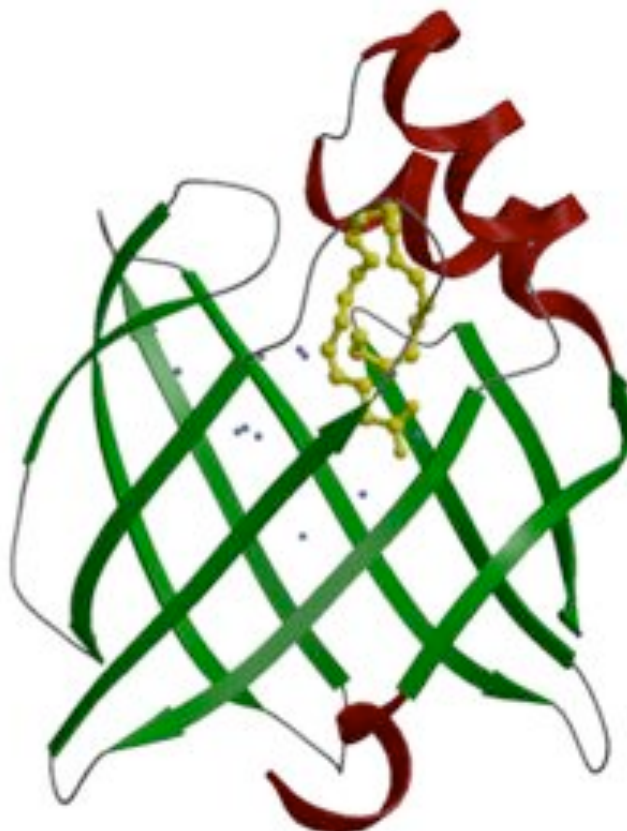
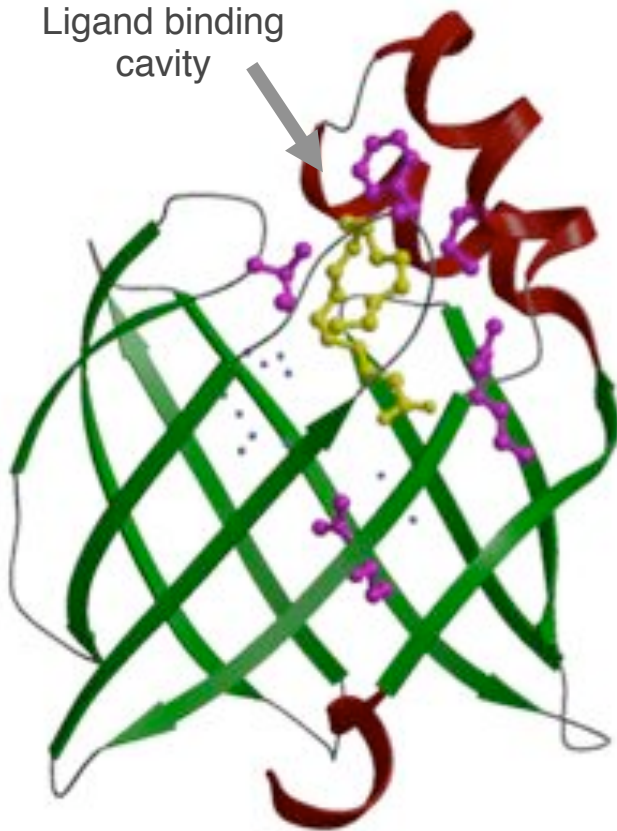
BLBP/oleic acid

BLBP/docosahexaenoic acid

Cavity is **not** filled

Cavity **is** filled

Ligand binding
cavity



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

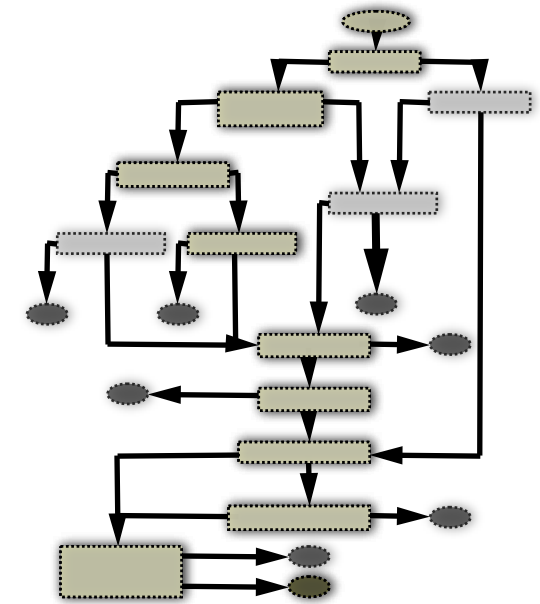
Cancer Research (June 2004). 64:3790-97

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.

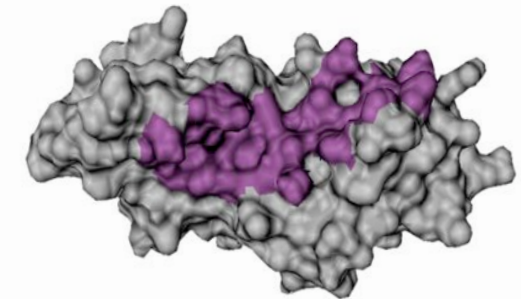
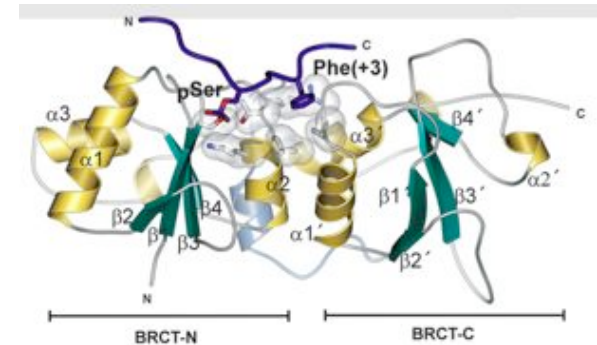
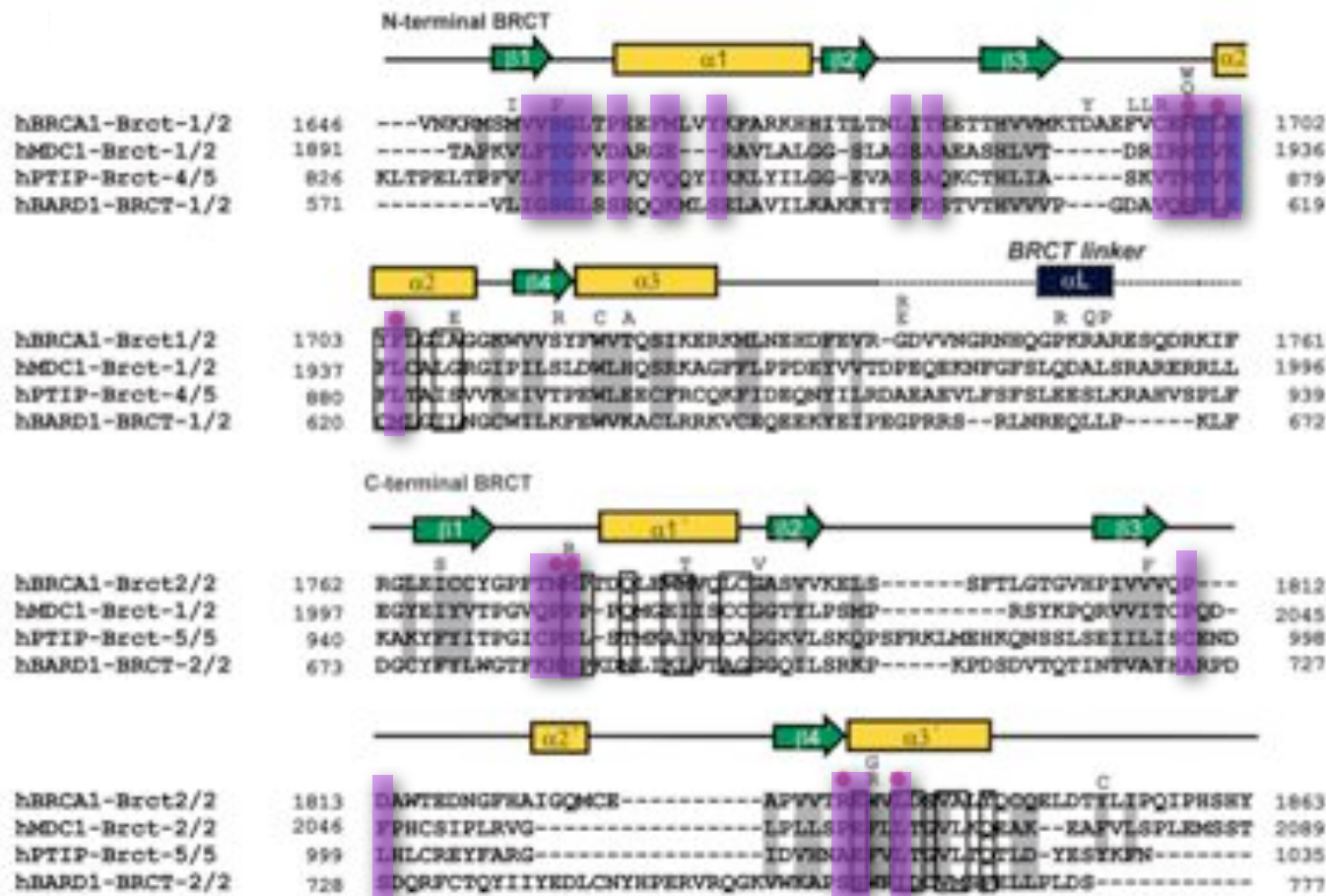


Missense mutations in BRCT domains by function

	cancer associated	not cancer associated	?				
no transcription activation	C1697R R1699W A1708E S1715R P1749R M1775R		M1652K L1657P E1660G H1686Q R1699Q K1702E Y1703HF1 704S	L1705PS1 715NS172 2FF1734L G1738EG 1743RA1 752PF176 1I	F1761S M1775E M1775K L1780P I1807S V1833E A1843T		
transcription activation		M1652I A1669S		V1665M D1692N G1706A D1733G M1775V P1806A			
?			M1652T V1653M L1664P T1685A T1685I M1689R D1692Y F1695L V1696L R1699L G1706E W1718C	W1718S T1720A W1730S F1734S E1735K V1736A G1738R D1739E D1739G D1739Y V1741G H1746N	R1751P R1751Q R1758G L1764P I1766S P1771L T1773S P1776S D1778N D1778G D1778H M1783T	C1787S G1788D G1788V G1803A V1804D V1808A V1809A V1809F V1810G Q1811R P1812S N1819S	A1823T V1833M W1837R W1837G S1841N A1843P T1852S P1856T P1859R



Putative binding site on BRCA1

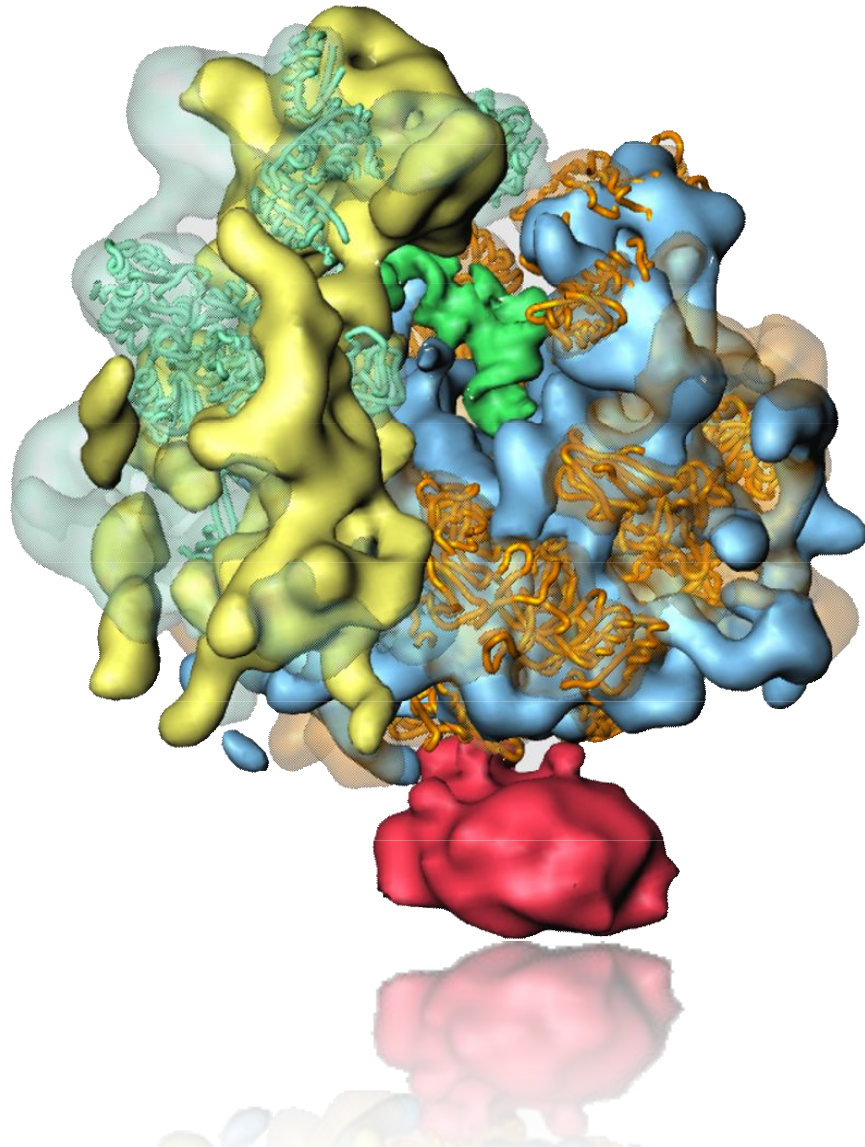


Putative binding site predicted in 2003
and accepted for publication on March 2004.

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

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

S. cerevisiae ribosome



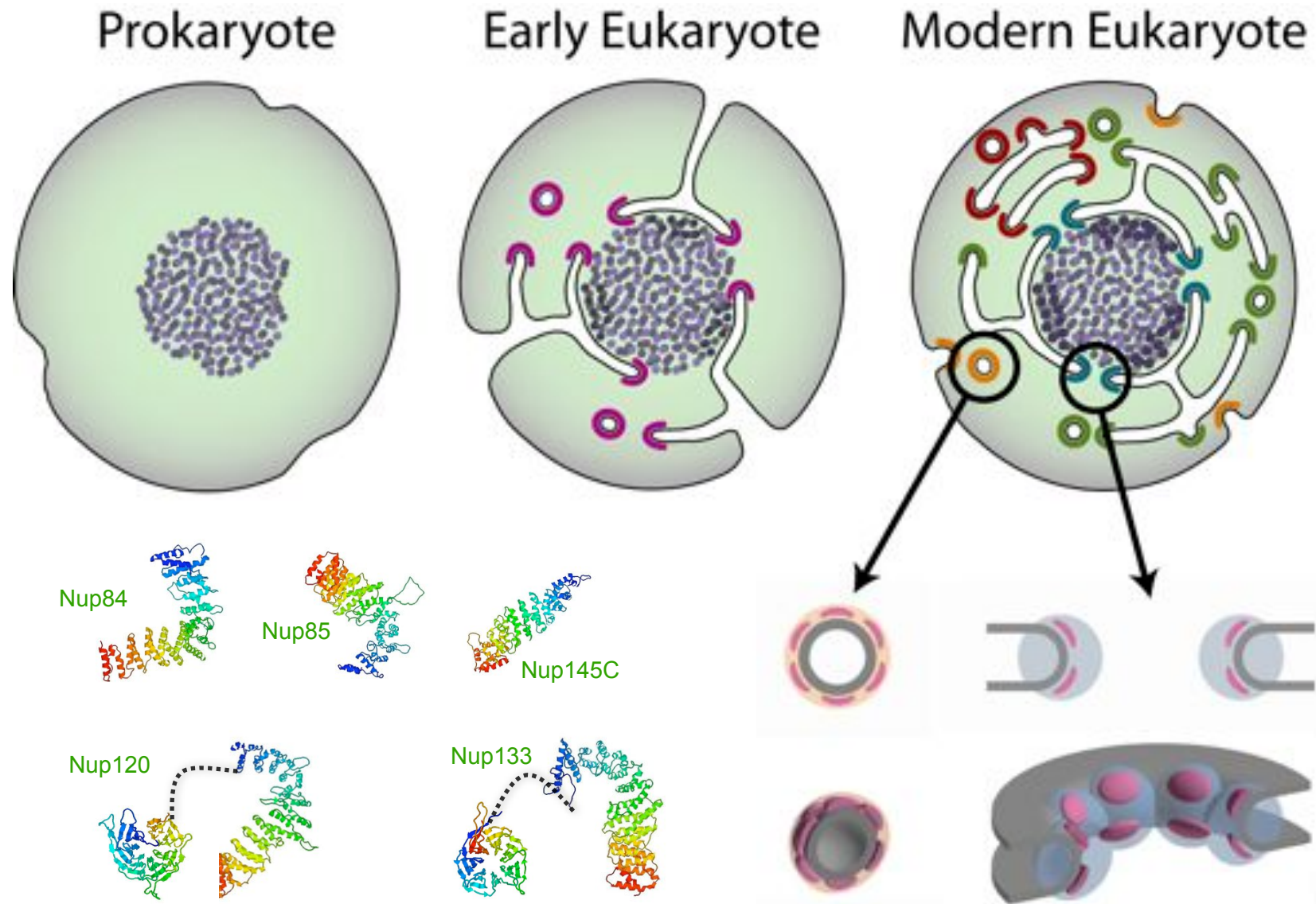
Fitting of comparative models into 15Å cryo-electron density map.

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 (?)



Tropical Disease Initiative (TDI)

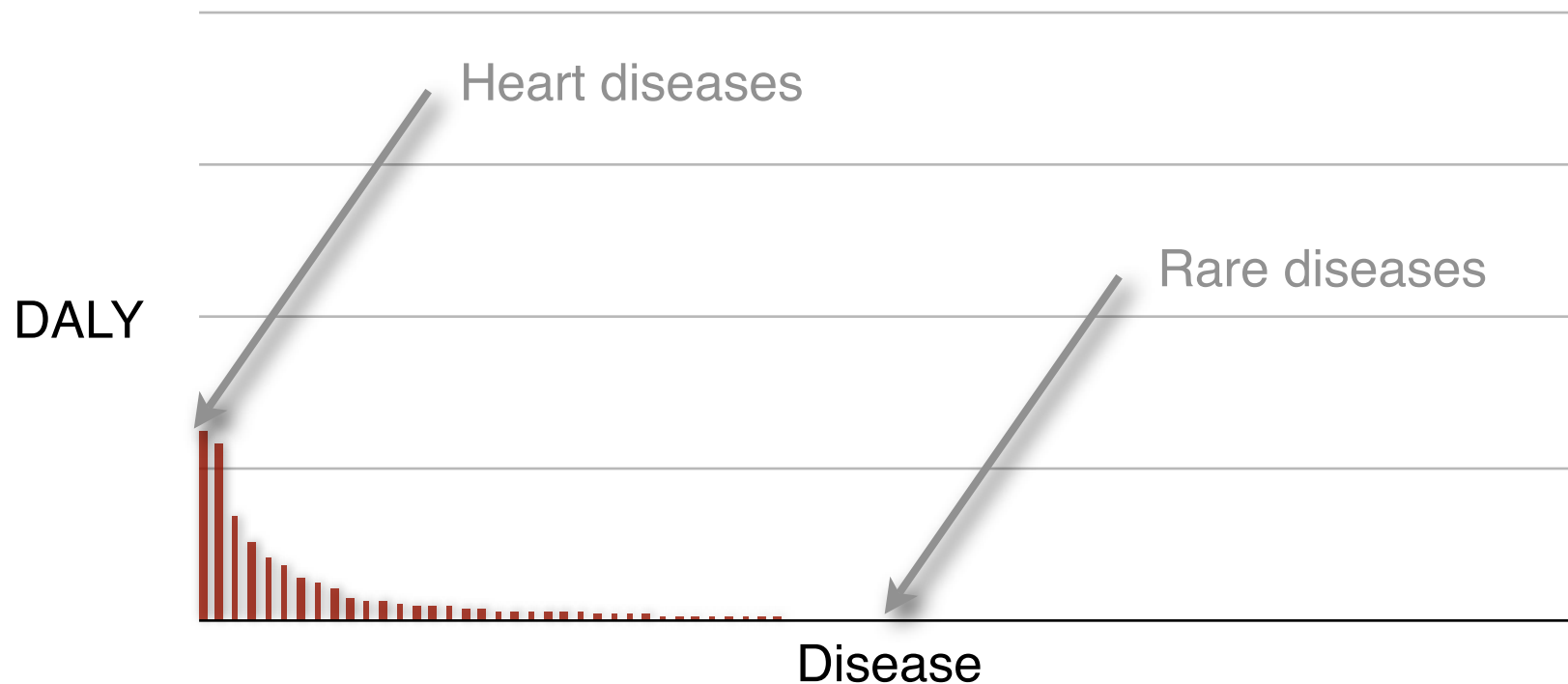
Predicting binding sites in protein structure models.



<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*

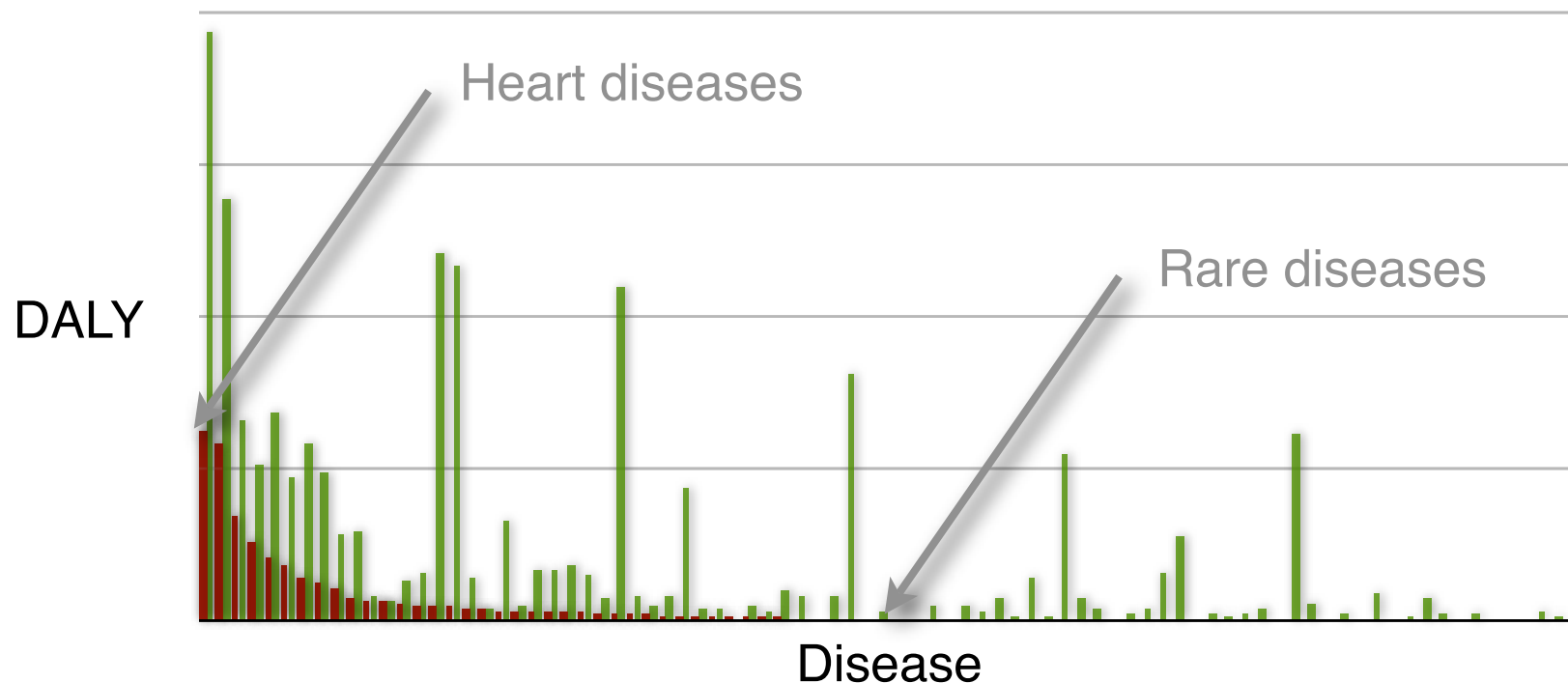
DALY - Disability adjusted life years

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*

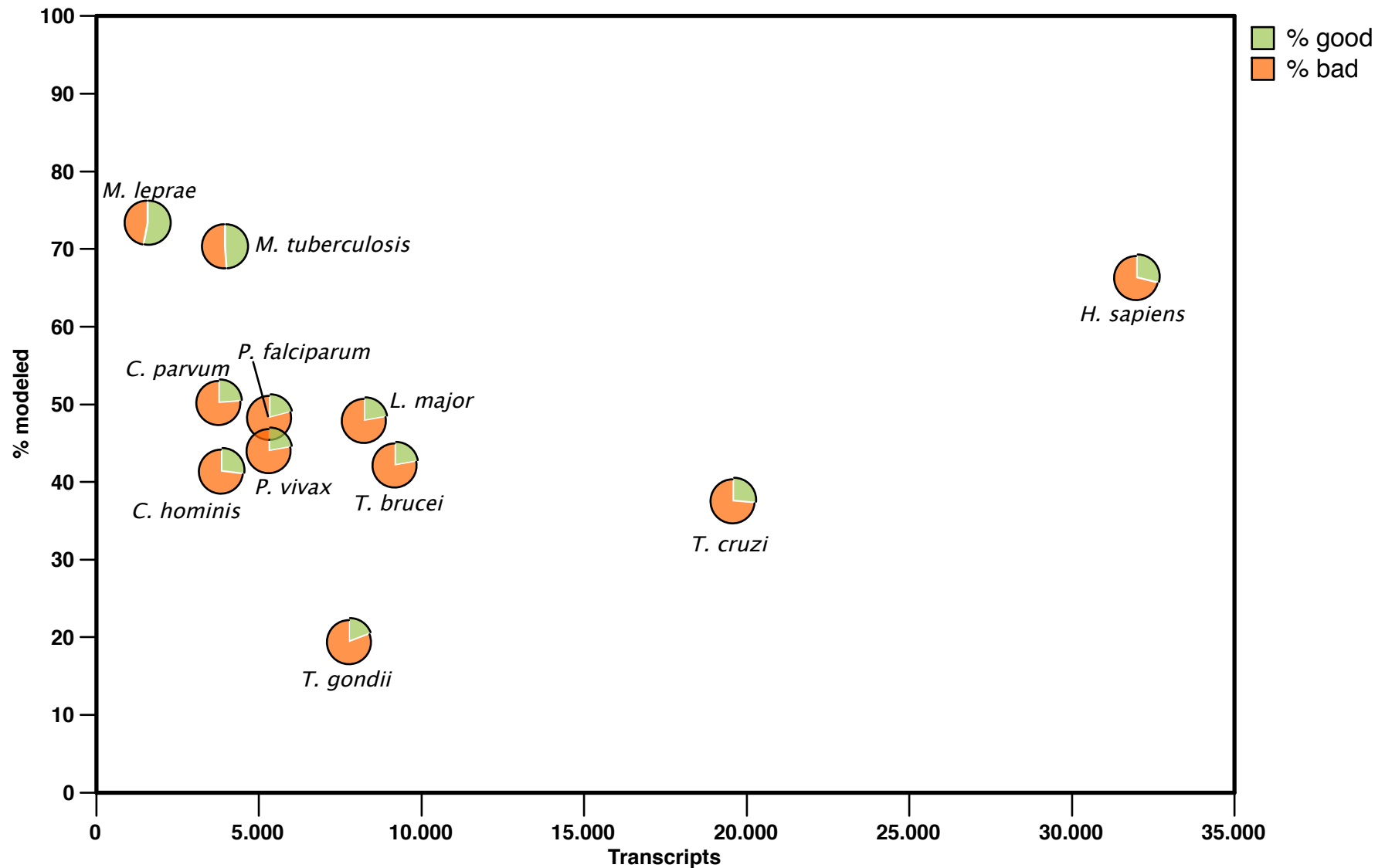
DALY - Disability adjusted life years

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.

Modeling Genomes

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

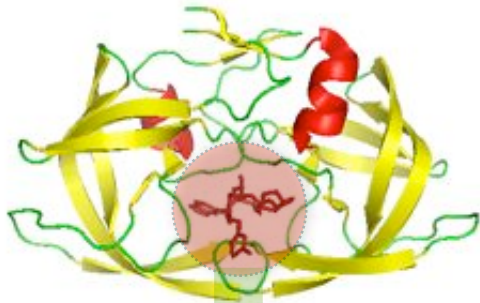


A good model has MPQS of 1.1 or higher

Comparative docking

1. Expansion

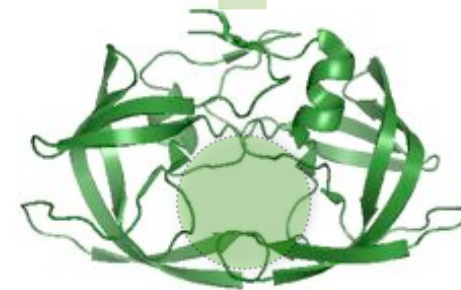
co-crystalized protein/ligand



crystallized 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

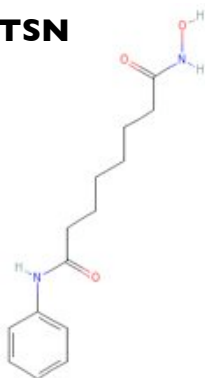
	Transcripts	Good	Ligands	Lipinski	Lipinski+ZINC	FDA+ZINC
<i>C. hominis</i>	3,886	886	183	131	28	12 (10)
<i>C. parvum</i>	3,806	949	219	145	30	12 (10)
<i>L. major</i>	8,274	1,845	488	334	84	44 (34)
<i>M. leprae</i>	1,605	1,321	286	189	39	29 (25)
<i>M. tuberculosis</i>	3,991	2,887	404	285	71	44 (37)
<i>P. 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)
TOTAL	60,588	16,284	3,462	2,378	586	295 (242)

Example of inheritance (inheritance)

LmjF2 1.0680 from L. major "Histone deacetylase 2" (model 1)

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

TSN



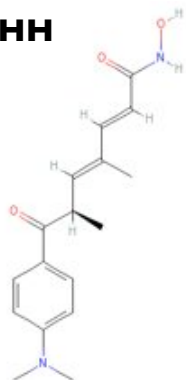
suberoylanilide hydroxamic acid

Pharmacological Action:

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

Inhibits histone deacetylase 1 and 3

SHH



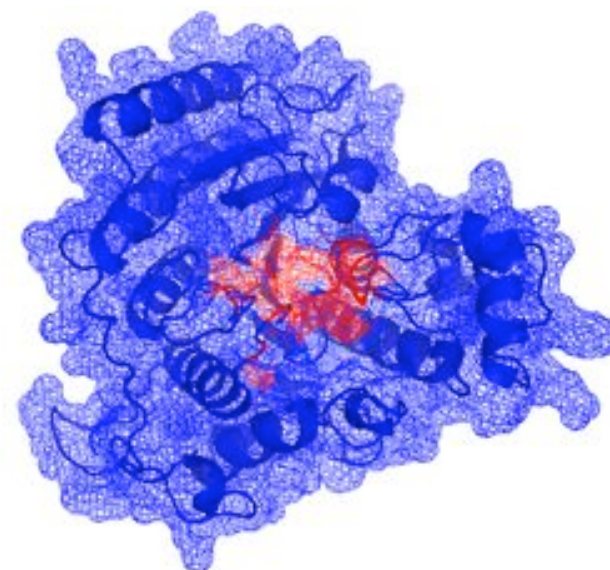
trichostatin A

Pharmacological Action:

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

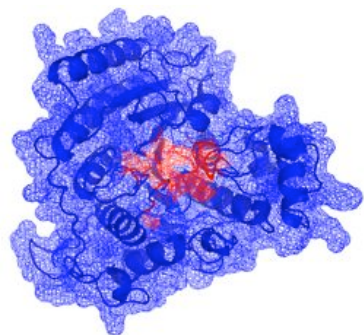
chelates zinc ion in the active site of histone deacetylases, resulting in preventing histone unpacking so DNA is less available for transcription

	LmjF2 1.0680.1.pdb
Template	1t64A
Seq. Id (%)	38.00
MPQS	1.47



Example of inheritance (CDD-Roos-literature)

LmjF2 1.0680 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^{*†}, ANNE M. GURNETT^{*}, ROBERT W. MYERS^{*}, PAULA M. DULSKI^{*},
TAMI M. CRUMLEY^{*}, JOHN J. ALLOCCO^{*}, CHRISTINE CANNOVA^{*}, PETER T. MEINKE[‡], STEVEN L. COLLETTI[‡],
MARIA A. BEDNAREK[‡], SHEO B. SINGH[§], MICHAEL A. GOETZ[§], ANNE W. DOMBROWSKI[§],
JON D. POLISHOOK[§], AND DENNIS M. SCHMATZ^{*}

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

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

Vol. 48, No. 4

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

Acknowledgments

Structural Genomics Unit (CIPF)

Marc A. Marti-Renom

Emidio Capriotti

Peio Ziarsolo Areitioaurtena

Comparative Genomics Unit (CIPF)

Hernán Dopazo

Leo Arbiza

Francisco García

Functional Genomics Unit (CIPF)

Joaquín Dopazo

Fátima Al-Shahrour

José Carbonell

Ignacio Medina

David Montaner

Joaquín Tárraga

Ana Conesa

Toni Gabaldón

Eva Alloza

Lucía Conde

Stefan Goetz

Jaime Huerta Cepas

Marina Marcet

Pablo Minguez

Jordi Burguet Castell

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)

Barri Bunin (CDD)

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>