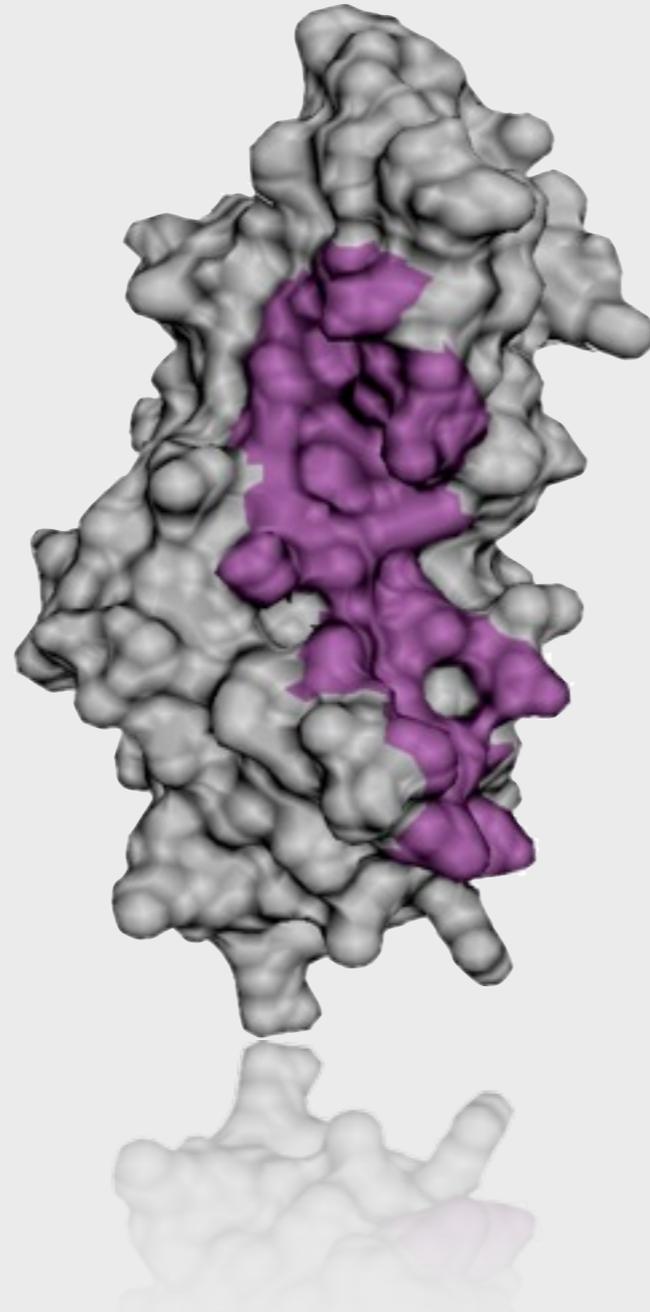


SNP analysis and binding site prediction



Marc A. Marti-Renom

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

Structural Genomics Unit
Bioinformatics Department

Prince Felipe Research Center (CIPF), Valencia, Spain



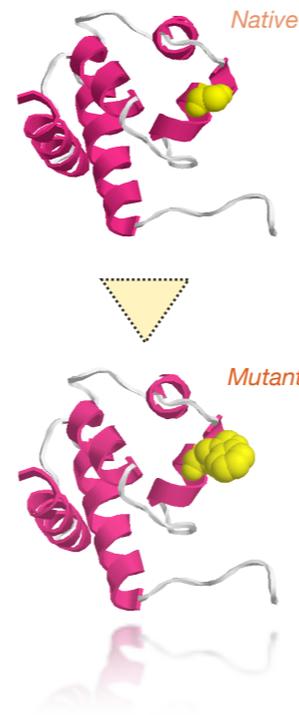
Program

SNP analysis
from sequence

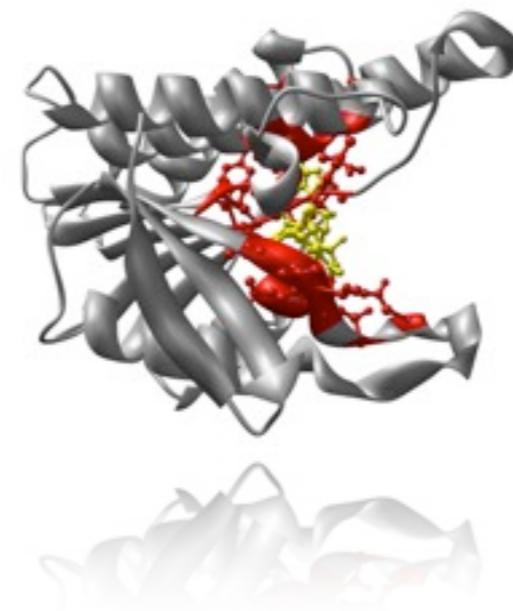
SNP analysis
from structure

Binding site
prediction

AutoDock



Disease?



Objective

TO **UNDERSTAND** THAT SNPs HAVE EFFECTS THAT CAN BE PREDICTED AND TO LEARN **HOW-TO** USE AutoDock FOR DOCKING SMALL MOLECULES IN THE SURFACE OF A PROTEIN

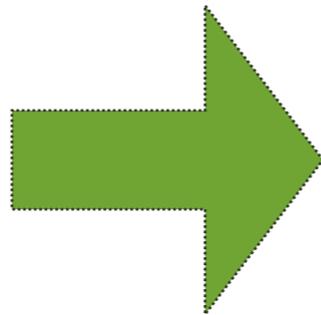
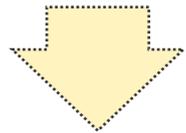
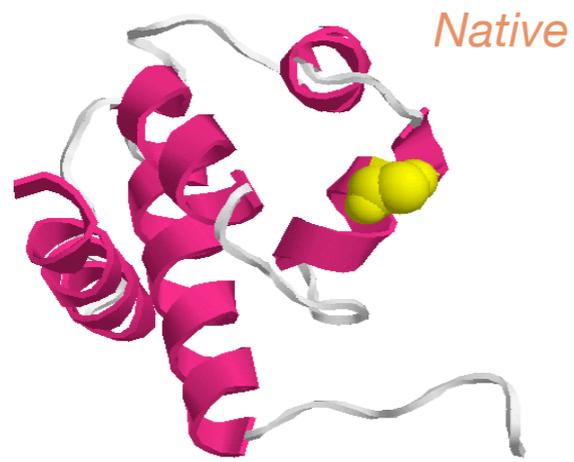
Nomenclature

SNP: Single Nucleotide Polymorphism. A single change in the DNA sequence, which may or may not result in a change in the protein sequence.

Ligand: Structure (usually a small molecule) that binds to the binding site.

Receptor: Structure (usually a protein) that contains the active binding site.

Binding site: Set of aminoacids (residues) that physically interact with the ligand (usually within 6 Ångstroms).



Disease?

Gene Sequence << +Protein Sequence << +Protein Structure

Single Nucleotide Polymorphism

Single Nucleotide Polymorphism or SNP

is a DNA sequence variation occurring when a single nucleotide - A, T, C, or G - in the genome differs between members of the species.

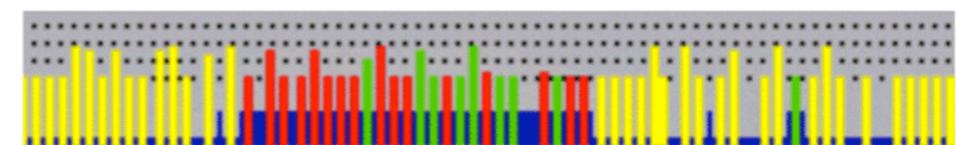
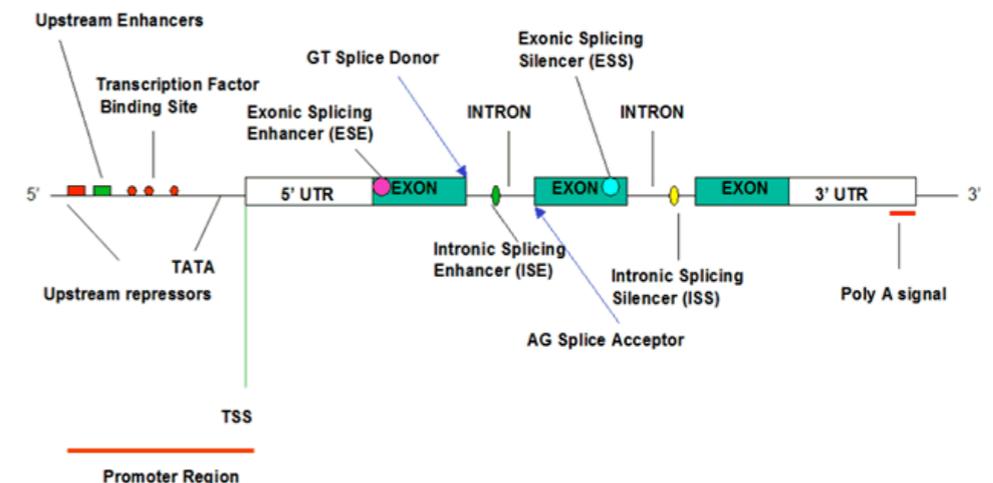
Usually one will want to refer to SNPs when the population frequency is $\geq 1\%$

SNPs occur at any position and can be classified on the base of their locations.

Coding SNPs can be subdivided into two groups:

Synonymous: when single base substitutions do not cause a change in the resultant amino acid

Non-synonymous: when single base substitutions cause a change in the resultant amino acid.



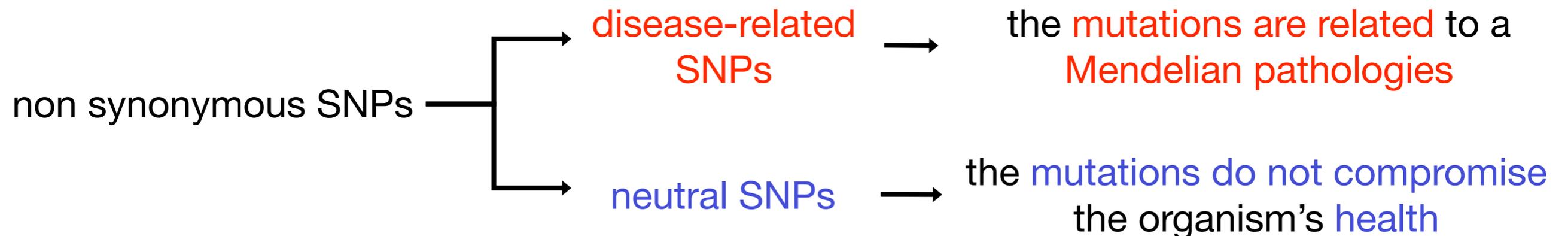
| Snp In Gene Model Legend: | |
|---------------------------|---------------------------------|
| ■ | - Region: exon |
| ■ | - Region: intron |
| | - snp: coding |
| | - snp: synonymous change |
| | - snp: nonsynonymous change |
| | - snp: untranslated region |
| | - snp: intron |
| | - snp: splice-site |
| | - snp: coding, synonymy unknown |

SNPs and disease

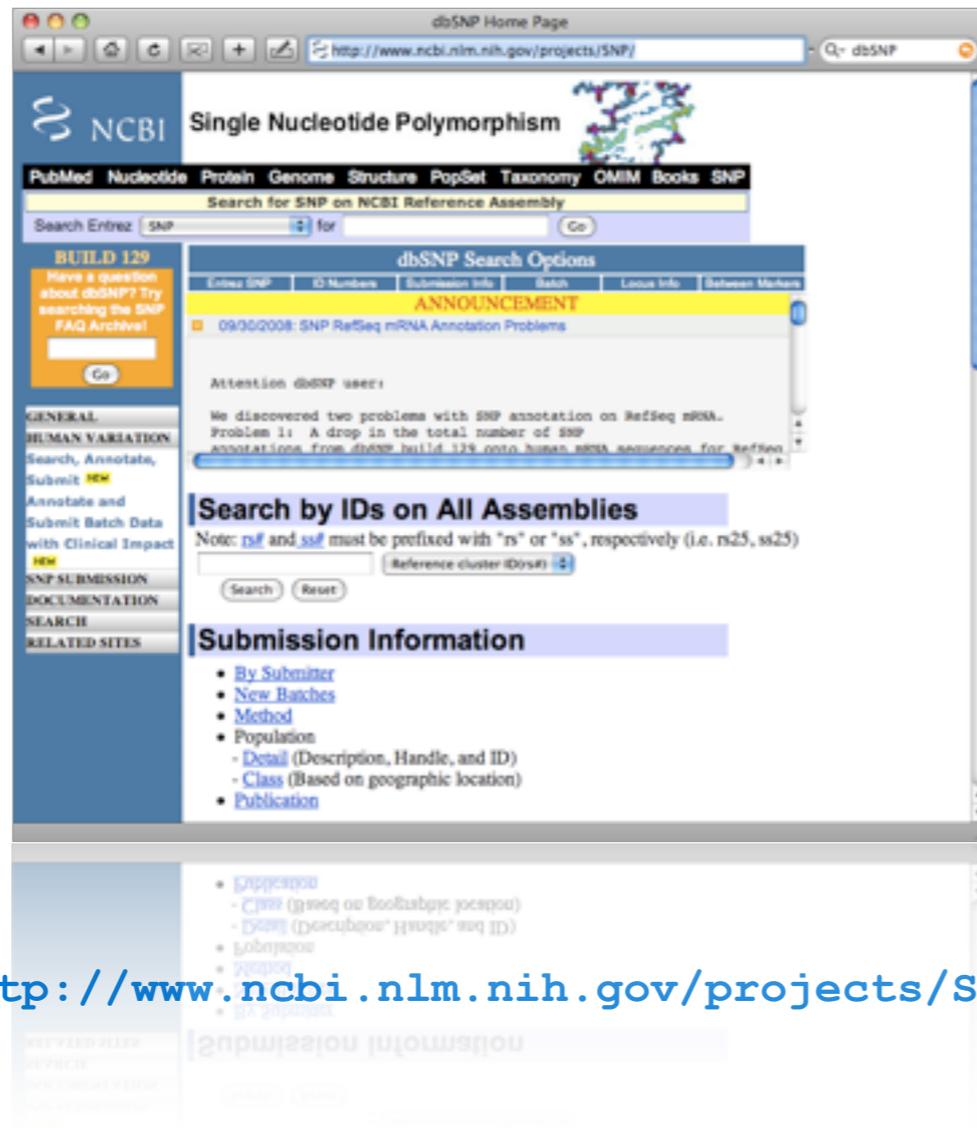
Single nucleotide polymorphism are the most common type of genetic variations in human accounting for about 90% of sequence differences (Collins et al., 1998).

Studying SNPs distribution in different human populations can lead to important considerations about the history of our species (Barbujani and Goldstein, 2004; Edmonds et al., 2004).

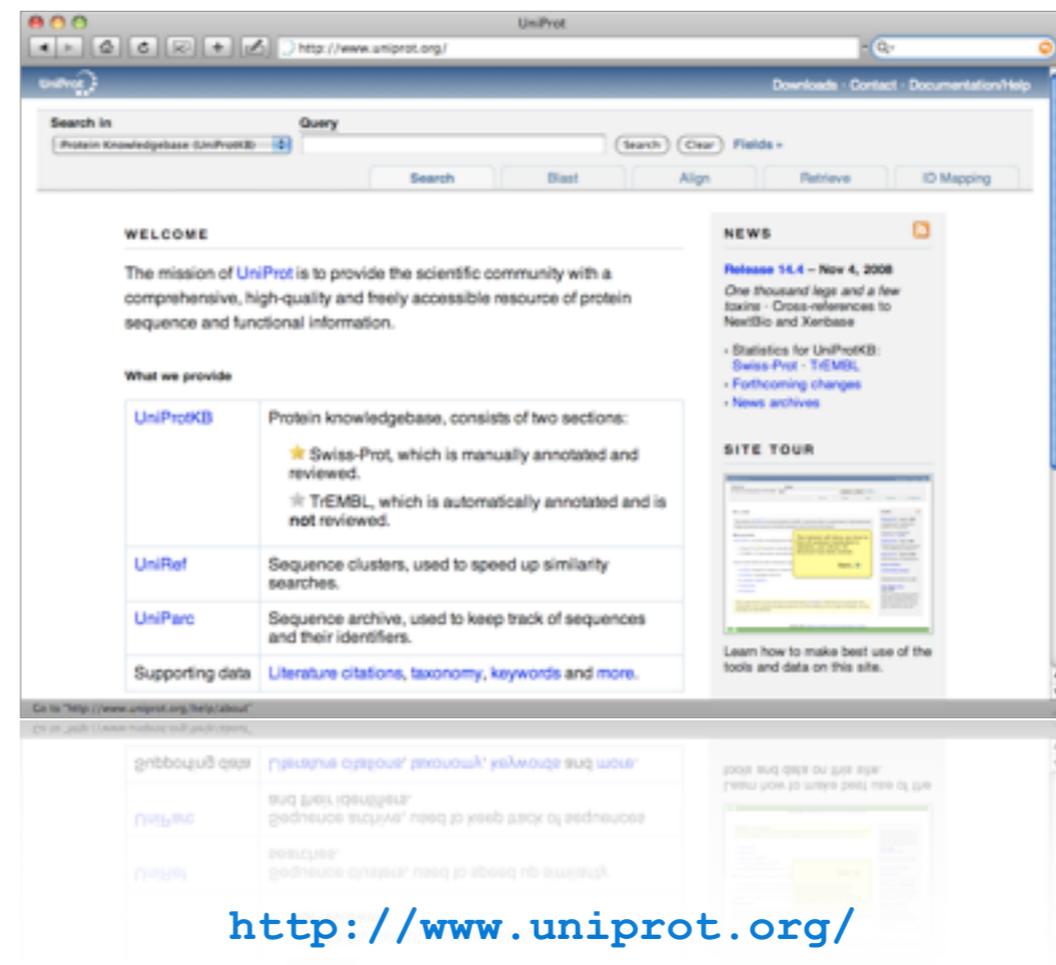
SNPs can also be responsible of genetic diseases (Ng and Henikoff, 2002; Bell, 2004).



SNP databases



<http://www.ncbi.nlm.nih.gov/projects/SNP/>



<http://www.uniprot.org/>

Evolutionary information for SNP analysis of p53 protein.

Arbiza et al. Selective pressures at a codon-level predict deleterious mutations in human disease genes. *J Mol Biol* (2006) vol. 358 (5) pp. 1390-404

doi:10.1016/j.jmb.2006.02.067 J. Mol. Biol. (2006) 358, 1390–1404

JMB Available online at www.sciencedirect.com  

Selective Pressures at a Codon-level Predict Deleterious Mutations in Human Disease Genes

Leonardo Arbiza¹, Serena Duchi¹, David Montaner², Jordi Burguet²
David Pantoja-Uceda³, Antonio Pineda-Lucena³, Joaquín Dopazo²
and Hernán Dopazo^{1*}

¹Pharmacogenomics and Comparative Genomics Unit Centro de Investigación Príncipe Felipe (CIPF). Autopista del Saler, 16-3, 46013 Valencia Spain

²Functional Genomics Unit Bioinformatics Department Centro de Investigación Príncipe Felipe (CIPF). Autopista del Saler, 16-3, 46013 Valencia Spain

³Structural Biology Laboratory Medicinal Chemistry Department, Centro de Investigación Príncipe Felipe (CIPF). Autopista del Saler 16-3, 46013 Valencia, Spain

***Corresponding author**

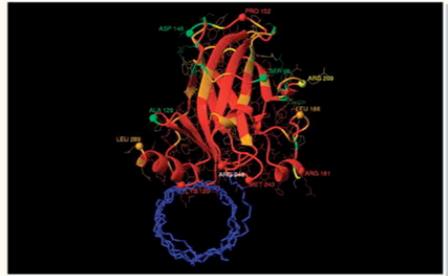
Deleterious mutations affecting biological function of proteins are constantly being rejected by purifying selection from the gene pool. The non-synonymous/synonymous substitution rate ratio (ω) is a measure of selective pressure on amino acid replacement mutations for protein-coding genes. Different methods have been developed in order to predict non-synonymous changes affecting gene function. However, none has considered the estimation of selective constraints acting on protein residues. Here, we have used codon-based maximum likelihood models in order to estimate the selective pressures on the individual amino acid residues of a well-known model protein: p53. We demonstrate that the number of residues under strong purifying selection in p53 is much higher than those that are strictly conserved during the evolution of the species. In agreement with theoretical expectations, residues that have been noted to be of structural relevance, or in direct association with DNA, were among those showing the highest signals of purifying selection. Conversely, those changing according to a neutral, or nearly neutral mode of evolution, were observed to be irrelevant for protein function. Finally, using more than 40 human disease genes, we demonstrate that residues evolving under strong selective pressures ($\omega < 0.1$) are significantly associated ($p < 0.01$) with human disease. We hypothesize that non-synonymous change on amino acids showing $\omega < 0.1$ will most likely affect protein function. The application of this evolutionary prediction at a genomic scale will provide an *a priori* hypothesis of the phenotypic effect of non-synonymous coding single nucleotide polymorphisms (SNPs) in the human genome.

© 2006 Elsevier Ltd. All rights reserved.

Keywords: comparative genomics; deleterious mutations; human diseases; purifying selection; codon-based models

Volume 358, Number 5, 19 May 2006 ISSN 0022-2836

 **JMB**
JOURNAL OF MOLECULAR BIOLOGY





© 2006 Elsevier Ltd. All rights reserved.

Επιτρέπεται η αναπαραγωγή αυτού του άρθρου σε οποιαδήποτε μορφή, με την προϋπόθεση να αναφέρεται η πηγή της προέλευσης και να μην γίνει αλλαγή στην ουσία του άρθρου.

© 2006 Elsevier Ltd. All rights reserved.

Επιτρέπεται η αναπαραγωγή αυτού του άρθρου σε οποιαδήποτε μορφή, με την προϋπόθεση να αναφέρεται η πηγή της προέλευσης και να μην γίνει αλλαγή στην ουσία του άρθρου.

Natural selection & human disease

SNPs can cause alterations of gene function by...

- Alterations at expression level
- Alternative splicing
- Alteration (or loss) of gene product function
 - Changes in the stability of the protein
 - Functionally important residues
 - Phylogenetic conservation

Natural selection working at codon level

nsSNP's functional prediction

JMB 2006; HM 2008, NatGen 2008



Selective Pressures at a Codon-level Predict Deleterious Mutations in Human Disease Genes

Leonardo Arbiza¹, Serena Duchi¹, David Montaner², Jordi Burguet², David Pantoja-Uceda³, Antonio Pineda-Lucena³, Joaquin Dopazo² and Hernán Dopazo^{1*}

¹Pharmacogenomics and Comparative Genomics Unit, Centro de Investigación Príncipe Felipe (CIPF), Valencia, Spain; ²Pharmacogenomics and Comparative Genomics Unit, Centro de Investigación Príncipe Felipe (CIPF), Valencia, Spain; ³Functional Genomics Unit, Bioinformatics Department, Centro de Investigación Príncipe Felipe (CIPF), Valencia, Spain; *Corresponding

Selective Constraints and Human Disease Genes: Evolutionary and Bioinformatics Approaches

Hernán Dopazo, Centro de Investigación Príncipe Felipe, Valencia, Spain

Advanced article

Article Contents

- Introduction
- Functional Prediction of nsSNPs
- Methods and Web Resources
- Natural Selection and Disease
- Bioinformatics Perspectives
- Conclusions
- Acknowledgments

Online posting date: 15th July 2008

Natural selection rejects with variable stringency mutations that reduce the ability to survive and reproduce. Evolutionary selection acting on protein likelihood models predict that the ratio of non-synonymous to synonymous substitutions (dN/dS) is much higher in disease genes than in other genes. In fact, we have noted that the ratio of non-synonymous to synonymous substitutions (dN/dS) in disease genes is significantly higher than in other genes. Conversely, those mutations that are not rejected by natural selection, were among the most conserved amino acids in the human genome.

METHODS

Use of Estimated Evolutionary Strength at the Codon Level Improves the Prediction of Disease-Related Protein Mutations in Humans

Emilio Capriotti,¹ Leonardo Arbiza,² Rita Casadio,⁴ Joaquin Dopazo,³ Hernán Dopazo,^{2*} and Marc A. Marti-Renom^{1*}

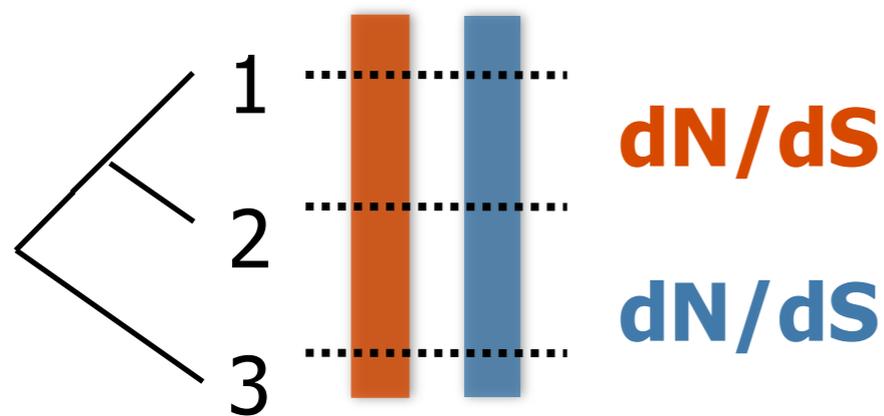
¹Structural Genomics Unit, Centro de Investigación Príncipe Felipe (CIPF), Valencia, Spain; ²Pharmacogenomics and Comparative Genomics Unit, Centro de Investigación Príncipe Felipe (CIPF), Valencia, Spain; ³Functional Genomics Unit, Bioinformatics Department, Centro de Investigación Príncipe Felipe (CIPF), Valencia, Spain; ⁴Laboratory of Bioinformatics, CIBERS Department of Biology, University of Bologna, Bologna, Italy

Main Question

- Could an estimator of the selective pressures acting at codon level (ω) be used as a predictor of the phenotype effect of SNP's ?

Detecting Positive & Negative Selection

Site-specific models average dN/dS over lineages but differentiate over sites



$$\omega = \frac{dN}{dS}$$

Bayes Empirical Bayes (BEB) analysis
Positively selected sites (*: P>95%; **: P>99%)

| | Pr(w>1) | post mean +- SE for w | |
|------|---------|-----------------------|-----------|
| 1 M | 0.007 | 0.156 +- 0.298 | |
| 2 E | 0.009 | 0.169 +- 0.353 | |
| 3 E | 0.009 | 0.169 +- 0.353 | |
| 4 P | 0.125 | 0.893 +- 1.263 | Neutral |
| 5 Q | 0.010 | 0.182 +- 0.370 | |
| 6 S | 0.015 | 0.212 +- 0.436 | |
| 7 D | 0.010 | 0.180 +- 0.375 | |
| 8 P | 0.368 | 2.207 +- 2.473 | Positive |
| 9 S | 0.007 | 0.160 +- 0.310 | |
| 10 V | 0.139 | 0.969 +- 1.480 | |
| 11 E | 0.009 | 0.169 +- 0.353 | Purifying |
| 12 P | 0.091 | 0.722 +- 1.043 | |
| 13 P | 0.014 | 0.208 +- 0.450 | |
| 14 L | 0.013 | 0.200 +- 0.411 | |
| 15 S | 0.009 | 0.178 +- 0.371 | |
| 16 Q | 0.010 | 0.182 +- 0.370 | |
| 17 E | 0.011 | 0.186 +- 0.405 | |

p53 evolutionary analysis

Many mutant forms are involved in different types of human cancer

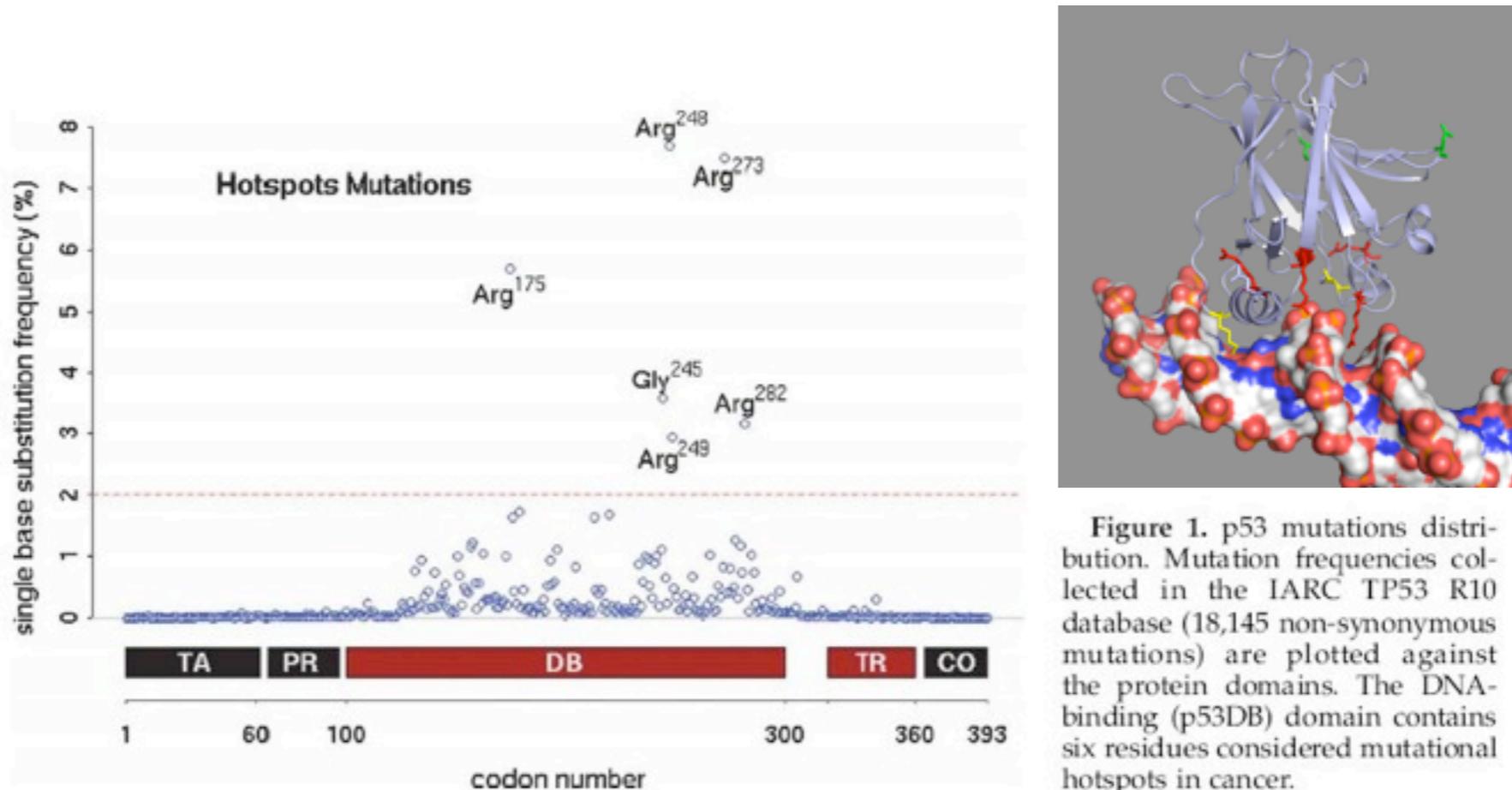


Figure 1. p53 mutations distribution. Mutation frequencies collected in the IARC TP53 R10 database (18,145 non-synonymous mutations) are plotted against the protein domains. The DNA-binding (p53DB) domain contains six residues considered mutational hotspots in cancer.

Table 1. Summary of p53 domains, mutations and ω statistics according to M8 and SLR models

| p53 alignment | | | Mutations | | | ω statistics | | | |
|---------------|---------|---------------------|-----------|------------------|-------|---------------------|--------|-------|-------|
| Domains | Codons | Indels ^a | Total | Mps ^b | Model | Min. | Median | Mean | Max. |
| TA | 1-60 | 38 | 96 | 1.6 | M8 | 0.030 | 0.334 | 0.379 | 1.747 |
| | | | | | SLR | 0.000 | 0.269 | 0.369 | 1.865 |
| PR | 61-97 | 22 | 151 | 4.2 | M8 | 0.029 | 0.314 | 0.376 | 1.338 |
| | | | | | SLR | 0.000 | 0.307 | 0.376 | 1.447 |
| DB | 100-300 | 5 | 17,389 | 87.0 | M8 | 0.027 | 0.039 | 0.116 | 1.423 |
| | | | | | SLR | 0.000 | 0.029 | 0.095 | 2.018 |
| TR | 325-355 | 0 | 178 | 5.1 | M8 | 0.028 | 0.067 | 0.126 | 0.456 |
| | | | | | SLR | 0.000 | 0.068 | 0.103 | 0.379 |
| CO | 361-393 | 11 | 18 | 1.6 | M8 | 0.027 | 0.216 | 0.255 | 0.878 |
| | | | | | SLR | 0.000 | 0.176 | 0.226 | 0.882 |

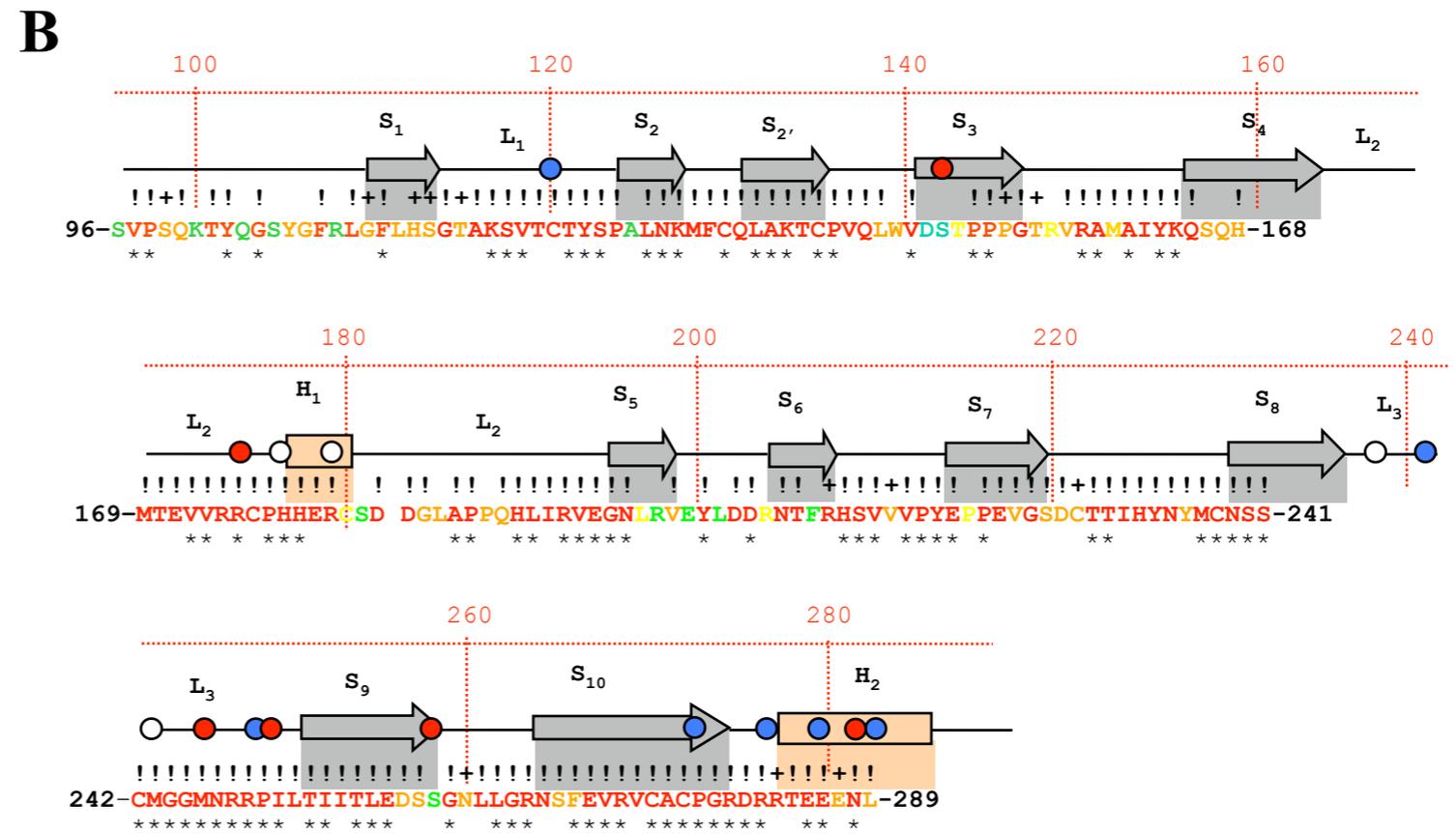
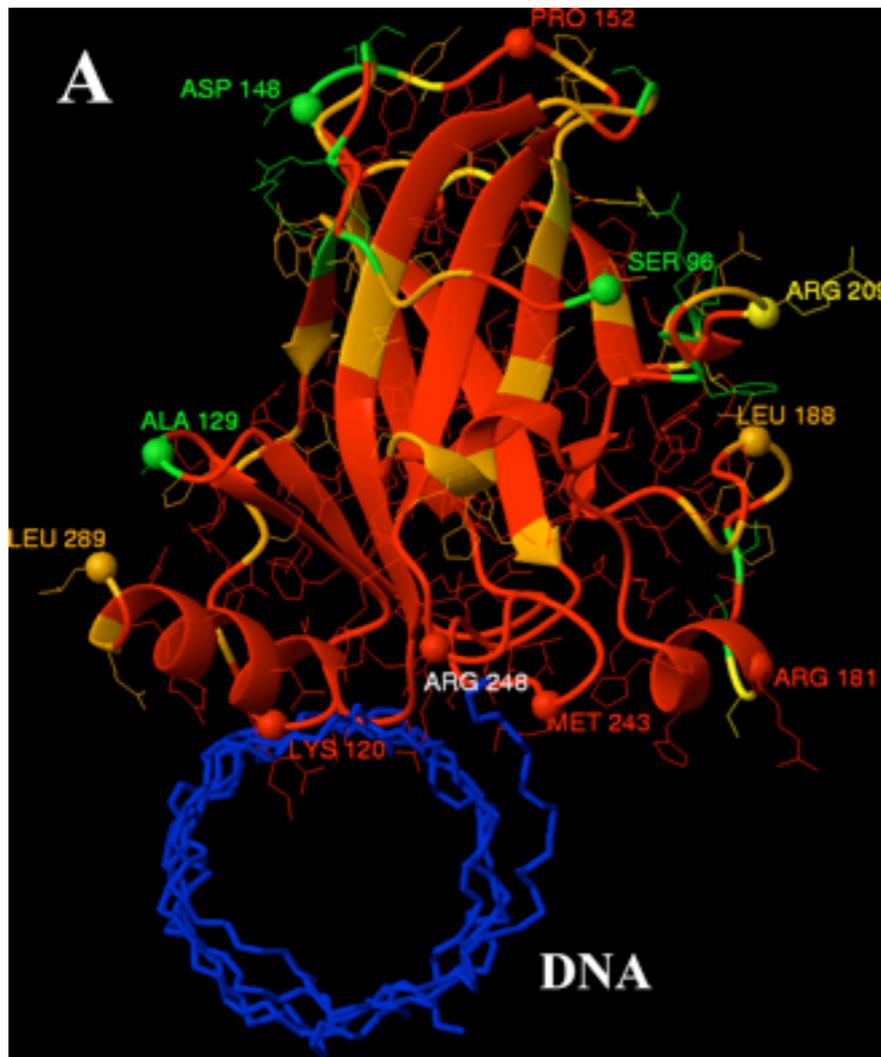
Mutations were deduced from the IARC TP53 database.

^a Insertions/deletions.

^b Mean number of mutations per site.

p53 evolutionary analysis

SLR: $\omega \leq 0.1$, $0.1 < \omega \leq 0.2$, $0.2 < \omega \leq 0.3$, $\omega > 0.3$



Beyond p53...

**Disease Proteins,
Immune, Cancer ~ 250 proteins**

SwissProt Database, ~3,000 proteins

Table 3. Evaluation of alternative $\omega_{\text{cut-off}}$ values and mutational frequencies in disease

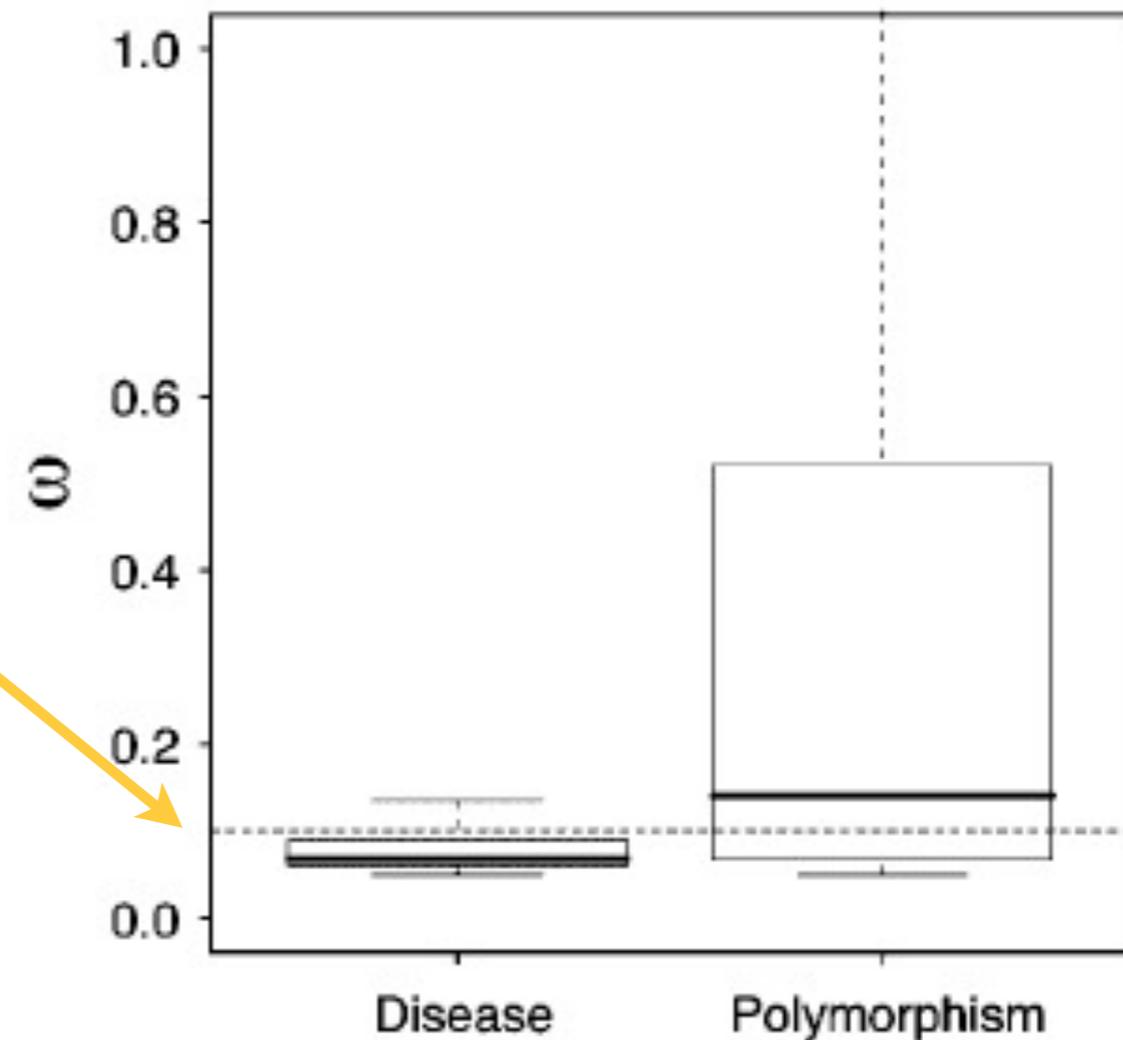
| $\omega_{\text{cut-off}}$ | Mammals | | Vertebrates | |
|---------------------------|-----------------------|--------|---------------|--------|
| | PAML | SLR | PAML | SLR |
| 0.03 | 0.9748 | 0.0095 | 0.0504 | 0.0061 |
| 0.05 | 0.0114 | 0.0075 | 0.0026 | 0.0008 |
| 0.10 | 3.0×10^{-05} | 0.0076 | 0.0016 | 0.0009 |
| 0.12 | 0.0007 | 0.0077 | 0.0010 | 0.0023 |
| 0.15 | 0.0025 | 0.0078 | 0.0012 | 0.0018 |
| 0.20 | 0.0715 | 0.0074 | 0.0019 | 0.0019 |
| 0.25 | 0.1938 | 0.0074 | 0.0044 | 0.0043 |
| 0.30 | 0.0188 | 0.0076 | 0.0035 | 0.0065 |
| 0.40 | 0.0486 | 0.0101 | 0.0176 | 0.0254 |
| 0.50 | 0.1849 | 0.0223 | 0.0534 | 0.1010 |
| G ^a | 43 | 43 | 43 | 43 |
| R ^b | 24,375 | 24,375 | 17,424 | 17,435 |
| M ^c | 8970 | 8970 | 8081 | 8083 |

One-tail K-S tests reject the null hypothesis, which considers that the frequency of mutations are not differentially distributed above and below the given $\omega_{\text{cut-off}}$. The alternative hypothesis, which considers that disease-associated mutations are preferentially associated with values below the $\omega_{\text{cut-off}}$ is accepted with the highest confidence using ω_{PAML} estimations on mammal ($\omega_{\text{cut-off}}=0.10$) and vertebrate ($\omega_{\text{cut-off}}=0.12$) datasets. The K-S test on SLR estimates reject the null hypothesis for all values of $\omega_{\text{cut-off}}$ evaluated. This is the consequence of the undesirable behaviour of the SLR method, which drops low values of ω to 0 (see the text and Figure 6 for explanation).

^a Number of genes evaluated.

^b Number of residues evaluated.

^c Number of mutations evaluated.



Evolution and disease.

Capriotti et al. Use of estimated evolutionary strength at the codon level improves the prediction of disease-related protein mutations in humans. *Hum Mutat* (2008) vol. 29 (1) pp. 198-204

HUMAN MUTATION 29(1), 198–204, 2008

METHODS

Use of Estimated Evolutionary Strength at the Codon Level Improves the Prediction of Disease-Related Protein Mutations in Humans

Emidio Capriotti,¹ Leonardo Arbiza,² Rita Casadio,⁴ Joaquín Dopazo,³ Hernán Dopazo,^{2*} and Marc A. Martí-Renom^{1*}

¹Structural Genomics Unit, Centro de Investigación Príncipe Felipe (CIPF), Valencia, Spain; ²Pharmacogenomics and Comparative Genomics Unit, Centro de Investigación Príncipe Felipe (CIPF), Valencia, Spain; ³Functional Genomics Unit, Bioinformatics Department, Centro de Investigación Príncipe Felipe (CIPF), Valencia, Spain; ⁴Laboratory of Biocomputing, CIRB/Department of Biology, University of Bologna, Bologna, Italy

Communicated by David N. Cooper

Predicting the functional impact of protein variation is one of the most challenging problems in bioinformatics. A rapidly growing number of genome-scale studies provide large amounts of experimental data, allowing the application of rigorous statistical approaches for predicting whether a given single point mutation has an impact on human health. Up until now, existing methods have limited their source data to either protein or gene information. Novel in this work, we take advantage of both and focus on protein evolutionary information by using estimated selective pressures at the codon level. Here we introduce a new method (SeqProfCod) to predict the likelihood that a given protein variant is associated with human disease or not. Our method relies on a support vector machine (SVM) classifier trained using three sources of information: protein sequence, multiple protein sequence alignments, and the estimation of selective pressure at the codon level. SeqProfCod has been benchmarked with a large dataset of 8,987 single point mutations from 1,434 human proteins from SWISS-PROT. It achieves 82% overall accuracy and a correlation coefficient of 0.59, indicating that the estimation of the selective pressure helps in predicting the functional impact of single-point mutations. Moreover, this study demonstrates the synergic effect of combining two sources of information for predicting the functional effects of protein variants: protein sequence/profile-based information and the evolutionary estimation of the selective pressures at the codon level. The results of large-scale application of SeqProfCod over all annotated point mutations in SWISS-PROT (available for download at <http://sgu.bioinfo.cipf.es/services/Omidios/>; last accessed: 24 August 2007), could be used to support clinical studies. *Hum Mutat* 29(1), 198–204, 2008. © 2007 Wiley-Liss, Inc.

KEY WORDS: SNP; nsSNP; disease; sequence profile; evolutionary strength; bioinformatics

INTRODUCTION

Studies characterizing the relationship between protein variants and human disease have grown rapidly over the past years, in part due to genomic-scale sequencing efforts [Krawczak et al., 2000; Sherry et al., 2001; Stenson et al., 2003]. For example, it is now known that single nucleotide polymorphisms (SNPs) constitute about the 90% of human protein sequence variability [Collins et al., 1998]. Synonymous and nonsynonymous SNPs (nsSNPs) may occur every ~350 bp in coding regions [Cargill et al., 1999] and about 50% of nsSNPs may be associated to pathologies of genetic origin. Therefore, predicting which nsSNPs are responsible for human disease is one of the major challenges in bioinformatics.

Recently, different methods have been developed for predicting the effect of single point mutations in humans [Arbiza et al., 2006; Bao and Cui, 2005; Bao et al., 2005; Capriotti et al., 2006; Chan et al., 2007; Karchin et al., 2005a; Ng and Henikoff, 2003; Ramensky et al., 2002; Santibáñez Koref et al., 2003; Thomas et al., 2003b; Yue and Moul, 2006]. In spite of the effort, however,

Received 30 May 2007; accepted revised manuscript 17 July 2007.
*Correspondence to: Marc A. Martí-Renom and Hernán Dopazo, Bioinformatics Department, Centro de Investigación Príncipe Felipe (CIPF), Av. Autopista del Saler, 16, 46013 Valencia, Spain. E-mail for Marc A. Martí-Renom: mmarti@cipf.es; E-mail for Hernán Dopazo: hdopazo@cipf.es

Grant sponsor: Ministerio dell'Università e della Ricerca, Italy; Grant: Fondo per gli Investimenti della Ricerca di Base 2003 LIBI-International Laboratory of Bioinformatics; Grant sponsor: Marie Curie International Reintegration Grant; Grant number: FP6-039722; Grant sponsor: Generalitat Valenciana; Grant numbers: GV/2007/065 and GV/06/080; Grant sponsor: Ministerio de Educación y Ciencia, Spain; Grant number: BFU2006-15413-CO2-02/BMC; Grant sponsor: European Union, (EU) Network of Excellence BIOSAPIENS; Grant number: LSHG-CT-2003-503265.
DOI 10.1002/humu.20628
Published online 12 October 2007 in Wiley InterScience (www.interscience.wiley.com).

© 2007 WILEY-LISS, INC.





ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA



PRINCIPE FELIPE
CENTRO DE INVESTIGACION

Classification results



| | Mutation | Disease | Neutral | Proteins |
|--|---------------|---------|---------|----------|
| Single point mutation with reported effect | 21,185 | 12,944 | 8,241 | 3,587 |
| Single point mutation with reported effect and profile | 8,718 | 3,852 | 4,866 | 2,538 |

SeqCod and SeqProf methods reach the **same level of accuracy** of about 79% and when the two different types of evolutive information are used the resulting predictor **Omidios** overcomes the others showing an overall accuracy of 82%

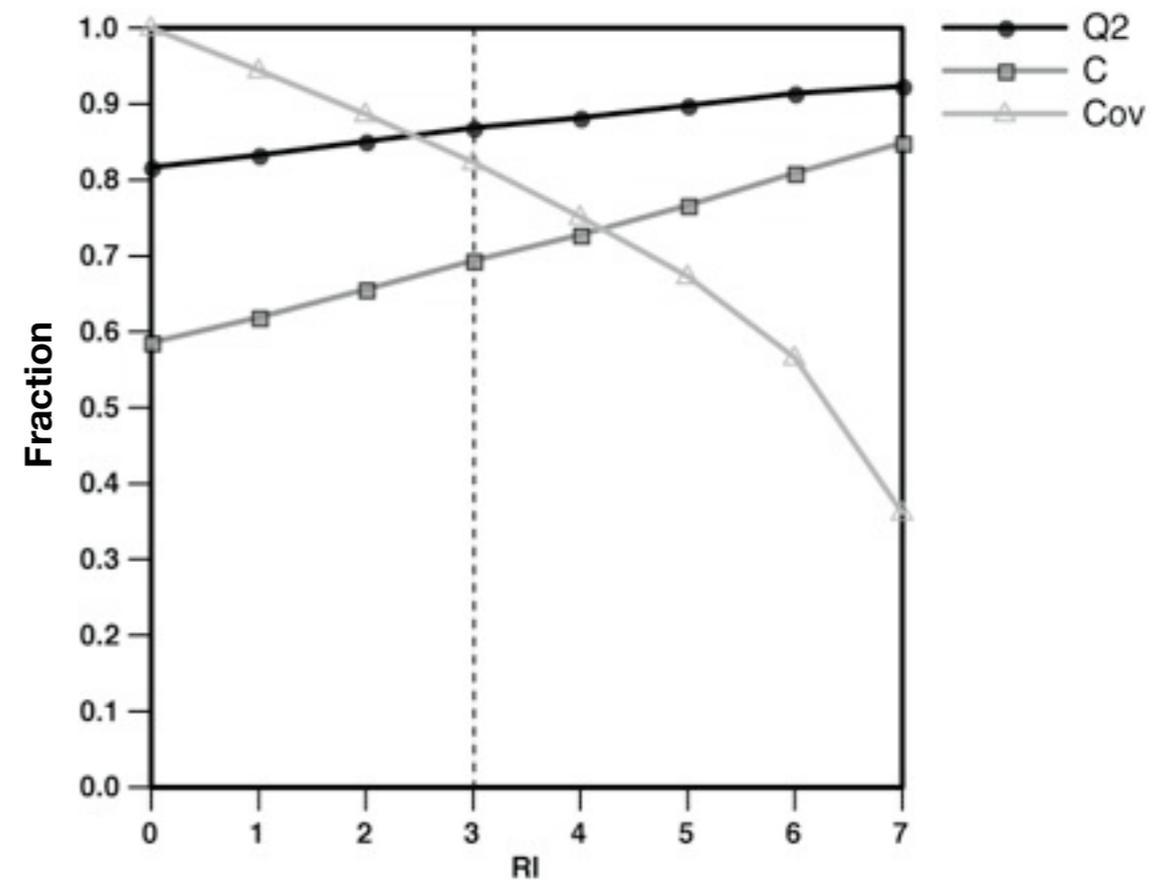
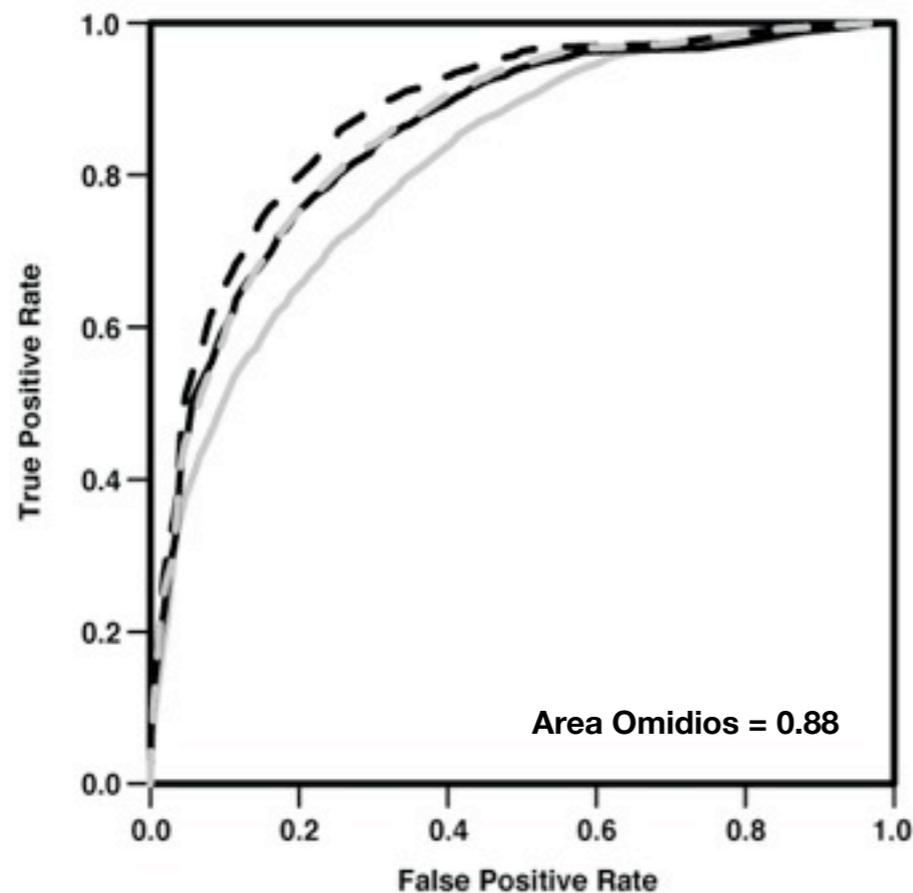
| | Q2 | P[D] | Q[D] | P[N] | Q[N] | C |
|----------------|----|------|------|------|------|------|
| Seq | 73 | 86 | 72 | 54 | 74 | 0.43 |
| SeqCod | 79 | 87 | 82 | 64 | 74 | 0.53 |
| SeqProf | 79 | 88 | 81 | 63 | 75 | 0.54 |
| Omidios | 82 | 89 | 84 | 68 | 76 | 0.59 |

D = Disease related N = Neutral

Omidios method

Omidios has higher accuracy than the previous two methods **increasing the accuracy** up to 82% **and the correlation coefficient** to 0.59.

| | Q2 | P[D] | Q[D] | P[N] | Q[N] | C |
|----------------|----|------|------|------|------|------|
| Omidios | 82 | 88 | 84 | 68 | 76 | 0.59 |



Q2: Overall Accuracy **C:** Correlation Coefficient **DB:** Fraction of database that are predicted with a reliability \geq the given threshold

Comparison

Omidios results in higher accuracy and correlation than the other available methods covering the 100% of the dataset (see column %PM).

Omidios results in **higher accuracy with respect to SIFT** and although the quality of **Omidios** is comparable to PANTHER, when our prediction are selected by RI index the accuracy of our method is higher than PANTHER.

| | Q2 | P[D] | Q[D] | P[N] | Q[N] | C | PM |
|----------------|----|------|------|------|------|----|-----|
| Omidios | 82 | 89 | 84 | 68 | 76 | 59 | 100 |
| SIFT | 71 | 84 | 72 | 51 | 69 | 38 | 97 |
| PANTHER | 74 | 87 | 75 | 53 | 72 | 43 | 83 |

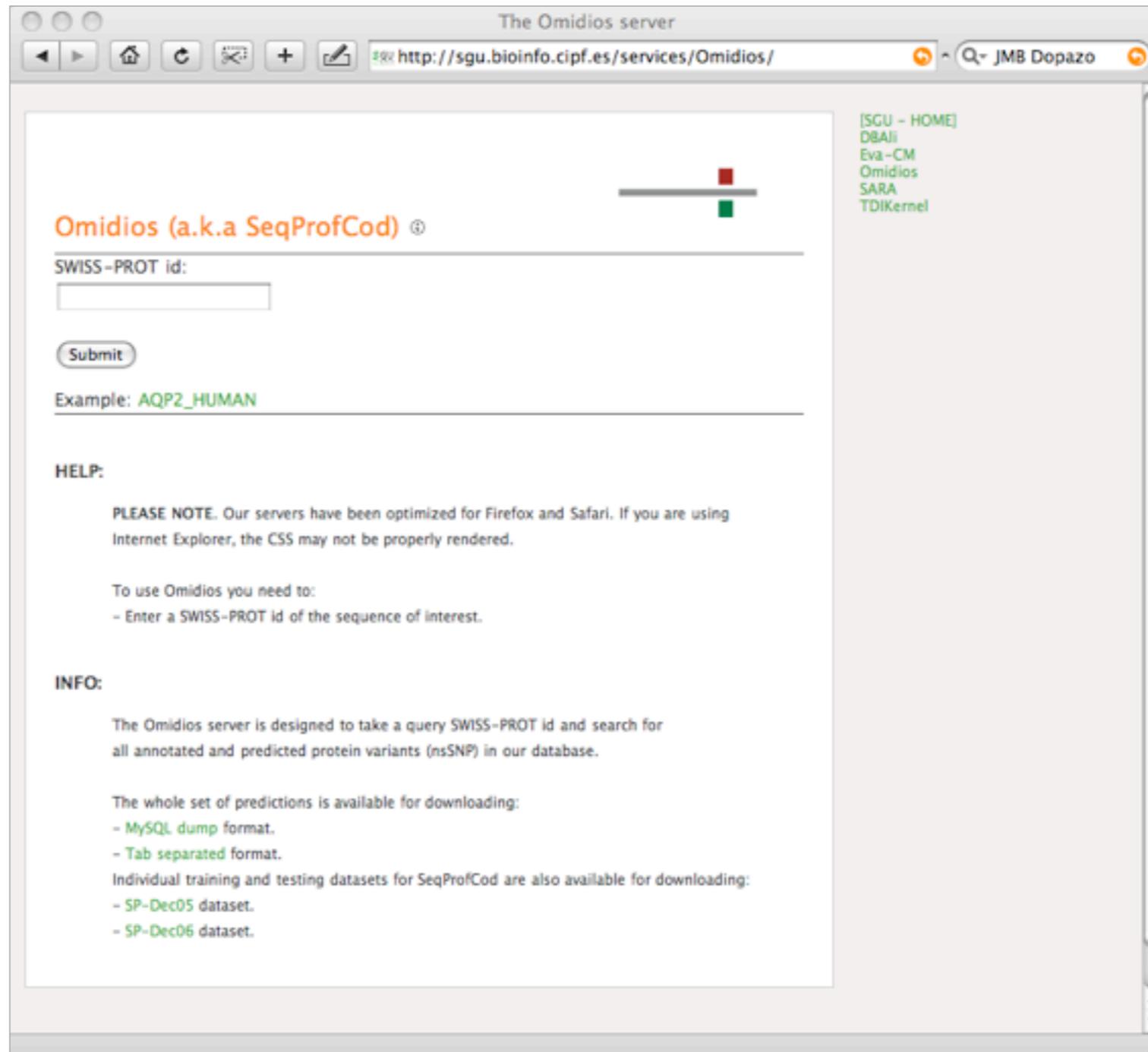
HM-Dic05: 8987 mutations

| | Q2 | P[D] | Q[D] | P[N] | Q[N] | C | PM |
|----------------|----|------|------|------|------|----|-----|
| Omidios | 74 | 65 | 79 | 83 | 72 | 48 | 100 |
| SIFT | 71 | 63 | 70 | 78 | 72 | 42 | 96 |
| PANTHER | 77 | 73 | 71 | 79 | 81 | 52 | 77 |

HM-Dic06: 2008 mutations

Omidios server

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



The screenshot shows a web browser window titled "The Omidios server" with the address bar containing "http://sgu.bioinfo.cipf.es/services/Omidios/". The user is identified as "JMB Dopazo". The main content area features a header "Omidios (a.k.a SeqProfCod) ©" with a logo consisting of a horizontal line and two small squares (red and green). Below the header is a form with a label "SWISS-PROT id:" and an input field. A "Submit" button is located below the input field. An example is provided: "Example: AQP2_HUMAN".

HELP:

PLEASE NOTE. Our servers have been optimized for Firefox and Safari. If you are using Internet Explorer, the CSS may not be properly rendered.

To use Omidios you need to:

- Enter a SWISS-PROT id of the sequence of interest.

INFO:

The Omidios server is designed to take a query SWISS-PROT id and search for all annotated and predicted protein variants (nsSNP) in our database.

The whole set of predictions is available for downloading:

- MySQL dump format.
- Tab separated format.

Individual training and testing datasets for SeqProfCod are also available for downloading:

- SP-Dec05 dataset.
- SP-Dec06 dataset.

On the right side of the page, there is a vertical menu with the following items: [SGU - HOME], DBAli, Eva-CM, Omidios, SARA, and TDIKernel.

Structural analysis of missense mutations in human BRCA1 BRCT domains

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

[CANCER RESEARCH 64, 3790-3797, June 1, 2004]

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

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

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

ABSTRACT

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

INTRODUCTION

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

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

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

MATERIALS AND METHODS

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

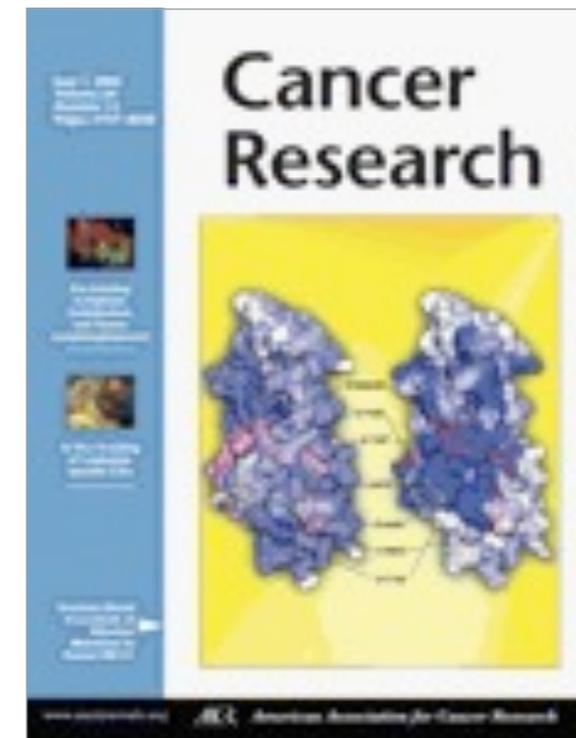
The multiple structure-based alignment of the native structures of the BRCT-like domains was obtained by the SALIGN command in MODELLER (Supplementary Fig. 2). It included the experimentally determined structures of the two human BRCA1 BRCT domains (Protein Data Bank code 1JNX; Refs. 8, 24), rat BRCA1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

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

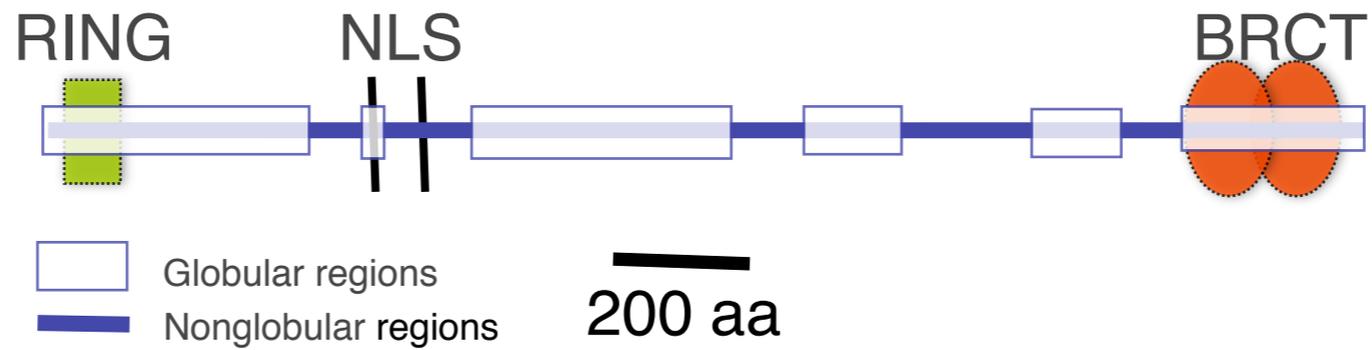
For the native structure of the human BRCT tandem repeat and each of the 94 mutant models, a number of sequence and structure features were calculated. These features were used in the classification tree in Fig. 3 (values for all 94 mutations are given in Supplementary Tables 1 and 2).

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

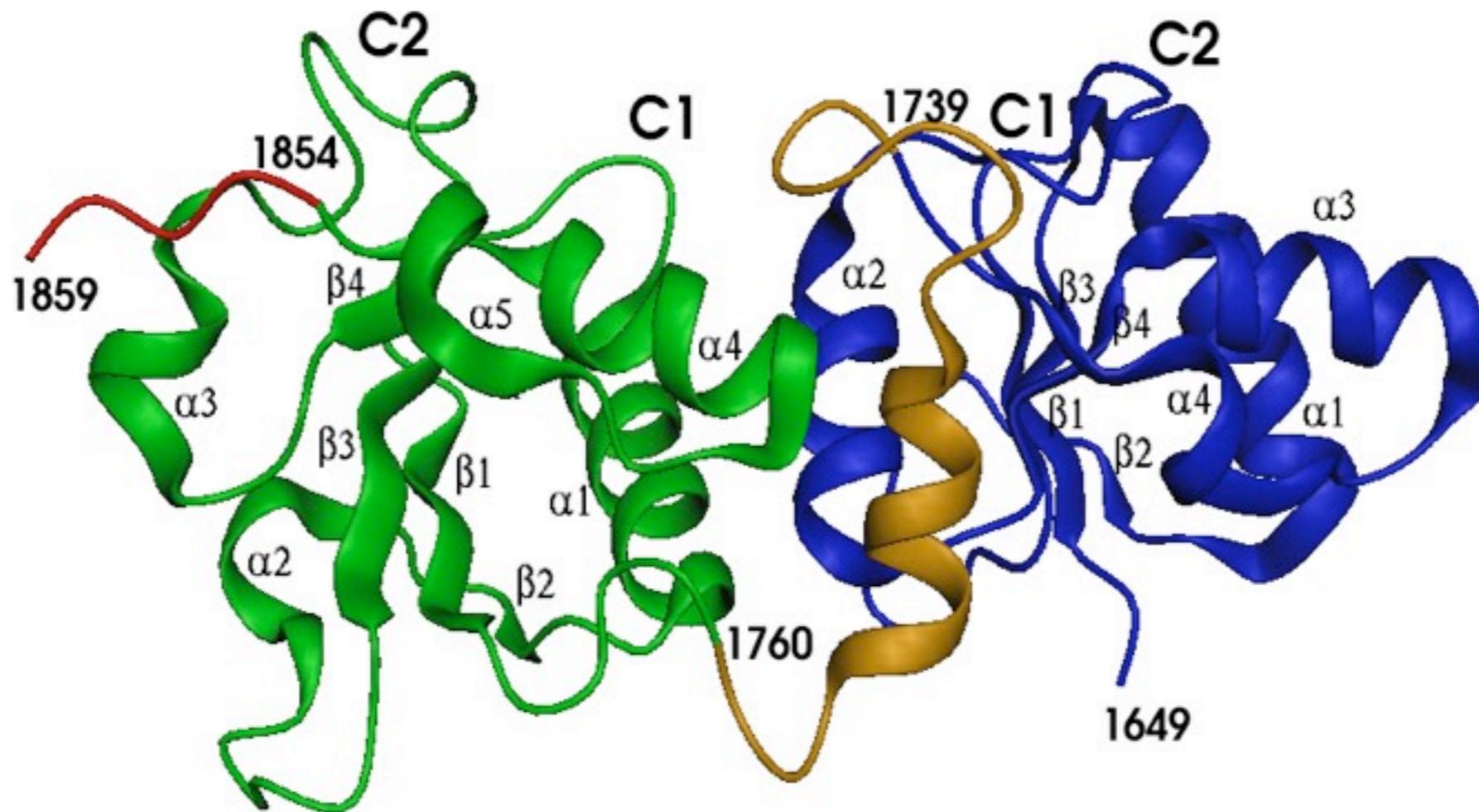
⁶ <http://salilab.org/modbase/>.



Human BRCA1 and its two BRCT domains



BRCA1 BRCT repeats, 1jnx



CONFIDENTIAL



MYRIAD

BRCAAnalysis™

Comprehensive BRCA1-BRCA2 Gene Sequence Analysis Result

| | | |
|---|---|--|
| Niecee Singer, MS Strang Cancer Prevention Center 428 E 72nd St New York, NY 10021 | SPECIMEN | PATIENT |
| | Specimen Type: Blood Draw Date: n/a Accession Date: Oct 27, 2000 Report Date: Nov 17, 2000 | Name: Date of Birth: Feb 02, 1953 Patient ID: Gender: Female Accession #: 00019998 Requisition #: 56694 |
| Physician: Fred Gilbert, MD | | |

Test Result

| Gene Analyzed | Specific Genetic Variant |
|---------------|--------------------------|
| BRCA2 | H2116R |
| BRCA1 | None Detected |

Interpretation

GENETIC VARIANT OF UNCERTAIN SIGNIFICANCE

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

Authorized Signature:

Brian E. Ward, Ph.D.
Laboratory Director

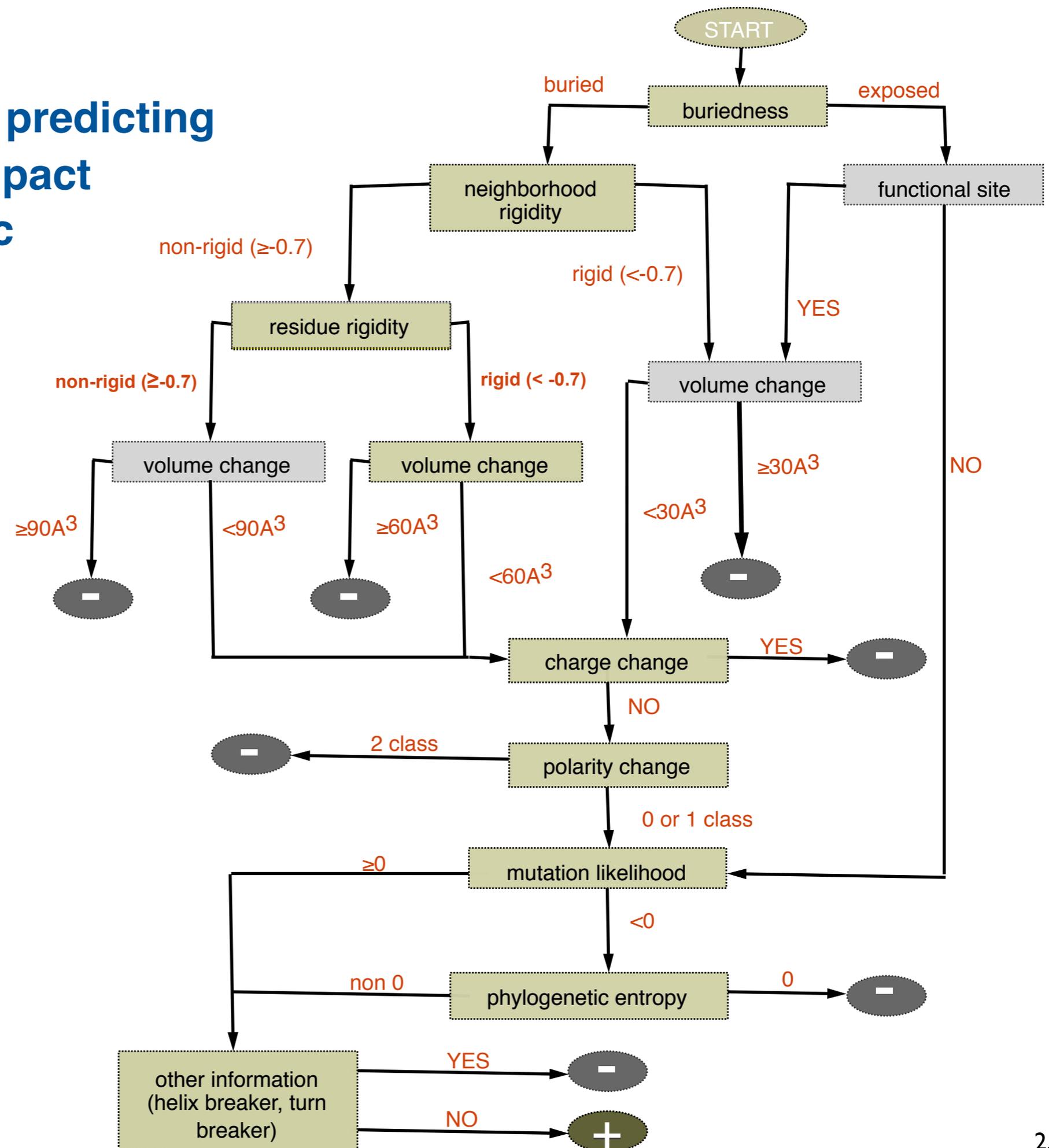
Thomas S. Frank, M.D.
Medical Director

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

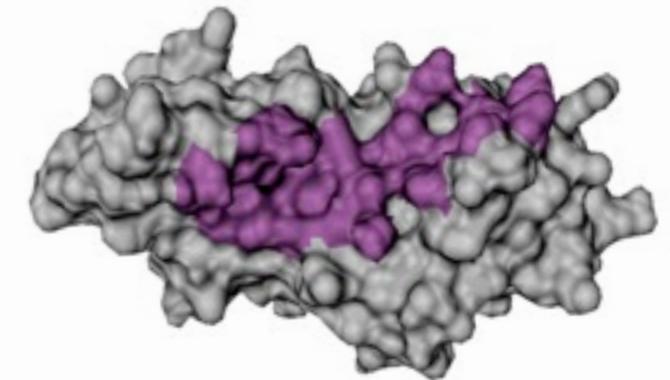
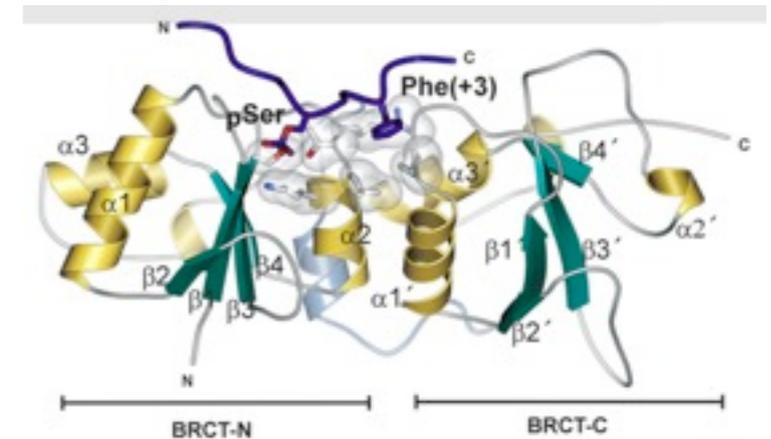
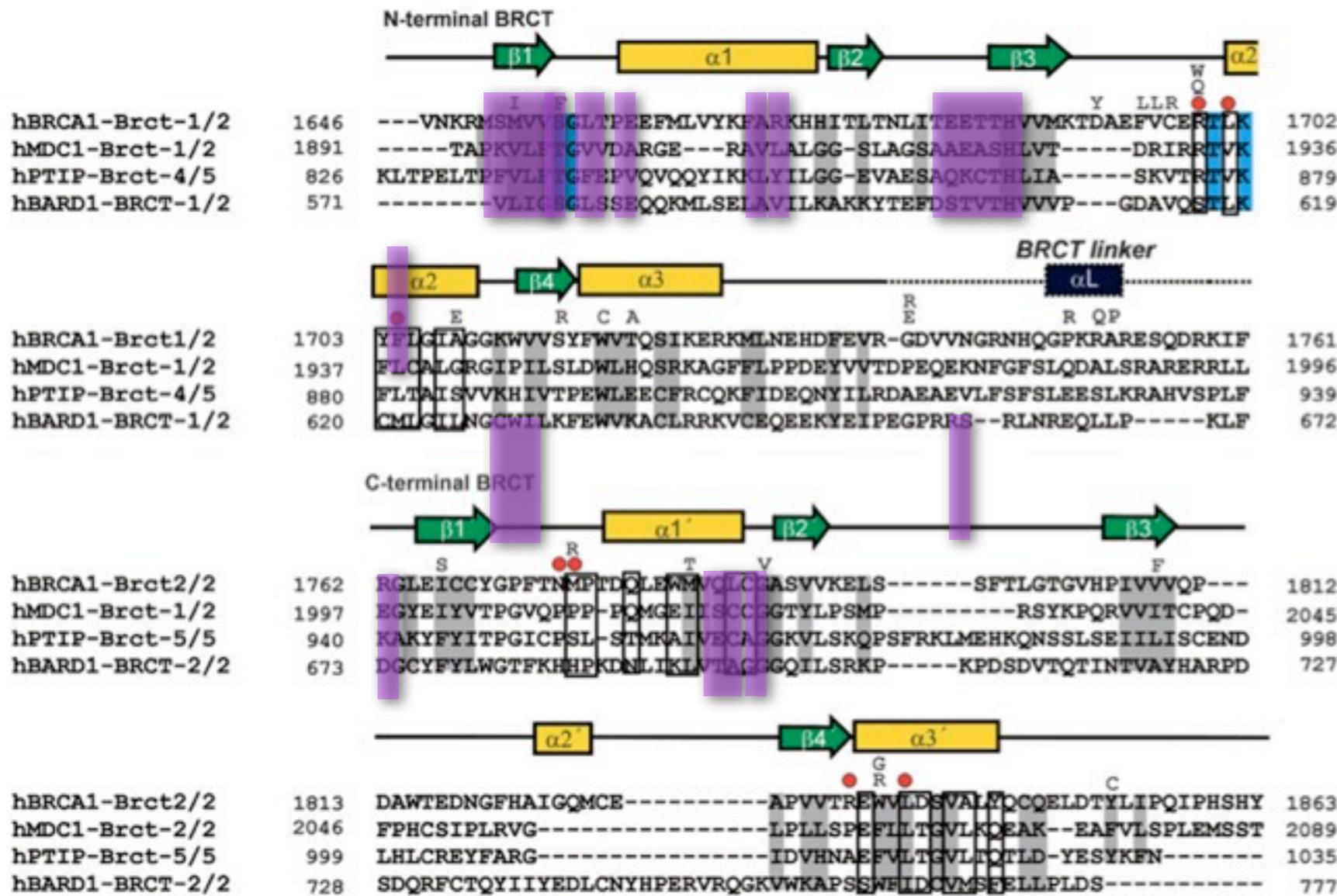
Missense mutations in BRCT domains by function

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

“Decision” tree for predicting functional impact of genetic variants



Putative binding site on BRCA1



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

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

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

Supervised learning approach

Karchin et al. Functional Impact of Missense Variants in BRCA1 Predicted by Supervised Learning. PLoS Comput Biol (2007) vol. 3 (2) pp. e26

OPEN ACCESS Freely available online

PLoS COMPUTATIONAL BIOLOGY

Functional Impact of Missense Variants in BRCA1 Predicted by Supervised Learning

Rachel Karchin^{1,2*}, Alvaro N. A. Monteiro³, Sean V. Tavtigian⁴, Marcelo A. Carvalho³, Andrej Sali^{5,6*}

1 Department of Biomedical Engineering, Johns Hopkins University, Baltimore, Maryland, United States of America, **2** Institute of Computational Medicine, Johns Hopkins University, Baltimore, Maryland, United States of America, **3** Risk Assessment, Detection, and Intervention Program, H. Lee Moffitt Cancer Center and Research Institute, Tampa, Florida, United States of America, **4** International Agency for Research on Cancer, Lyon, France, **5** Department of Pharmaceutical Chemistry, University of California San Francisco, San Francisco, California, United States of America, **6** California Institute for Quantitative Biomedical Research, University of California San Francisco, San Francisco, California, United States of America

Many individuals tested for inherited cancer susceptibility at the *BRCA1* gene locus are discovered to have variants of unknown clinical significance (UCVs). Most UCVs cause a single amino acid residue (missense) change in the *BRCA1* protein. They can be biochemically assayed, but such evaluations are time-consuming and labor-intensive. Computational methods that classify and suggest explanations for UCV impact on protein function can complement functional tests. Here we describe a supervised learning approach to classification of *BRCA1* UCVs. Using a novel combination of 16 predictive features, the algorithms were applied to retrospectively classify the impact of 36 *BRCA1* C-terminal (BRCT) domain UCVs biochemically assayed to measure transactivation function and to blindly classify 54 documented UCVs. Majority vote of three supervised learning algorithms is in agreement with the assay for more than 94% of the UCVs. Two UCVs found deleterious by both the assay and the classifiers reveal a previously uncharacterized putative binding site. Clinicians may soon be able to use computational classifiers such as those described here to better inform patients. These classifiers can be adapted to other cancer susceptibility genes and systematically applied to prioritize the growing number of potential causative loci and variants found by large-scale disease association studies.

Citation: Karchin R, Monteiro ANA, Tavtigian SV, Carvalho MA, Sali A (2007) Functional impact of missense variants in *BRCA1* predicted by supervised learning. PLoS Comput Biol 3(2): e26. doi:10.1371/journal.pcbi.0030026

Introduction

The *BRCA1* gene encodes a large multifunction protein involved in cell-cycle and centrosome control, transcriptional regulation, and in the DNA damage response [1–3]. Inherited mutations in this gene have been associated with an increased lifetime risk of breast and ovarian cancer (6–8 times that of the general population) [4]. There are several thousand known deleterious *BRCA1* mutations that result in frame-shifts and/or premature stop codons, producing a truncated protein product [5]. In contrast, the functional impact of most missense variants that result in a single amino acid residue change in *BRCA1* protein is not known. The Breast Cancer Information Core database (<http://research.nhgri.nih.gov/bic/>), a central repository of *BRCA1* and *BRCA2* mutations identified in genetic tests, currently contains 487 unique missense *BRCA1* variants (April 2006), of which only 17 have sufficient genetic/epidemiological evidence to be classified as deleterious (Clinically Important) and 33 as neutral or of little clinical importance (Not Clinically Important). As genetic testing for inherited disease predispositions becomes more commonplace, predicting the clinical significance of missense variants and other UCVs will be increasingly important for risk assessment.

Because most UCVs in *BRCA1* and *BRCA2* occur at very low population frequencies (<0.0001) [6], direct epidemiological measures, such as familial cosegregation with disease, are often not sufficiently powerful to identify the variants associated with cancer predisposition. A promising approach is to supplement epidemiological and clinical analysis of UCVs with indirect approaches such as biochemical studies of

protein function and bioinformatics analysis [6–8]. In the future, physicians and genetic counselors may be able to rely on all these sources of information about UCVs when counseling their patients.

Previous bioinformatics analysis of *BRCA1* UCVs has depended primarily on measures of evolutionary conservation in multiple sequence alignments of human *BRCA1* and related proteins from other organisms [9–11]. Two groups have attempted to include information about *BRCA1* protein structure. Williams et al. predicted the impact of 25 missense variants in *BRCA1*'s C-terminal BRCT domains by considering both conservation and location of variant amino acid residues in an X-ray crystal structure [12]. Variants were predicted deleterious if their properties were similar to

Editor: Greg Tucker-Kellogg, Lilly Systems Biology, Singapore

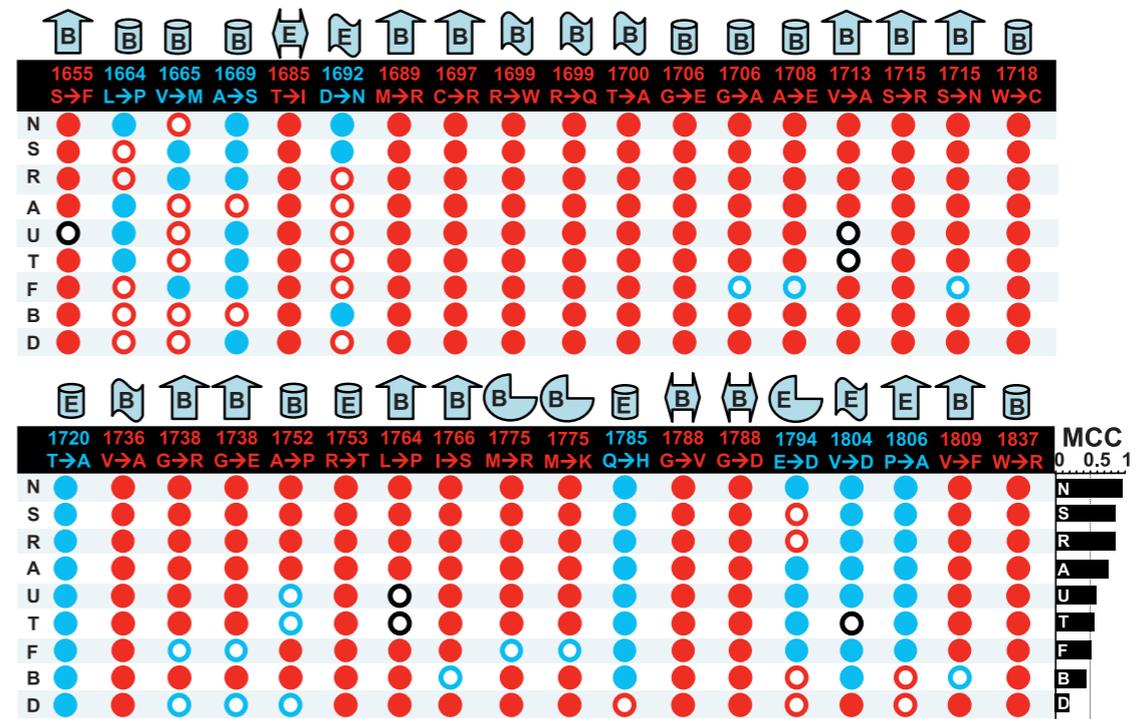
Received September 5, 2006; Accepted December 27, 2006; Published February 16, 2007

A previous version of this article appeared as an Early Online Release on December 28, 2006 (doi:10.1371/journal.pcbi.0030026.eor).

Copyright: © 2007 Karchin et al. This is an open-access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Abbreviations: Align-GVGD, Align Grantham Variation Grantham Deviation; AUC, area under the ROC curve; BIC, breast information core database; BRCT, *BRCA1* C-terminal domain; BRCT-C, BRCT C-terminal domain; BRCT-N, BRCT N-terminal domain; GD, Grantham Deviation; GV, Grantham Variation; ROC, receiver operating characteristic; Rule-based decision tree, empirically derived rules encoded in a decision tree; SIFT, Sorting Intolerant from Tolerant; UCV, variant of unknown clinical significance

* To whom correspondence should be addressed. E-mail: karchin@karchinlab.org (RK); sali@sallilab.org (AS)



- β-sheet
- bend
- H-bonded turn
- loop
- α-helix
- B** buried
- E** exposed

● ○ Correctly or incorrectly classified as deleterious
● ○ Correctly or incorrectly classified as neutral
○ Insufficient confidence to classify

Predictors are combined in support vector machine supervised learning

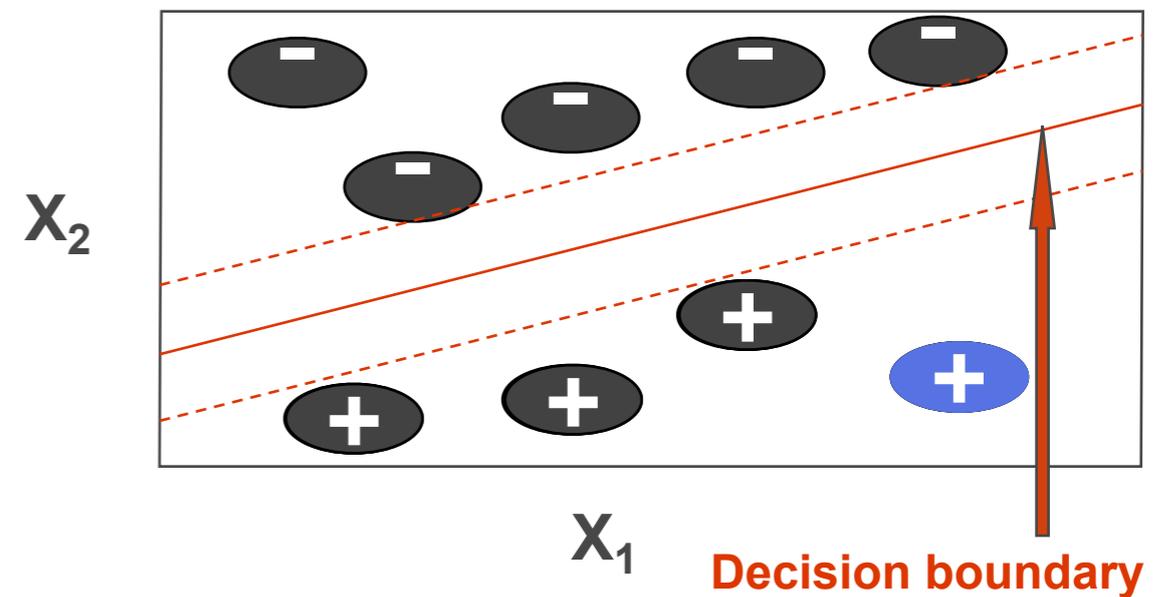
$X_1 \dots X_k = \text{TRAINING}$

| |
|-----------------------|
| relative entropy |
| Grantham distance |
| solvent accessibility |
| methyl(ene) groups |
| volume change |

-
-
-

| |
|------|
| + |
| 0.12 |
| 21 |
| 7 |
| 45 |
| -3.2 |

| |
|-----|
| - |
| 3.1 |
| 120 |
| 30 |
| 8 |
| 1.5 |



| |
|-----|
| 0.7 |
| 26 |
| 0 |
| 17 |
| 1.5 |

?

TESTING? PREDICTING...

Features

| Feature Category | Feature Description |
|---|--|
| Structural | Solvent Accessibility of wild-type amino acid residue (\AA^2) |
| | Solvent Accessibility of wild-type residue normalized by maximum exposed Solvent Accessibility of that residue type in a GLY-X-GLY tripeptide, using values given by Rose et al. [80] |
| | Solvent Accessibility of variant residue |
| | Normalized Solvent Accessibility of variant residue |
| | Number of methyl(ene) groups within 6 \AA of the variant sidechain [81] |
| | Number of unsatisfied spatial restraints in the MODELLER objective function after in silico mutation and simulated annealing refinement of the variant ^a |
| Physiochemical differences between wild-type and variant amino acid residues | Φ and Ψ backbone dihedral angles at the mutated position |
| | Whether the mutation results in buried charge |
| | Change in formal charge |
| | Change in volume (\AA^3) [82] |
| | Change in polarity [83] |
| Evolutionary conservation of amino acid residues in protein orthologs | Grantham difference [37] |
| | Relative entropy estimated by amino acids in the variant's alignment column [84] |
| | Positional hidden Markov model conservation score based on the probabilities of the wild-type, variant, and most probable amino acid residue in the variant's alignment column ^b [24] |

^aViolated restraints suggest that the mutated sidechain introduced steric clashes or unusual geometries into the protein model. Examples of violated restraints include extreme values of the Lennard-Jones 6–12 potential [85], bond angle potential, bond length potential, sidechain dihedral angle restraints, and nonbonded restraints. Two thresholds are used to identify violated restraints yielding two features.

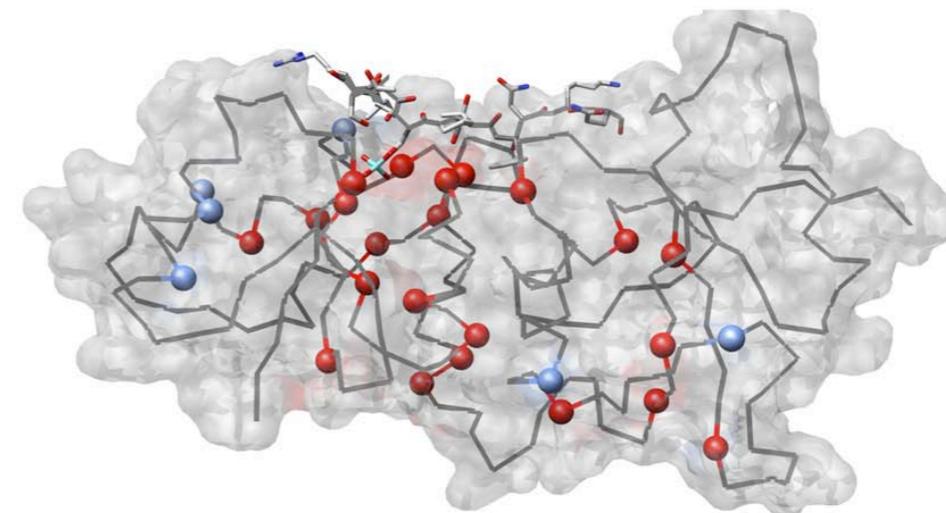
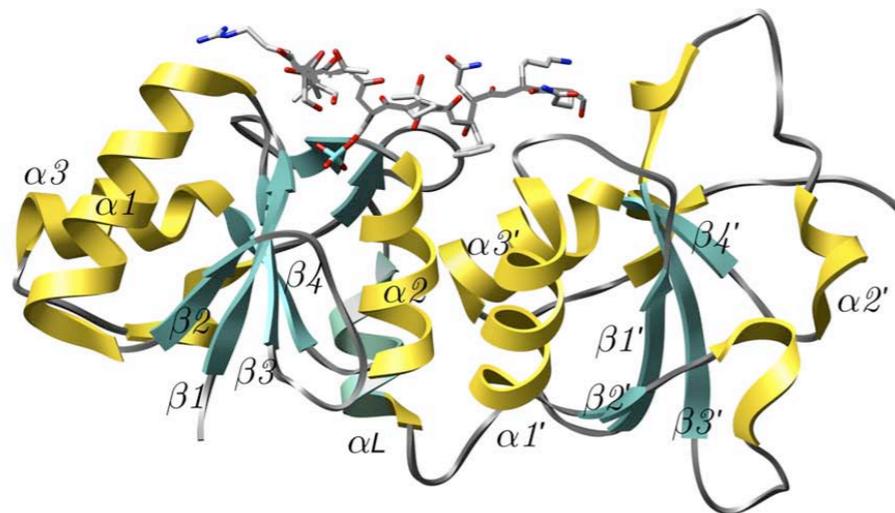
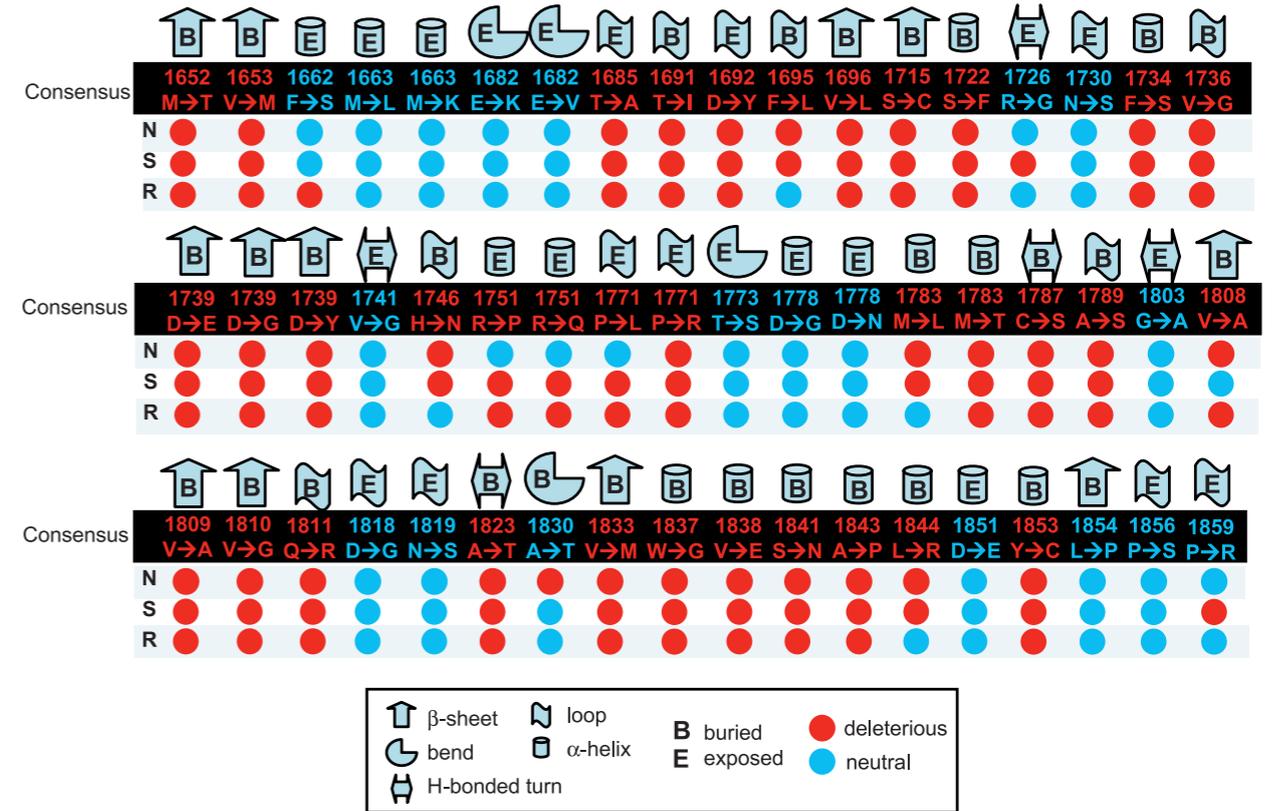
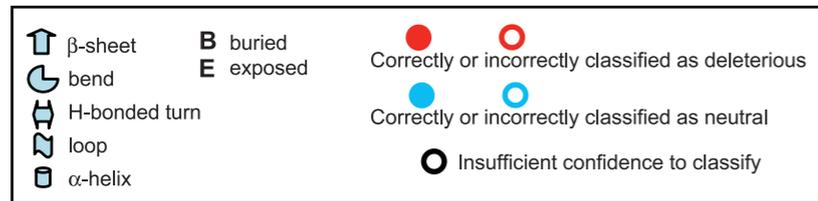
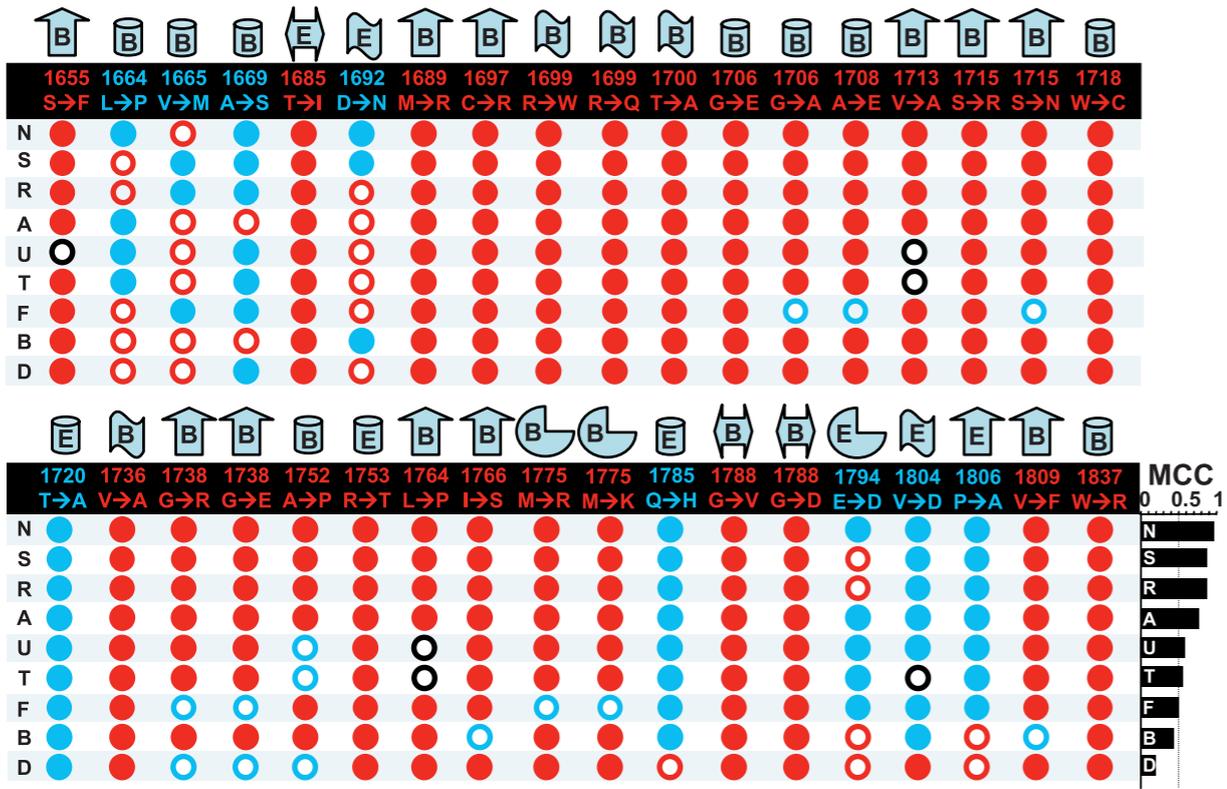
^bThe probabilities are estimated by a hidden Markov model built with SAM-T2K and the w0.5 script [23].

$\text{PHC} = \log(|p(\text{Wild-type}) - p(\text{Variant})|) + \log(p(\text{Wild-type})) + \log(P(\text{Most Probable})) - \log(p(\text{Variant}))$

The features were computed for 618 TP53 missense variants, 36 BRCA1 BRCT missense variants biochemically characterized in our companion paper [14], and 54 BRCA1 BRCT UCVs found in BIC.

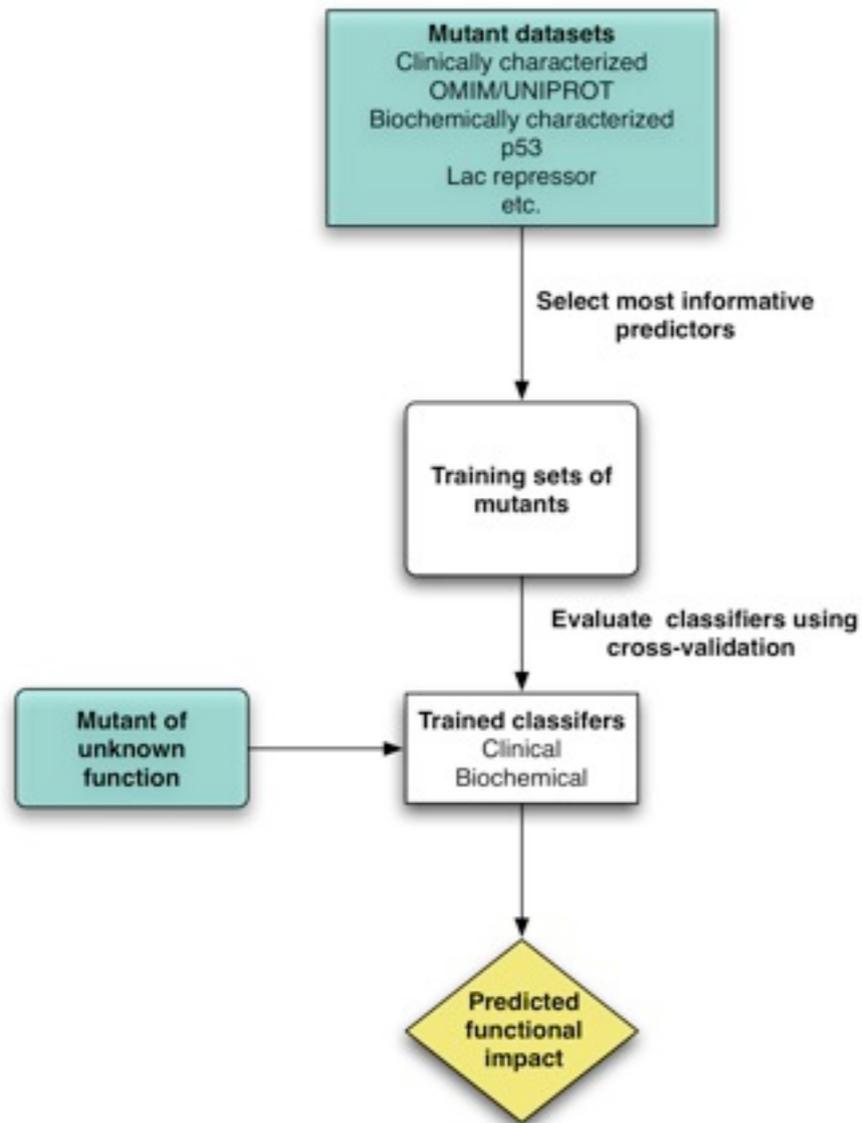
doi:10.1371/journal.pcbi.0030026.t002

Results



LS-SNP Large Scale SNP analysis

<http://salilab.org/LS-SNP/>



The screenshot shows the 'LS-SNP annotated database' web interface. The browser address bar displays <http://modbase.compbio.ucsf.edu/LS-SNP/Queries.html>. The page features a navigation menu on the left with options: Home, About, Background, Queries (selected), Downloads, Help, and Citation. The main content area is titled 'Queries' and includes a sub-section 'Query LS-SNP for SNP annotations'. This section offers two query methods: 'Query by ID type' and 'Query by genomic range'. Under 'Query by ID type', the 'ID type' is set to 'SwissProt ID' and the 'Enter ID(s):' field contains 'P19801,Q8WU19,P78549'. A note states: 'To query multiple ids of a single type, enter them in a comma-separated list.' Under 'Query by genomic range', the 'chr10:77300000-82600000' is entered. Below these fields, 'Accepted formats' are listed as 'chr11:59196382-68215218, 22q11.2, Xp21.2, 19p12'. The 'Display Information Type' section has radio buttons for 'Genomic sequence', 'Protein sequence', 'Protein structure' (selected), and 'Functional'. A checkbox for 'Validated SNPs only' is checked. 'Submit' and 'Reset' buttons are located at the bottom of the query form.

Protein function from structure

ab-initio localization of binding sites

Rossi. Localization of binding sites in protein structures by optimization of a composite scoring function. *Protein Science* (2006) vol. 15 (10) pp. 2366-2380

Downloaded from www.proteinscience.org on September 18, 2006

Localization of binding sites in protein structures by optimization of a composite scoring function

ANDREA ROSSI, MARC A. MARTI-RENO, AND ANDREJ SALI

Departments of Biopharmaceutical Sciences and Pharmaceutical Chemistry, California Institute for Quantitative Biomedical Research, University of California, San Francisco, California 94143-2552, USA

(RECEIVED March 28, 2006; FINAL REVISION July 10, 2006; ACCEPTED July 11, 2006)

Abstract

The rise in the number of functionally uncharacterized protein structures is increasing the demand for structure-based methods for functional annotation. Here, we describe a method for predicting the location of a binding site of a given type on a target protein structure. The method begins by constructing a scoring function, followed by a Monte Carlo optimization, to find a good scoring patch on the protein surface. The scoring function is a weighted linear combination of the z-scores of various properties of protein structure and sequence, including amino acid residue conservation, compactness, protrusion, convexity, rigidity, hydrophobicity, and charge density; the weights are calculated from a set of previously identified instances of the binding-site type on known protein structures. The scoring function can easily incorporate different types of information useful in localization, thus increasing the applicability and accuracy of the approach. To test the method, 1008 known protein structures were split into 20 different groups according to the type of the bound ligand. For nonsugar ligands, such as various nucleotides, binding sites were correctly identified in 55%–73% of the cases. The method is completely automated (<http://salilab.org/patcher>) and can be applied on a large scale in a structural genomics setting.

Keywords: protein function annotation; small ligand binding-site localization

Many protein targets of structural biologists are no longer chosen because of their function, but rather by their location in the protein sequence-structure space (Burley et al. 1999; Brenner 2000, 2001; Sali 2001; Vitkup et al. 2001; Chance et al. 2002; Goldsmith-Fischman and Honig 2003). Therefore, the number of functionally uncharacterized protein structures is growing. Of the 36,606 entries in the Protein Data Bank (PDB) (Kouranov et al. 2006) as of February 23, 2006, 1407 structures were deposited by structural genomics consortia, 985 (70%)

of which had an unknown function according to the HEADER record of their PDB files. In contrast, only 174 (0.5%) of the 35,199 protein structures solved outside of structural genomics had no functional annotations in their PDB files.

To classify the functions of thousands of uncharacterized protein structures that will become available over the next few years and millions of comparative models based on the known structures, automated structure-based functional annotation is required (Wallace et al. 1996, 1997; Kleywegt 1999; Thornton et al. 2000; Babbitt 2003; Laskowski et al. 2003). In particular, we need to be able to identify the locations and types of binding sites on a given structure, because the binding sites define the molecular function of a protein.

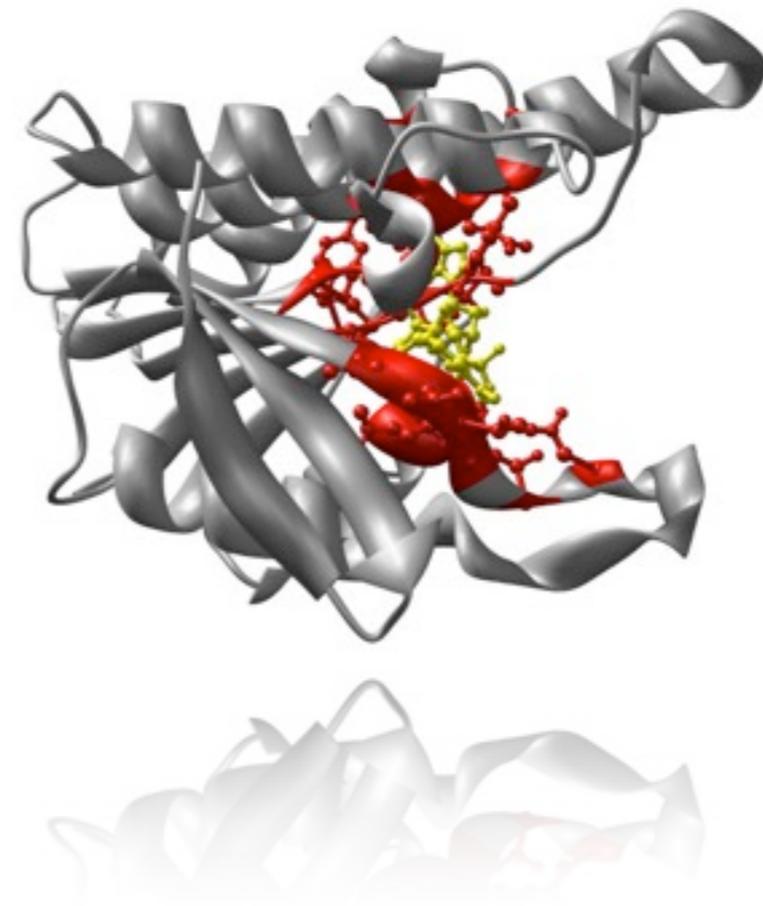
The most principled computational approach to predicting the molecular function is to dock a large library of potential ligands against the surface of the protein. In

Reprint requests to: Andrea Rossi or Andrej Sali, Departments of Biopharmaceutical Sciences and Pharmaceutical Chemistry, California Institute for Quantitative Biomedical Research, University of California, San Francisco Byers Hall, Office 503B, 1700 4th Street, San Francisco, CA 94143-2552, USA; e-mail: andrea@salilab.org or sali@salilab.org; fax: (415) 514-4231.

Article published online ahead of print. Article and publication date are at <http://www.proteinscience.org/cgi/doi/10.1110/ps.062247506>.

Protein Science (2006), 15:1–15. Published by Cold Spring Harbor Laboratory Press. Copyright © 2006 The Protein Society

1



15:1-15 (2006) doi:10.1110/ps.062247506

15:1-15 (2006) doi:10.1110/ps.062247506

15:1-15 (2006) doi:10.1110/ps.062247506

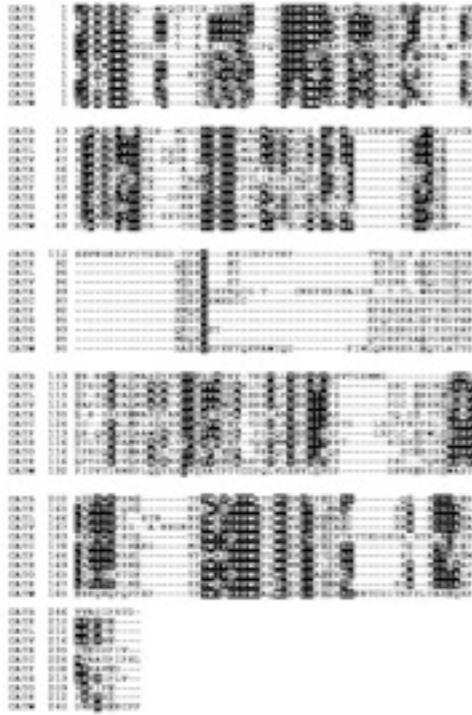
For **20%** protein structures function is *unknown*

| | Structural Genomics* | Traditional methods |
|------------------------|-----------------------------|----------------------------|
| Annotated** | 654 | 28,342 |
| Not Annotated | 506 (43.6%) | 6,815 (19.4%) |
| Total deposited | 1,160 | 35,157 |

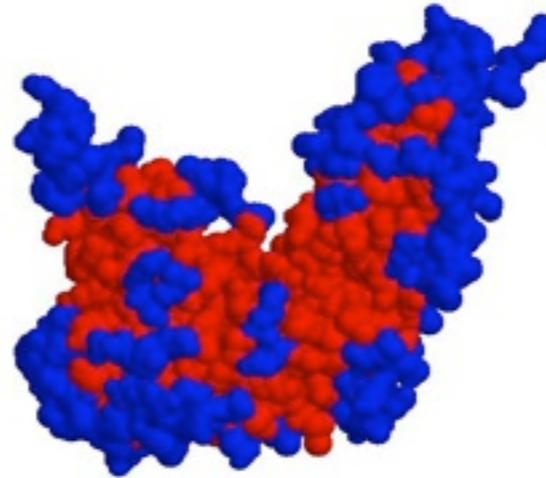
* annotated as *STRUCTURAL GENOMICS* in the header of the PDB file
**annotated with either *CATH*, *SCOP*, *Pfam* or *GO* terms in the MSD database
36,317 protein structures, as of August 8th, 2006

Representation

Sequence conservation



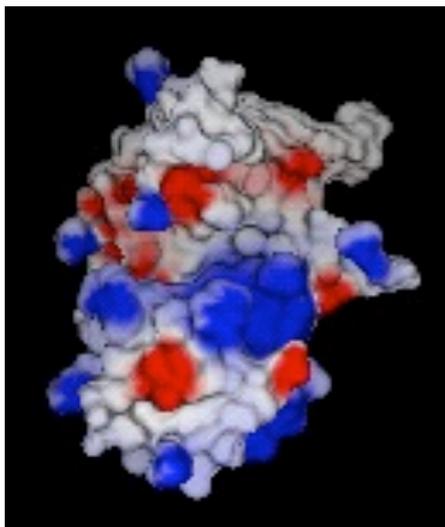
Surface geometry



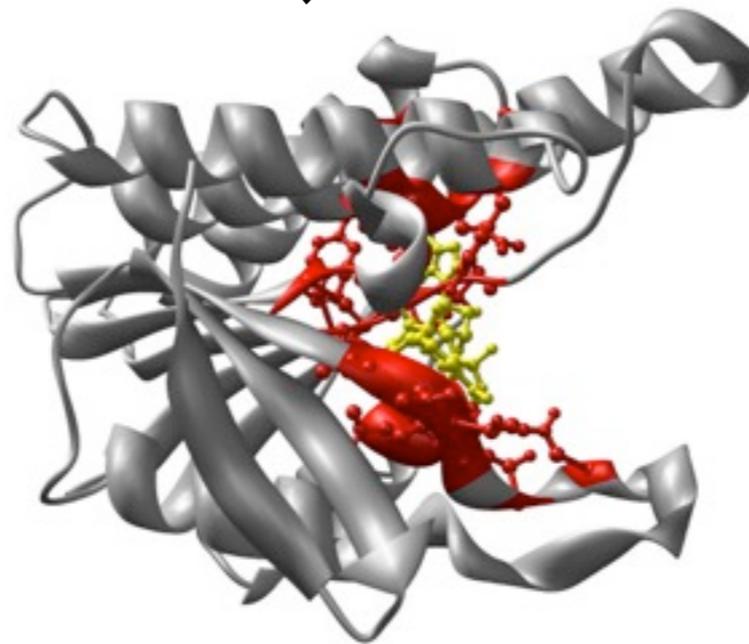
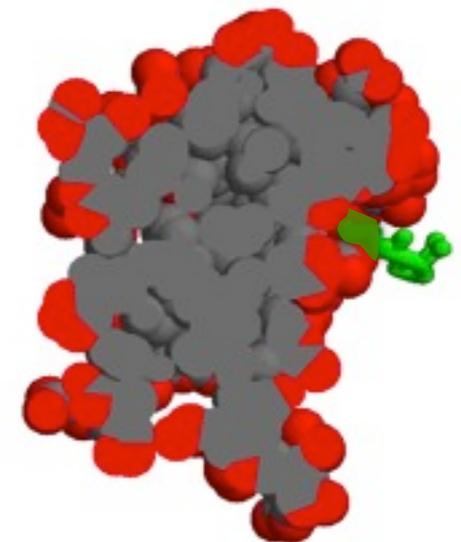
Structure conservation



Electrostatics

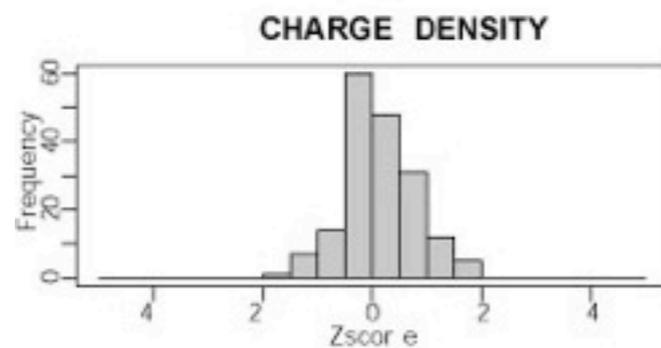
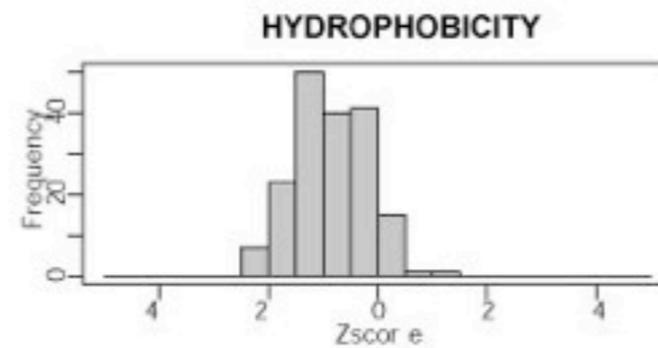
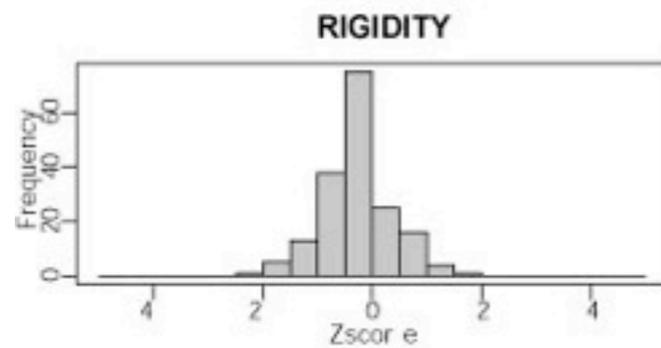
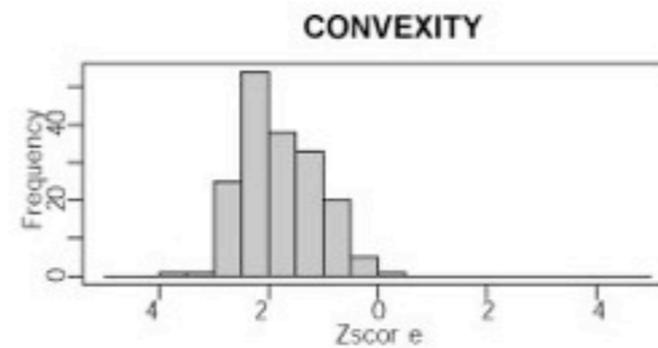
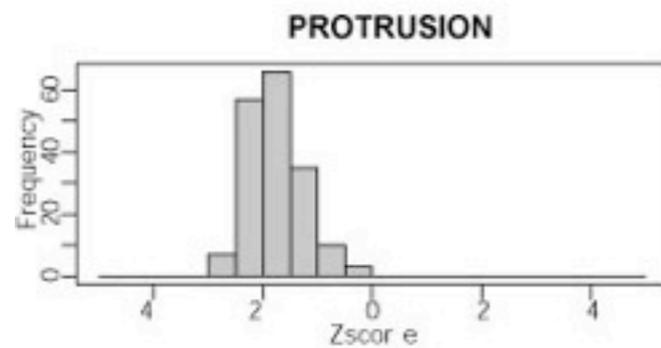
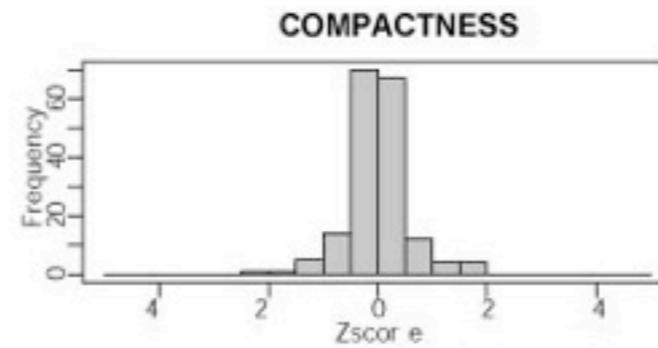
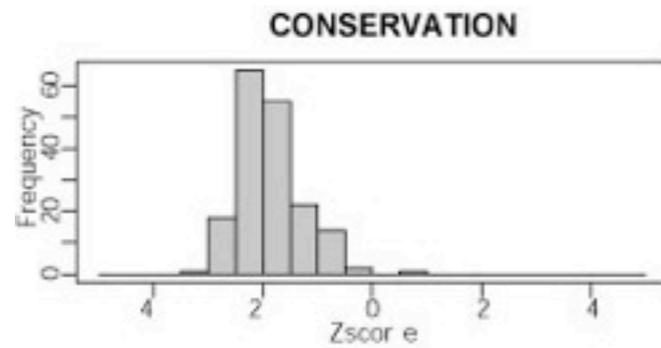


Solvent accessibility



Scoring

NAD



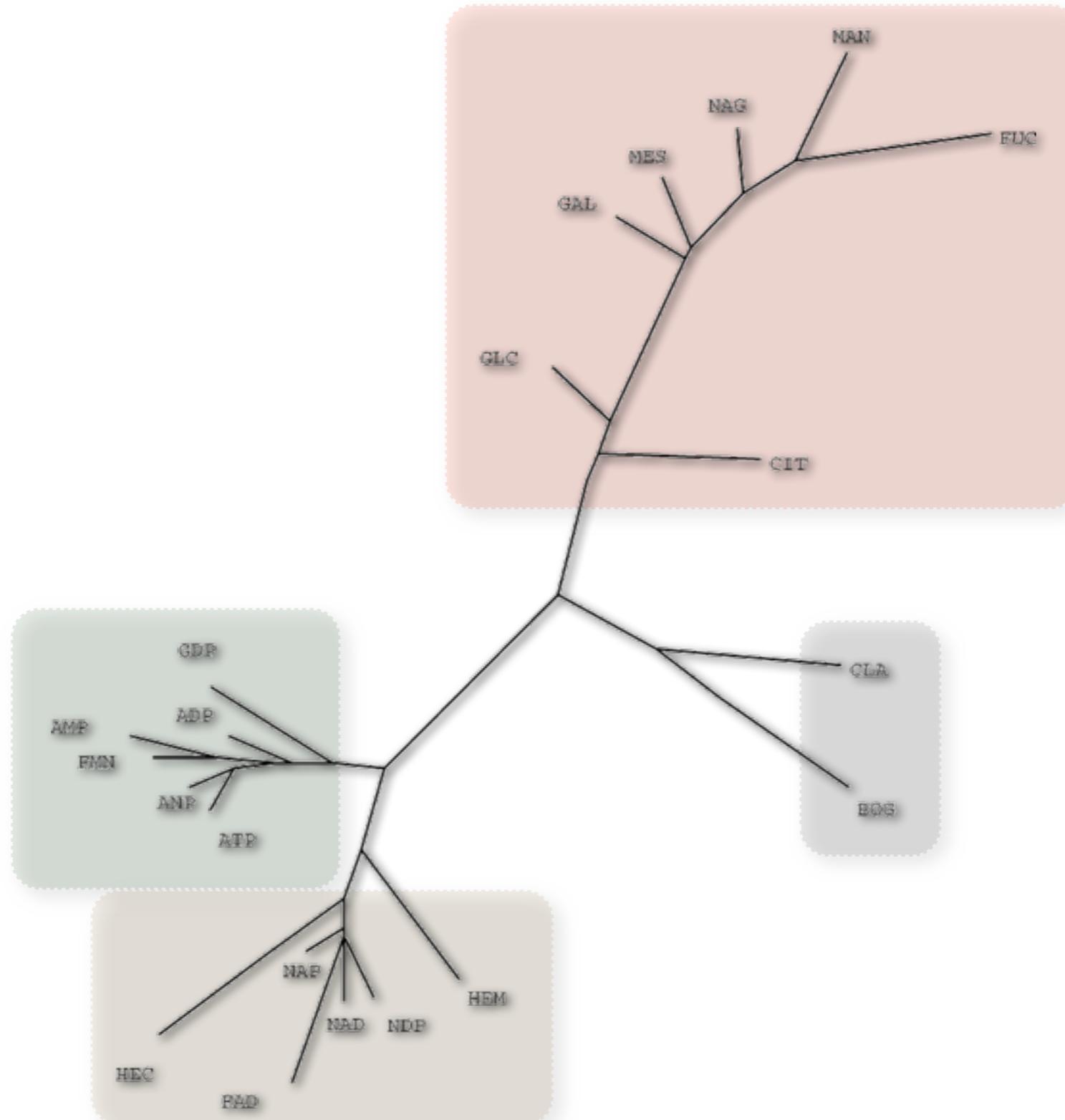
$$\rightarrow w_k = \frac{1}{M} \sum_{\alpha=1}^M \tilde{f}_k^{(\alpha)}$$

M = number of proteins in training set

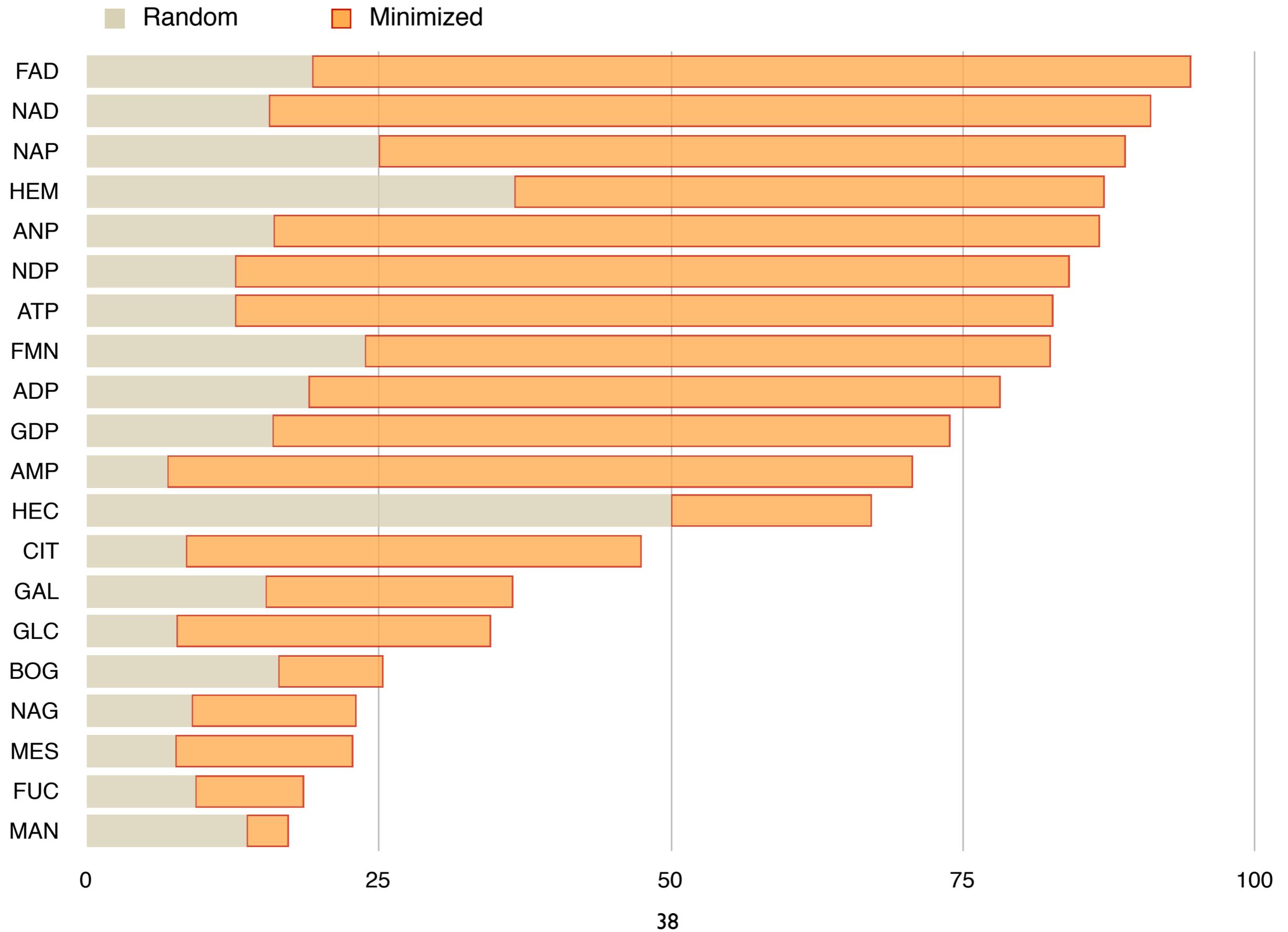
Ligand fingerprints

| | Compactness | Conservation | Charge density | B-factor | Protrusion coefficient | Convexity score | Hydrophobicity |
|-----|-------------|--------------|----------------|----------|------------------------|-----------------|----------------|
| ADP | -1.266 | -2.009 | 0.447 | -0.414 | -1.521 | -1.388 | -0.118 |
| AMP | -1.62 | -1.962 | 0.341 | -0.381 | -1.909 | -1.944 | -0.518 |
| ANP | -1.007 | -2.227 | 0.176 | -0.392 | -1.706 | -1.595 | -0.14 |
| ATP | -1.122 | -2.156 | 0.228 | -0.274 | -1.845 | -1.768 | 0.038 |
| BOG | -2.067 | -0.012 | 0.552 | -0.465 | -0.356 | -0.49 | -0.781 |
| CIT | -2.948 | -1.58 | 0.563 | -0.527 | -0.922 | -0.838 | -0.113 |
| FAD | 0.505 | -2.108 | 0.366 | -0.702 | -1.735 | -1.725 | -0.75 |
| FMN | -1.132 | -1.98 | 0.382 | -0.387 | -1.803 | -1.886 | -0.695 |
| FUC | -3.43 | 0.016 | -0.295 | -0.123 | 0.002 | 0.132 | 0.459 |
| GAL | -3.186 | -0.538 | -0.234 | -0.068 | -0.906 | -0.987 | 0.298 |
| GDP | -1.061 | -1.471 | 0.409 | -0.81 | -1.472 | -1.423 | 0.182 |
| GLC | -2.813 | -1.247 | -0.207 | -0.399 | -1.247 | -1.337 | -0.089 |
| HEC | -0.172 | -0.912 | 0.286 | -0.325 | -1.153 | -1.27 | -1.282 |
| HEM | -0.651 | -1.571 | 0.683 | -0.51 | -1.797 | -1.937 | -1.47 |
| MAN | -3.72 | 0.131 | 0.105 | -0.52 | -0.605 | -0.509 | 0.405 |
| MES | -3.049 | -0.24 | -0.338 | -0.479 | -0.714 | -0.926 | 0.296 |
| NAD | -0.005 | -1.852 | 0.156 | -0.232 | -1.775 | -1.804 | -0.858 |
| NAG | -3.419 | -0.46 | -0.126 | -0.154 | -0.341 | -0.523 | -0.078 |
| NAP | -0.009 | -1.898 | 0.612 | -0.321 | -1.587 | -1.656 | -0.336 |
| NDP | 0.217 | -1.741 | 0.535 | -0.312 | -1.463 | -1.562 | -0.498 |

Ligand fingerprints



Prediction accuracy



Protein function from structure

Comparative annotation. AnnoLite and AnnoLyze.

Marti-Renom et al. The AnnoLite and AnnoLyze programs for comparative annotation of protein structures.
BMC Bioinformatics (2007) vol. 8 (Suppl 4) pp. S4

BMC Bioinformatics



Proceedings

Open Access

The AnnoLite and AnnoLyze programs for comparative annotation of protein structures

Marc A Marti-Renom*¹, Andrea Rossi², Fátima Al-Shahrour³, Fred P Davis², Ursula Pieper², Joaquín Dopazo³ and Andrej Sali²

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

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

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

Published: 22 May 2007

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

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

© 2007 Marti-Renom et al; licensee BioMed Central Ltd.

This is an open access article distributed under the terms of the Creative Commons Attribution License (<http://creativecommons.org/licenses/by/2.0>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Abstract

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

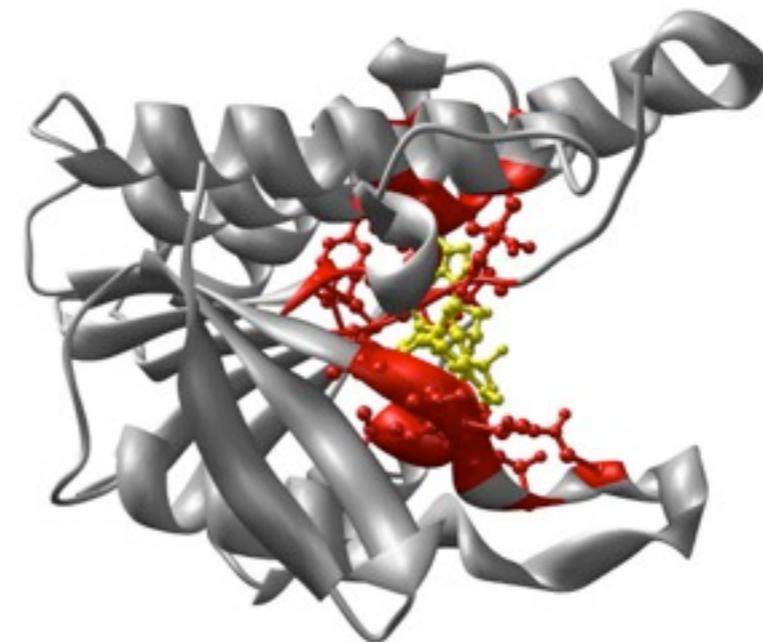
Description: AnnoLite predicts the SCOP, CATH, EC, InterPro, PfamA, and GO terms with an average sensitivity of ~90% and average precision of ~80%. AnnoLyze predicts ligand binding site and domain interaction patches with an average sensitivity of ~70% and average precision of ~30%, correctly localizing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

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

Background

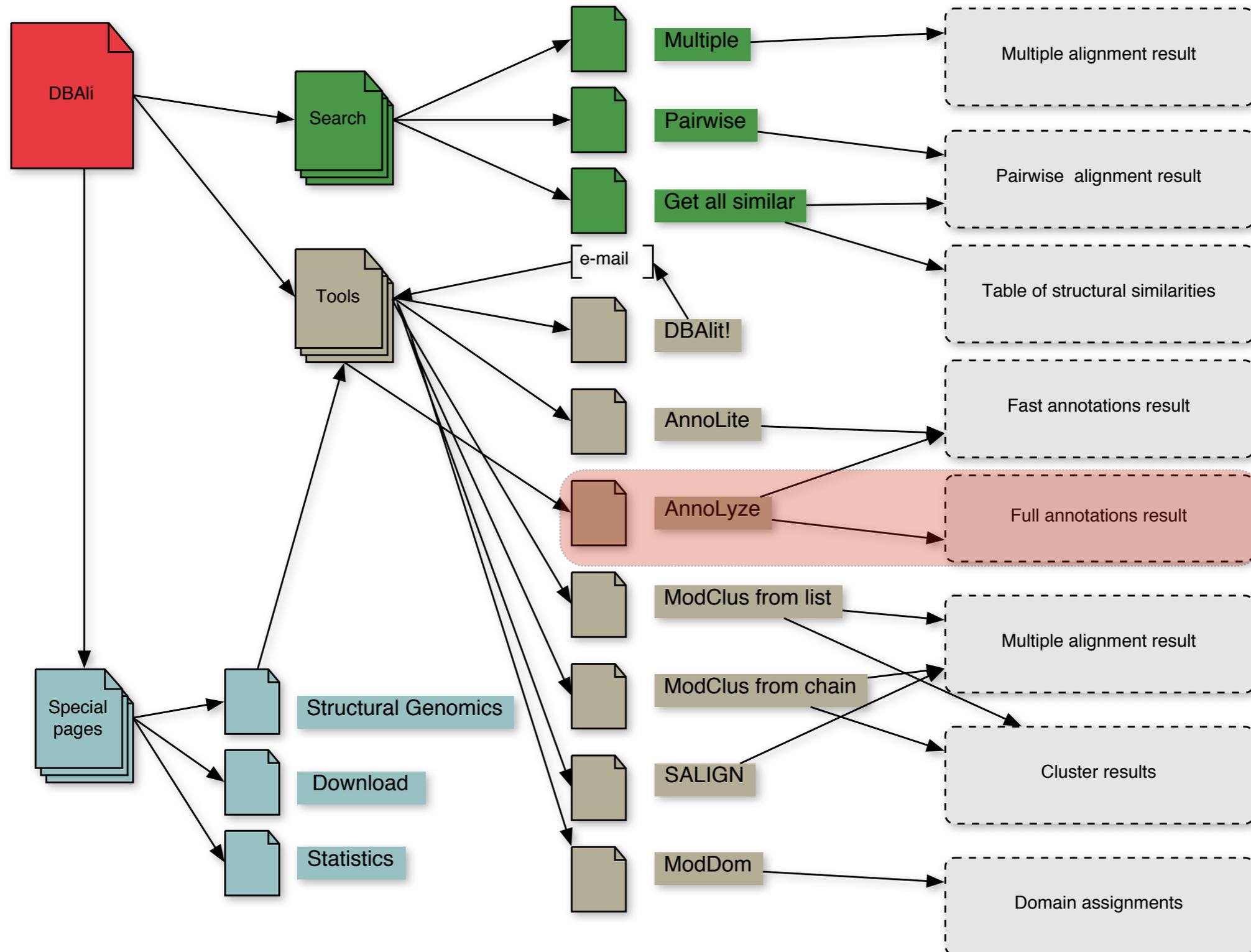
Genomic efforts are providing us with complete genetic blueprints for hundreds of organisms, including humans.

We are now faced with assigning, understanding, and modifying the functions of proteins encoded by these genomes. This task is generally facilitated by protein 3D



DBAli_{v2.0} database

<http://www.dbali.org>



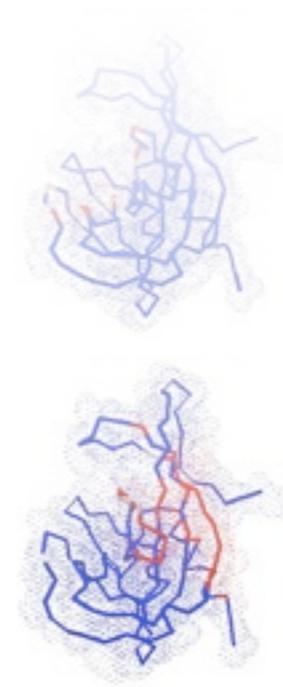
AnnoLyze

Inherited ligands: 4

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

Inherited partners: 1

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



Benchmark

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

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

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

****not two chains can be structurally aligned within 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*

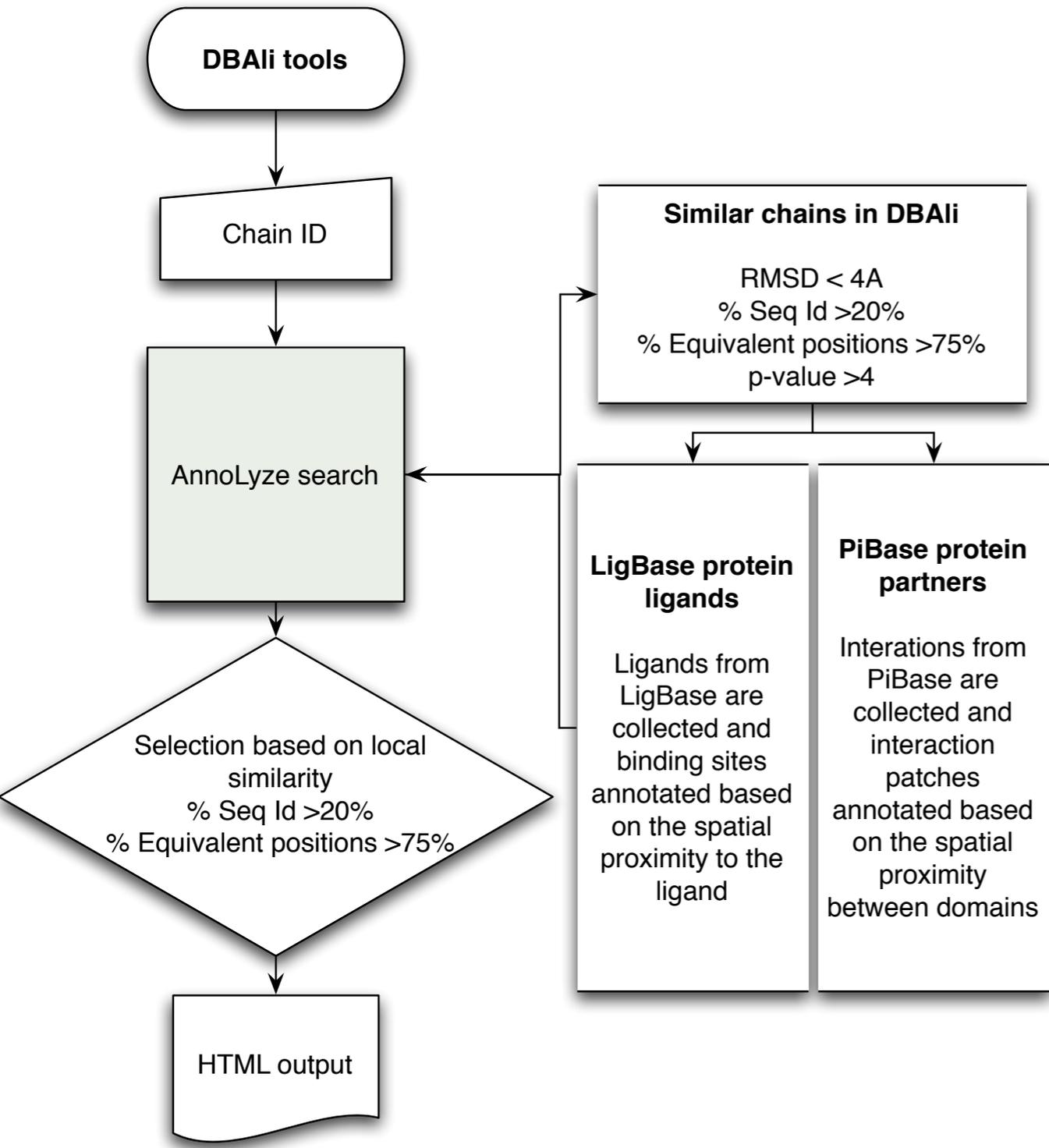
| | Number of chains |
|-------------------------------|------------------------------------|
| Initial set* | 78,167 |
| πBase** | 30,425 |
| Non-redundant set*** | 4,613 (11,641 partnerships) |

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

***annotated with at least one partner in the π Base database*

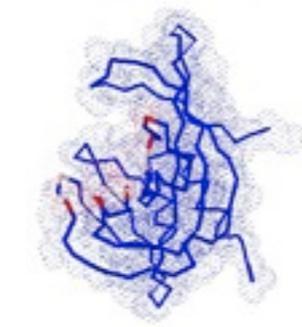
****not two chains can be structurally aligned within 3Å, superimposing more than 75% of their Ca atoms, result in a sequence alignment with more than 30% identity, and have a length difference inferior to 50aa*

Method



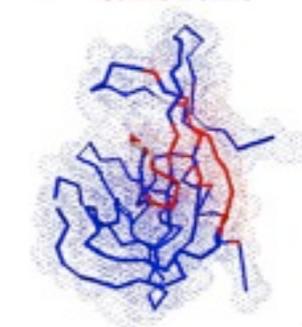
Inherited ligands: 4

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



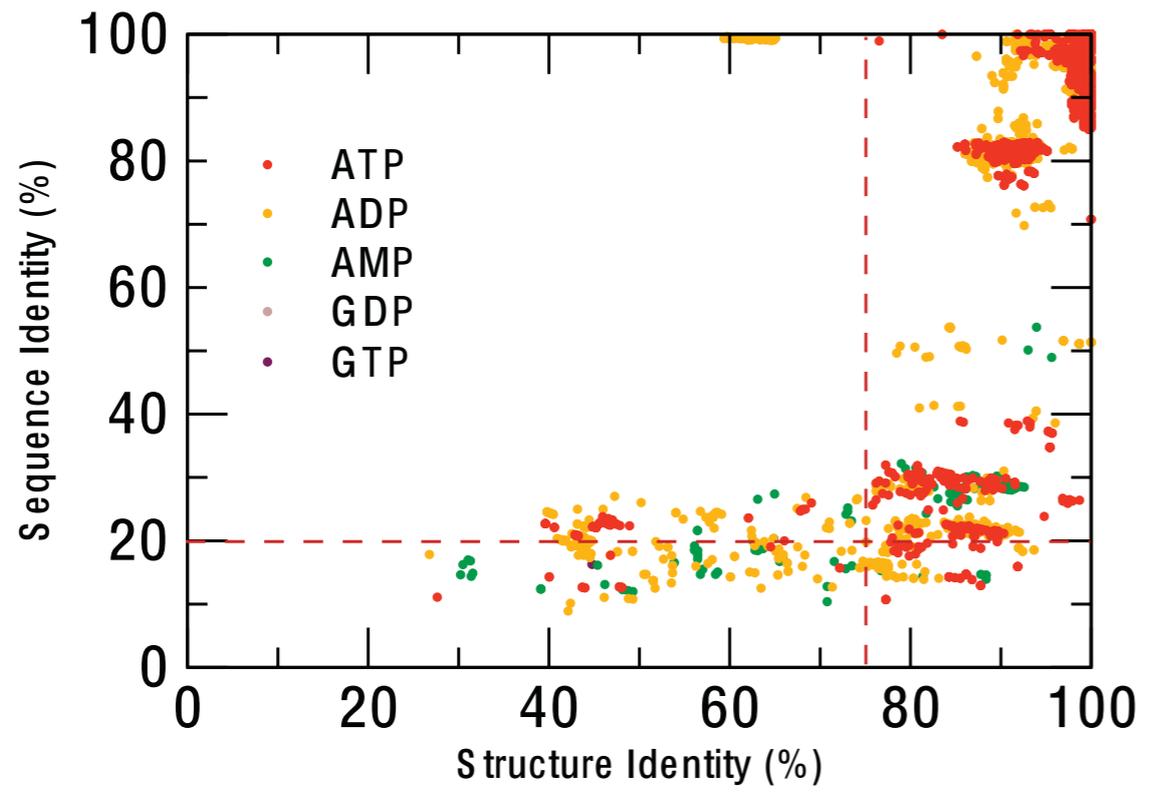
Inherited partners: 1

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

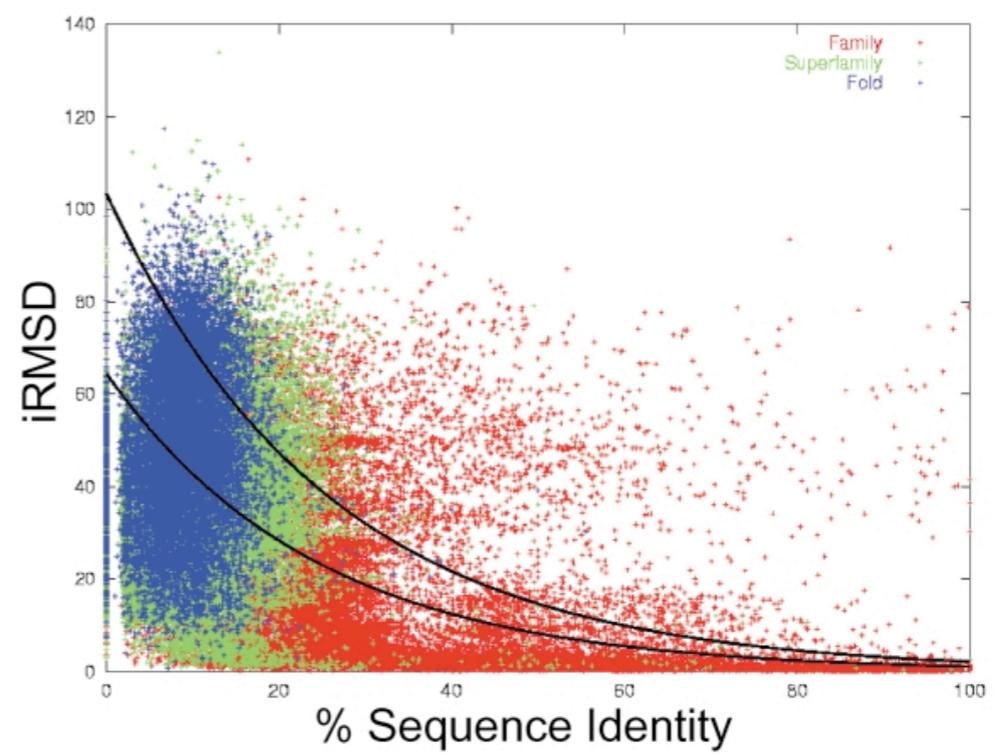


Scoring function

Ligands



Partners



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

Sensitivity .vs. Precision

| | Optimal cut-off | Sensitivity (%) 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:
Md5: 09b13d23ceae01dfcaddec636e2dda6KTPTAAAS
Length: 145

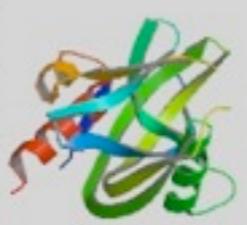
KTPFPKREE TLTYQTRYAA YIIIVSKPENN TMVLQAPNG AYFLPGGEIE
GTETKEEAIH REVLEELGIS VEIGCYLGEA DEYFYNSHRQ TAYYNPCYFY
VANTWRQLSE PLRNTLHWV APEEAVRLK RGSRWAVEK WLAAS

SCOP:
none

CATH:
none

Ligands:
none

Interacting partners:
none

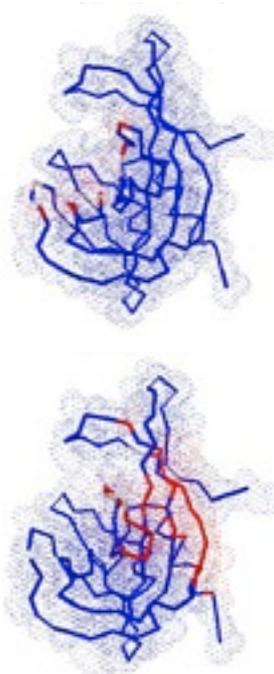


Inherited ligands: 4

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

Inherited partners: 1

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



Similar structures: [20](#)

Similar sequences: 890

Most similar structure in DBAll:

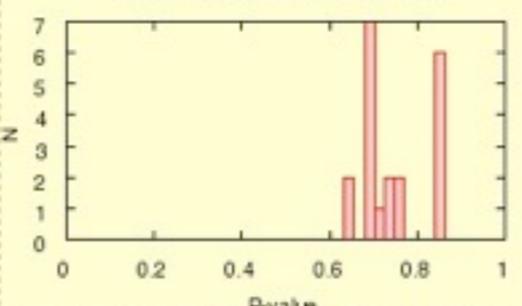
| Code | SeqId(%) | EqPos | RMSD | P-Value | See |
|------------------------|----------|-------|------|---------|---------------------|
| 1vc9:A | 22.76 | 123 | 3.57 | 17.28 | ali |

Most similar sequence in DBAll:

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

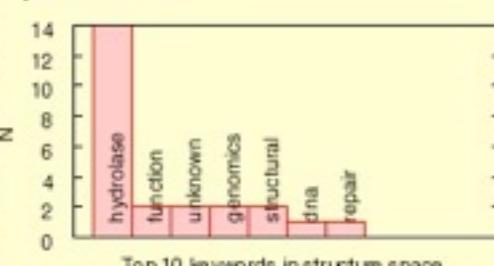
P-value distribution:

P-value distribution for similar chains

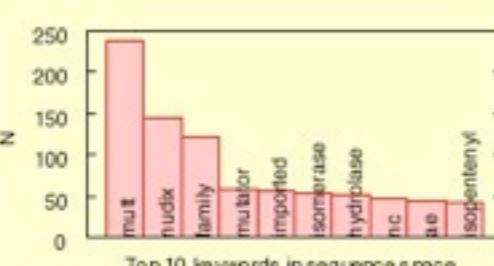


Keyword distribution:

Top 10 keywords in structure space

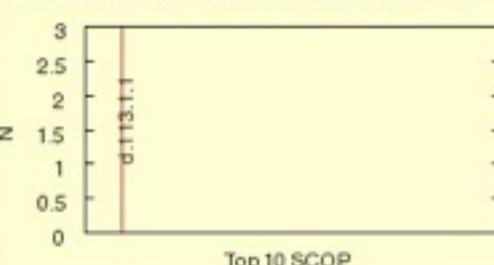


Top 10 keywords in sequence space

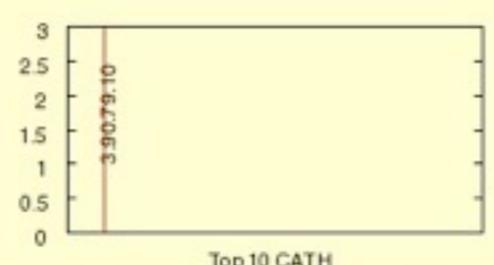


SCOP and CATH distribution for similar structures:

Top 10 SCOP

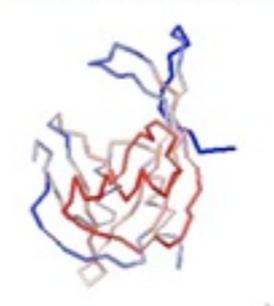


Top 10 CATH



Chain: 2azwA
Equivalences selected: 6793
Coverage: 100.00 %
Number of domains: 1
Assignment score: 0.3470
Alternative assignments: [none](#)

highest score assignment:
Domain 1:
2-113,115-147
Length: 145
Conservation: 308.49

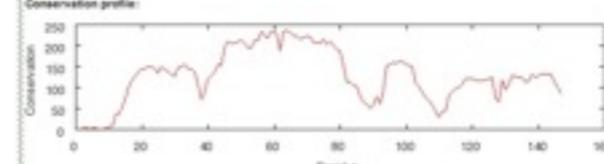


Color by conservation
Color by domains

Protein sequence colored by conservation and underlined by domain color:

KTPFPKREE TLTYQTRYAA YIIIVSKPENN TMVLQAPNG AYFLPGGEIE GTETKEEAIH REVLEELGIS VEIGCYLGEA DEYFYNSHRQ TAYYNPCYFY VANTWRQLSE PLRNTLHWV APEEAVRLK RGSRWAVEK WLAAS

Conservation profile:



| | Conf. P-value | Link | Description |
|-------------------------------|---------------|----------------------------|--|
| CATH: | 1.1e-20 | 3.90.79.10 | Nucleoside Triphosphate Pyrophosphohydrolase |
| SCOP: | 4.2e-29 | d.113.1.1 | MutT-like |
| PFAM: | 2.0e-74 | PF00293 | NUDIX domain |
| InterPro: | 1.9e-65 | IPR000086 | NUDIX hydrolase |
| | 2.7e-20 | IPR003561 | Mutator MutT |
| | 2.9e-14 | IPR002667 | isopentenyl-diphosphate delta-isomerase |
| EC Number: | 1.7e-4 | 3.6.1.17 | Bis(5'-nucleosyl)-tetraphosphatase (asymmetric) |
| GO Molecular Function: | 4.5e-19 | 0008413 | 8-oxo-7,8-dihydroguanine triphosphatase activity ↓ |
| | 3.8e-13 | 0004452 | isopentenyl-diphosphate delta-isomerase activity ↓ |
| | 1.9e-6 | 0016787 | hydrolase activity ↓ |
| | 5.4e-3 | 0004081 | Bis(5'-nucleosyl)-tetraphosphatase (asymmetric) activity ↓ |
| | 1.9e-2 | 0000287 | magnesium ion binding ↓ |
| GO Biological Process: | 7.7e-11 | 0008299 | isoprenoid biosynthesis ↓ |
| | 1.5e-5 | 0008974 | response to DNA damage stimulus ↓ |
| | 1.7e-5 | 0006260 | DNA replication ↓ |
| | 2.4e-5 | 0006281 | DNA repair ↓ |

AnnoLyze

<http://www.dbali.org>

DBAli v2.0 tools page

http://salilab.org/DBAli/?page=tools&action=f_annotatechain

CIPF | SGU Lab | UCSF | Sall Lab | DBAli | MAMMOTH

DBAli v2.0

last update: Oct 6th, 2007

Home
Search
Tools
Structural Genomics
Help

DBAli ALERT!
05/17/08 - Due to changes of disks, some DBAli tools may not be properly functioning. We are working to solve such problems.

DBAli. Tools associated to the database.

- **DBAli!** Compare your own structure to the whole PDB (temporarily not available)
- [AnnoLite: Fast annotation of a chain](#)
- [AnnoLyze: Annotate a chain](#)
- [ModClus: Cluster a list of chains](#)
- [ModClus: Cluster from a chain](#)
- [ModDom: Define domains from a chain](#)
- [SALIGN: Get a multiple structure alignment of a list of chains](#)

Annotate a given chain using the DBAli, LigBase, PiBase and ModBase databases.

Chain:

Mn Seq. Id.: Max Seq. Id.:

Min RMSD: Max RMSD:

Min % Eqpos: Max % Eqpos:

Mn P-value: Max P-value:

Type of annotation:

- General data
- Homology based data
- Inherited data
- Domain data

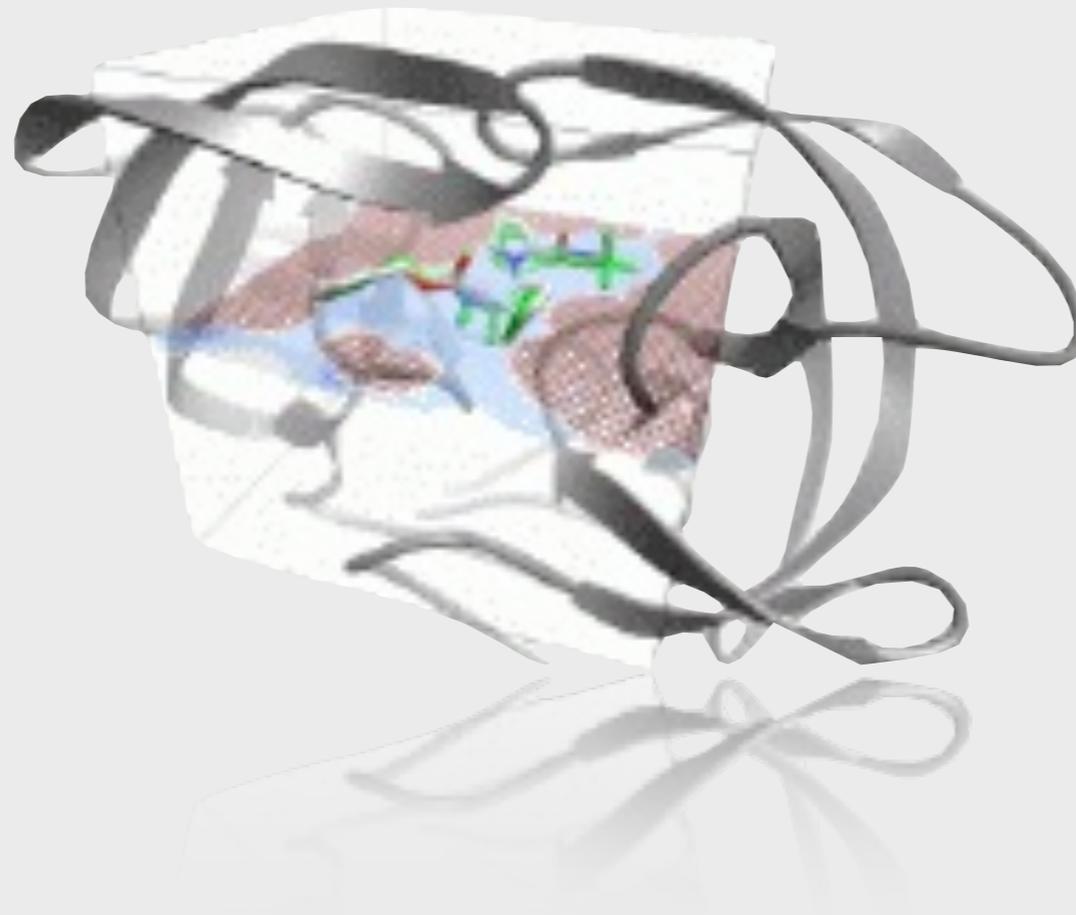
Please note:

- A permissive selection may result in significant server delay and incorrect annotation.
- Running ModDom to obtain domain based data may result in significant server delay.
- The annotation of a chain takes significant CPU time. Expect delays of about 2 minutes when selecting all available options.

Site Map - Reference - Download - Statistics - Suggestions - Report a problem - Visitors: 546 - © 2003 - 2008 Merit Discovery



Docking of small molecules. Vina.



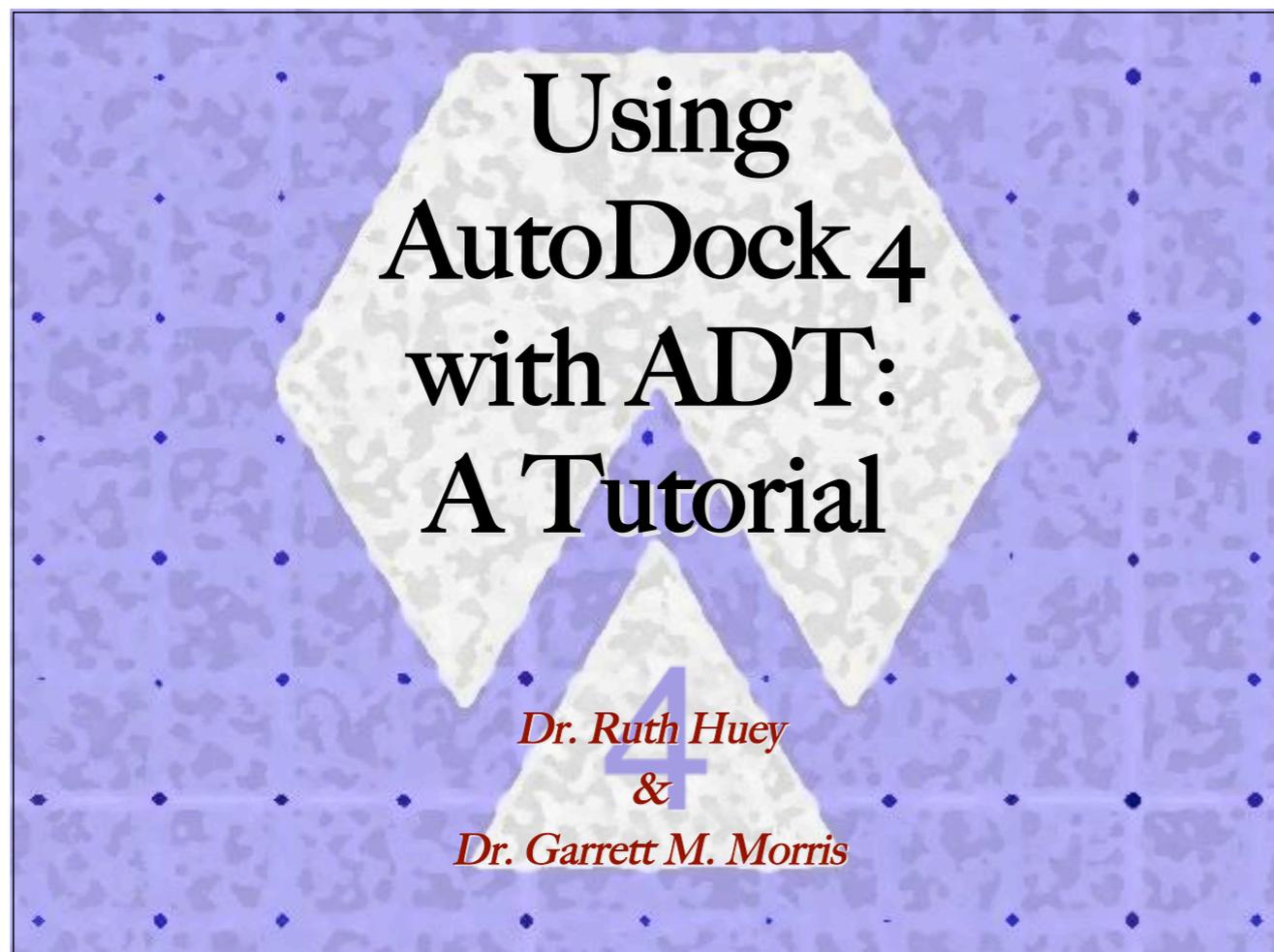
Marc A. Marti-Renom
<http://bioinfo.cipf.es/squ/>

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



DISCLAIMER!

Credit should go to Dr. Oleg Trott, Dr. Ruth Huey and Dr. Garret M. Morris



<http://autodock.scripps.edu>

<http://vina.scripps.edu>



O. Trott, A. J. Olson, *Journal of Computational Chemistry* (2009)

Summary

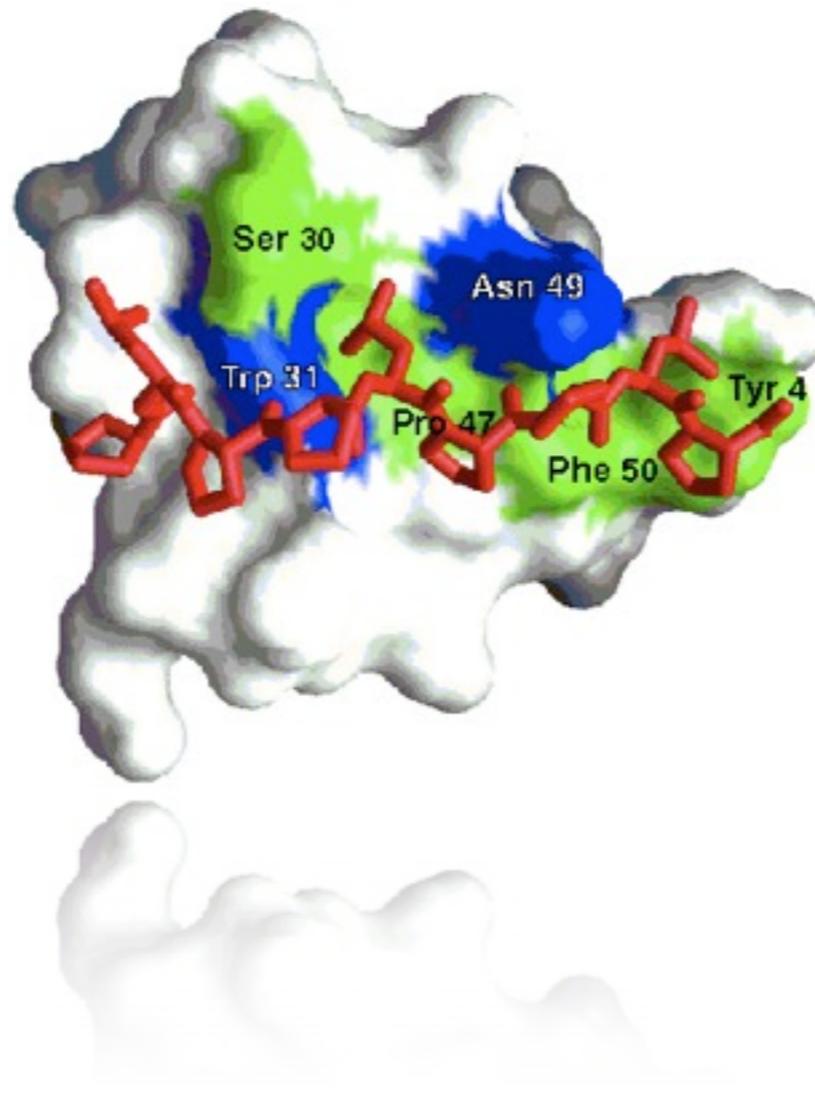
- **INTRO**
- **DOCKING**
- **SEARCH METHODS**
- **EXAMPLE**

- **Vina 1.0 with ADT**

What is docking?

Predicting the best ways two molecules interact.

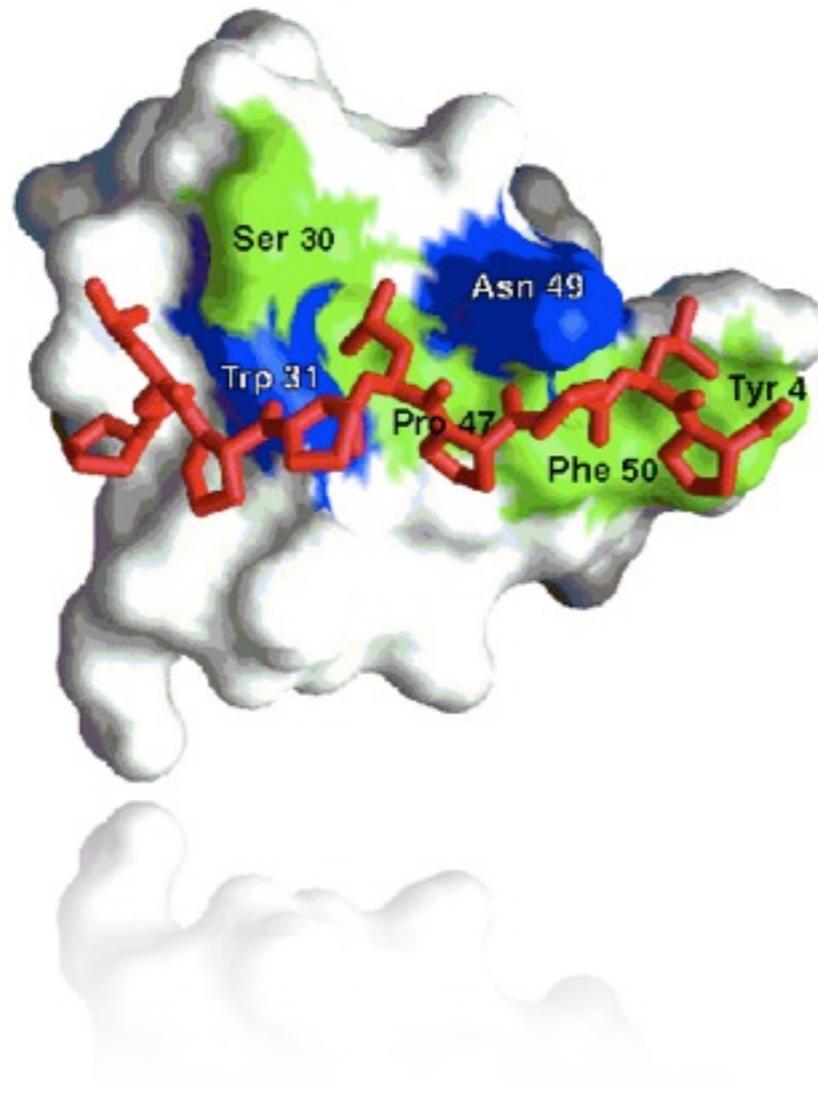
- ◆ Obtain the 3D structures of the two molecules
- ◆ Locate the best binding site (**Remember AnnoLyze?**)
- ◆ Determine the best binding mode.



What is docking?

Predicting the **best** ways two molecules interact.

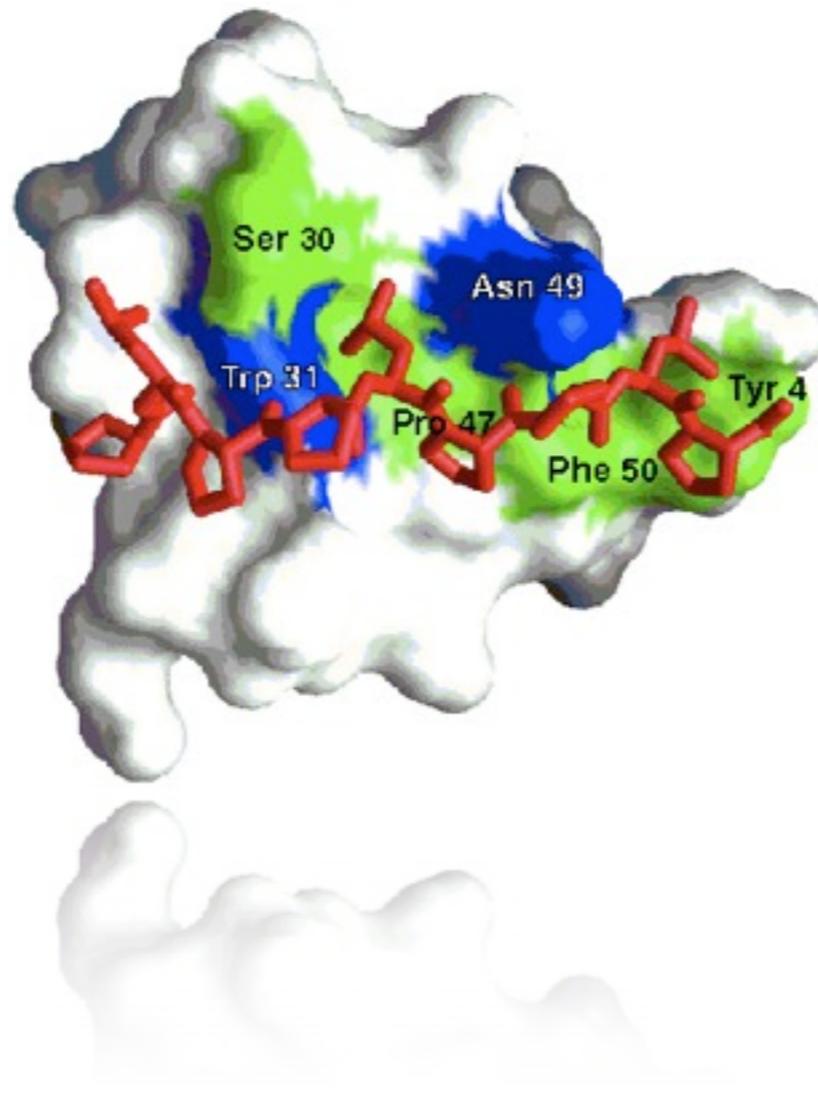
- ◆ We need to quantify or rank solutions
- ◆ We need a good scoring function for such ranking



What is docking?

Predicting the best **ways** two molecules interact.

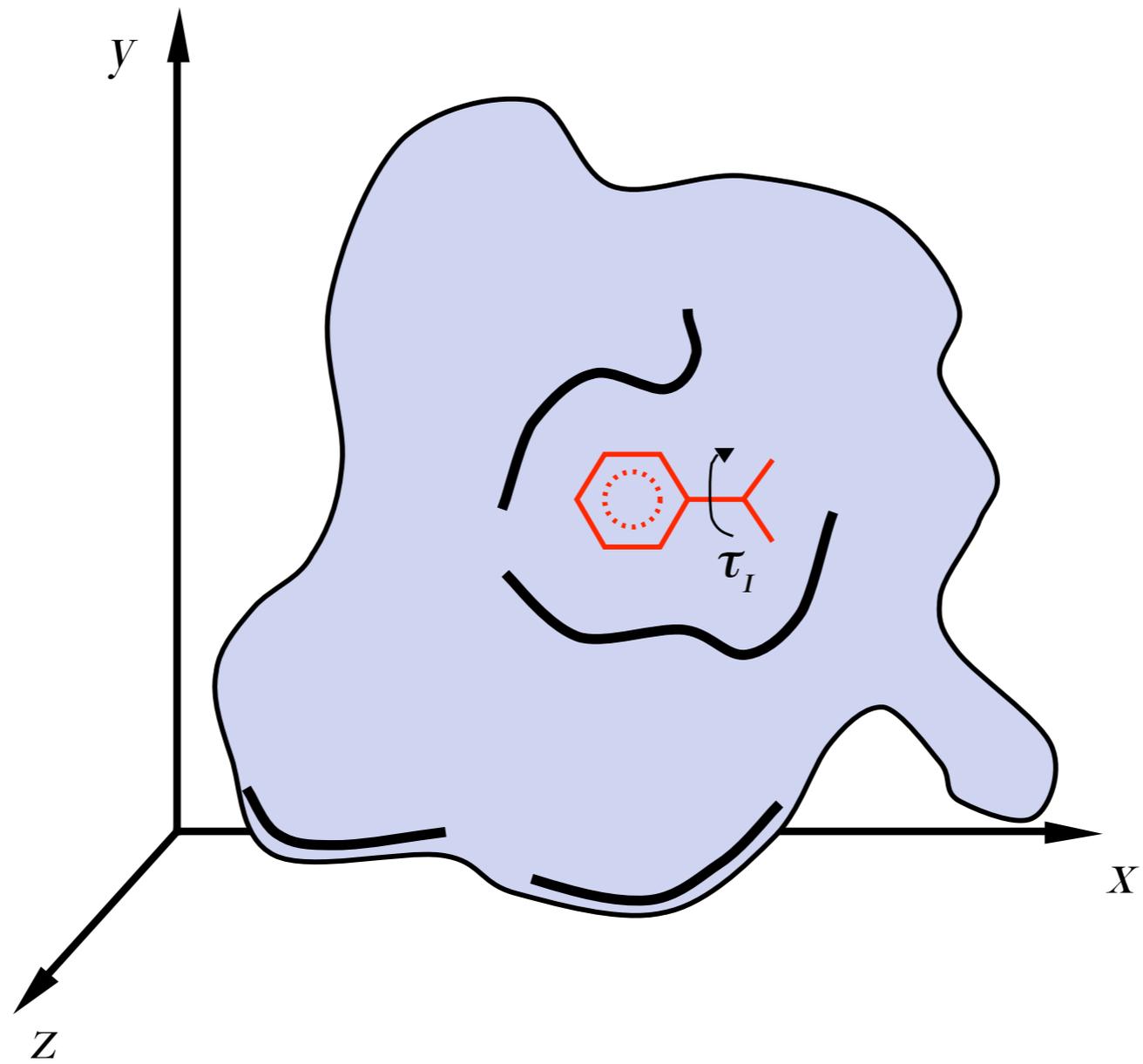
- ◆ X-ray and NMR structures are just ONE of the possible solutions
- ◆ There is a need for a search solution



BIOINFORMATICS

REPRESENTATION
SCORING
SAMPLING

REPRESENTATION

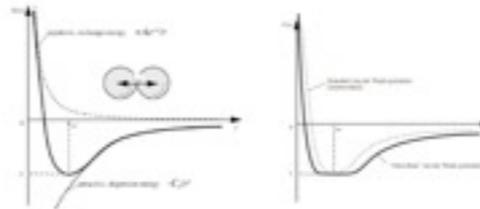


SCORING

AutoDock Vina

$$\Delta G_{binding} = \Delta G_{vdW} + \Delta G_{elec} + \Delta G_{hbond} + \Delta G_{desolv} + \Delta G_{tors}$$

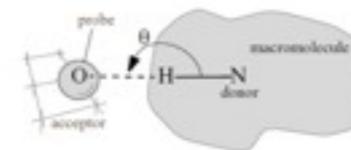
- ΔG_{vdW}
12-6 Lennard-Jones potential



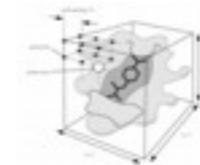
- ΔG_{elec}
Coulombic with Solmajer-dielectric

$$\epsilon(r) = A + \frac{B}{1 + ke^{-\lambda Br}}$$

- ΔG_{hbond}
12-10 Potential with Goodford Directionality



- ΔG_{desolv}
Stouten Pairwise Atomic Solvation Parameters



- ΔG_{tors}
Number of rotatable bonds



SAMPLING

AutoDock Vina

- ◆ **Global search algorithms**
 - ◆ Simulated annealing (Goodsell et al. 1990)
 - ◆ Distributed SA (Morris et al. 1996)
 - ◆ Genetic Algorithm (Morris et al. 1998)
- ◆ **Local search algorithms**
 - ◆ Solis & Wets (Morris et al. 1998)
- ◆ **Hybrid global-local search**
 - ◆ Lamarckian GA (Morris et al. 1998)

PROBLEM!

Very CPU time consuming...



$$N = T^{360/i}$$

N: number of conformations

T: number of rotatable bonds

i: incremental degrees

Metotrexato

10 rotatable bonds

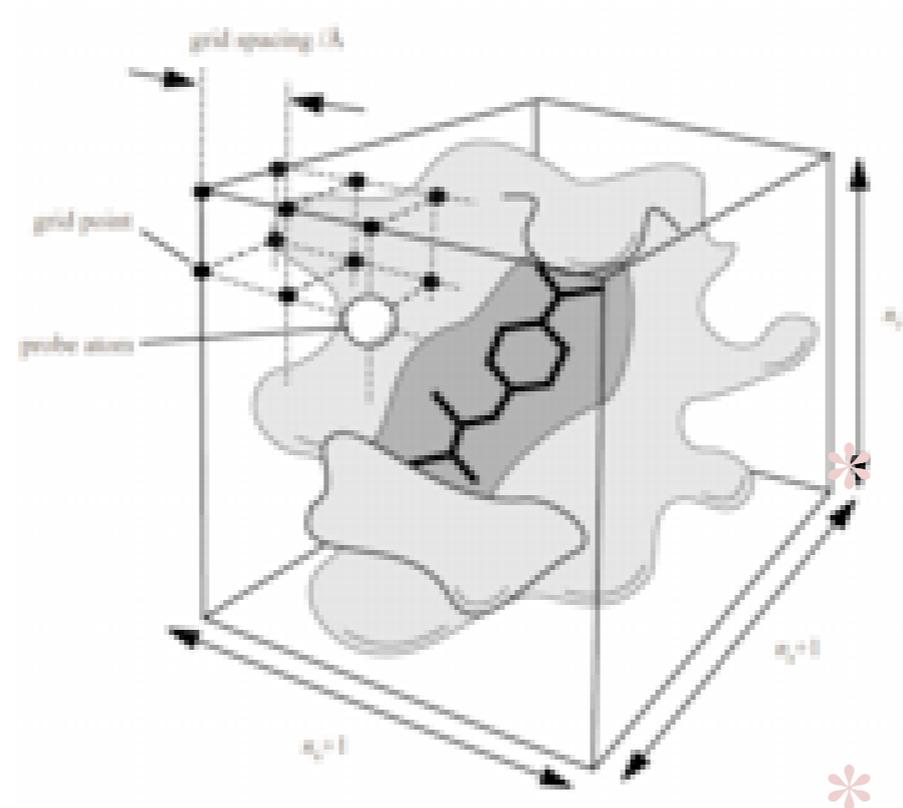
30° increments (discrete)

10¹² plausible conformations!

Dihydrofolate reductase with a metotrexate (4dfr.pdb)

SOLUTION

Use of grid maps!

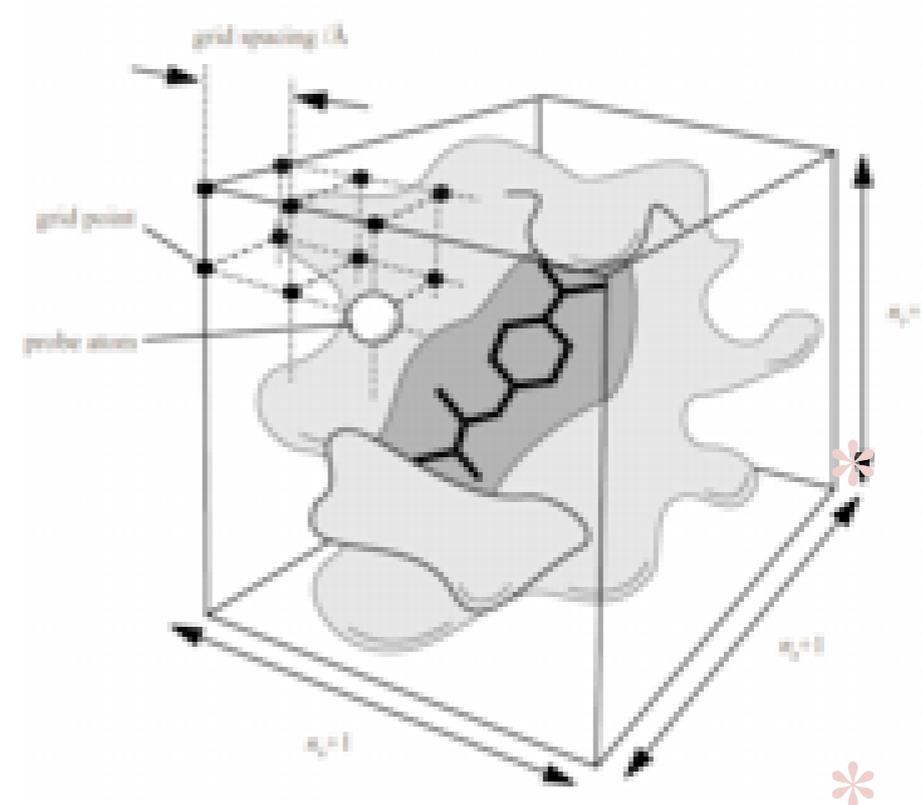


- ◆ Saves lots of time (compared to classical MM/MD)
- ◆ Need to map each atom to a grid point
- ◆ Limits the search space!

AutoGrid Vina

Use of grid maps!

- ◆ Center of grid *
 - ◆ center of ligand
 - ◆ center of receptor
 - ◆ a selected atom or coordinate
- ◆ Box dimension *
- ◆ Grid resolution (spacing)
 - ◆ default 0.375 Angstroms
- ◆ Number of grid points (dimension)
 - ◆ use ONLY even numbers
- ◆ MAKE SURE ALL LIGAND IS INSIDE GRID AND CAN MOVE!



With VINA much simplified (*)

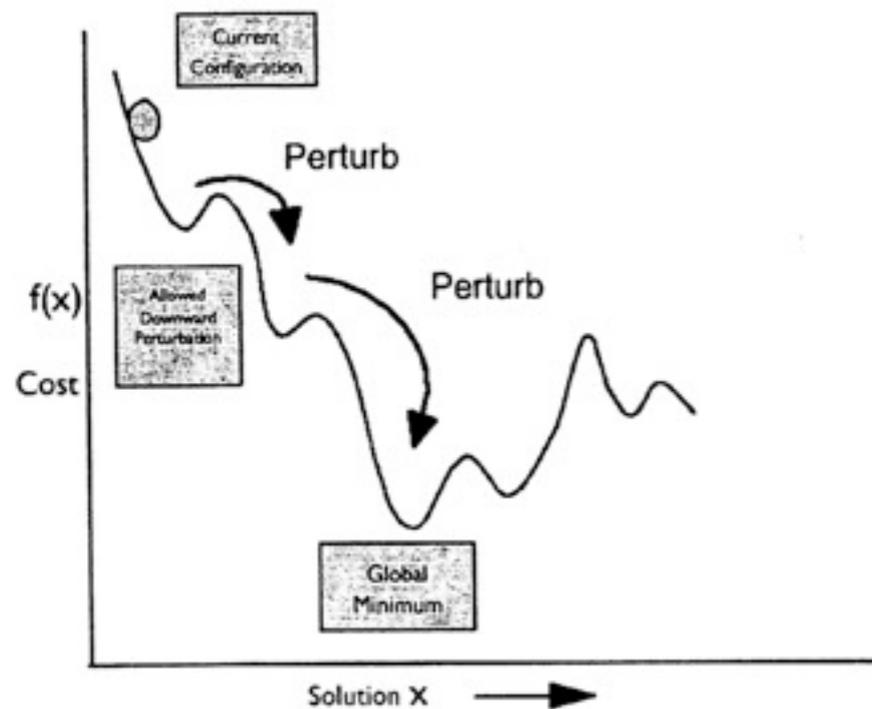
Search algorithms

Simulated Annealing

Ligand starts at initial state (random or user-defined)

The temperature of the system is reduced with time and the moves of the atoms are accepted depending on its energy compared to previous energy (with a probability proportional to the temperature!)

Repeat until reaching final solution.



Search algorithms

Genetic Algorithm

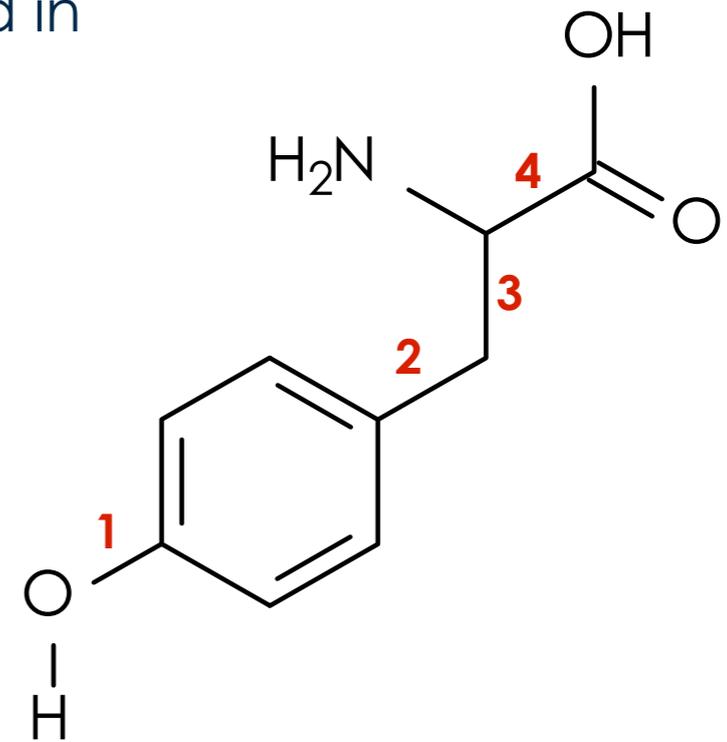
Use of a Genetic Algorithm as a sampling method

- Each conformation is described as a set of rotational angles.
- 64 possible angles are allowed to each of the bond in the ligand.
- Each plausible dihedral angle is codified in a set of binary bits ($2^6=64$)
- Each conformation is codified by a so called chromosome with 4×6 bits (0 or 1)

111010.010110.001011.010010

$\underbrace{\hspace{1.5cm}}_{\Phi_1}$ $\underbrace{\hspace{1.5cm}}_{\Phi_2}$...

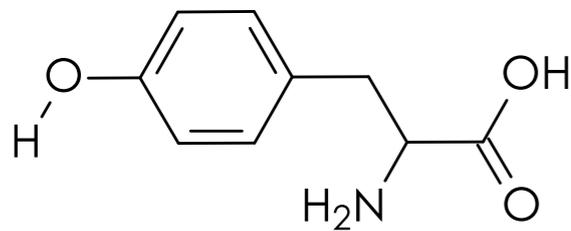
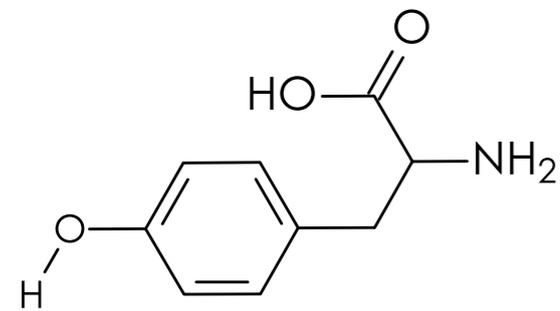
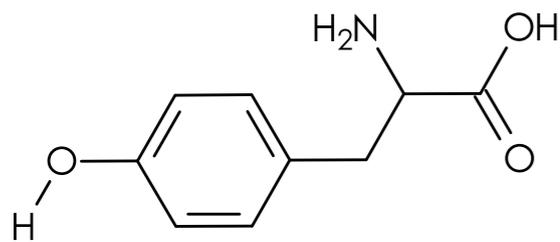
$$\Phi_1 = 1 \times 2^5 + 1 \times 2^4 + 1 \times 2^3 + 0 \times 2^2 + 1 \times 2^1 + 0 \times 2^0 = 58^\circ$$



Search algorithms

Genetic Algorithm

Population (ie, set of chromosomes or configurations)



011010.010110.011010.010111
111010.010110.001011.010010
001010.010101.000101.010001
101001.101110.101010.001000
001010.101000.011101.001011

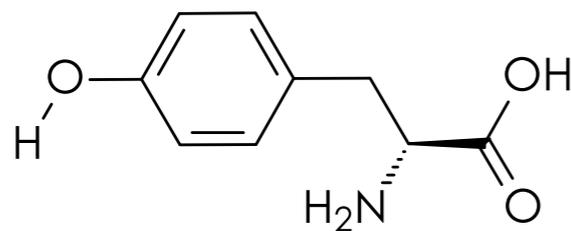
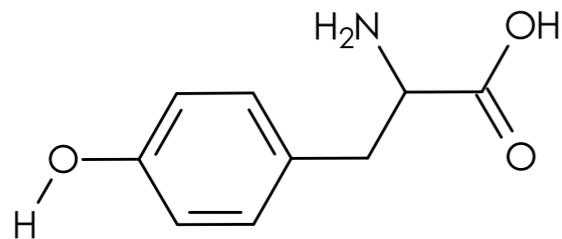
← Chromosome

↑
Gene

Search algorithms

Genetic Algorithm

Genetic operators...



011010.010110.011010.010111

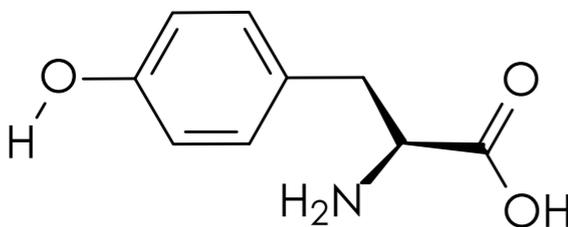
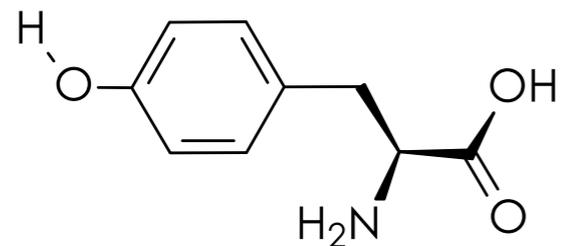
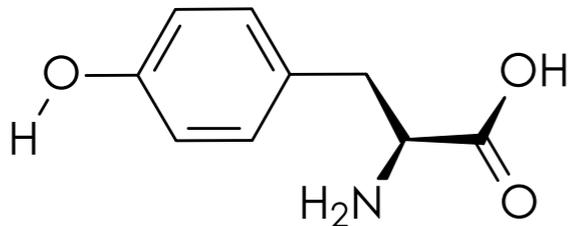
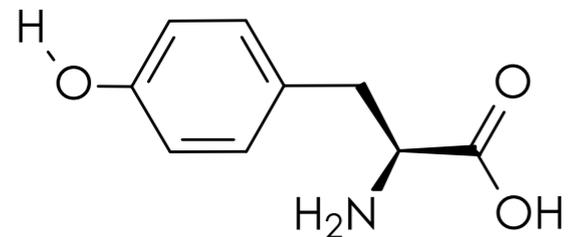
Single
mutation

011010.01**1**110.011**1**10.010111

Search algorithms

Genetic Algorithm

Genetic operators...



001010.010101.000101.010001

011010.010110.011010.010111

Recombination

001010.010101.011010.010111

011010.010110.000101.010001

Search algorithms

Genetic Algorithm

Genetic operators...

011010.010110.011010.010111
111010.010110.001011.010010
001010.010101.000101.010001
101001.101110.101010.001000
001010.101000.011101.001011

Migration



111110.010010.011110.010101
101010.110110.011011.011010
001010.010101.000101.010001
101101.101010.101011.001100
011010.100000.011001.101011

Search algorithms

Default parameters in AutoDock Vina

Simulated annealing

- ◆ Initial temperature
 - ◆ `rt0 = 61600 K`
- ◆ Temperature reduction factor
 - ◆ `rtrf = 0.95 K/cycle`
- ◆ Termination criteria
 - ◆ accepted moves (`accs = 25,000`)
 - ◆ rejected moves (`rejs = 25,000`)
 - ◆ annealing cycles (`cycles = 50`)

Genetic algorithm

- ◆ Population size
 - ◆ `ga_pop_size = 300`
- ◆ Crossover rate
 - ◆ `ga_crossover_rate = 0.8`
- ◆ Mutation rate
 - ◆ `ga_mutation_rate = 0.02`
- ◆ Solis and Wets local search (LGA only)
 - ◆ `sw_max_its = 300`
- ◆ Termination criteria
 - ◆ `ga_num_evals = 25,000 (short)`
 - ◆ `ga_num_evals = 250,000 (medium)`
 - ◆ `ga_num_evals = 2,500,000 (large)`
 - ◆ `ga_num_generations = 27,000`

AutoDock Example

Discovery of a novel binding trench in HIV Integrase

Schames, J.R., R.H. Henchman, J.S. Siegel, C.A. Sotriffer, H. Ni, and J.A. McCammon, Discovery of a novel binding trench in HIV integrase. J Med Chem, 2004. 47(8): 1879-81

The screenshot shows the Merck website's newsroom page. At the top, there is a navigation bar with the Merck logo and the slogan "Where patients come first". Below this is a search bar and a main navigation menu. The "Newsroom" section is highlighted, and a sidebar on the left lists various news categories. The main content area features a "Product News" section with a headline about FDA approval for ISENTRESS. The text below the headline provides details about the approval and clinical studies. A right-hand sidebar contains links for "ABOUT ISENTRESS", including "Full Prescribing Information" and "Patient Product Information".

Where patients come first **MERCK** Patients & Caregivers | Healthcare Professionals | Worldwide Quick Find Search

HOME | ABOUT MERCK | PRODUCTS | NEWSROOM | INVESTOR RELATIONS | CAREERS | RESEARCH | LICENSING | THE MERCK MANUALS

Newsroom

- Product News
- Research & Development News
- Corporate News
- Financial News
- Corporate Responsibility News
- Fact Sheet
- Executive Speeches
- Webcasts
- VIOXX® (rofecoxib) Information Center

Contact Newsroom
Podcast
RSS

Product News

FDA Approves ISENTRESS™ (raltegravir) Tablets, First-in-Class Oral HIV-1 Integrase Inhibitor

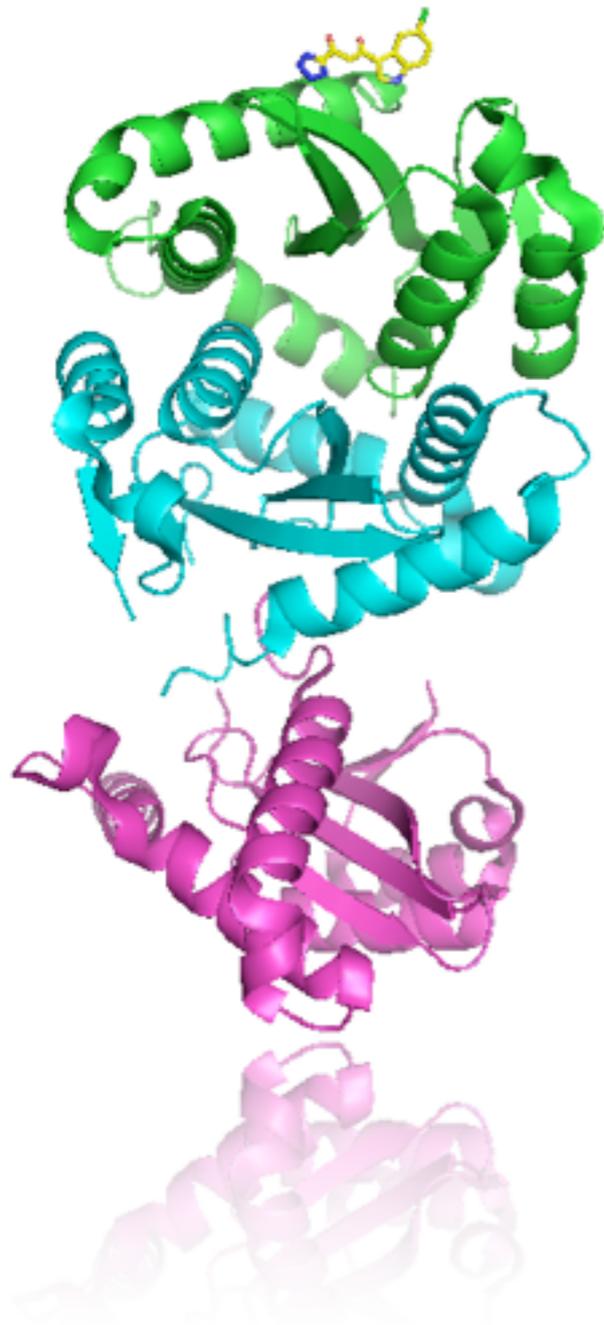
WHITEHOUSE STATION, N.J., Oct. 12, 2007 - Merck & Co., Inc., announced today that the U.S. Food and Drug Administration (FDA) granted ISENTRESS™ (raltegravir) tablets accelerated approval for use in combination with other antiretroviral agents for the treatment of HIV-1 infection in treatment-experienced adult patients who have evidence of viral replication and HIV-1 strains resistant to multiple antiretroviral agents.

This indication is based on analyses of plasma HIV-1 RNA levels up through 24 weeks in two controlled studies of ISENTRESS [pronounced i-sen-tris]. These studies were conducted in clinically advanced, three-class antiretroviral [nucleoside reverse transcriptase inhibitors (NRTIs), non-nucleoside reverse transcriptase inhibitors (NNRTIs) and protease inhibitors (PIs)] treatment-experienced adults. The use of other active agents with ISENTRESS is associated with a greater likelihood of treatment response. The safety and efficacy of ISENTRESS have not been established in treatment-naïve adult patients or pediatric patients. There are no study results demonstrating the effect of ISENTRESS on clinical progression of HIV-1 infection. Longer term data will be required before the FDA can consider traditional approval for ISENTRESS.

ABOUT ISENTRESS

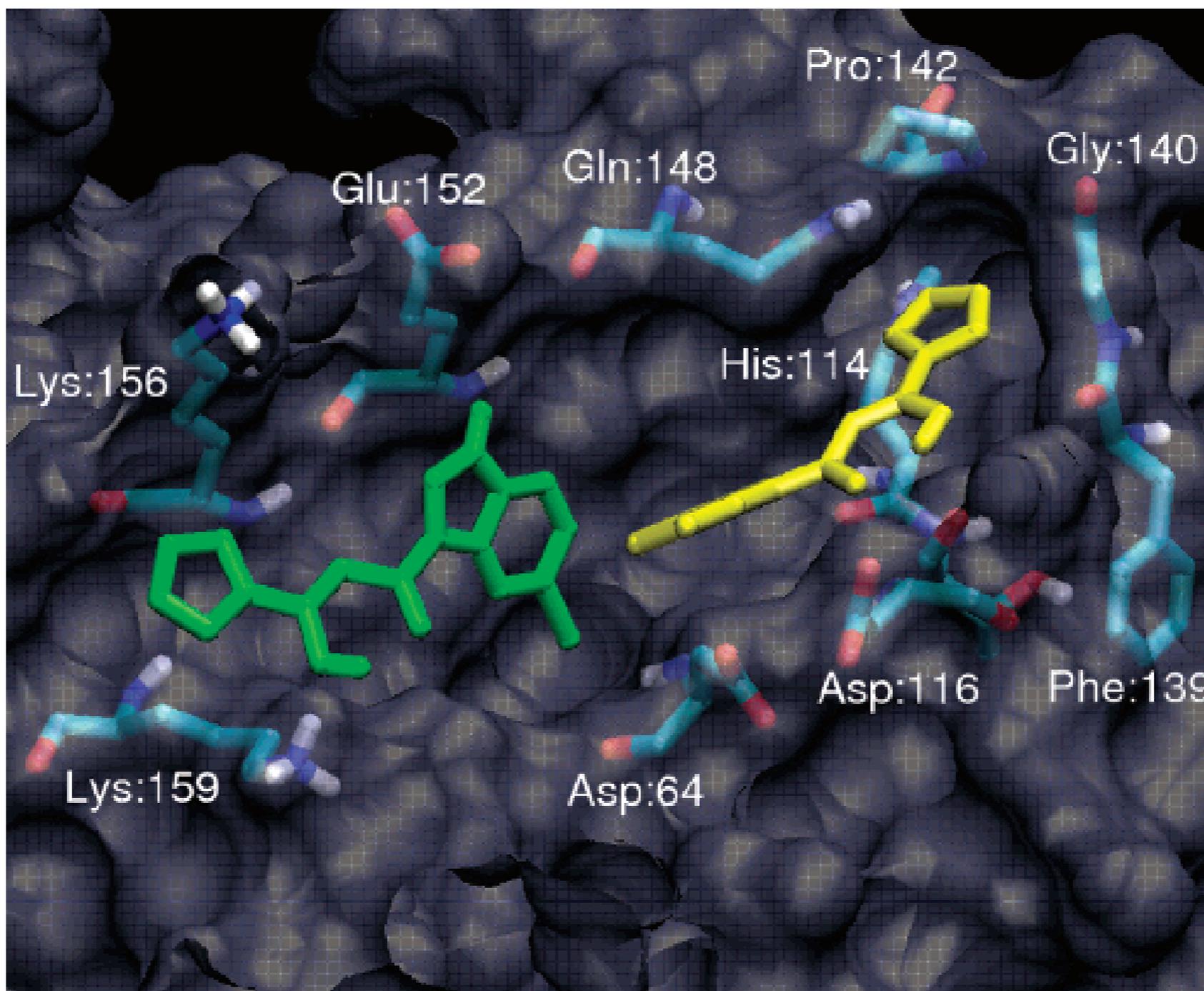
- Full Prescribing Information
- Patient Product Information

ISENTRESS example

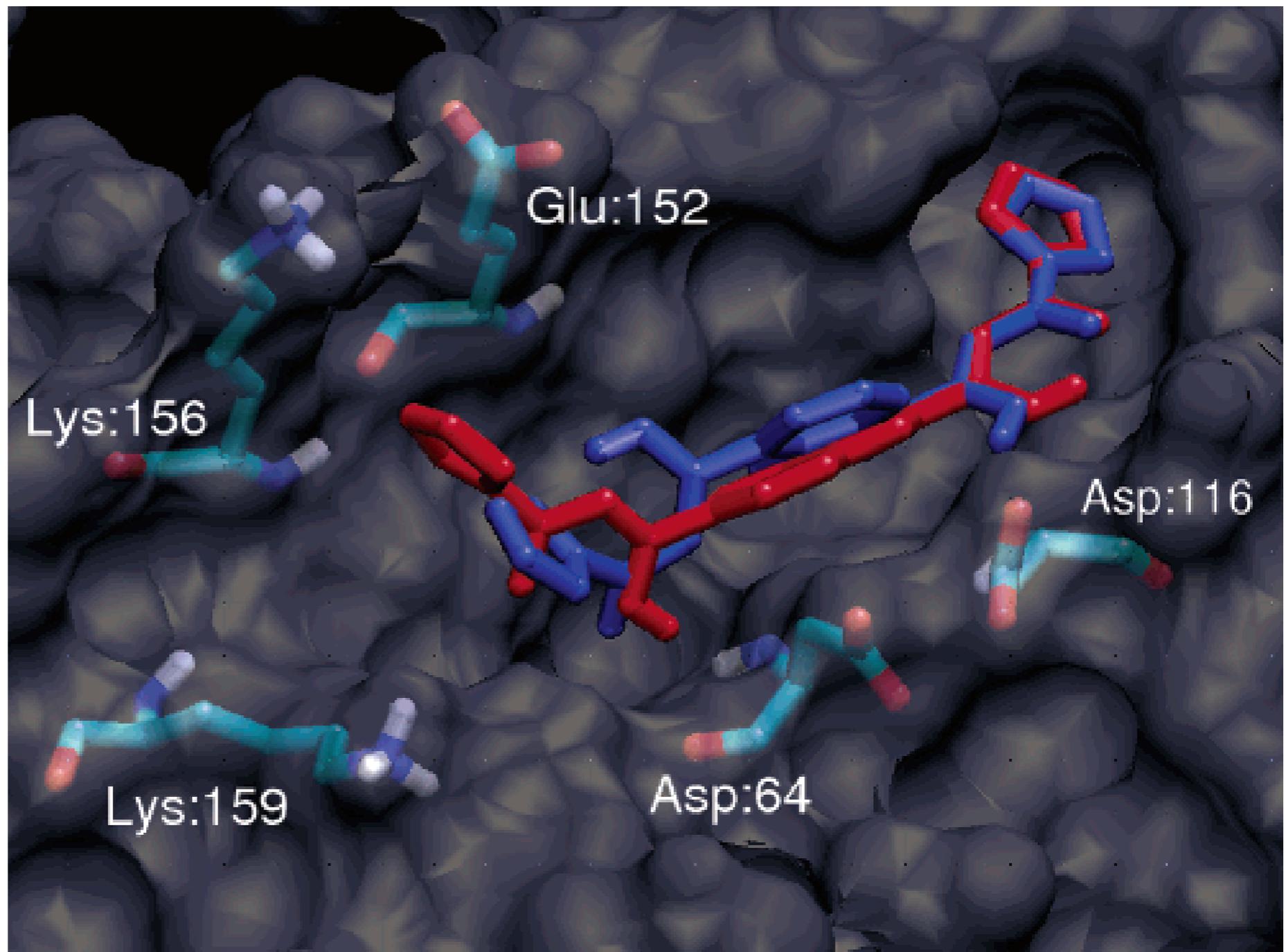
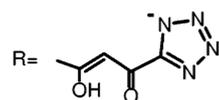
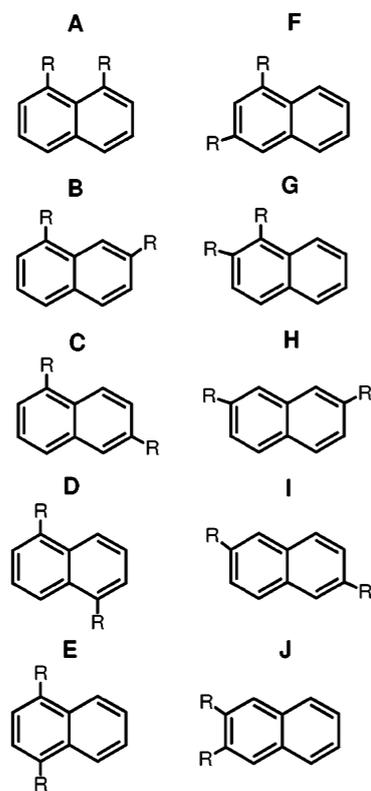
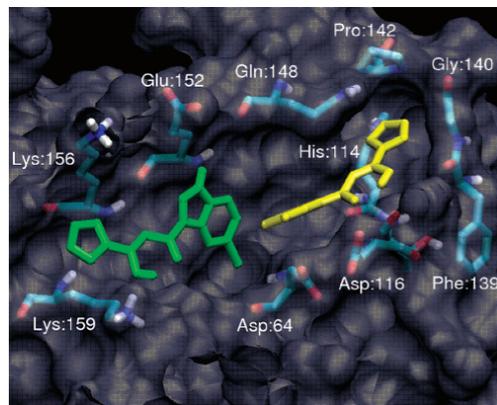


- › One structure known with 5CITEP
 - ◆ Not clear (low resolution)
 - ◆ Binding near to DNA interacting site
 - ◆ Loop near the binding
- › Docking + Molecular Dynamics
 - ◆ AMBER snapshots
 - ◆ AutoDock flexible torsion thetetrazolering and indole ring.

ISENTRESS example



ISENTRESS example



ISENTRESS example

Where patients come first  **MERCK** Patients & Caregivers | Healthcare Professionals | Worldwide

Quick Find Search

[HOME](#) | [ABOUT MERCK](#) | [PRODUCTS](#) | [NEWSROOM](#) | [INVESTOR RELATIONS](#) | [CAREERS](#) | [RESEARCH](#) | [LICENSING](#) | [THE MERCK MANUALS](#)

Newsroom

- Product News**
- Research & Development News
- Corporate News
- Financial News
- Corporate Responsibility News
- Fact Sheet
- Executive Speeches
- Webcasts
- VIOXX® (rofecoxib) Information Center**

Product News



FDA Approves ISENTRESS™ (raltegravir) Tablets, First-in-Class Oral HIV-1 Integrase Inhibitor

WHITEHOUSE STATION, N.J., Oct. 12, 2007 - Merck & Co., Inc., announced today that the U.S. Food and Drug Administration (FDA) granted ISENTRESS™ (raltegravir) tablets accelerated approval for use in combination with other antiretroviral agents for the treatment of HIV-1 infection in treatment-experienced adult patients who have evidence of viral replication and HIV-1 strains resistant to multiple antiretroviral agents.

This indication is based on analyses of plasma HIV-1 RNA levels up through 24 weeks in two controlled studies of ISENTRESS [pronounced i-sen-tris]. These studies were conducted in clinically advanced, three-class antiretroviral [nucleoside reverse transcriptase inhibitors (NRTIs), non-nucleoside reverse transcriptase inhibitors (NNRTIs) and protease inhibitors (PIs)] treatment-experienced adults. The use of other active agents with ISENTRESS is associated with a greater likelihood of treatment response. The safety and efficacy of ISENTRESS have not been established in treatment-naïve adult patients or pediatric patients. There are no study results demonstrating the effect of ISENTRESS on clinical progression of HIV-1 infection. Longer term data will be required before the FDA can consider traditional approval for ISENTRESS.

ABOUT ISENTRESS

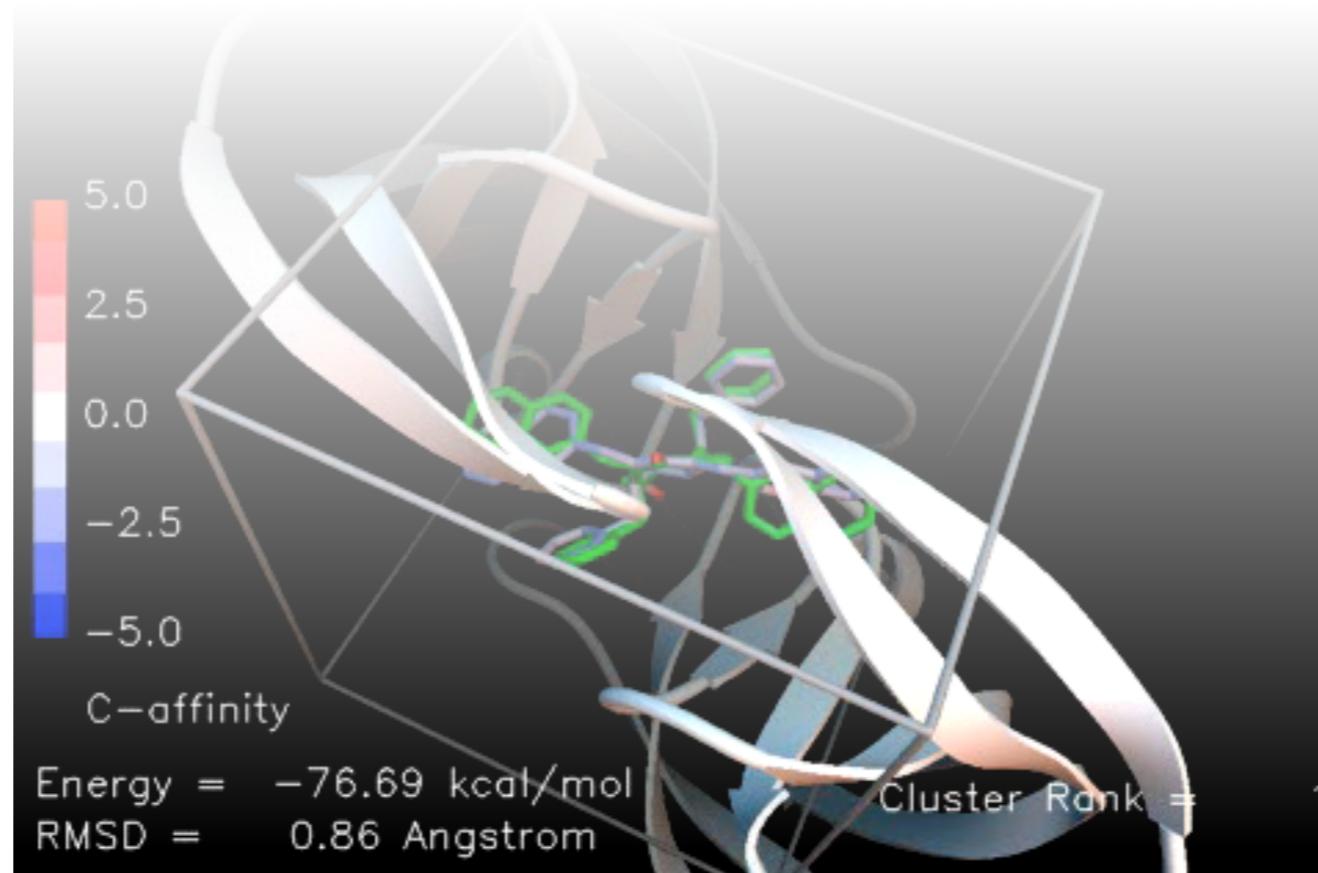
-  [Full Prescribing Information](#)
-  [Patient Product Information](#)

-  [Contact Newsroom](#)
-  [Podcast](#)
-  [RSS](#)

ISENTRESS®

data will be required before the FDA can consider traditional approval for effect of ISENTRESS on clinical progression of HIV-1 infection. Longer term patients or pediatric patients. There are no study results demonstrating the

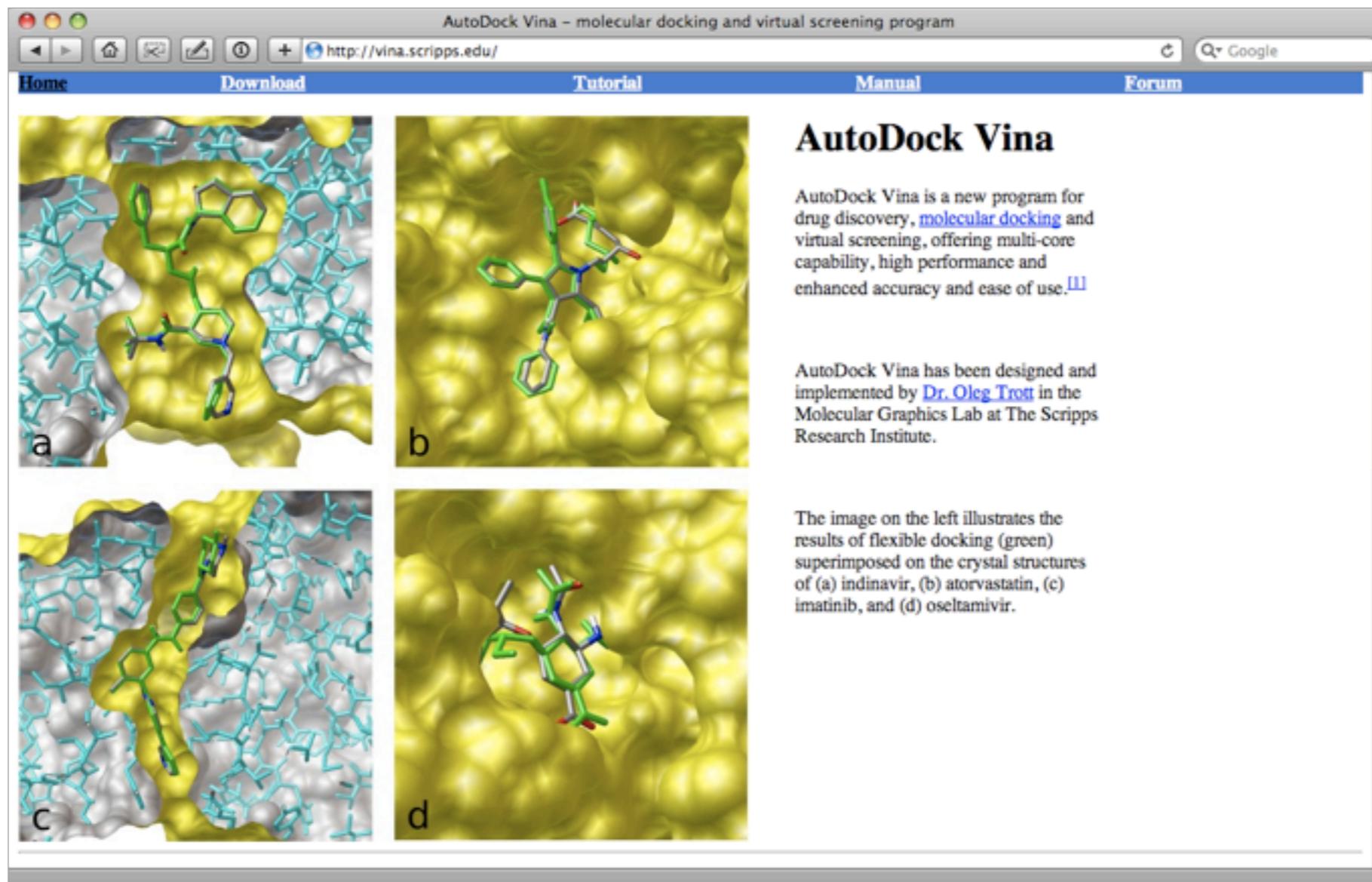
Vina 1.0



Goodsell, D. S. and Olson, A. J. (1990), Automated Docking of Substrates to Proteins by Simulated Annealing Proteins:Structure, Function and Genetics., 8: 195-202.
Morris, G. M., et al. (1996), Distributed automated docking of flexible ligands to proteins: Parallel applications of AutoDock 2.4 J. Computer-Aided Molecular Design, 10: 293-304.
Morris, G. M., et al. (1998), Automated Docking Using a Lamarckian Genetic Algorithm and an Empirical Binding Free Energy Function J. Computational Chemistry, 19: 1639-1662.
Huey, R., et al. (2007), A Semiempirical Free Energy Force Field with Charge-Based Desolvation J. Computational Chemistry, 28: 1145-1152.

Vina 1.0

Where to get help...



The screenshot shows a web browser window with the title "AutoDock Vina - molecular docking and virtual screening program". The address bar shows the URL "http://vina.scripps.edu/". The page has a navigation menu with "Home", "Download", "Tutorial", "Manual", and "Forum". The main content area features a 2x2 grid of images labeled a, b, c, and d. Each image shows a green molecular structure docked into a yellow and grey protein binding pocket. To the right of the images is the heading "AutoDock Vina" followed by a paragraph describing the program's capabilities and its developer, Dr. Oleg Trott. Below this is another paragraph explaining that the images illustrate flexible docking results for indinavir, atorvastatin, imatinib, and oseltamivir.

AutoDock Vina - molecular docking and virtual screening program

http://vina.scripps.edu/

Home Download Tutorial Manual Forum

AutoDock Vina

AutoDock Vina is a new program for drug discovery, [molecular docking](#) and virtual screening, offering multi-core capability, high performance and enhanced accuracy and ease of use. [\[1\]](#)

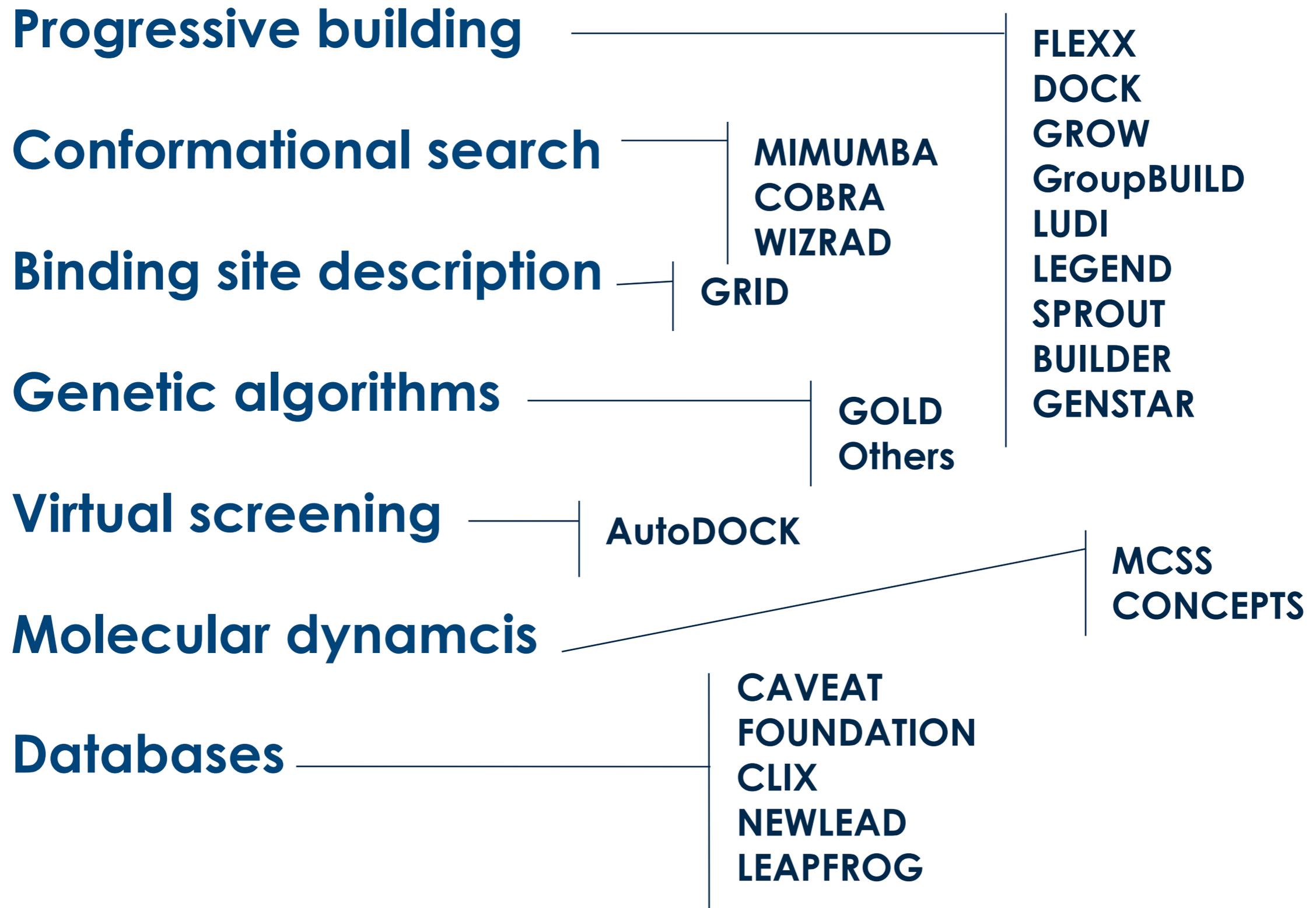
AutoDock Vina has been designed and implemented by [Dr. Oleg Trott](#) in the Molecular Graphics Lab at The Scripps Research Institute.

The image on the left illustrates the results of flexible docking (green) superimposed on the crystal structures of (a) indinavir, (b) atorvastatin, (c) imatinib, and (d) oseltamivir.

<http://vina.scripps.edu>

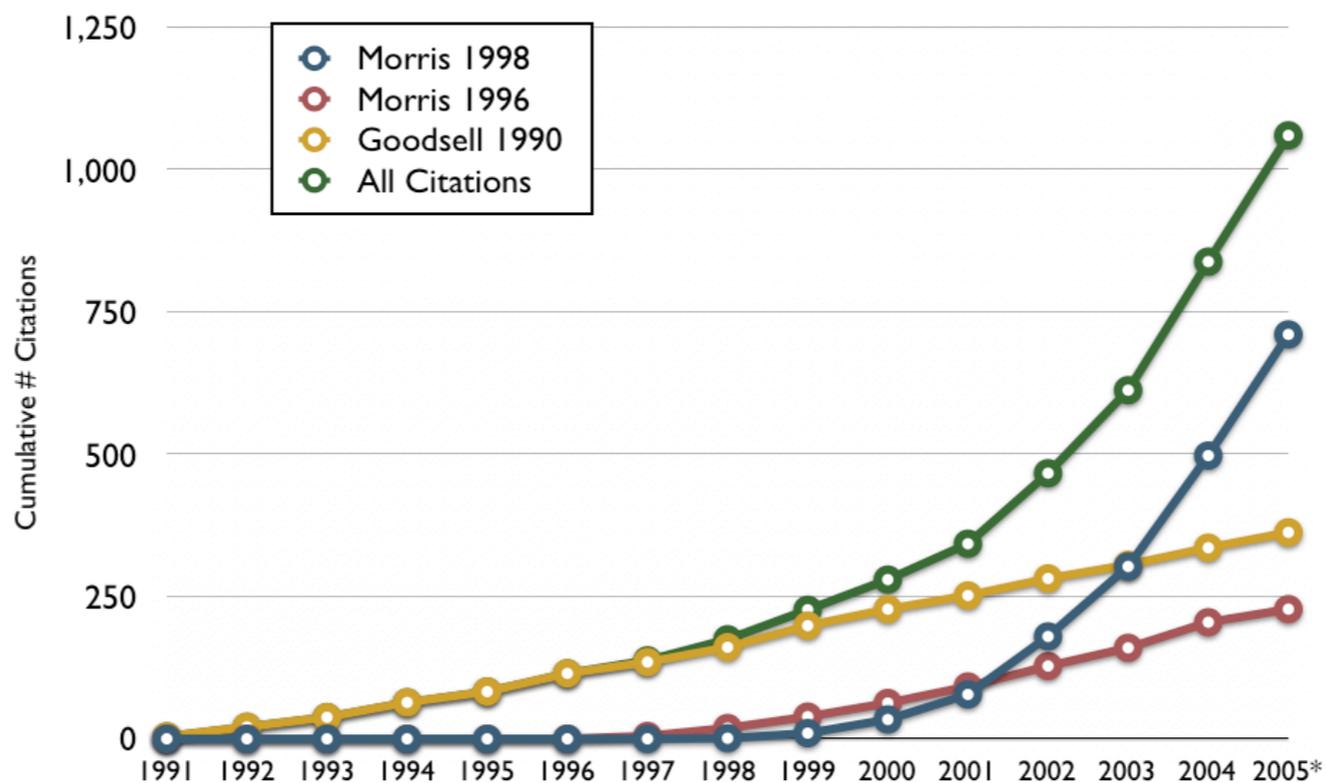
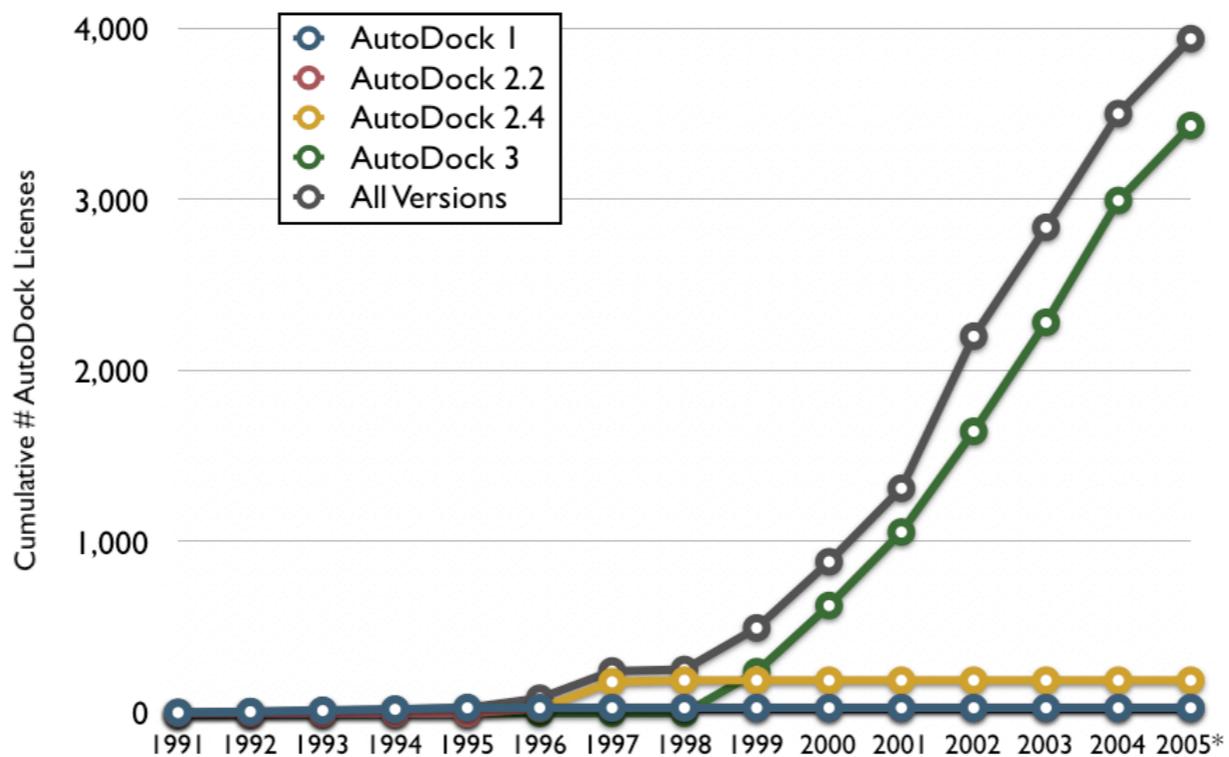
Vina 1.0

Alternatives



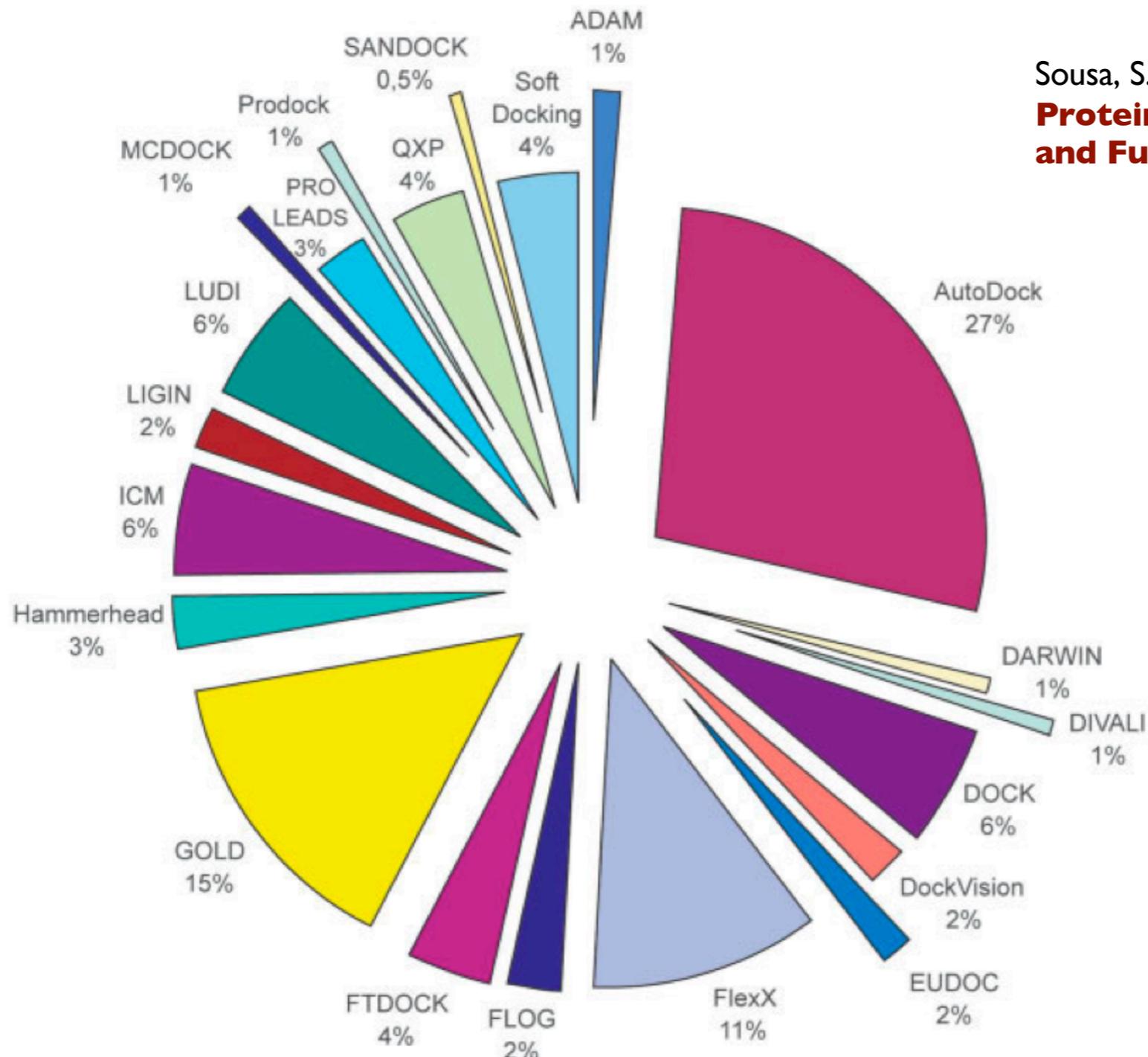
AutoDock 4.0

Why AutoDock over others



AutoDock 4.0

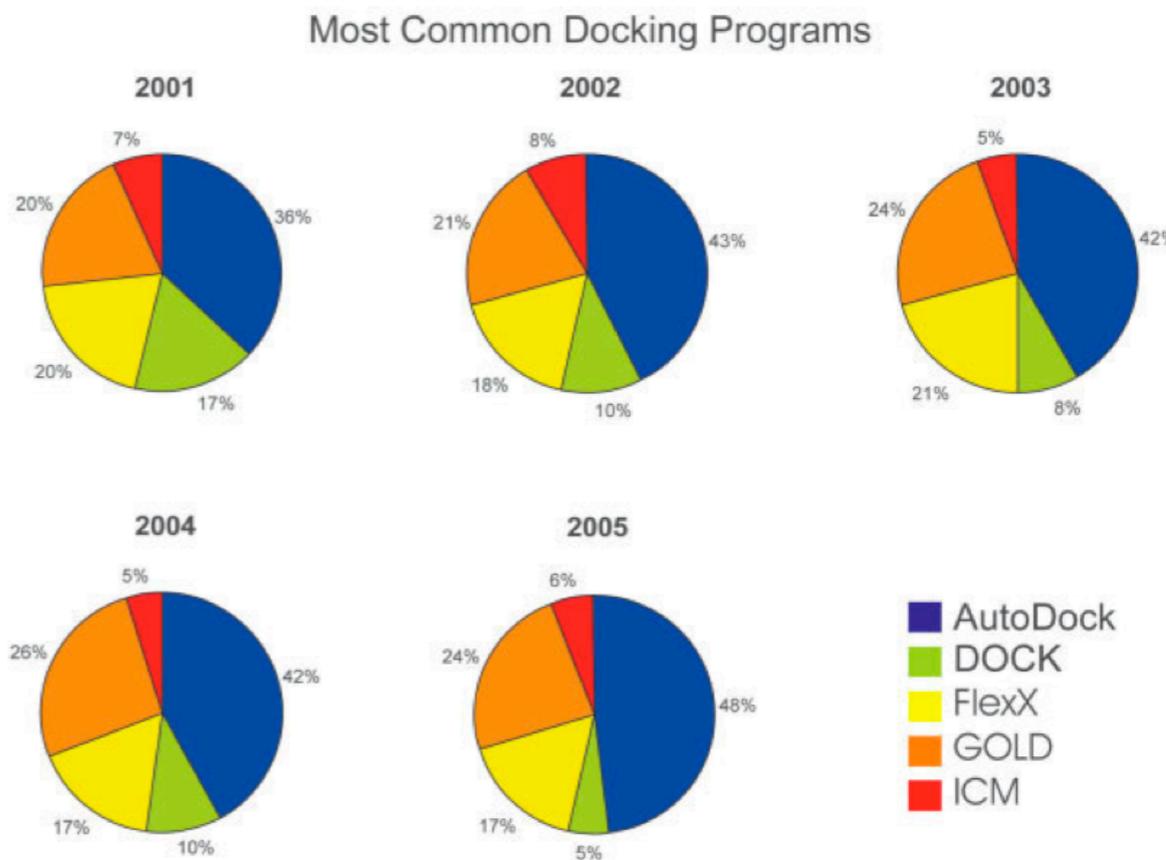
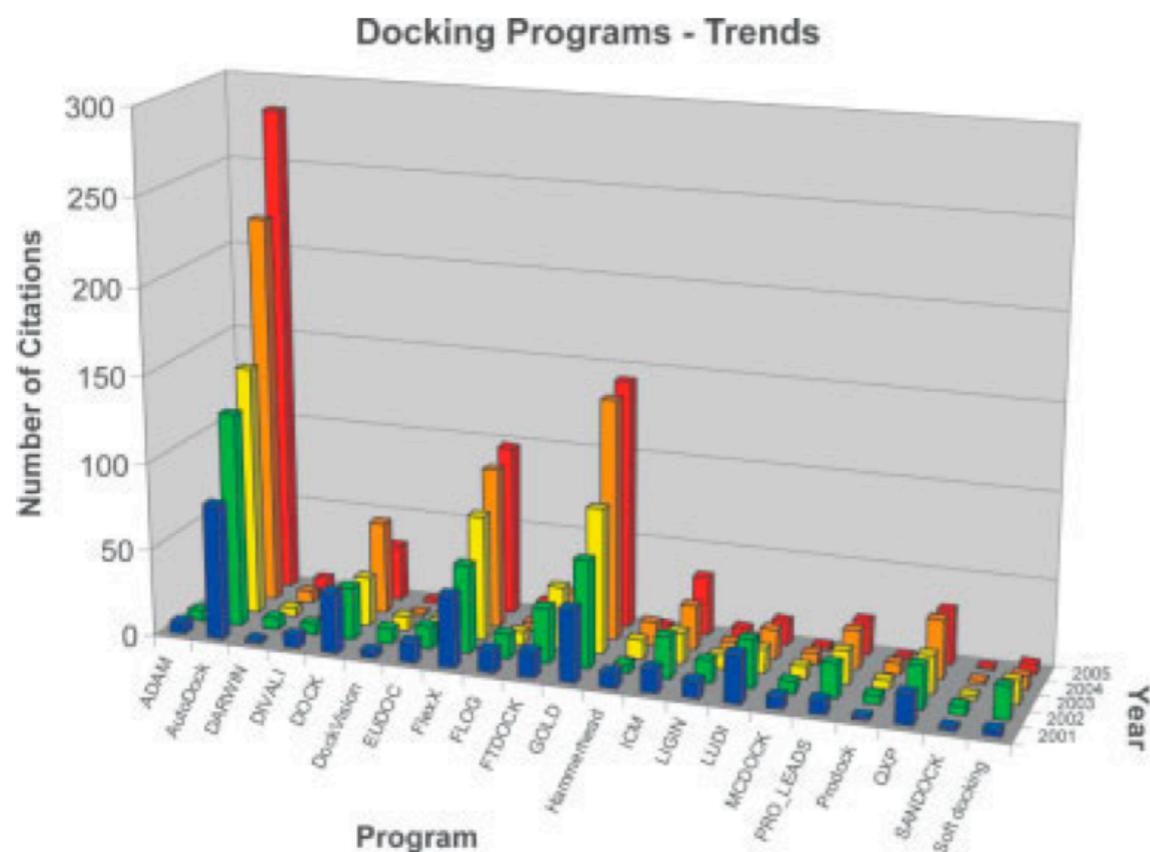
Why AutoDock over others



Sousa, S.F., Fernandes, P.A. & Ramos, M.J. (2006)
**Protein-Ligand Docking: Current Status
and Future Challenges** *Proteins*, **65**:15-26

AutoDock 4.0

Why AutoDock over others



Sousa, S.F., Fernandes, P.A. & Ramos, M.J. (2006)
**Protein-Ligand Docking: Current Status
 and Future Challenges** *Proteins*, **65**:15-26

Vina 1.0

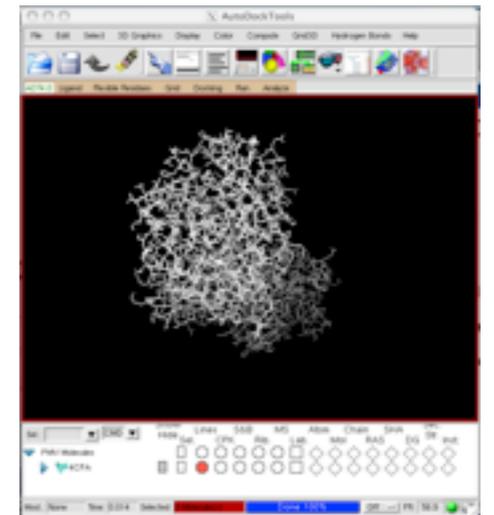
Vina and ADT

Vina

- ◆ 1990 (AutoDock)
- ◆ Number crunching (CPU expensive)
- ◆ Command-line!
- ◆ C & C++ compiled

AutoDock Tools

- ◆ 2000
- ◆ Visualizing set-up
- ◆ Graphical user interphase
- ◆ Python interpreter



AutoDock / Vina

Practical considerations

- * What problem does AutoDock solve?
 - * *Flexible* ligands (4.0 *flexible* protein).
- * What range of problems is feasible?
 - * Depends on the search method:
 - * **LGA** > **GA** >> **SA** >> **LS**
 - * **SA** : can output trajectories, $D <$ about 8 torsions.
 - * **LGA** : $D <$ about 8-32 torsions.
- * When is AutoDock not suitable?
 - * No 3D-structures are available;
 - * Modelled structure of poor quality;
 - * Too many (32 torsions, 2048 atoms, 22 atom types);
 - * Target protein too flexible.

AutoDock 4.0

Using AutoDock step-by-step

- * Set up ligand PDBQT—using ADT’s “Ligand” menu
- * *OPTIONAL*: Set up flexible receptor PDBQT—using ADT’s “Flexible Residues” menu
- * Set up macromolecule & grid maps—using ADT’s “Grid” menu
- * Pre-compute AutoGrid maps for all atom types in your set of ligands—using “autogrid4”
- * Perform dockings of ligand to target—using “autodock4”, and in parallel if possible.
- * Visualize AutoDock results—using ADT’s “Analyze” menu
- * Cluster dockings—using “analysis” DPF command in “autodock4” or ADT’s “Analyze” menu for parallel docking results.

AutoDock 4.0

Things to know before using AutoDock

Ligand:

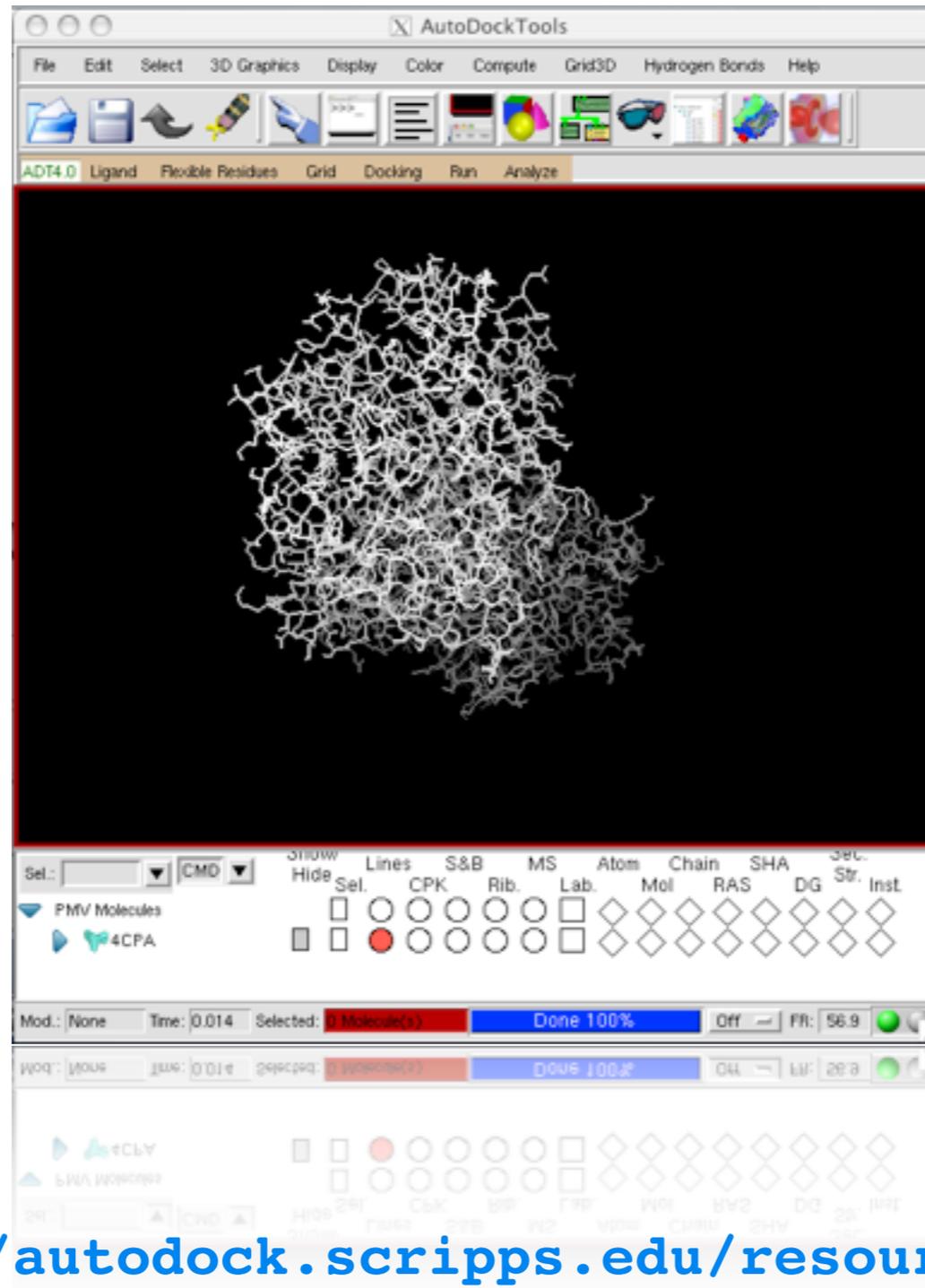
- * Add all hydrogens, compute Gasteiger charges, and merge non-polar H; also assign AutoDock 4 atom types
- * Ensure total charge corresponds to tautomeric state
- * Choose torsion tree root & rotatable bonds

Macromolecule:

- * Add all hydrogens, compute Gasteiger charges, and merge non-polar H; also assign AutoDock 4 atom types
- * Assign Stouten atomic solvation parameters
- * Optionally, create a flexible residues PDBQT in addition to the rigid PDBQT file
- * Compute AutoGrid maps

AutoDock 4.0

Good that we have AutoDock Tools (ATD)



<http://autodock.scripps.edu/resources/adt>

Vina 1.0

Good we have a nice tutorial

The screenshot shows a Windows desktop environment. The desktop background is blue. In the center, a window titled "AutoDockTools" is open, displaying two protein structures (one in green and yellow, the other in pink and yellow) and the name "Michel Sanner". Below the structures is a play button icon and the text "(c) 1999-2008 Molecular Graphics Laboratory, The Scripps Research Institute ALL RIGHTS RESERVED". At the bottom of the window, there is a progress bar labeled "Loading Modules... 31%". The desktop has several icons, including "Recycle Bin", "PMV-1.5.1", "Cygwin", "soft", "vina", "vina_tutorial...", "Adobe Reader 9", "Shortcut to MWED", "AutoDockTo...", "PyMOL", "Microsoft Visual Stu...", and "Notepad". The taskbar at the bottom shows the Start button, several open applications, and system tray icons.

<http://vina.scripps.edu/tutorial.html>

Acknowledgements

This presentation was based on
“Using AutoDock 4 with ADT. A tutorial”
by Dr. Ruth Huey and Dr. Garret M. Morris

