Informatics and Visualization Tools for Structural Genomics Research

Tom Ferrin, Ph.D.

Departments of Pharmaceutical Chemistry and Biopharmaceutical Sciences

Resource for Biocomputing, Visualization, and Informatics

University of California, San Francisco

Resource for Biocomputing, Visualization, and Informatics

The RBVI is a NIH/NCRR Biomedical Technology Research Center

We create innovative computational and visualization-based data analysis methods and algorithms, turns these into easy-to-use software tools, and apply these tools for solving a wide range of genomic and molecular recognition problems within the complex sequence → structure → function triad

Application areas

Gene characterization and interpretation

Drug design

Variation in drug response due to genetic factors

Protein engineering

Biomaterials design

Bioremediation

Prediction of protein function from sequence and structure

"It's sink or swim as a tidal wave of data approaches"

Petabyte (1,000 terabytes)

Exabyte (1,000 petabytes)

Zettabyte (1,000 exabytes)

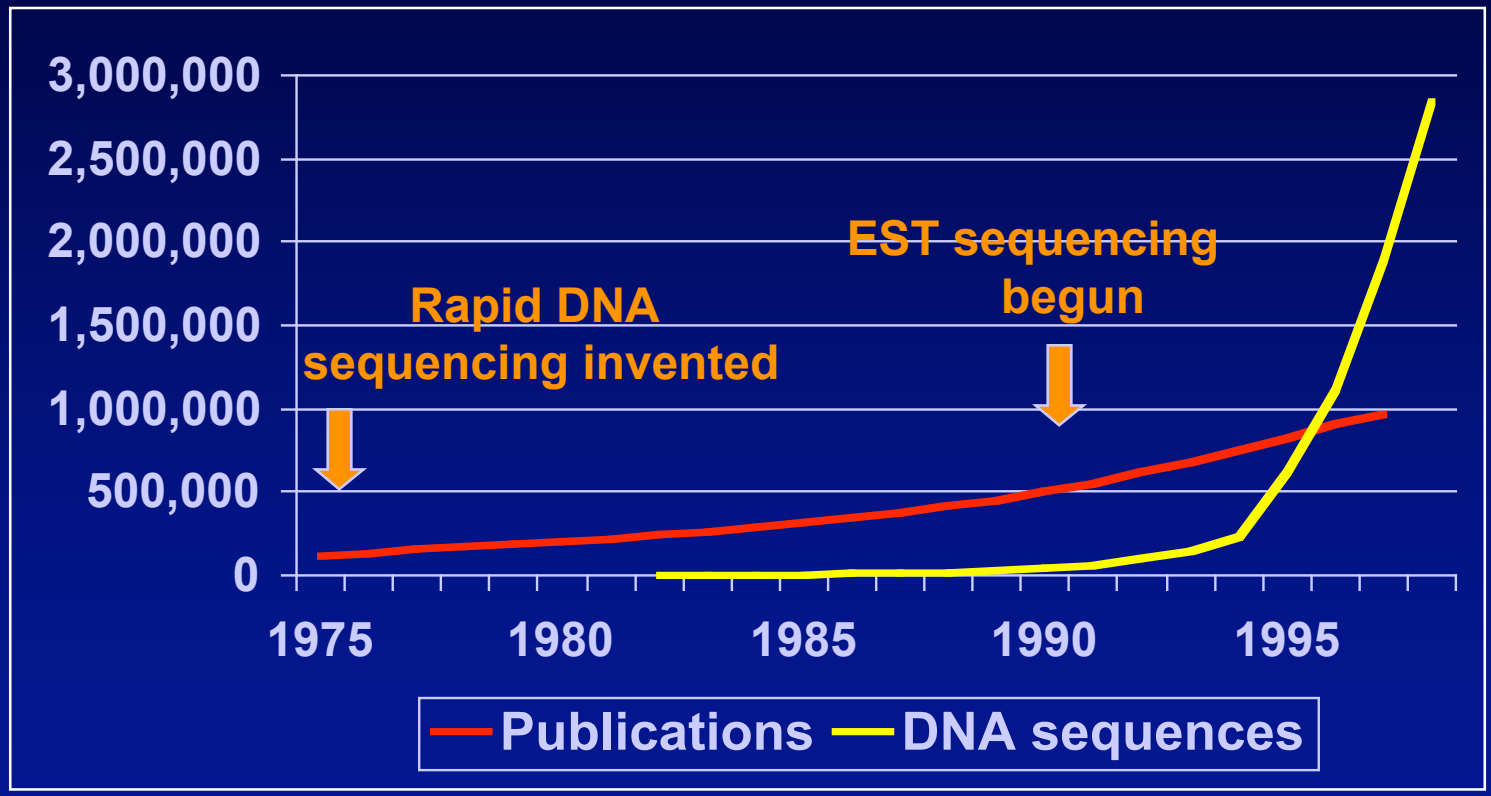
Yottabyte (1,000 zettabytes)

Tony Reichhardt
Nature 399:517-520 10 June 1999

"Many biologists are still in denial, never having faced the amount of information now pouring into databases such as Genbank and SwissProt… They haven't really thought about how they're going to use all this data..."

Ibid.

The Growing Gap in Functional Knowledge



Sample RVBI projects

- New methods for large-scale data collection, storage, analysis, and presentation for polymorphism (SNP) genotyping project

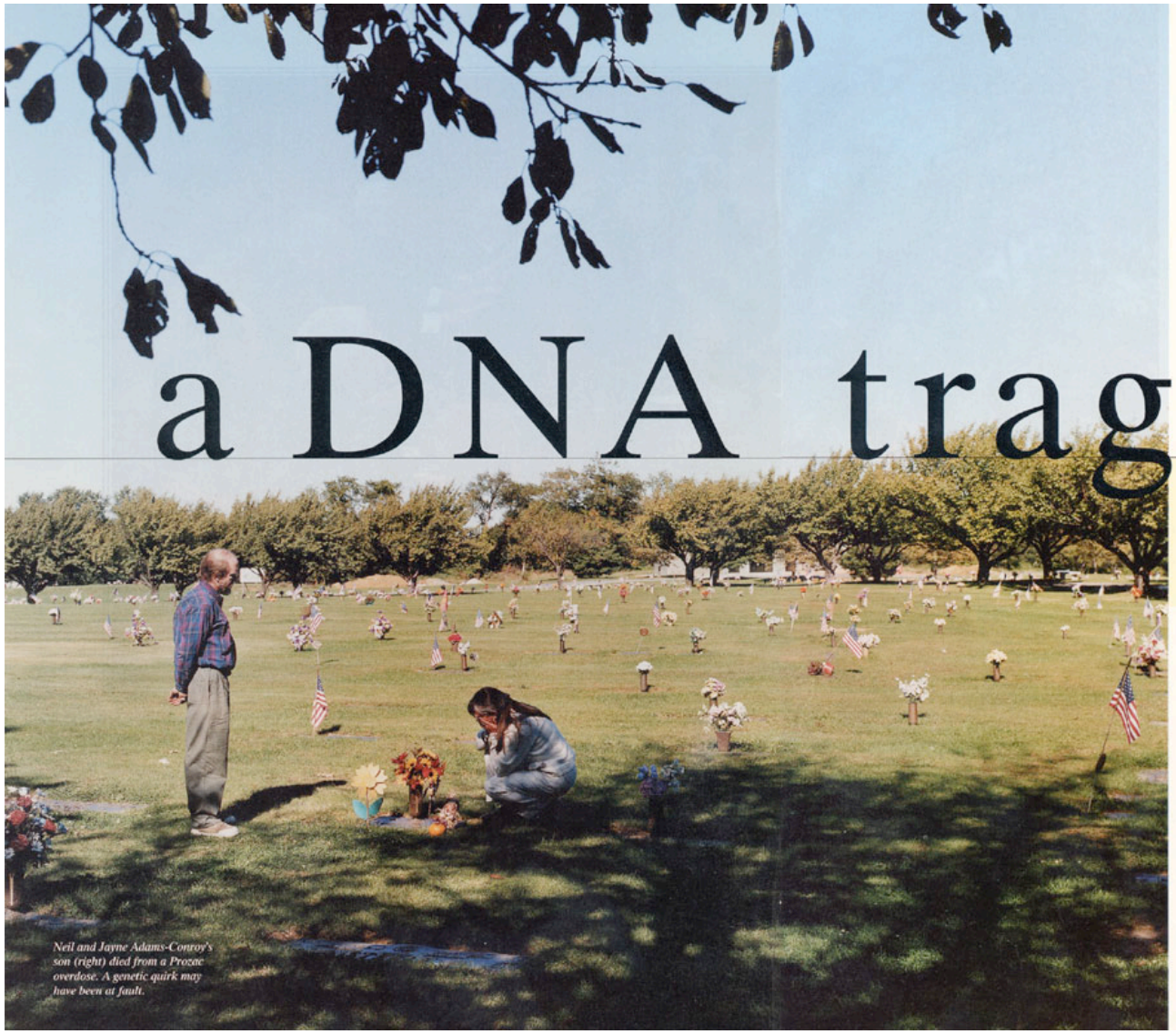
- Extensible visualization tools for comparative studies of protein sequence, structure, and function

a DNA tragedy

Genetic tests to prevent adverse drug reactions may save tens of thousands of lives a year, but for a troubled boy named Michael they came too late.

By David Stipp

Photographs by Suzanne Opton



Neil and Jayne Adams-Conroy's son (right) died from a Prozac overdose. A genetic quirk may have been at fault.

THE DEATH OF NINE-YEAR-OLD MICHAEL ADAMS-CONROY didn't seem at first like a signal event in medicine. It seemed like homicide.

Michael's short life was an uphill struggle from the start. Malnourished as an infant, he was taken from an abusive mother and placed in a temporary foster home before his first birthday. By the time he was 6, his medical record bulged with bad news: Michael was cognitively blunted and violently moody, and appeared to be afflicted with the brain damage of fetal alcohol syndrome, as well as with obsessive-compulsive disorder, tic-inducing Tourette's syndrome, and attention-deficit hyperactivity disorder.

Over the next few years he achieved a semblance of normalcy, thanks to the steadying hands of the resolutely affectionate couple who adopted him at age 3 and to daily doses of drugs to check his tics and obsessions. Small for his age, he took pride at finally being able to fling his coat up onto the grownups' pegs at his home in Martins Creek, Pa., a one-stoplight town two hours north of Philadelphia. He was learning to bowl in a league for handicapped kids and help his dad tend the garden.

Fortune - Oct 30, 2000

Case Report #1: Michael Adams-Conroy

Young child born to abusive mother, adopted at age 3, with signs of fetal alcohol syndrome, obsessive-compulsive disorder, Tourette's syndrome, and attention-deficit hyperactivity disorder. Prescribed Prozac to help control emotional outbursts.

Child dies suddenly; toxicology tests show massive overdose of Prozac. Adoptive parents investigated for homicide and their other two children put into protective custody.

Michael Adams-Conroy (continued)

Sharp-eyed psychiatrist notices unusually high levels of other metabolites in toxicology report, indicating child may have had an enzyme deficiency inhibiting Prozac from being metabolized normally.

Subsequent genetic testing showed child had defect in 2D6 gene which resulted in abnormal liver enzyme that metabolizes antidepressants.

Adoptive parents exonerated.

Case Report #2

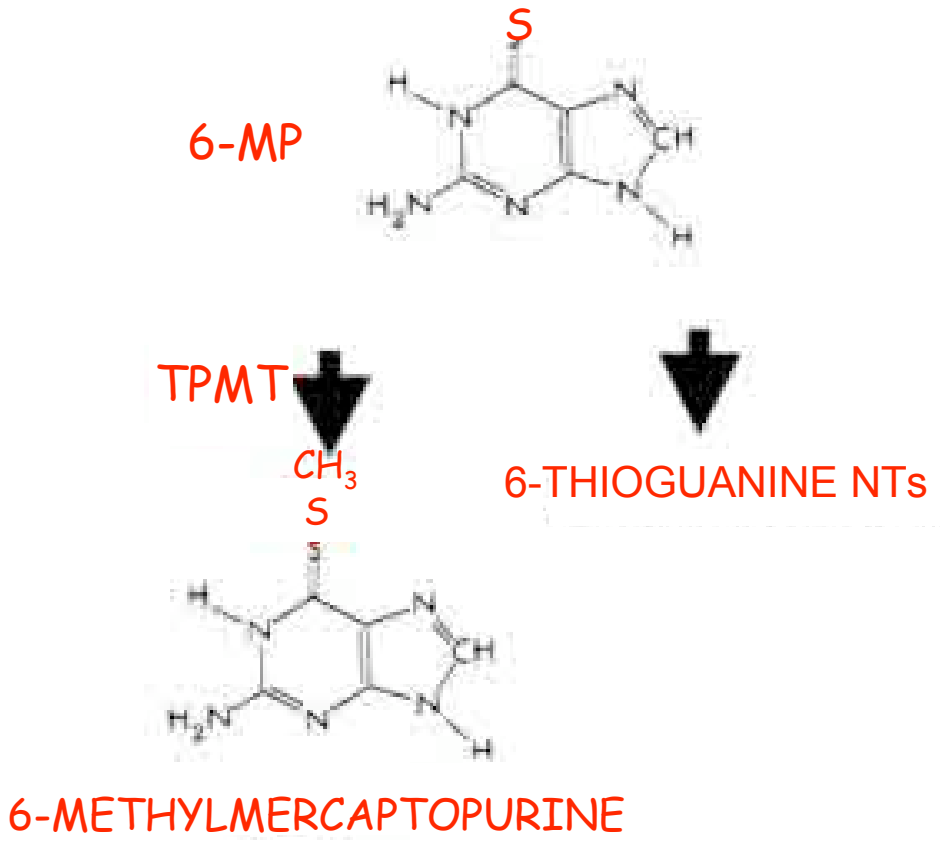
Patient: 3-year old boy

Diagnosis: Acute Lymphoblastic Leukemia (ALL)

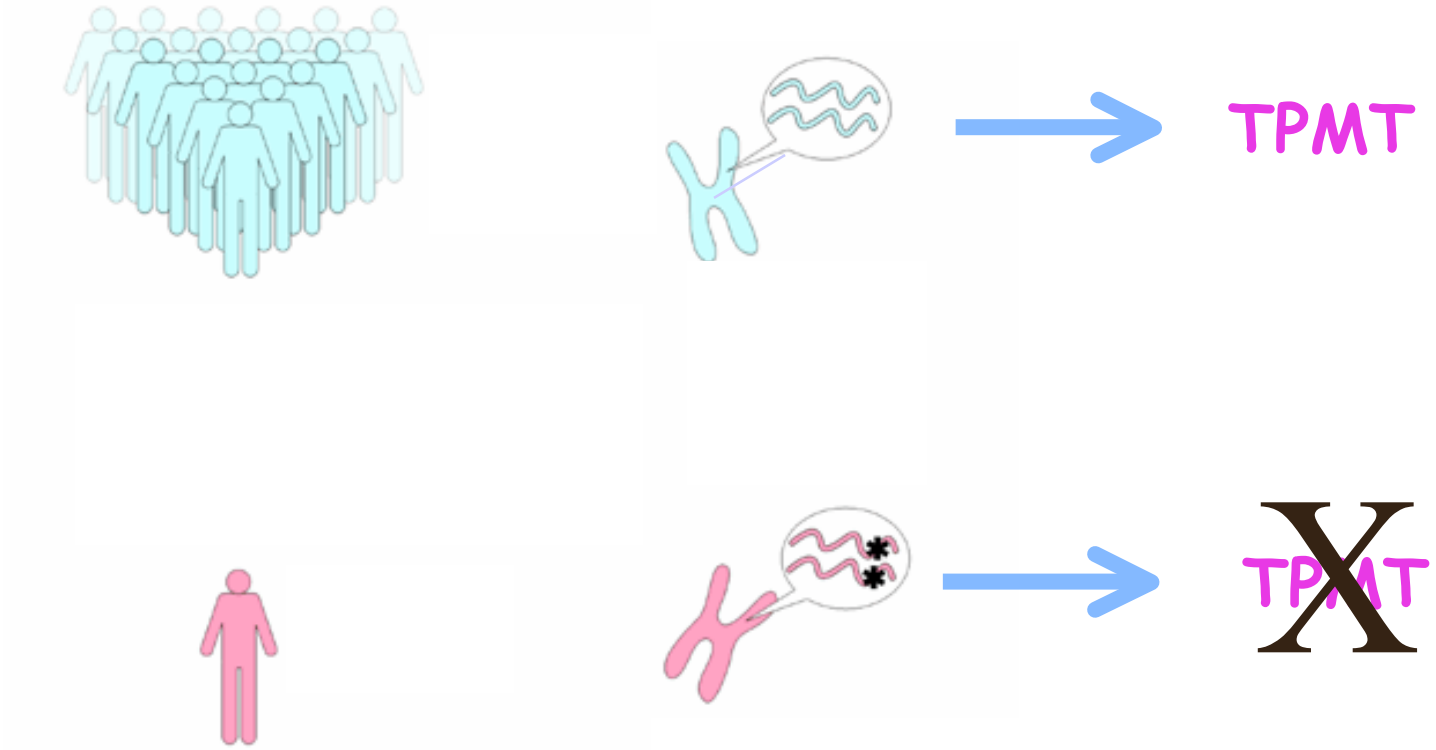
Standard therapy: 6-mercaptopurine (6-MP)

Result: Adverse Drug Reaction leading to acute bone marrow suppression

Normal Mechanism of Action

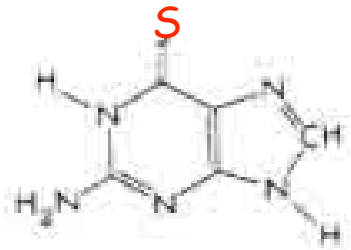


THIOPURINE METHYLTRANSFERASE (TPMT) GENES ARE DEFECTIVE IN 1:300 PEOPLE



This leads to elevated levels of Thioguanine Nucleotides

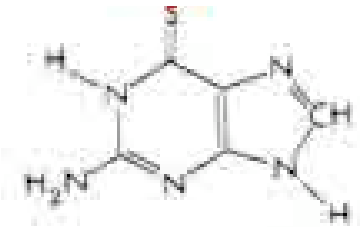
6-MP



TPMT

X

6-THIOGUANINE NTs ↑ ↑ ↑



6-METHYLMERCAPTOPURINE

PEOPLE DIFFER IN THEIR RESPONSE TO DRUGS

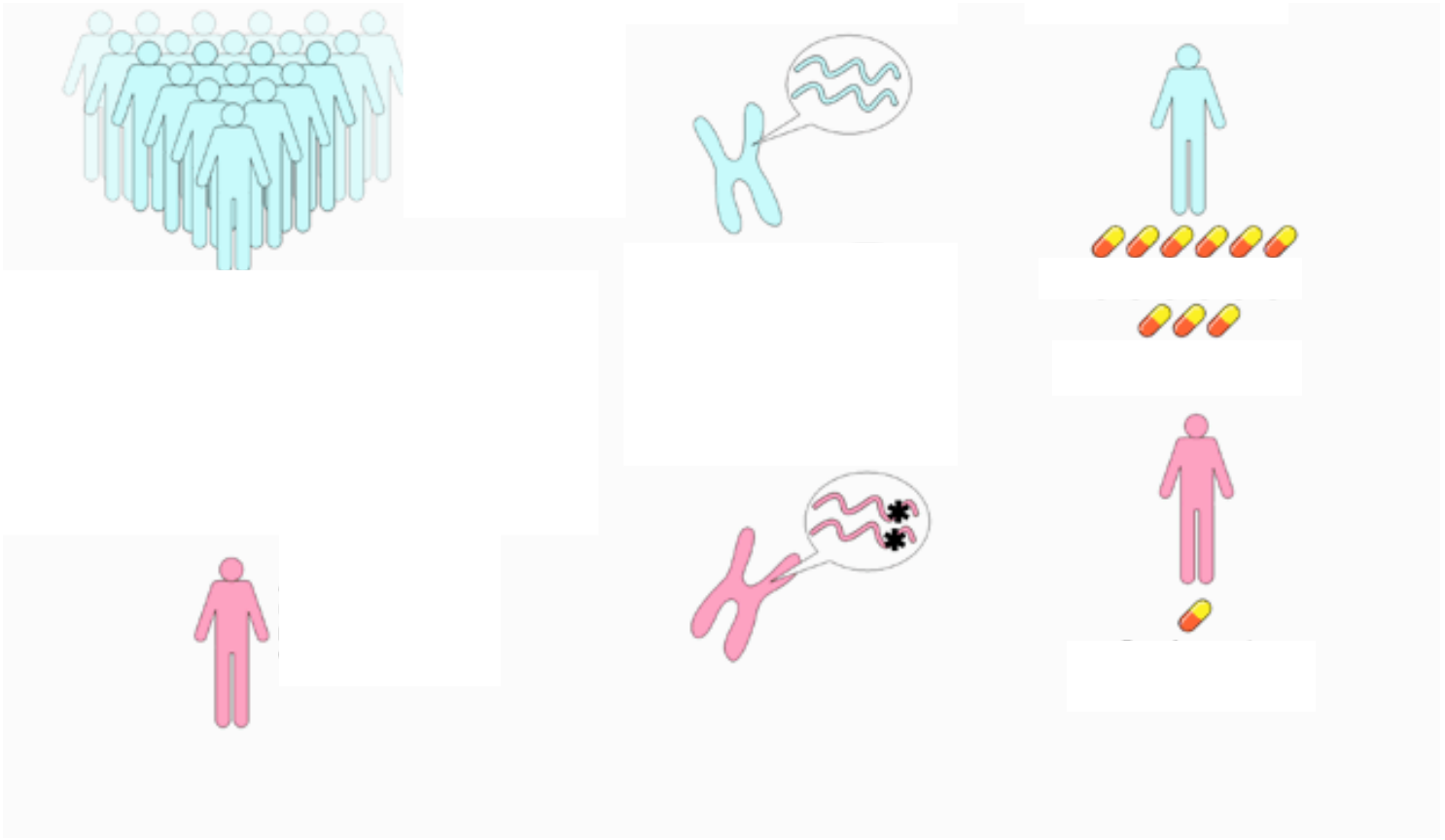
NO RESPONSE

THERAPEUTIC RESPONSE

ADVERSE DRUG REACTION (ADR)

TESTING FOR TPMT GENES IS NOW AVAILABLE

CHILDREN WITH DEFECTIVE TPMT GENES SHOULD RECEIVE A LOWER DOSE OF 6-MP



Adverse Drug Reactions

ADRs may kill 30,000 - 40,000 Americans each year and cause 2,200,000 serious nonfatal reactions. JAMA 1998 June 3;279(21):1684

Drugs with known genetically-linked potential for fatal adverse reactions (partial list):

Drug (Brand Name)	Perscribed For...	Adverse Reaction	Gene at Cause
Imipramine (Tofrannil)	Depression, ATD	Heartbeat irregularity	CYP2D6
Isoniazid (Laniazid)	Tuberculosis	Liver toxicity	NAT2
Warfarin (Coumadin)	Preventation of blood clots	Internal bleeding	CYP2C9
5-fluorouracil (Adrucil)	Cancer	Severe immune suppression	DPD
Clarithromycin (Biaxin)	Antibiotic	Heartbeat irregularity	KCNE2
Azathioprine (Imuran)	Rheumatoid arthritis	Severe immune suppression	TPMT

Pharmacogenetics of Membrane Transporters

$12-million, 4-year NIH grant

- Kathleen Giacomini and Ira Herskowitz, co-PIs, plus ~20 other UCSF researchers

Major Project Goal:

- Understand the genetic basis for variation in response to drugs which interact with membrane transporters. This class of proteins is of great pharmacological importance, as it provides the target for about 30% of the most commonly used prescription drugs and is a major determinant of the absorption, distribution and elimination of many others.

PMT project goals - continued

- Determine the amount of genetic variation (single-nucleotide polymorphisms) in at least 40 transporter genes by examining the DNA from an ethnically diverse sample of 250 people.
- Test the performance of these transporter variants in cell cultures and determine, through clinical phenotype studies, if people with those variants respond differently to drugs in a clinically significant way.
- Provide access to the data from these studies to the general scientific community through the World Wide Web to facilitate collaborative research and to speed development of new drug treatments.

The Corriel Cell Collection

African American (AA) - 100

Caucasian (CA) - 100

Asian American (AS) - 30

Mexican American (ME) - 10

Pacific Islander (PA) - 7

TOTAL - 247

Pharmacogenetics.ucsf.edu - Mozilla {Build ID: 2002031104}

File Edit View Search Go Bookmarks Tasks Help Debug QA

Back Forward Reload Stop http://pharmacogenetics.ucsf.edu/ Search Print

Home Bookmarks FaxReader Google Web Development

Pharmacogenetics.UCSF.edu

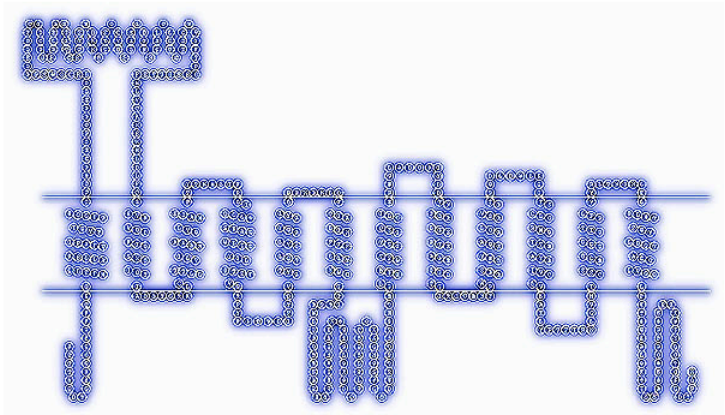
Contents

- home
- goals
- people
- institutions
- publications
- available data
- training
- relevant links
- contact us

pmt intranet
(password protected)

UCSF Pharmacogenetics of Membrane Transporters

The UCSF Pharmacogenetics of Membrane Transporters (PMT) Project is sponsored by the National Institutes of Health's National Institute of General Medical Sciences (grant U01 GM61390). The Project is part of the Pharmacogenetics Research Network and Knowledgebase. Information about the entire Network can be found at PharmGKB. Pharmacogenetics is the study of the genetic basis for variation from person to person in response to drugs. Membrane transporters play a major role in drug response in two ways. First, many drugs work by affecting function of transporters. Second, transporters determine the level of drugs within the body and thus determine whether drug levels are adequately high for therapeutic effect. The goal of the UCSF PMT Project is to understand the genetic basis for variation in drug response for drugs which interact with membrane transporters.



Model of organic cation transporter (OCT2)

Document: Done (0.23 secs)

PMT Intranet Website

Used by ~100 researchers at UCSF

Effective data analysis and display driven by iterative design/refinement cycle, successful because the bioinformatics team works closely with the molecular biologists

- Jill Mesirov, Whitehead: "Bioinformatics needs to be tightly integrated with the scientific research, not a service function"

Flexibility key!

- Multiple ways to display same data
- Simple download mechanism for scientists who want to load raw data into Excel spreadsheets

PMT Scientist-Users Are a Demanding Bunch…

OCT1 Transporter

Variant Results

PMT Project Investigator
Dr. Kathleen Giacomini

HGNC symbol
SLC22A1

Chromosome
6q26

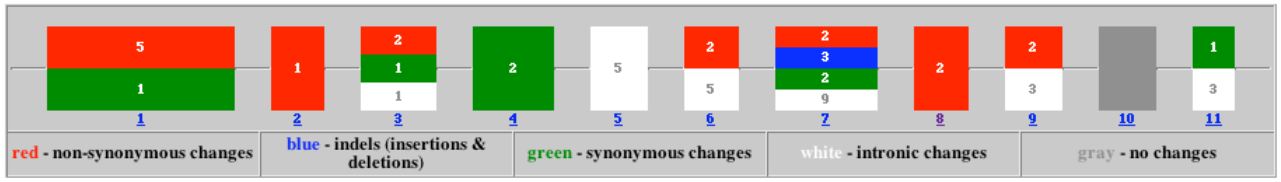
Aliases
ORGANIC CATION TRANSPORTER; OCT1
SOLUTE CARRIER FAMILY 22, MEMBER 1; SLC22A1

Background information
NCBI data
Omim data
LocusLink data

Reference Entry
U77086.1
Homo sapiens organic cation transporter 1 (hOCT1) mRNA, complete cds.

Exons
A schematic representation of the gene showing its exons with the exons scaled in size relative to one another. Variant results are indicated by color(s) of the exon block.

Click on number to view desired individual exon results.

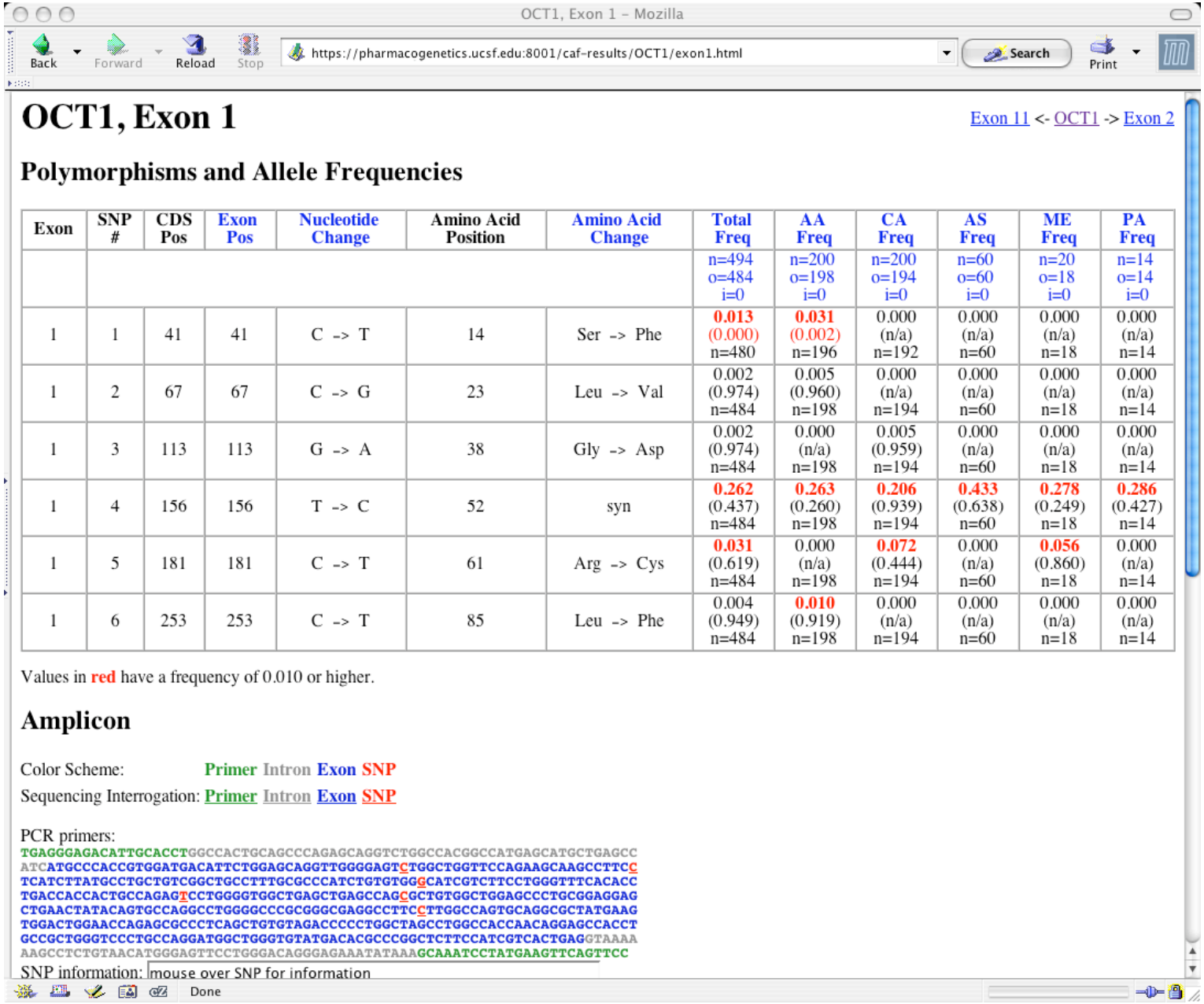


Amplicons
UCSC Blat search of all amplicons

Variants

Version 19 of data analysis results, generated on May 22, 2003.

Variant identification summary (tab-delimited text version)
Variant identification summary by ethnicity
Variant identification per-sample data (list) (tab-delimited text version)
Variant identification per-sample data (plot)



OCT1, Exon 1

Exon 11 <- OCT1 -> Exon 2

Polymorphisms and Allele Frequencies

Exon	SNP #	CDS Pos	Exon Pos	Nucleotide Change	Amino Acid Position	Amino Acid Change	Total Freq	AA Freq	CA Freq	AS Freq	ME Freq	PA Freq
							n=494 o=484 i=0	n=200 o=198 i=0	n=200 o=194 i=0	n=60 o=60 i=0	n=20 o=18 i=0	n=14 o=14 i=0
1	1	41	41	C -> T	14	Ser -> Phe	**0.013** (0.000) n=480	**0.031** (0.002) n=196	0.000 (n/a) n=192	0.000 (n/a) n=60	0.000 (n/a) n=18	0.000 (n/a) n=14
1	2	67	67	C -> G	23	Leu -> Val	0.002 (0.974) n=484	0.005 (0.960) n=198	0.000 (n/a) n=194	0.000 (n/a) n=60	0.000 (n/a) n=18	0.000 (n/a) n=14
1	3	113	113	G -> A	38	Gly -> Asp	0.002 (0.974) n=484	0.000 (n/a) n=198	0.005 (0.959) n=194	0.000 (n/a) n=60	0.000 (n/a) n=18	0.000 (n/a) n=14
1	4	156	156	T -> C	52	syn	**0.262** (0.437) n=484	**0.263** (0.260) n=198	**0.206** (0.939) n=194	**0.433** (0.638) n=60	**0.278** (0.249) n=18	**0.286** (0.427) n=14
1	5	181	181	C -> T	61	Arg -> Cys	**0.031** (0.619) n=484	0.000 (n/a) n=198	**0.072** (0.444) n=194	0.000 (n/a) n=60	**0.056** (0.860) n=18	0.000 (n/a) n=14
1	6	253	253	C -> T	85	Leu -> Phe	0.004 (0.949) n=484	**0.010** (0.919) n=198	0.000 (n/a) n=194	0.000 (n/a) n=60	0.000 (n/a) n=18	0.000 (n/a) n=14

Values in **red** have a frequency of 0.010 or higher.

Amplicon

Color Scheme: **Primer** Intron **Exon** **SNP**

Sequencing Interrogation: **Primer** Intron **Exon** **SNP**

PCR primers:

TGAGGGAGACATTGCACCTGGCCACTGCAGCCCAGAGCAGGTCTGGCCACGGCCATGAGCATGCTGAGCC
ATCATGCCCACCGTGGATGACATTCTGGAGCAGGTTGGGGAGTCTGGCTGGTTCCAGAAGCAAGCCTTCC
TCATCTTATGCCTGCTGTCGGCTGCCTTTGCGCCCATCTGTGTGGGCATCGTCTTCCTGGGTTTCACACC
TGACCACCACTGCCAGAGTCCTGGGGTGGCTGAGCTGAGCCAGCGCTGTGGCTGGAGCCCTGCGGAGGAG
CTGAACTATACAGTGCCAGGCCTGGGGCCCGCGGGCGAGGCCTTCCTTGGCCAGTGCAGGCGCTATGAAG
TGGACTGGAACCAGAGCGCCCTCAGCTGTGTAGACCCCCTGGCTAGCCTGGCCACCAACAGGAGCCACCT
GCCGCTGGGTCCCTGCCAGGATGGCTGGGTGTATGACACGCCCGGCTCTTCCATCGTCACTGAGGTAAAA
AAGCCTCTGTAACATGGGAGTTCCTGGGACAGGGAGAAATATAAAGCAAATCCTATGAAGTTCAGTTCC

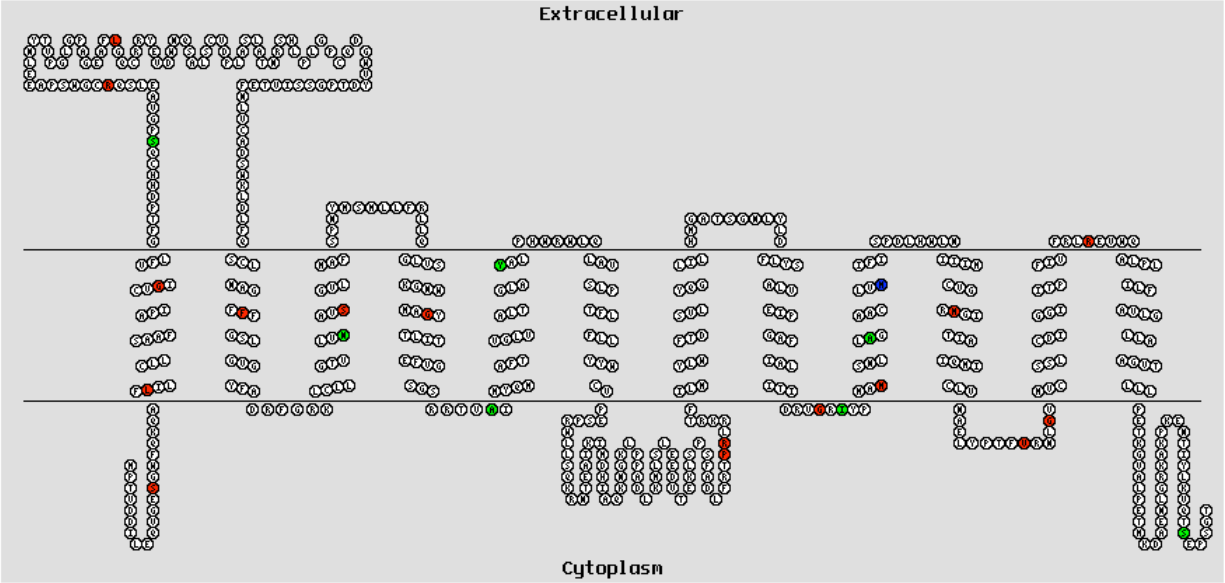
SNP information: mouse over SNP for information

Variants

Version 19 of data analysis results, generated on May 22, 2003.

Variant identification summary (tab-delimited text version)
Variant identification summary by ethnicity
Variant identification per-sample data (list) (tab-delimited text version)
Variant identification per-sample data (plot)
Population genetics statistics (tab-delimited text version)
Consensus sequences with mammalian species

Transmembrane prediction for OCT1 protein. Non-synonymous amino acid changes shown in **red,** indels (insertions and deletions) in **blue,** and synonymous changes in **green.**

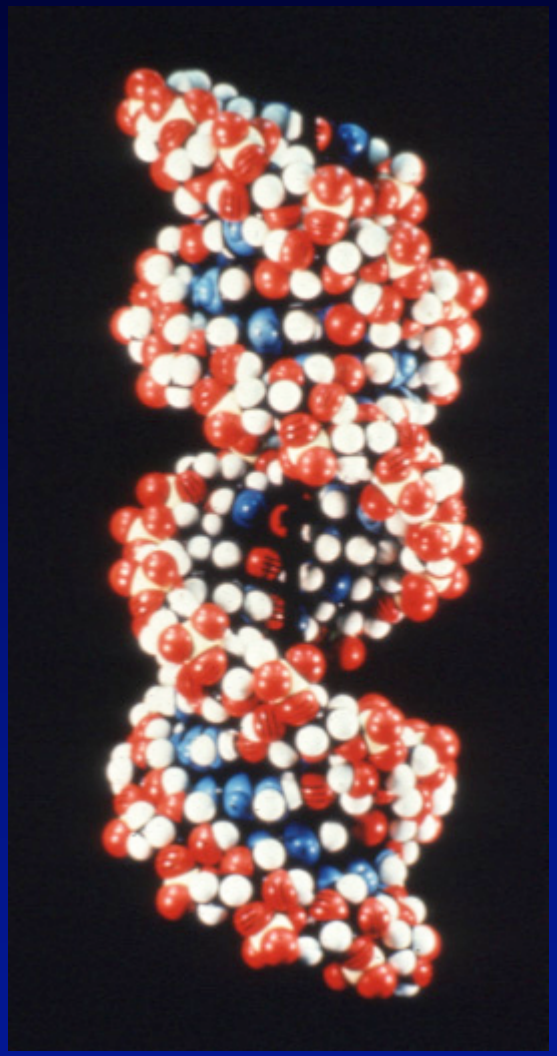


Cellular Phenotyping Results

Clinical Studies

Influence of hOCT1 Genotypes on 123I-MIBG Distribution to the Liver
Effect of Genetic Variation in the Liver Transporter, OCT1, on Response to Metformin in Healthy Subjects

You've now seen DNA and AA sequences



What about structure?

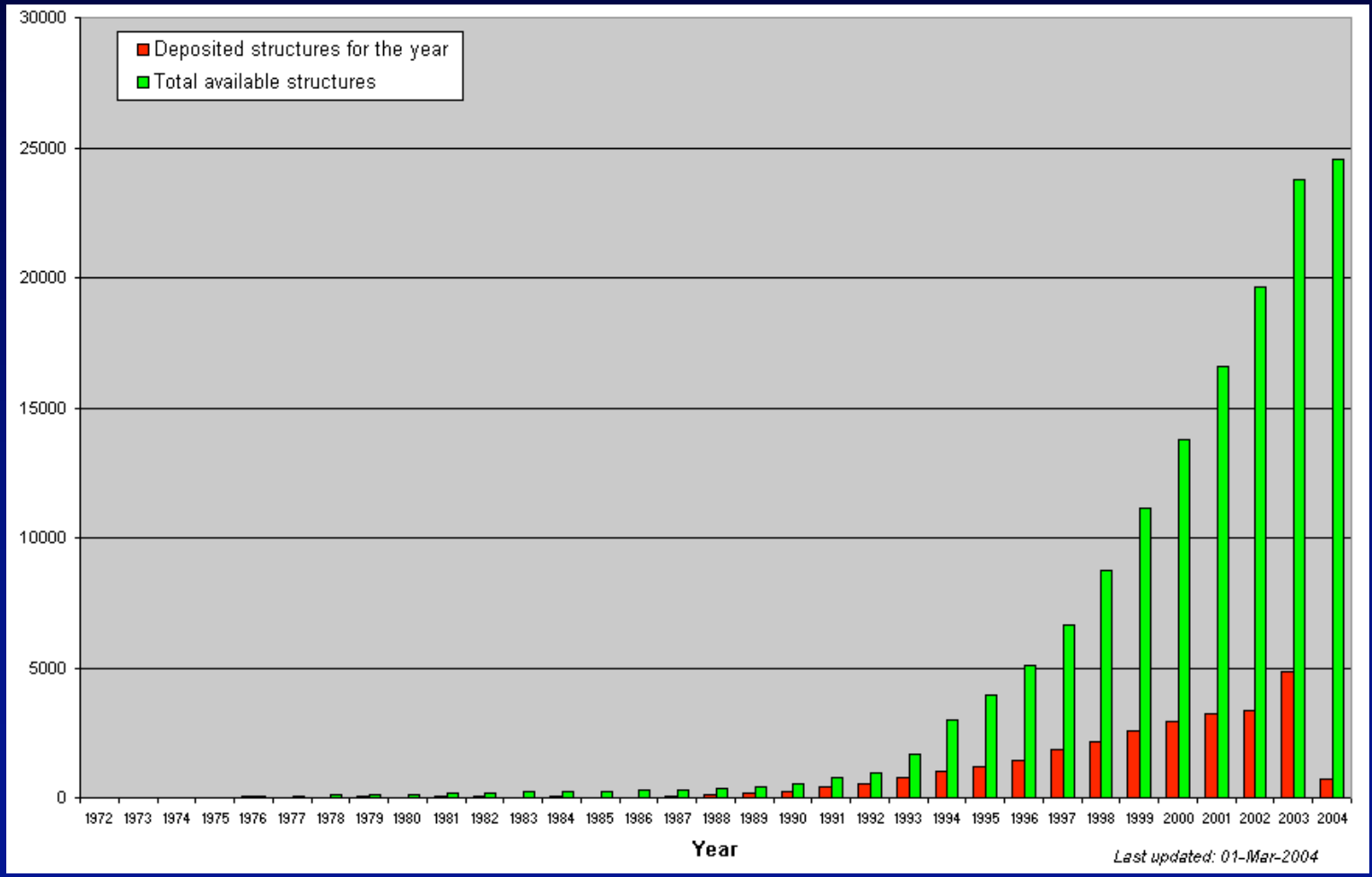
Why is Structure Important?

Sequence → Structure → Function

- Current research areas:
  - Prediction of structure from sequence
  - Prediction of function from sequence and structure
  - Understanding evolutionary changes
  - Engineering proteins for specialized function

- Applications in pharmacogenomics …
  - Improvements in drug discovery and development process
  - Prediction of drug response
  - Avoidance of toxic side effects

Growth in Protein Structures



The Structural Genomics Initiatives

"The next step beyond the human genome project"

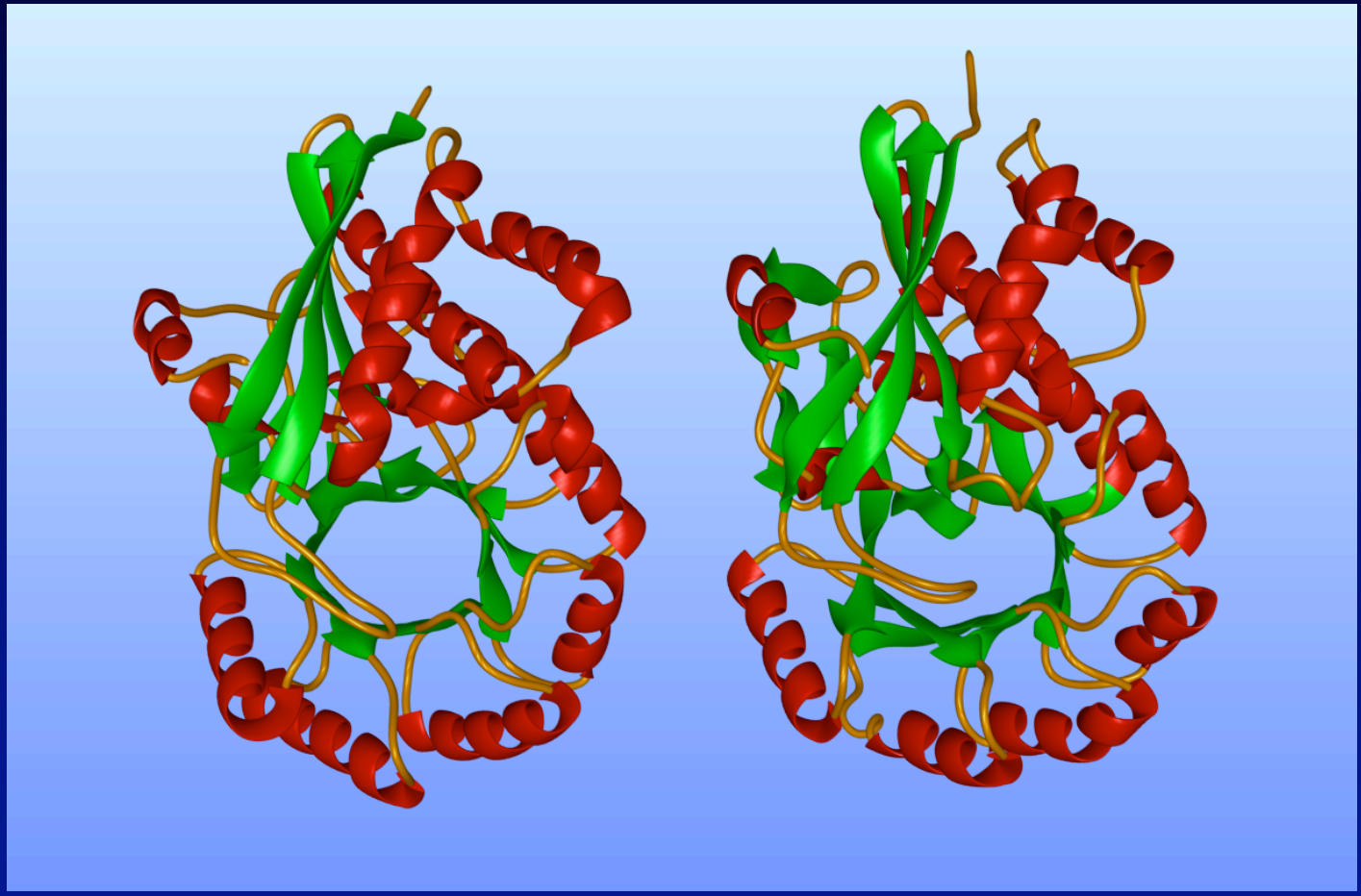
$150 million in NIH grants to establish 9 U.S. centers

- Goals:
  - Speed the determination of three-dimensional atomic-scale maps of proteins
  - 35,000 structures by 2005
  - Identify all proteins expressed in an organism - "proteomics"

Center	Lead Institution	Target
NY Struct. Genomics Res. Consortium	Rockefeller Univ.	Bacteria/yeast/human
Northeast Struct. Gen. Consort.	Rutgers Univ.	Roundworm/fly/human
Southeast Collab. for Struct. Gen.	Univ. of Georgia	Bacteria/roundworm/human
Berkeley Struct. Genomic Center	Lawrence Berkeley Lab.	Bacteria
Joint Ctr. for Struc. Genomics	Scripps Research Inst.	Roundworm/human
TB Struct. Genomics Consortium	Los Alamos Nat. Lab.	Tuberculosis
Midwest Ctr. for Struct. Genomics	Argonne National Lab.	Archaea/bacteria/eukarya
Ctr. for Eukaryotic Struct. Genomics	Univ. of Wisconsin	*Arabidopsys thaliana*
Struct. Gen. of Pathogenic Protozoa	Univ. of Washington	Protozoans

See http://www.nigms.nih.gov/funding/psi.html for additional information

Stereo pairs ?



Visualizing 3D Structure: The Chimera Molecular Modeling System

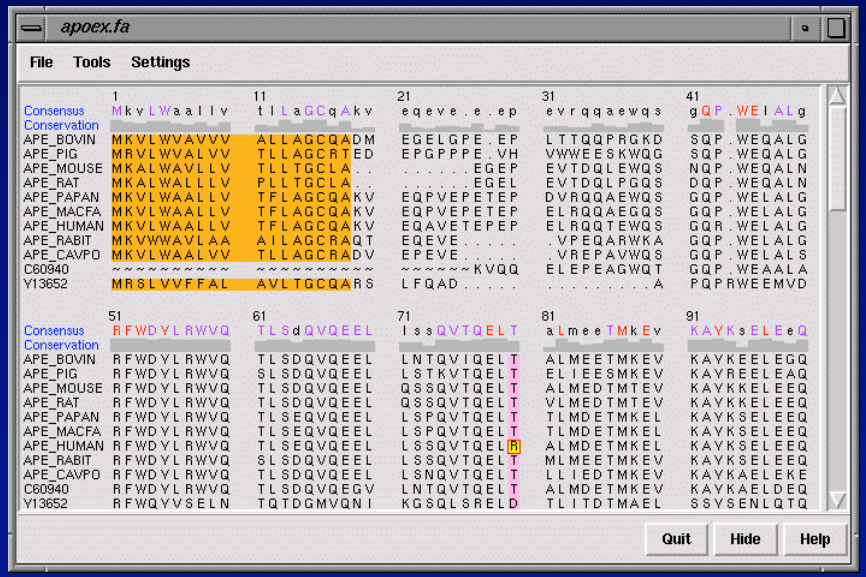
Chimera is an extensible interactive 3-D modeling system designed to allow developers to quickly incorporate novel algorithms and analysis tools

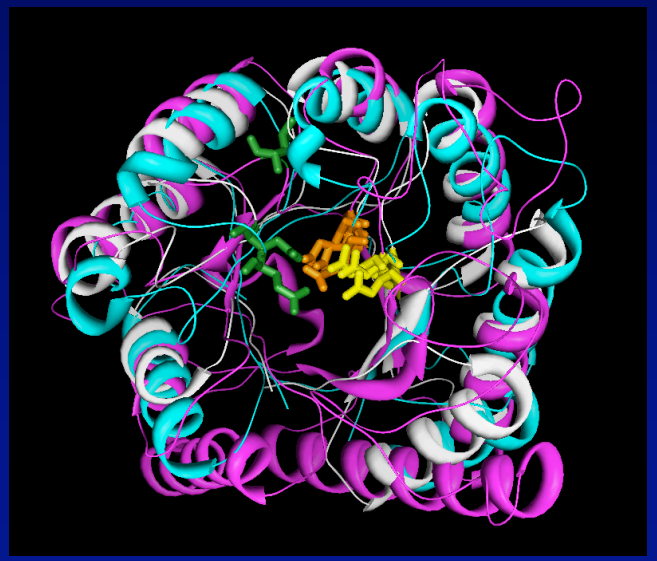
- ~30 extensions written to date
- Extensions are written in the Python programming language
  - Easy to learn, even for novice programmers
  - Offers object-oriented language features
- Extensions can control standard user interface features (e.g. camera, help, menus, toolbar) as well as their own custom interfaces

Sample Chimera Extension

Multalign Viewer

- simultaneous display of protein sequence and structure





Chimera Demo

Tools for Comparative Protein Studies

MinRMS - exhaustive search for all plausible structural alignments of two proteins

AlignPlot - interactive exploration of structural alignments

MultAlign Viewer - integrates sequence and structure space

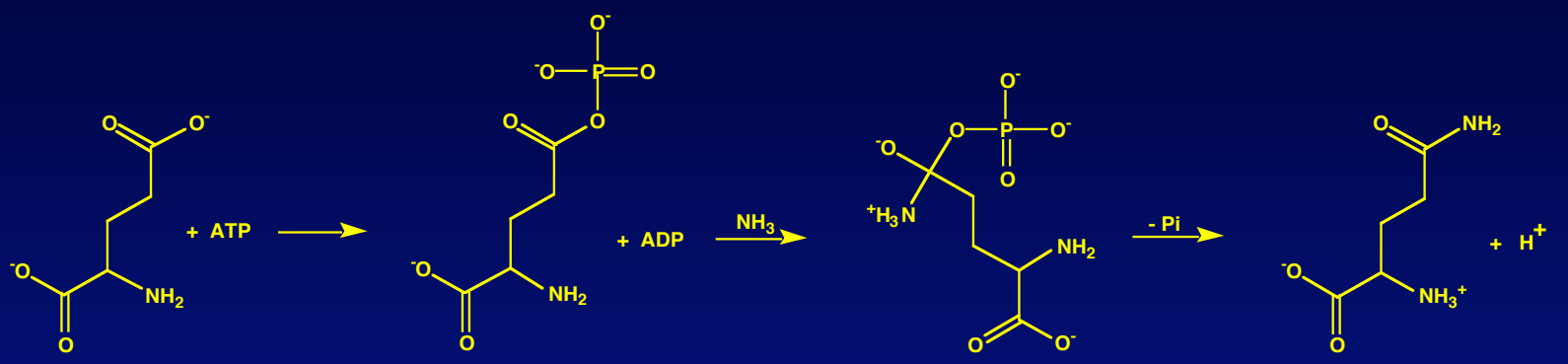
Chimera - extensible 3-D molecular modeling system

Example Study

Structural comparison of glutamine synthetase (GS) and creatine kinase (CK)

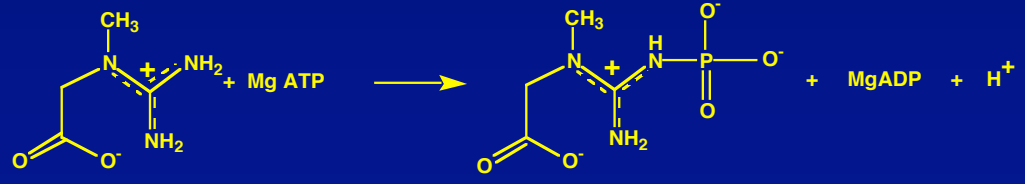
- GS: 468 residues, PDB entry 2gls
- CK: 380 residues, PDB entry 1crk
- No significant sequence similarity, both have multimeric forms, proposed similar tertiary structures, and catalyze similar reactions

GS and CK catalysis



Glutamate

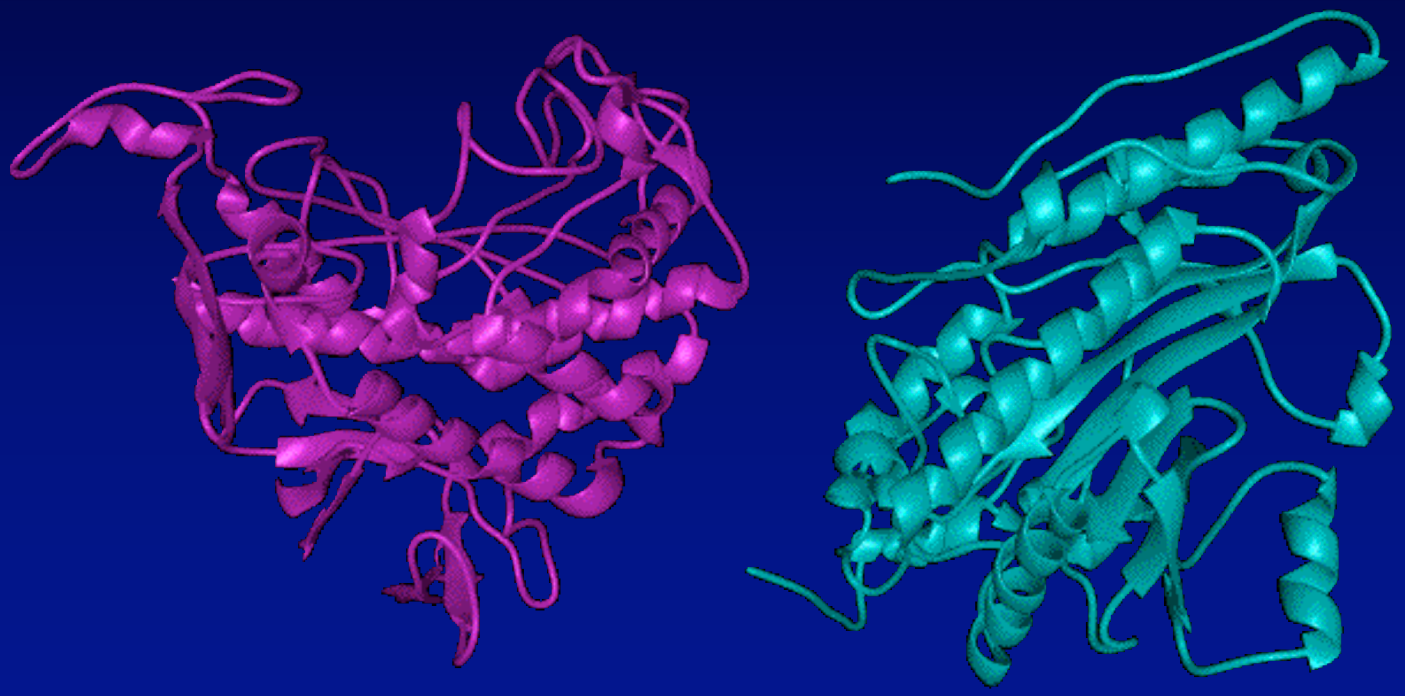
Glutamine



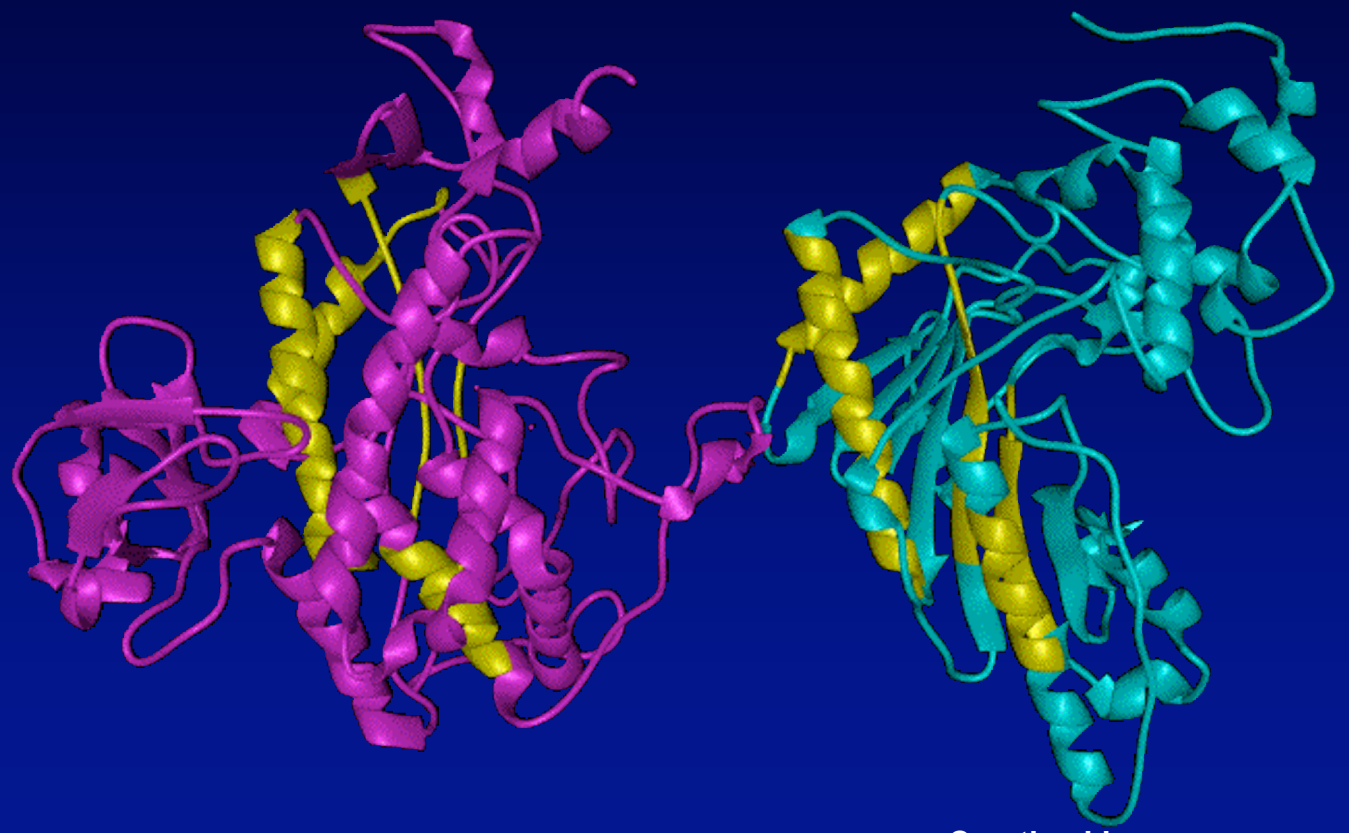
Creatine

Phosphocreatine

Glutamine synthetase and creatine kinase

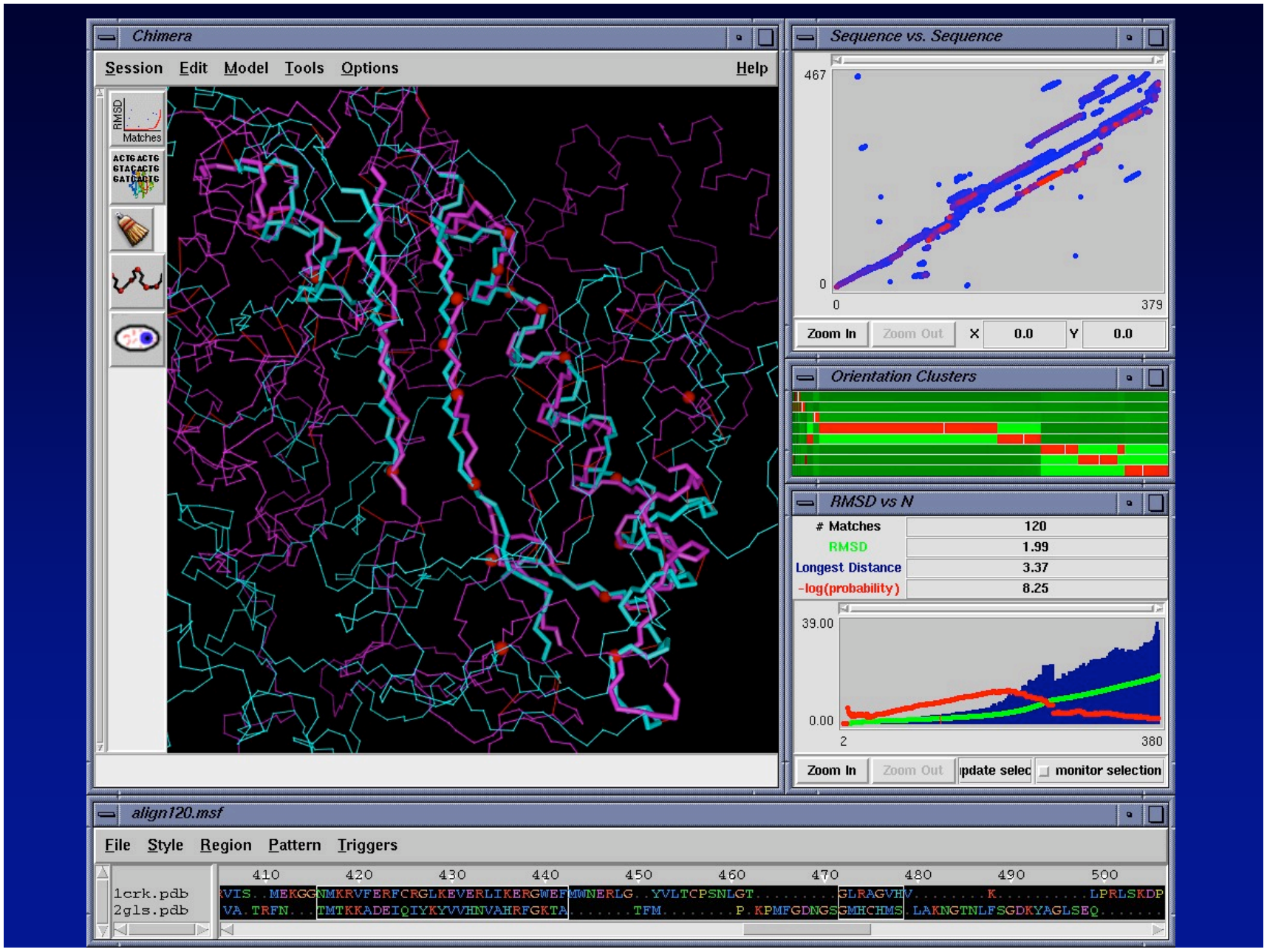


After MinRMS alignment

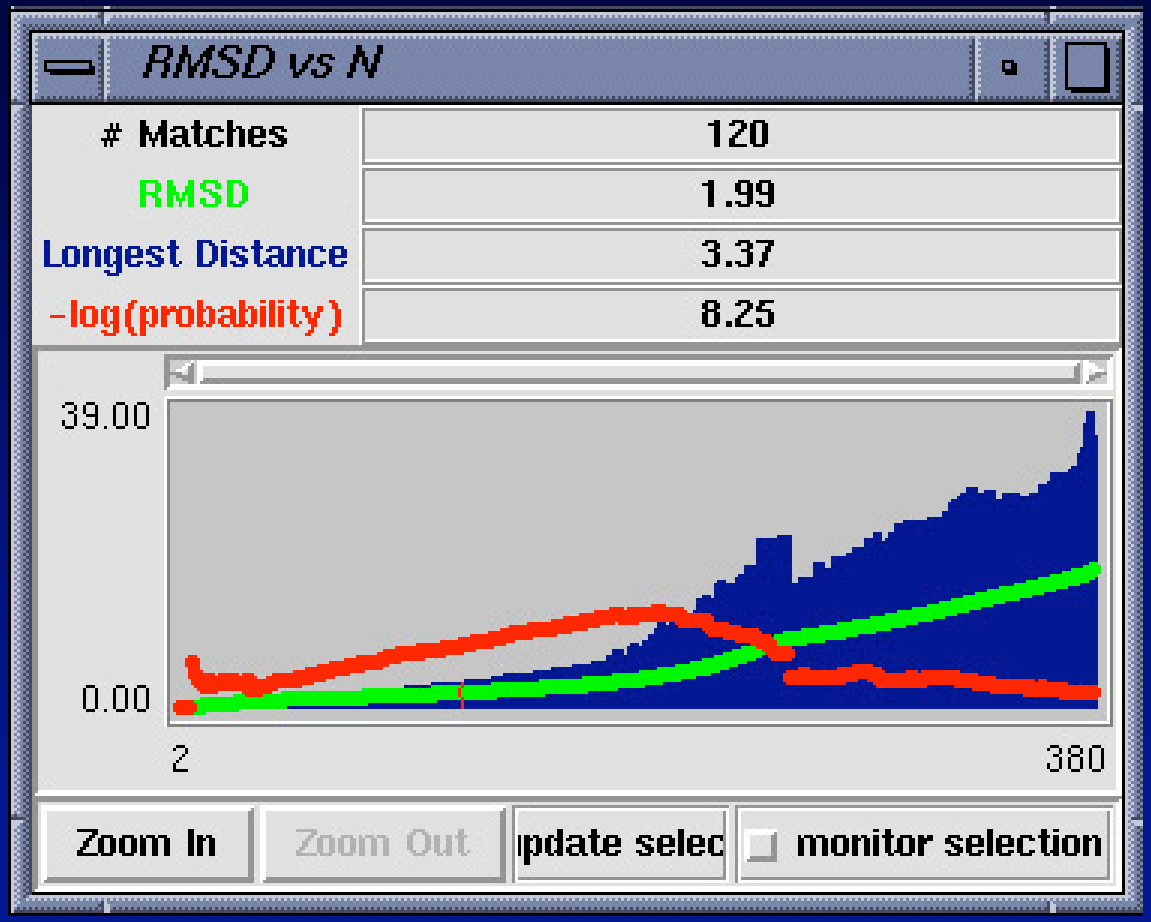


Glutamine synthetase

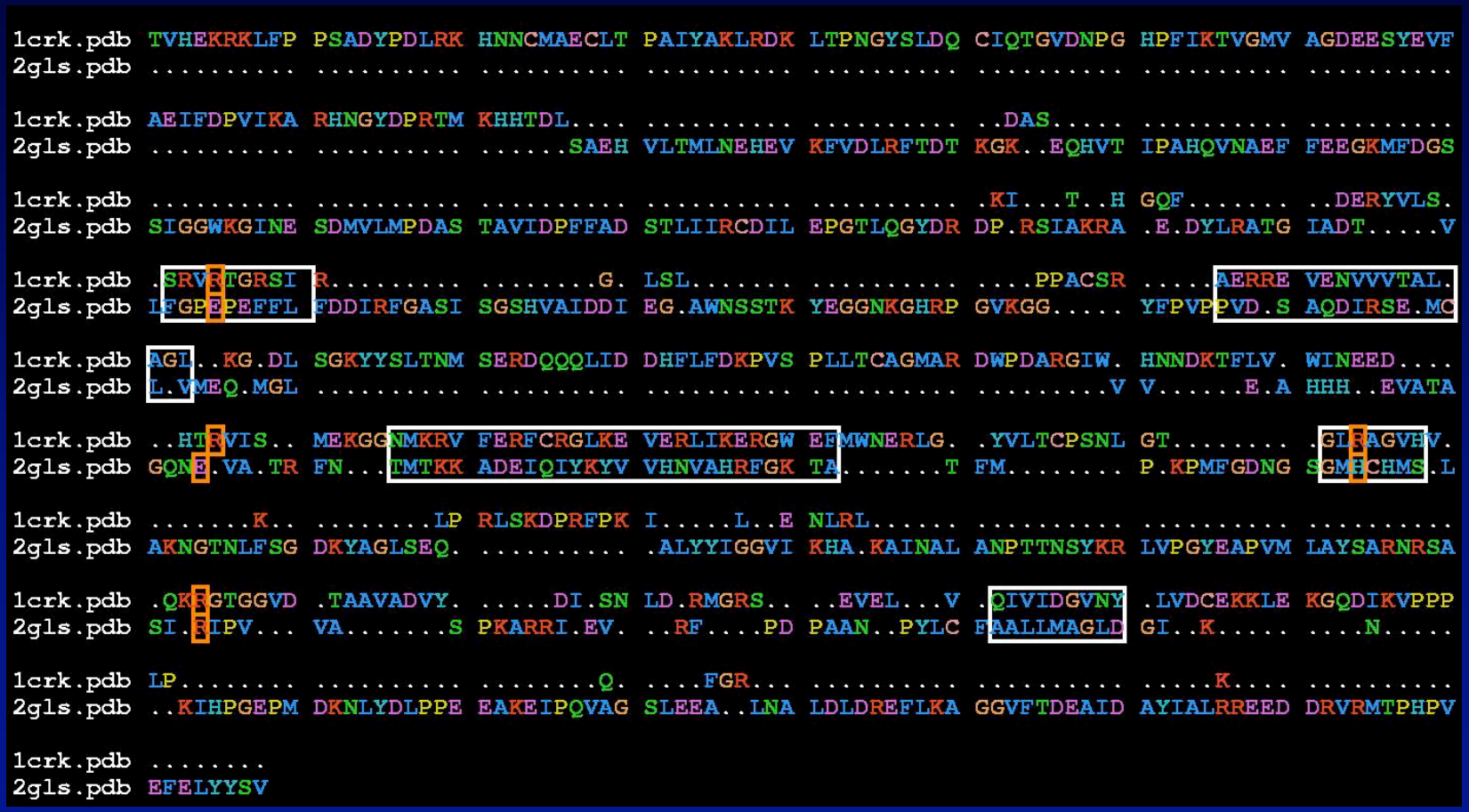
Creatine kinase

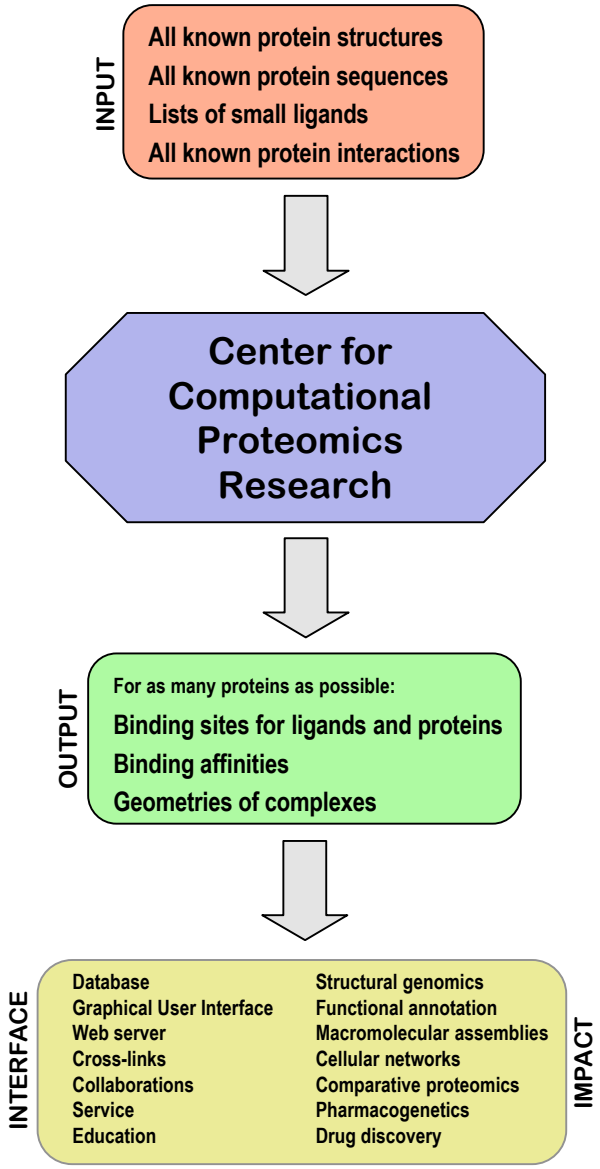

Chimera
Session Edit Model Tools Options Help
RMSD
Matches
Sequence vs. Sequence
467
0
0
379
Zoom In
Zoom Out
X
0.0
Y
0.0
Orientation Clusters
RMSD vs N
Matches
120
RMSD
1.99
Longest Distance
3.37
-log(probability)
8.25
39.00
0.00
2
380
Zoom In
Zoom Out
pdate selec
monitor selection
align120.msf
File Style Region Pattern Triggers
410 420 430 440 450 460 470 480 490 500
1crk.pdb
2gls.pdb

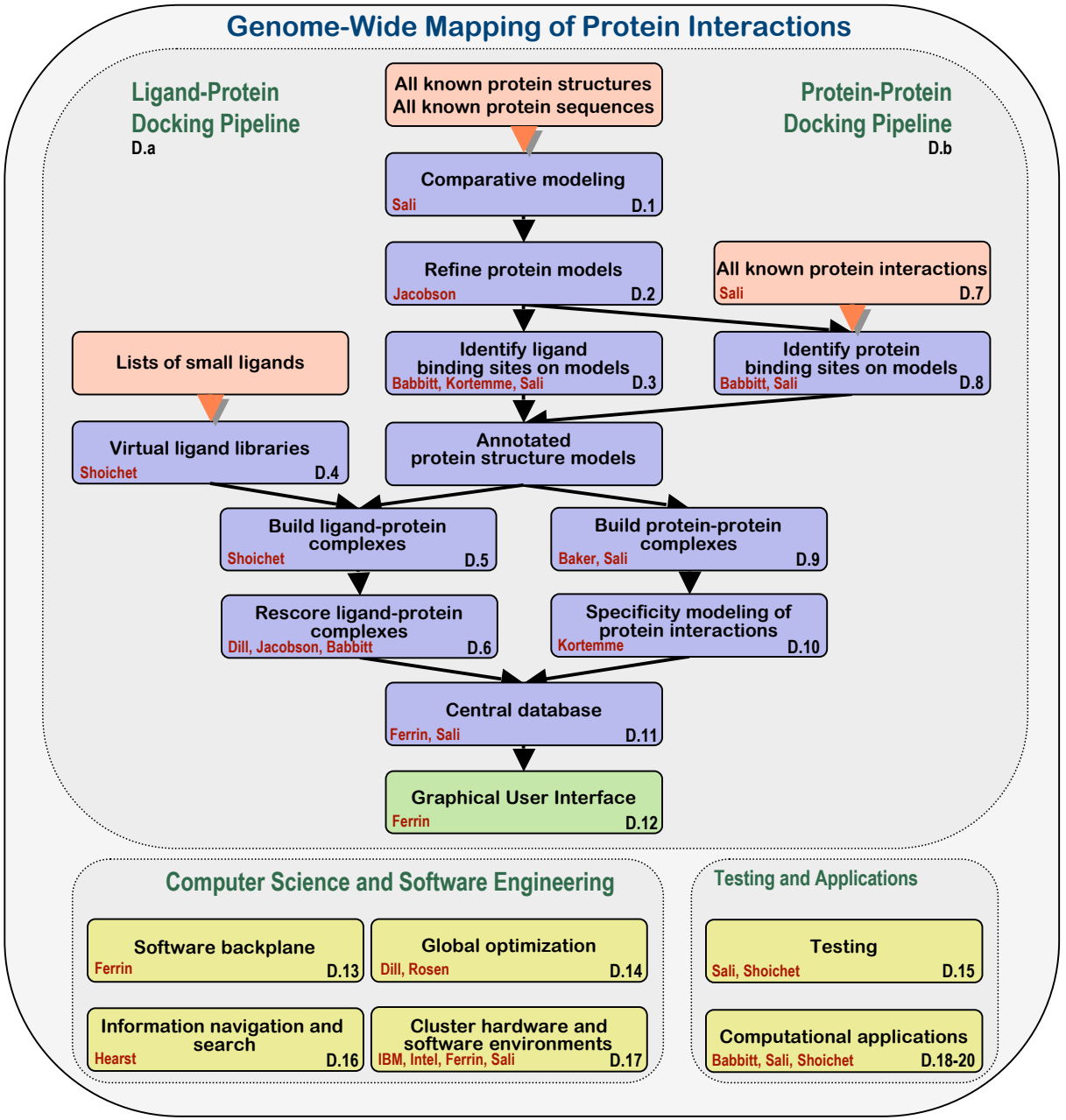
AlignPlot GUI



Resulting structure-based sequence alignment




INPUT
All known protein structures
All known protein sequences
Lists of small ligands
All known protein interactions
Center for Computational Proteomics Research
OUTPUT
For as many proteins as possible:
Binding sites for ligands and proteins
Binding affinities
Geometries of complexes
INTERFACE
Database
Graphical User Interface
Web server
Cross-links
Collaborations
Service
Education
IMPACT
Structural genomics
Functional annotation
Macromolecular assemblies
Cellular networks
Comparative proteomics
Pharmacogenetics
Drug discovery


Genome-Wide Mapping of Protein Interactions
Ligand-Protein Docking Pipeline
D.a
Protein-Protein Docking Pipeline
D.b
All known protein structures
All known protein sequences
Comparative modeling
Sali
D.1
Refine protein models
Jacobson
D.2
All known protein interactions
Sali
D.7
Lists of small ligands
Identify ligand binding sites on models
Babbitt, Kortemme, Sali
D.3
Identify protein binding sites on models
Babbitt, Sali
D.8
Virtual ligand libraries
Shoichet
D.4
Annotated protein structure models
Build ligand-protein complexes
Shoichet
D.5
Build protein-protein complexes
Baker, Sali
D.9
Rescore ligand-protein complexes
Dill, Jacobson, Babbitt
D.6
Specificity modeling of protein interactions
Kortemme
D.10
Central database
Ferrin, Sali
D.11
Graphical User Interface
Ferrin
D.12
Computer Science and Software Engineering
Software backplane
Ferrin
D.13
Global optimization
Dill, Rosen
D.14
Information navigation and search
Hearst
D.16
Cluster hardware and software environments
IBM, Intel, Ferrin, Sali
D.17
Testing and Applications
Testing
Sali, Shoichet
D.15
Computational applications
Babbitt, Sali, Shoichet
D.18-20

Summary

We are in the midst of a profound and exciting new era in bioinformatics and computational biology

The data made available by the various genome and structural genomics projects will occupy researchers for decades to come

High performance computing and the internet play a critical role in the navigation, analysis, and dissemination of this data and the resulting scientific knowledge

The tremendous volume of data makes for a critical need for tools and techniques that make information navigation easy

The potential impact on drug development and treatment of human disease is enormous

Acknowledgements

Collaborators & Staff

- Dr. Conrad Huang, Dr. Elaine Meng, Prof. Patricia Babbitt, Prof. Kathy Giacomini, Greg Couch, Eric Pettersen, Al Conde, Tom Goddard, Susan Johns, Doug Stryke, Michiko Kawamoto

NIH National Center for Research Resources

- P41-RR01081

National Institute of General Medical Sciences

- GM61390

Additional information

RBVI:

www.rbvi.ucsf.edu

PMT project:

www.pharmacogenetics.ucsf.edu

Chimera:

www.cgl.ucsf.edu/chimera

CCPR:

www.computationalproteomics.org