Visualization Tools for Integrating Sequence and Structural Information

Thomas Ferrin

University of California, San Francisco

“It’s sink or swim as a tidal wave of data approaches”

Petabyte (1,000 terabytes)

Exabyte (1,000 petabytes)

Zettabyte (1,000 exabytes)

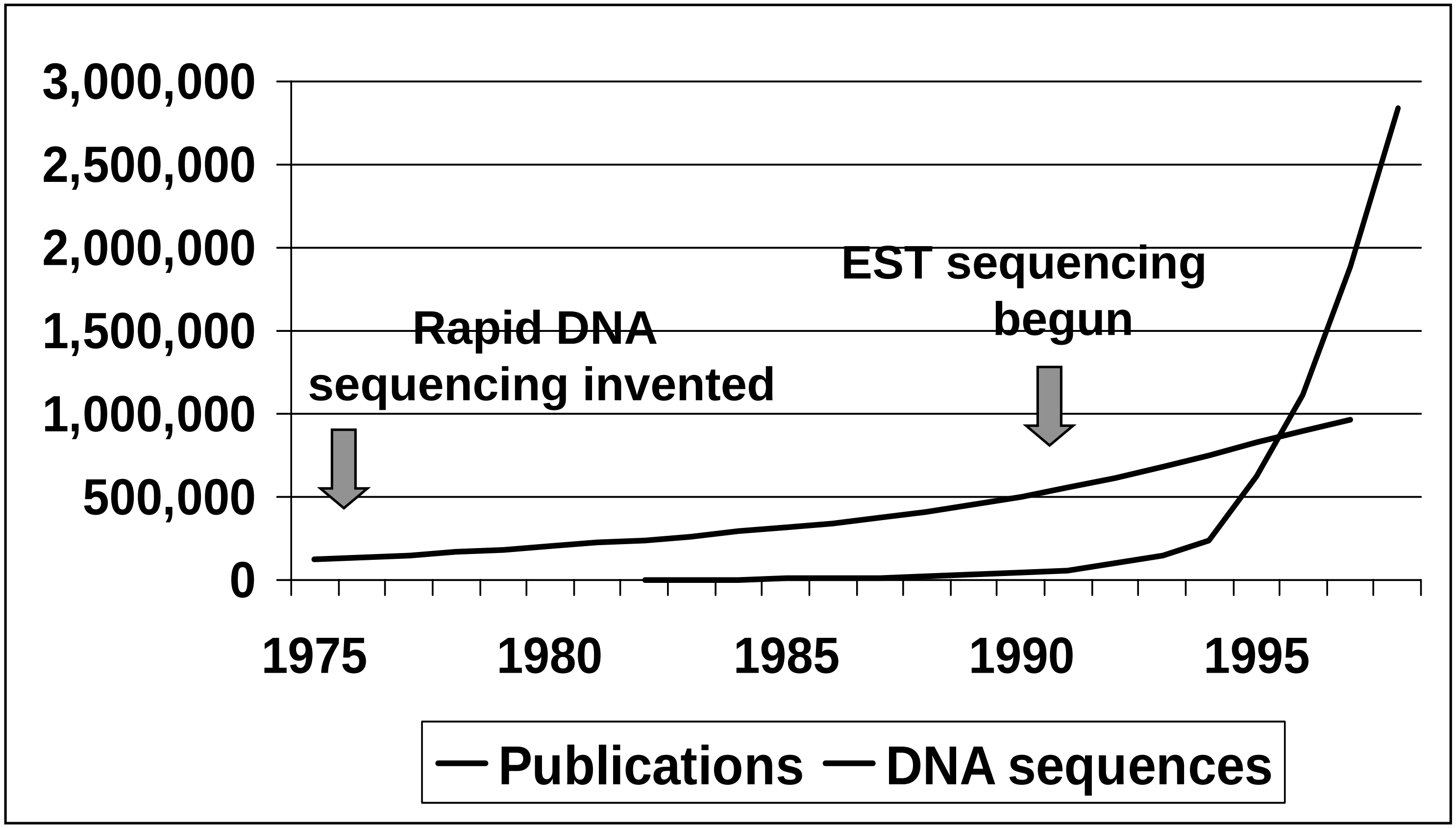
Yottabyte (1,000 zettabytes)

Tony Reichhardt
Nature 399:517-520 10 June 1999

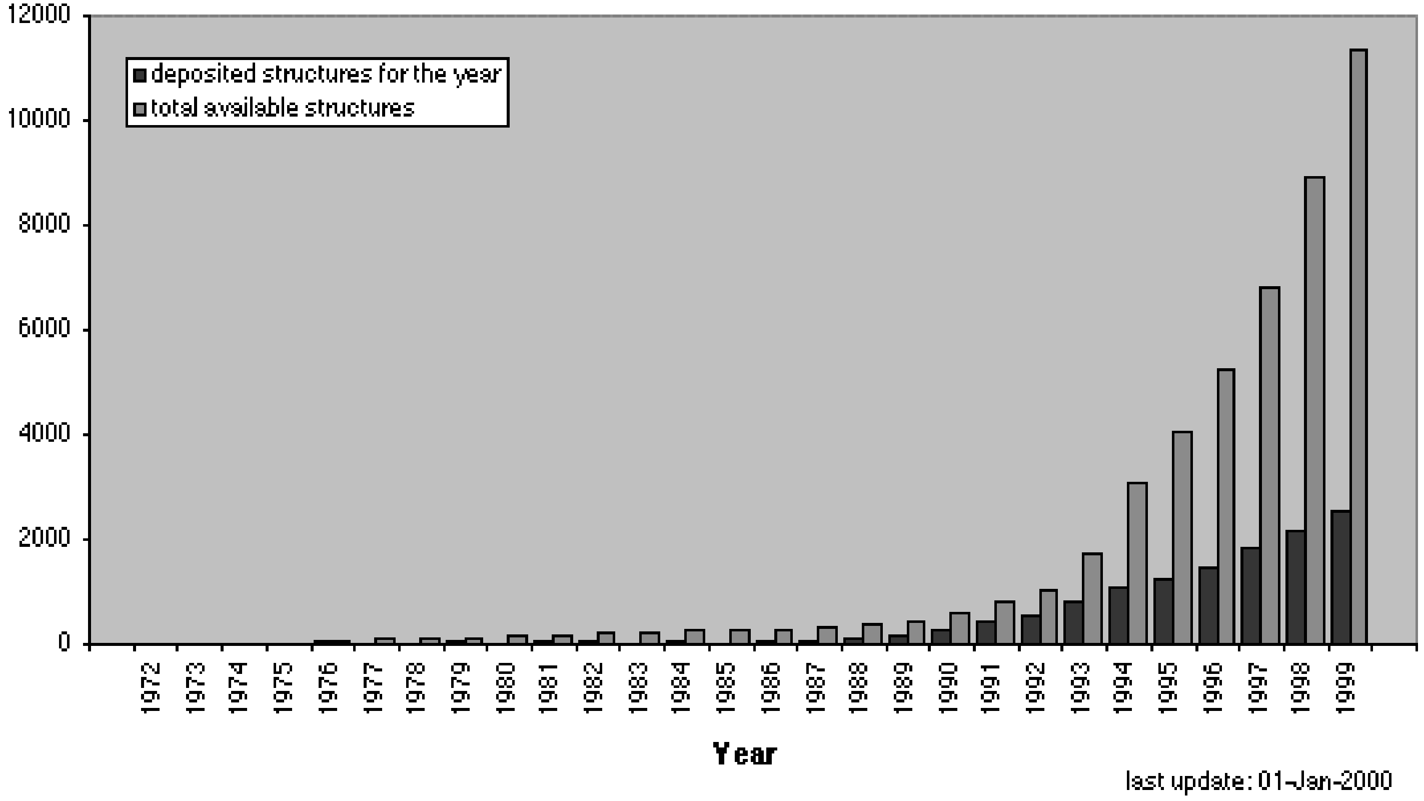
“Many biologists are still in denial, never having faced the amount of information now pouring into databases such as Genbank and SwissProt… They haven’t really thought about how they’re going to use all this data...”

***Ibid*.**

The Growing Gap in Functional Knowledge



Growth in Protein Structures



Sequence -> Structure -> Function

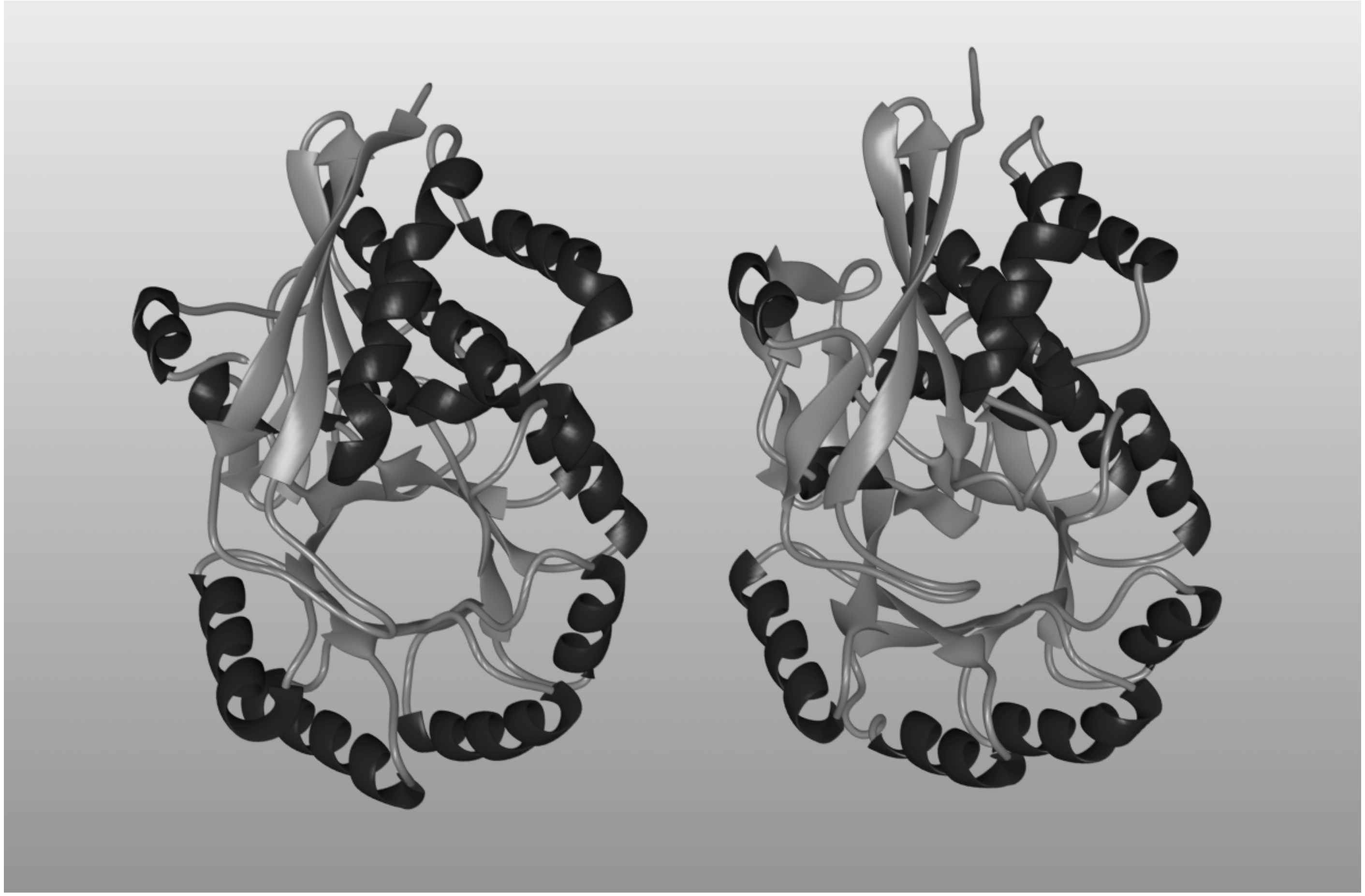
Challenges:

- Prediction of structure from sequence
- Prediction of function from sequence
- Understanding of evolutionary changes
- Engineering proteins for specialized function
- Applications in pharmacogenomics and pharmacogenetics

Potential for major impact on…

- Drug discovery
- Prediction of drug response
- Avoidance of toxic effects in many individuals

Stereo pairs ?



[redacted]

Tools for Comparative Protein Studies

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

AlignPlot - interactive exploration of structural alignments

MSFviewer - integrates sequence and structure space

Chimera - extensible 3-D molecular modeling system

MinRMS

Find all plausible alignments between two protein structures (experimentally-determined or modeled) using root-mean-square difference of coordinates of alpha-carbons.

- **RMSD metric easy to interpret**
- **Avoids “single best alignment” problem**
- **Avoids need for parameters**
- **Finds reasonable alignments even for apparently dis-similar structures**

MinRMS Algorithm

Two step process:

- 1. Rotate & translate the two structures to bring similarly shaped regions into close proximity;
- 2. With the two proteins fixed at a particular relative position, select corresponding alpha-carbon atoms between the proteins which minimizes the intermolecular RMSD.

Apply a dynamic programming algorithm to find best matches for different numbers of amino acid residues

Algorithm runs in O(n^5) time

- For two 300-residue proteins requires ~1 hour on a fast workstation

MinRMS Output

Large table containing, for each structure alignment:

- Number of matched residues
- RMSD for the alignment
- Longest distance between any pair of matched residues
- Levitt & Gerstein similarity score, -log(P)
- Transformation matrix for aligning the structures

AlignPlot

Used to examine MinRMS output for alignments of interest

- RMSD vs. Number of matched residue pairs
  - Useful for examining trade-off between number of matched residues and global superposition
- Orientation clusters
  - Reduces hundreds of alignments into a few representative groups
- Sequence vs. sequence histogram
  - Provides easy identification of patterns such as secondary structure

MSFviewer

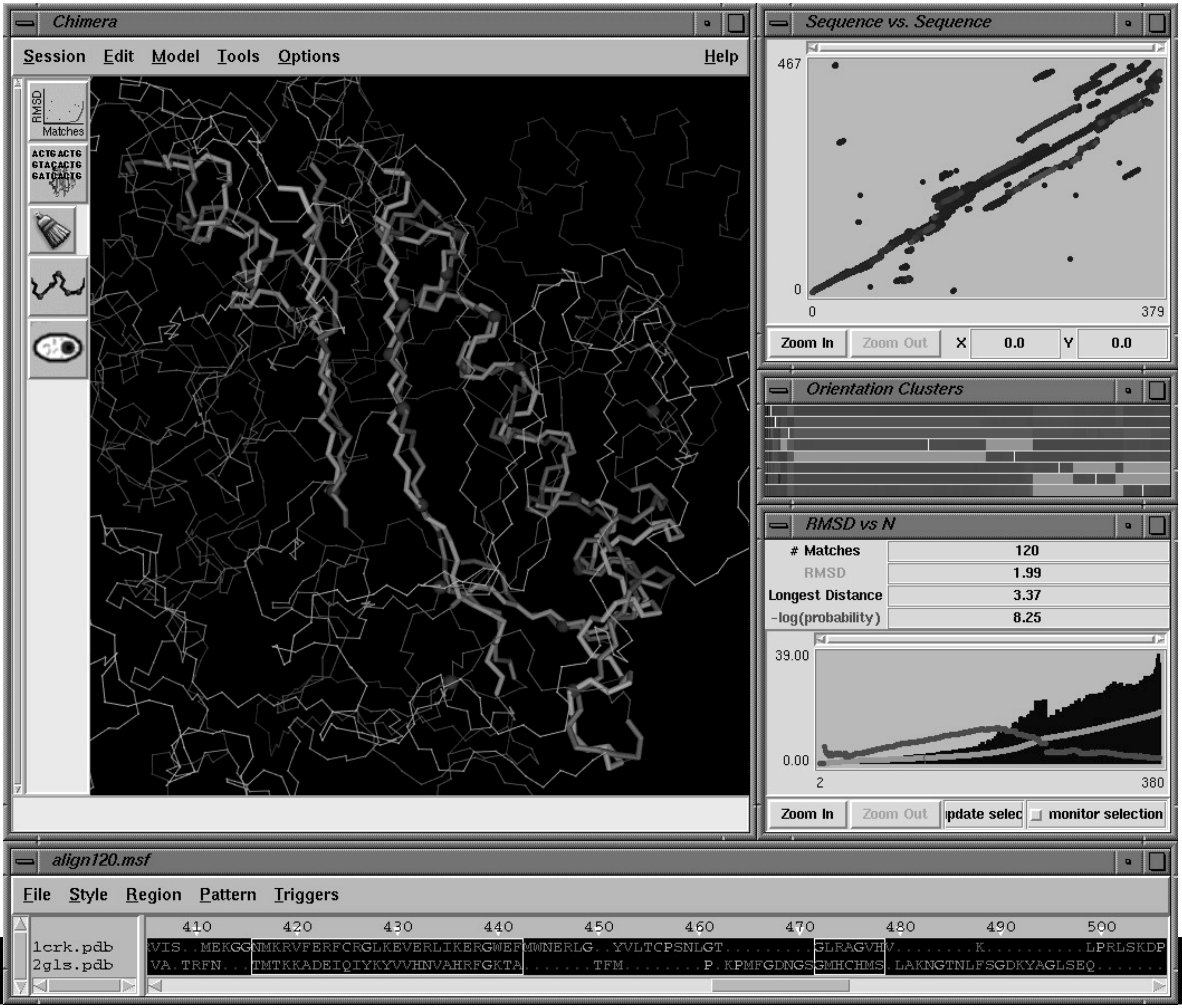
Displays multiple sequence alignments from common alignment programs

- Groups of residues in the alignment can be selected
- Corresponding residues in the structure also get highlighted
- Allows user to facile interface to sequence space

Chimera

Molecular visualization system providing:

- Interactive manipulation of multiple molecular structures
- Real-time rendering of models in several formats
  - e.g. ball-and-stick, ribbons, molecular surfaces
- Support for non-molecular objects
  - e.g. points, vectors, markers, spheres, cylinders, polygons
- Command line compatibility with MidasPlus
- Extensible functionality without access to source code
- Use of standard APIs ensure portability to many platforms
  - Windows 95/98/NT/2000, Compaq, SGI, Linux, …



# Matches	120
RMSD	1.99
Longest Distance	3.37
-log(probability)	8.25

```
          410       420       430       440       450       460       470       480       490       500
1crk.pdb  VIS..MEKGGNMKRVFERFCRGLKEVERLIKERGWEFMWNERLG..YVLTCPSNLGT.........GLRAGVHV........K..........LPRLSKDP
2gls.pdb  VA.TRFN...TMTKKADEIQIYKYVVHNVAHRFGKTA......TFM........P.KPMFGDNGSGMHCHMS.LAKNGTNLFSGDKYAGLSEQ.......
```

Chimera’s Extensibility

Use of Python programming language as Chimera’s command language provides for both complex command “scripts” and user-written extensions

- **True programming language allows for user commands to contain such constructs as iterative loops and conditional execution with full access to internal data structures**
- **Widely available Python libraries provide for custom GUIs**
  - **e.g. menus, dialog boxes, custom graphics**
- **Python’s interpreted language provides for dynamic run-time linking**
  - **Don’t need access to source code to add new features**
  - **New modules “linked in” when Chimera executes**

Chimera Extensions

Extensions are just groups of one or more cooperating processes

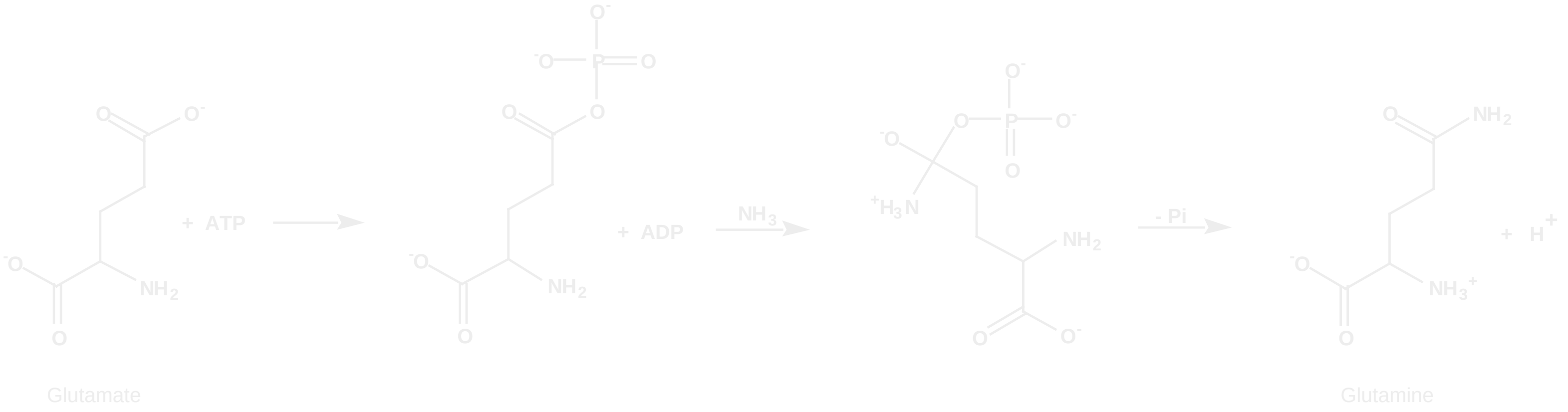
- AlignPlot, MSFviewer, MidasPlus Command Interpreter are all implemented as extensions
- Extensions can maintain their won state and have their own graphical user interface
- Extensions can be ancillary to Chimera or Chimera can be invoked by another program to provide interactive graphical output

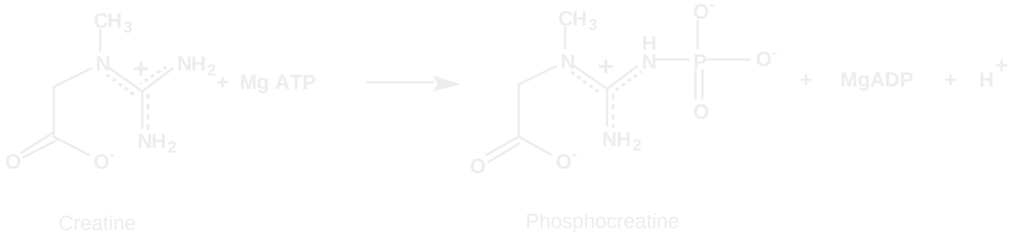
Example Study

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

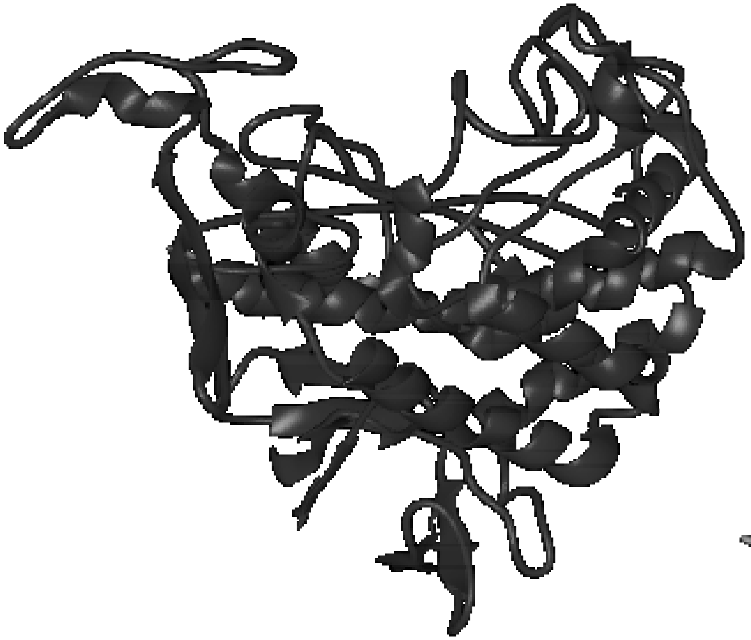
- GS: 468 residues, PDB entry 2gls
- CK: 380 residues, PDB entry 1crk
- No significant sequence similarity, both have multimeric forms, proposed similar tertiary structures, and catalyze similar reactions

GS and CK catalysis



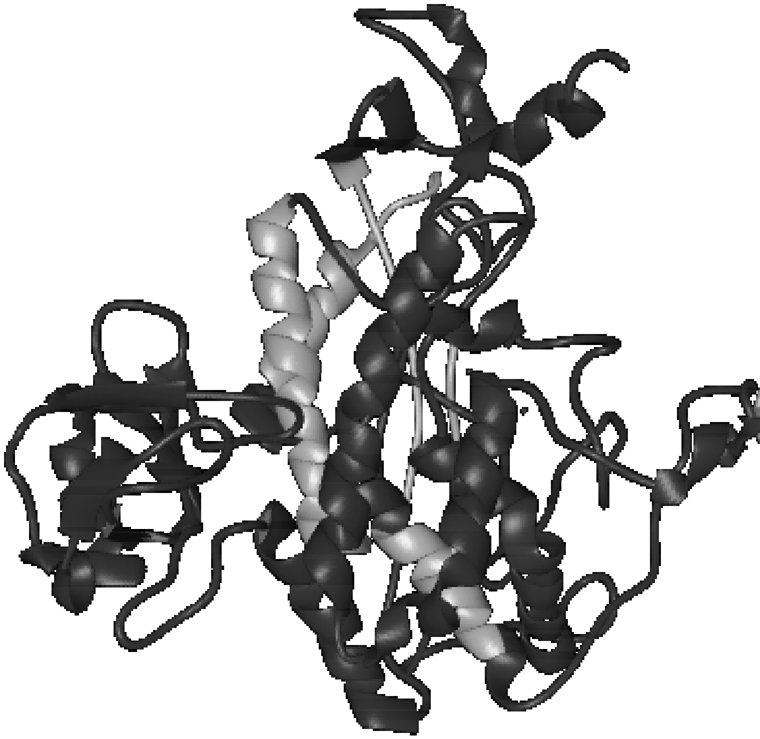


Glutamine synthetase and creatine kinase

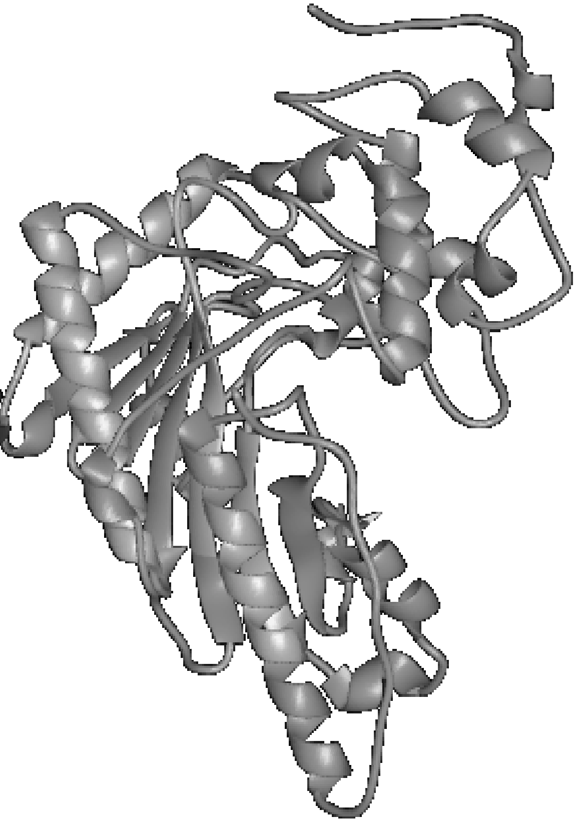


[redacted]

After MinRMS alignment

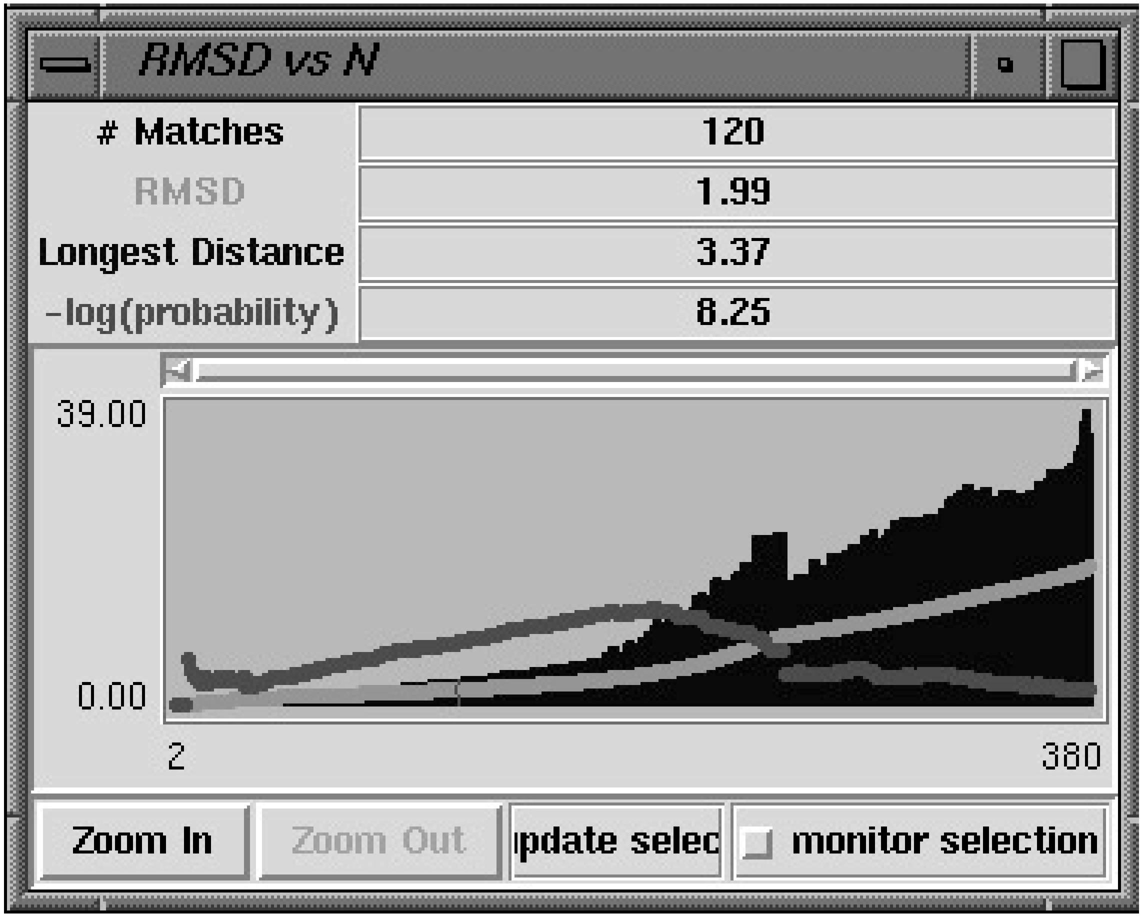


Glutamine synthetase



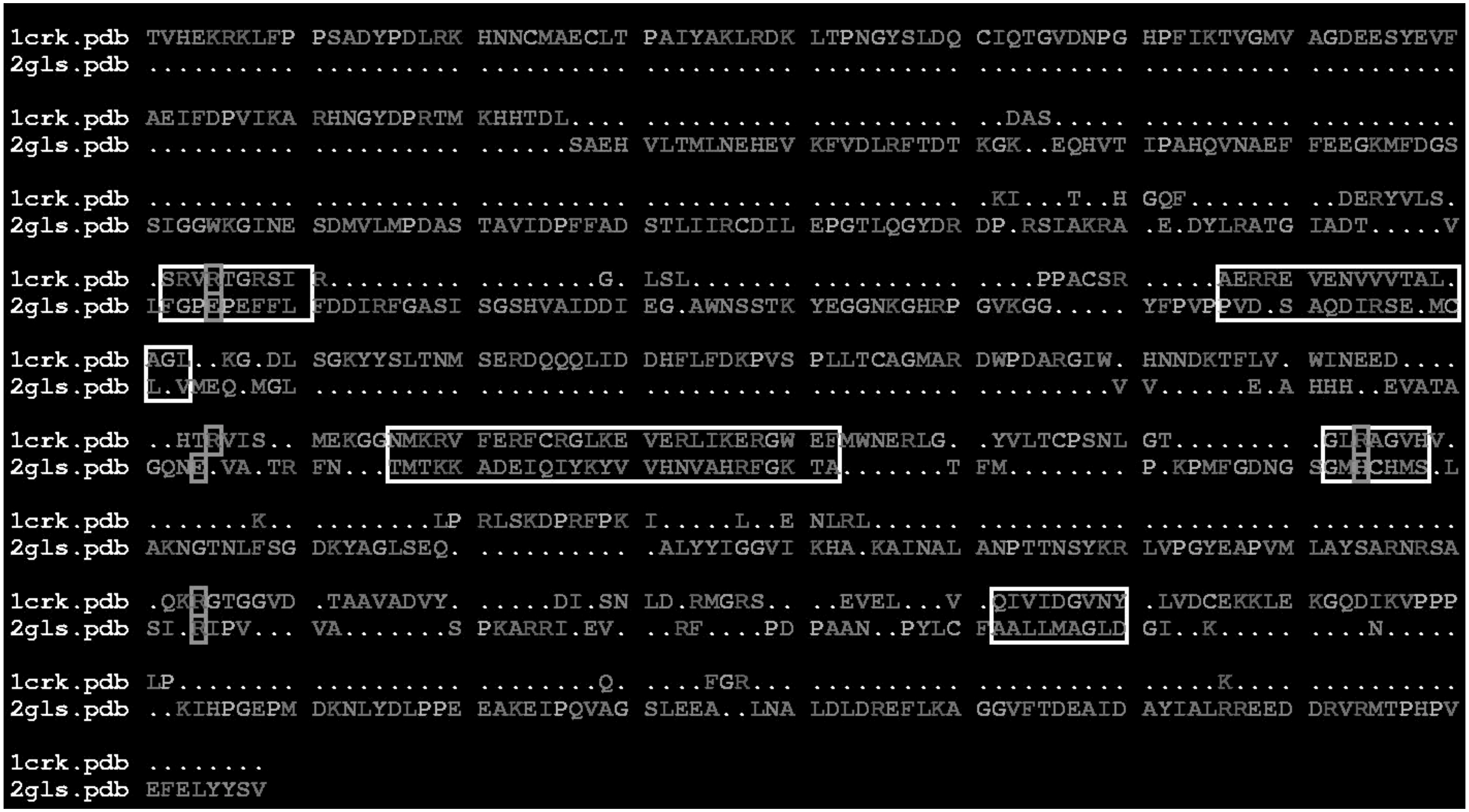
Creatine kinase

AlignPlot GUI



# Matches	120
RMSD	1.99
Longest Distance	3.37
-log(probability)	8.25

Resulting structure-based sequence alignment



```
1crk.pdb TVHEKRKLFP PSADYPDLRK HNNCMAECLT PAIYAKLRDK LTPNGYSLDQ CIQTGVDNPG HPFIKTVGMV AGDEESYEVF
2gls.pdb .......... .......... .......... .......... .......... .......... .......... ..........

1crk.pdb AEIFDPVIKA RHNGYDPRTM KHHTDL.... .......... .......... ..DAS..... .......... ..........
2gls.pdb .......... .......... ......SAEH VLTMLNEHEV KFVDLRFTDT KGK..EQHVT IPAHQVNAEF FEEGKMFDGS

1crk.pdb .......... .......... .......... .......... .......... .KI...T..H GQF....... ..DERYVLS.
2gls.pdb SIGGWKGINE SDMVLMPDAS TAVIDPFFAD STLIIRCDIL EPGTLQGYDR DP.RSIAKRA .E.DYLRATG IADT.....V

1crk.pdb .SRVRTGRSI R......... ........G. LSL....... .......... ....PPACSR .....AERRE VENVVVTAL.
2gls.pdb LFGPEPEFFL FDDIRFGASI SGSHVAIDDI EG.AWNSSTK YEGGNKGHRP GVKGG..... YFPVPPVD.S AQDIRSE.MC

1crk.pdb AGL..KG.DL SGKYYSLTNM SERDQQQLID DHFLFDKPVS PLLTCAGMAR DWPDARGIW. HNNDKTFLV. WINEED....
2gls.pdb L.VMEQ.MGL .......... .......... .......... .......... .........V V......E.A HHH..EVATA

1crk.pdb ..HTRVIS.. MEKGGNMKRV FERFCRGLKE VERLIKERGW EFMWNERLG. .YVLTCPSNL GT........ .GLRAGVHV.
2gls.pdb GQNE.VA.TR FN...TMTKK ADEIQIYKYV VHNVAHRFGK TA.......T FM........ P.KPMFGDNG SGMHCHMS.L

1crk.pdb .......K.. ........LP RLSKDPRFPK I.....L..E NLRL...... .......... .......... ..........
2gls.pdb AKNGTNLFSG DKYAGLSEQ. .......... .ALYYIGGVI KHA.KAINAL ANPTTNSYKR LVPGYEAPVM LAYSARNRSA

1crk.pdb .QKRGTGGVD .TAAVADVY. .....DI.SN LD.RMGRS.. ..EVEL...V .QIVIDGVNY .LVDCEKKLE KGQDIKVPPP
2gls.pdb SI.RIPV... VA.......S PKARRI.EV. ..RF....PD PAAN..PYLC FAALLMAGLD GI..K..... ....N.....

1crk.pdb LP........ .......... ........Q. ....FGR... .......... .......... .....K.... ..........
2gls.pdb ..KIHPGEPM DKNLYDLPPE EAKEIPQVAG SLEEA..LNA LDLDREFLKA GGVFTDEAID AYIALRREED DRVRMTPHPV

1crk.pdb ........
2gls.pdb EFELYYSV
```

Live Demonstration

Disclaimer: Anything that can go wrong will do so in direct proportion to the number of people in the room.

Hardware:

- **Compaq AlphaStation DS10 (466Mhz EV6)**
- **PowerStorm 350 graphics accelerator**

Recent developments

Re-engineering of a natural enzyme with new catalytic function

- **Alan Fersht & coworkers at Cambridge Centre for Protein Engineering**
- **Converted activity of indole-3-glycerol phosphate synthase (IGPS) into that of phosphoribosylanthranilate isomerase (PRAI)**
- **See C&E News February 21, 2000**
- ***Nature* 403,** 617 (2000)

Significant Challenges

Robust methods for predicting function from sequence

Ways to represent biological function, including detailed chemistry, in databases

Facile access for ordinary biologists to the wealth of available sequence, structure, and function data

Training for students for the breadth of knowledge required in biology today

UCSF’s Program in Quantitative Biology

SCIENTIFIC DISCIPLINES	BIOINFORMATICS	STRUCTURAL BIOLOGY/BIOPHYSICS	COMPLEX SYSTEMS
COMPONENT FIELDS	Genomics Proteomics	Structure Determination Molecular Recognition	Cell Biology Pharmacology Neuroscience
RESEARCH AREAS	pharmacogenomics molecular evolution whole-genome analysis (e.g. gene prediction, sequence comparison) function prediction biocomputing biological databases	structure prediction crystallography NMR spectroscopy imaging (large structures, in-vivo methods, confocal and electron microscopy drug design, development & delivery molecular modeling	complex networks population genetics pharmacokinetics and pharmcodynamics modeling chaos theory statistical genetics neural information theory neural networks (analytic & computational) microarray data analysis
CLINICAL CORRELATES	genetic basis for disease clinical information systems	diagnostic & therapeutic imaging	functional imaging clinical outcomes research & epidemiology
GRADUATE & POSTDOCTORAL TRAINING PROGRAMS	**Medical Information Science** (MIS), Genetics, Pharmaceutical Sciences & Pharmacogenomics (PSPG)	**Biophysics**, **Chemistry & Chemical Biology**,	**Neuroscience**, PSPG, MIS, Bioengineering,

Acknowledgements

Collaborators

- Prof. Patricia Babbitt, Prof. Teri Klein, Dr. Conrad Huang

National Center for Research Resources

- P41-RR01081

Department of Energy

- DE-FG03-96ER62269

National Institutes of Health

- AR17323

Additional Information

See UCSF Computer Graphics Laboratory web site:

http://www.cgl.ucsf.edu/chimera