Pacific Symposium on Biocomputing 5:227-238 (2000)

Integrated Tools for Structural and Sequence Alignment and Analysis

Conrad C. Huang, Walter R. Novak, Patricia C. Babbitt, Andrew I. Jewett, Thomas E. Ferrin, and Teri E. Klein*

Departments of Pharmaceutical Chemistry and Biopharmaceutical Sciences
University of California, San Francisco
San Francisco, California 94143-0446

We have developed new computational methods for displaying and analyzing members of protein superfamilies. These methods (*MinRMS*, *AlignPlot* and *MSFviewer*) integrate sequence and structural information and are implemented as separate but cooperating programs to our *Chimera* molecular modeling system. Integration of multiple sequence alignment information and three-dimensional structural representations enable researchers to generate hypotheses about the sequence-structure relationship. Structural superpositions can be generated and easily tuned to identify similarities around important characteristics such as active sites or ligand binding sites. Information related to the release of *Chimera*, *MinRMS, AlignPlot* and *MSFviewer* can be obtained at http://www.cgl.ucsf.edu/chimera.

1. Introduction

By July of 1999, the number of non-redundant protein sequences in the Genbank database had reached ~400,000 and included completed genome sequences for 23 organisms. These data provide an opportunity to explore the evolution of functional diversity by probing the entire repertoire of many organisms. One powerful approach to this study is the comparative analyses of large numbers of protein structures and their associated functions through primary sequence analysis and computer-assisted modeling of three-dimensional structures.

For example, discovery of a large enzyme superfamily whose members represent a surprising range of different chemical functions extended the insight that the evolution of new functions is linked to chemical capabilities associated with a given tertiary fold.[1,2] Because it illuminates the constraints built into the evolution of protein structure, this focus on chemistry is a crucial element for learning how new enzyme functions evolve from pre-existing structural scaffolds. This observation provides the conceptual framework for developing computational tools that integrate sequence and structure, and provides the basis for formulating hypotheses about

*Corresponding Author: klein@cgl.ucsf.edu

function. The function of an unknown reading frame is rather easily deduced from sequence similarity when the function is the same as that of its homologs. For more divergent proteins, the predictions can be much more difficult because the function of each unknown enzyme may have no apparent relationship to that of its homologs. In each case, the crucial clues are provided by hidden similarities in chemical mechanisms that can be inferred from the structural comparisons. Because the most interesting insights come from relationships among such highly dissimilar proteins, we have developed methods to identify these distant sequence relationships[3] and to interpret them using tools designed to integrate sequence and structural information.

Aspects of this problem have been solved by a number of investigators. There are several examples of homology modeling packages such as the Swiss-Model[4] web server, Molecular Applications Group's LOOK[5] and Molecular Simulations Inc.[6] Homology and Insight II. There are also tools such as DINAMO[7], CINEMA[8], and PROTALIGN[9] and PROMUSE[10] which are useful in analyzing structure-sequence alignments. However, these tools have limitations such as extensibility, interactive real-time three-dimensional graphics display and analysis, and/or cost.

2. New Computational and Analysis Tools

The set of tools, *MinRMS*[11], *AlignPlot* and *MSFviewer* were developed for sequence and structural alignment and analysis. These methods were easily integrated with *Chimera*[12] using Python[13], *Wrappy*[14], the *Object Technology Framework*[15], C and C^{++}. *MinRMS,* written in C^{++}, generates a family of structural alignments, allowing the user to explore the similarities between two proteins, including highly divergent structures (Figure 1). The unique ability to examine the optimum RMSD (Root Mean Square Distance) superpositions generated from the α-carbons of the structures being compared provides a much richer environment for exploring structural similarities than methods that produce a single pairwise alignment[16, 17]. Details of *MinRMS* and *Chimera* are published elsewhere[11, 12].

The focus of this paper is on new tools for structural and sequence analysis and visualization. *AlignPlot*, written in C^{++} and Python, provides a graphical representation of the RMSD values for each alignment in the set, allowing the user to quickly identify the regions of two structures that are most similar. Particularly important, it provides a user-friendly way to display specific alignments on the screen and navigate among them. *MSFviewer*, written in Python, provides an integrated link to sequence space, displaying multiple alignments of related sequences on the screen and providing for interactive highlighting of a selected structural align-

ment and the associated multiple sequence alignment.

2.1 MinRMS

Holm and Sander[16], Godzik[17], Fenz and Sippl[18], and Orengo *et. al.*[19] have suggested that determining the single best structural alignment may not be possible. Given two proteins with experimentally-determined or modeled three-dimensional coordinates, *MinRMS*[11] solves this issue by generating multiple structural alignments and their corresponding sequence alignments. The *MinRMS* algorithm performs an exhaustive analysis of all plausible shape similarities between two proteins using RMSD between aligned α-carbon atoms. This method generates alignments containing all possible amino acid residues in a single pass without the need of parameters.

MinRMS uses intermolecular RMSD as the metric for comparing protein structures. The appropriateness of RMSD as a metric for comparing protein structures has been discussed elsewhere.[20-22] The main advantage of the RMSD in that it is easy to interpret. The *MinRMS* algorithm is a two-step process: (1) Two proteins are rotated and translated to bring similarly shaped regions into close proximity; and, (2) With the two proteins fixed at a particular relative position, corresponding residues are chosen between the two proteins which minimize RMSD. Candidate superpositions are generated by selecting every fragment of 4 consecutive residues for each of the proteins and superimposing them by least-squared distance between α-carbons using the method described by Diamond[23]. Given the relative positions of the two structures, we developed a new dynamic programming algorithm to choose the matching residues between the proteins that minimizes RMSD. Similar to the Needleman and Wunsch[24] algorithm, our algorithm is recursive and blind to "non-topological" similarities[25]. For each candidate superposition, the algorithm generates multiple alignments containing different numbers of corresponding residues which minimize the intermolecular RMSD.[11] The dynamic programming algorithm is applied to all candidate superpositions between the proteins with small local regions well matched. Typical execution time for aligning two proteins with an average length of 300 residues is less than 1 hour on an SGI Onyx 2.

The output of *MinRMS* is a large table of data that contains, for each structural alignment, the number of matched residues for the two proteins, the RMSD for the alignment, and the longest distance between any pair of matched residues. For comparison purposes, the $-log(P)$ is calculated where P is the probability that a better alignment is found between two unrelated proteins occurring in the SCOP[26] database as described by Levitt and Gerstein[22]. Structural alignment is presented in sequence alignment form as MSF (Multiple Sequence Format) files (Table 1). Relative positions are stored as comments in the MSF file. The program *Chimera*, in cooperation

with *AlignPlot* and *MSFViewer*, is used to view the volumes of data produced from *MinRMS*.

2.2 Chimera

*Chimer*a is a molecular visualization graphics package developed at the UCSF Computer Graphics Laboratory. *Chimera* is written in the Python programming language with C^{++} extensions and uses standard multi-platform libraries such as the Tk toolkit for it's graphical user interface and OpenGL for three-dimensional graphics primitives. *Chimera* also uses the *Object Technology Framework* object class library for manipulating molecular data.

A major design goal for *Chimera* is program extensibility. By choosing Python as the *Chimera* command language, users can create complex command "scripts" (*e.g.*, with iterative loop and conditional execution) which in turn allow for sophisticated operations to be performed on multiple molecular models. Python has an extensive library[13] that include interfaces to Tk. This means that users are easily able to create their own custom graphical user interface (GUI) elements such as menus and dialog boxes as part of their extensions. *Chimera* itself is implemented with a small set of core functionalities, including graphical display, Protein Data Bank (PDB) input, and basic user interface elements (menu bars, tool bar, command line, reply window and status line). More advanced features are constructed on top of the core using Python extension modules. This results in a program architecture in which new functionality is easily added when needed. The separate applications *AlignPlot* and *MSFviewer* are example extensions of *Chimera*.

2.3 AlignPlot

AlignPlot displays three different representations that summarize the data from MinRMS. The bottom graph (Figure 1: RMSD *vs*. *N*) displays three numerical quantities as a function of matched residue pairs (*N*): RMSD, *-log(P)* of Levitt and Gerstein[22] and the longest pairwise distance between matched residues. *MinRMS* and Levitt and Gerstein scores are displayed to provide multiple evaluation criteria. Levitt and Gerstein favor matching more residues over better geometric fit. Thus, their method is less distance sensitive than *MinRMS*. The user can easily select a particular alignment by point and click with the mouse in the graph. The corresponding three-dimensional superposition is visualized in *Chimera*. Matched residues closer than one angstrom are denoted by a small sphere. Matched residues with a distance greater than one angstrom have a line drawn between them. This plot allows the user to discern patterns over the entire set of solutions.

The middle representation (Figure 1: Orientation Clusters) in *AlignPlot* uses a genetic algorithm (GA) to condense the data from *MinRMS* by selecting a small set of orientations to represent the entire data set. For any given set of representative orientations, a structure in a non-representative orientation contributes a penalty proportional to the RMSD from the most similar representative orientation. The GA "fitness" metric is the sum of penalties of all non-representative orientations. The GA uses the fitness metric to find a good representative set, which is then used to divide the data set into clusters. The clustering results are displayed as a table where the columns represent alignments and the rows represent clusters. The cells of the table are color-coded and the brightness of each cell is proportional to RMSD from the representative of that cluster. The cluster plot classifies the solutions into a small number of groups which reduce the amount of information that the user needs to process.

The top representation (Figure 1: Sequence vs. Sequence) is a two-dimensional histogram of residue pairs. Each cell of the histogram represents a pair of residues, one from each structure. The value of the cell is the number of *MinRMS* alignments that match the two residues. The value is converted to color. The color scale is blue to red representing values that range from 1 to the maximum cell value. If there is no match, the cell is colored like the background. Information displayed in this manner provides easy identification of matching patterns (*e.g.*, secondary structure matches appear as diagonal runs).

Using these three tools together, one can identify structural alignments of interest. The orientation cluster plot reduces hundreds of alignments into tens of alignments. The RMSD vs. *N* plot illustrates the trade-off between the number of matched residues and closeness of global superpositioning. Lastly, the Sequence vs. Sequence plot typically identifies secondary structural elements important to the alignment. These tools used in combination facilitates the analysis of a large data set.

2.4 MSFviewer

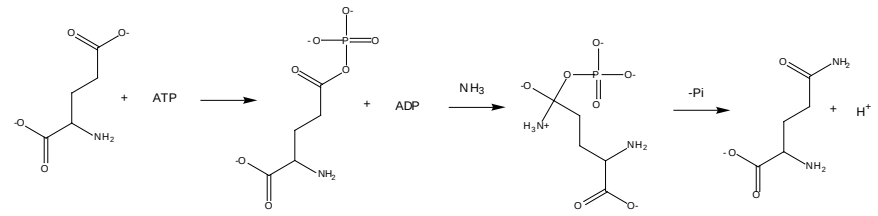
MSFviewer was developed independently of *AlignPlot* to view multiple sequence alignment in MSF format (*e.g.*, an output option of commonly used multiple alignment programs). Fragments of the sequence can be selected and highlighted on the structure, allowing the user to focus on secondary structure elements, active site residues, monitoring of residues of interest and filtering of data (Figure 1). The alignment can be edited interactively, saved in MSF format or printed for presentation purposes (Figure 2).

MSFviewer cooperates with *AlignPlot* through *Chimera* for the selection and mapping of sequences to their structures. Selecting a specific alignment in *AlignPlot* will highlight the matched residues in *MSFviewer*. Upon selecting specific residues in *MSFviewer*, *AlignPlot* displays the matching statistics of these residues

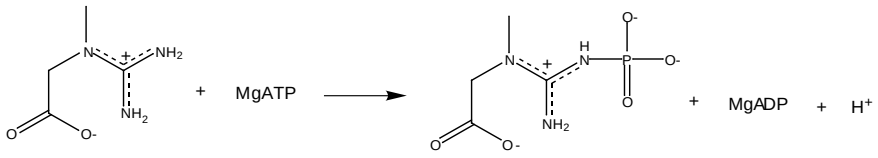
for each alignment. *Chimera* serves as the data repository and communication channel between *AlignPlot* and *MSFviewer.*

3. Example: Structural Comparisons of Glutamine Synthetase with Creatine Kinase and other Guanidino Kinases

The study of structure-function relationships is important to the understanding of proteins and provides guidance for protein engineering. We have attempted to better understand structure-function relationships through the structural comparison of glutamine synthetase (GS) with creatine kinase (CK) and other guanidino kinases. While GS and CK have no significant sequence similarity, they both have multimeric forms, have been proposed to have similar tertiary structures[27] (Figure 3), and catalyze similar reactions. GS catalyzes the reaction of glutamate and ammonia to form glutamine through a phosphorylated intermediate, while CK catalyzes the transfer of a phosphate group from ATP to creatine to yield phosphocreatine.



Glutamate Glutamine



Creatine Phosphocreatine

Liaw and Eisenberg[28] solved a series of crystal structures of GS to elucidate the mechanism of glutamine synthesis and to identify residues possibly involved in ATP binding and in transfer of the γ phosphate. This structure of GS was superimposed with an available CK structure [29] using *MinRMS*. A family of several hundred structural superpositions resulted reflecting many possible orientations of GS and CK (Figure 1). Simultaneous viewing of the three-dimensional and sequence alignments and interactive editing of the sequence alignments allowed for comparison of catalytic residues and binding domains using all of the sequence and structural infor-

mation available (Figures 4 and 5). These tools allowed us to examine the ATP-binding residues of GS and CK using sequence alignments informed by the structural superpositions. While crystal structures of CK bound with MgATP or substrate are not available, our studies indicate that many of the ATP binding residues in GS have potential homologs in CK.

The information gained from this analysis supports the previous suggestion that a similar scaffold is used in both GS and CK[27]. The analyses of this work suggest that this scaffold also utilizes potentially homologous residues to bind ATP and assist in the transfer of the γ-phosphate group. Use of the tools described here have provided a useful model to continue the study of structure-function relationships in the guanidino kinases. Prior to using these tools, it was difficult to obtain a useful structural alignment.

4. Concluding Remarks

Superfamily analysis frequently involves proteins whose sequence similarities may fall below the level of *statistical significance* but whose relationships are nontheless *biologically significant*. *MinRMS*, *AlignPlot*, *MSFviewer* along with Shotgun[3], in cooperation with *Chimera*, provide a set of tools for generating and testing hypotheses about sequence, structure and functional relationships of such proteins.

Initial testing of this software has suggested additional functionalities to allow users to choose the subsets of alignments that provide the best overlap over specific residues such as active site residues. More extensive editing capabilities will be added to facilitate correcting the registration between (1) sub-group multiple alignments of very distantly related sequences based on the structural alignments; and (2) very distantly related sequences based on the structural alignments of representative sub-group members. Lastly, we are exploring non-distance methods for comparing more than two proteins at one time.

Information on the availability of the software tools described here can be found at http://www.cgl.ucsf.edu/chimera.

Acknowledgments

This work is supported by the Department of Energy (DE-FG03-96ER62269), NIH National Center for Research Resources (P41-RR01081) and NIH (AR17323).

References

1. P.C. Babbitt, G.T. Mrachko, M.S. Hasson, G.W. Huisman, R. Kolter, D. Ringe, G.A. Petsko, G.L. Kenyon. and J.A. Gerlt, "A Functionally Diverse

Enzyme Superfamily that Abstracts the α-protons of Carboxylic Acids." *Science* **267**: 1159-1161, 1995.

2. P.C. Babbitt, M. Hasson, J.E. Wedekind, D.J. Palmer, M.A. Lies, G.H. Reed, I. Rayment, D. Ringe, G.L. Kenyon, and J.A. Gerlt., "The Enolase Superfamily: A General Strategy for Enzyme-Catalyzed Abstraction of the α-protons of Carboxylic Acids." *Biochem.* **35**: 16489-16501, 1996.
3. S.C.-H. Pegg and B.C. Babbitt, "Shotgun: Getting More from Sequence Similarity Searches." *Bioinformatics* (in press).
4. N. Guex and M.C. Peitsch, "SWISS-MODEL and the Swiss-PdbViewer: An Environment for Comparative Protein Modeling," *Electrophoresis* **18**:2714-2723, 1997.
5. Molecular Applications Group, 607 Hansen Way, Building One, Palo Alto, California 94304. See http://www.mag.com/.
6. Molecular Simulations Inc., 9685 Scranton Road, San Diego, Caliofrnia 92121. See http://www.msi.com/.
7. M. Hansen, J. Bentz, A. Baucom and L. Gregoret, "DINAMO: A Coupled Sequence Alignment Editor/Molecular Graphics Tool for Interactive Homology Modeling of Proteins", *PSB* 106-117, 1998.
8. T.K. Attwood, A.W.R. Payne, A.D. Michie and D.J. Parry-Smith, "A Colour INteractive Editor for Multiple Alignments - CINEMA," *EMBnet.news* **3**, 1997.
9. D. Meads, M.D. Hansen and A. Pang, "PROTALIGN: A 3-Dimensional Protein Alignment Assessment Tool," *Pacific Symposium on Biocomputing* 354-367, 1999.
10. M.D. Hansen, E. Charp, S. Lodha, D. Meads and A. Pang, "PROMUSE: A System for Multi-Media Data Presentation of Protein Structural Alignments," *Pacific Symposium on Biocomputing* 368-379, 1999.
11. A.I Jewett, C. C. Huang, C. and T.E. Ferrin, "MinRMS: An Efficient Algorithm for Determining Protein Structure Similarity." (submitted)
12. C.C. Huang, G.S. Couch, E.F. Pettersen and T.E. Ferrin, "Chimera: An Extensible Molecular Modeling Application Constructured using Standard Components", *Pacific Symposium on Biocomputing*, 724, 1996.
13. See http://www.python.org/.
14. G.S. Couch, "Wrappy -- A Python Wrapper Generator for C^{++} Classes," in O'Reilly Open Source Convention Python Conference Proceedings, 1999, http://conferences.oreilly.com/.
15. C.C. Huang, E.F. Pettersen, G.S. Couch, T.E. Ferrin, A.E. Howard and T.E. Klein, "The Object Technology Framework (OTF): An Object-Oriented

Interface to Molecular Data and Its Application to Collagen." *Pacific Symposium on Biocomputing*, 349-361, 1998.

16. L. Holm and C. Sander, "Protein Structure Comparison by Alignment of Distance Matrices", *J. Mol. Biol.* **233**:123-138, 1993.
17. A. Godzik, "The Structural Alignment Between Two Proteins: Is there a Unique Answer?", *Protein Science* **5**:1325-1338, 1996.
18. Z.K. Feng and M.J. Sippl, "Optimal Superimposition of Protein Structures: Ambiguities and Implications," *Folding & Design* **1**:123-132, 1996.
19. C.A. Orengo, M.B. Swindells, A.D. Michie, M.J. Zvelebil, P.C. Driscoll, M.D. Waterfield and J.M. Thornton, "Structural Similarity Between the Pleckstrin Homology Domain and Verotoxin: The Problem of Measuring and Evaluating Structural Similarity," *Protein Science* **4**:1977-1983, 1995.
20. F.E. Cohen and M.J.E. Sternberg, "On the Prediction of Protein Structure: The Significance of the Room Mean Squared Deviation," *J. Mol. Biol.* **138**:321-333, 1980a.
21. A. Falicov and F.E. Cohen, "A Surface of Minimum Area Metric for the Structural Comparison of Proteins," *J. Mol. Biol.* **258**:871-892, 1996.
22. M. Levitt and M. Gernstein, "A Unified Statistical Framework for Sequence Comparison and Structure Comparison," *PNAS* **95**:5913-5920, 1998.
23. R. Diamond, "A Note on the Rotational Superposition Problem," *Acta. Cryst.* **A44**:211-216, 1988.

24 S.B. Needleman and C.D. Wunsch, "A General Method Applicable to the Search for Similarities in the Amino Acid Sequence of Two Proteins," *J. Mol. Biol.* **48**:443-453, 1970.

25. I.N. Shindyalov and P.E. Bourne, "Protein Structure Alignment by Incremental Combinatorial Extension (CE) of the Optimal Path," *Protein Engineering* **11**:739-747, 1998.
26. T.J. Hubbard, B. Ailey, S.E. Brenner, A.G. Murzin and C. Chothia, "SCOP, Structural Classification of Proteins Database: Applications to Evaluation of the Effectiveness of Sequence Alignment Methods and Statistics of Protein Structural Data," *Acta Crysta*. **D54**:1147-1154, 1998.
27. W. Kabsch and K. Fritz-Wolf, "Mitochondrial Creatine Kinase--A Square Protein," *Curr. Op. in Struct. Bio.*, **7**:811-818, 1997.
28. S-H Liaw and D. Eisenberg, "Structural Model for the Reaction Mechanism of Glutamine Synthetase, Based on Five Crystal Structures of Enzyme-Substrate Complexes," *Biochemistry*, **33**:675-681, 1994.
29. K. Fritz-Wolf, T. Schnyder, T. Wallimann and W. Kabsch, "Structure of Mitochondrial Creatine Kinase," *Nature* **381**:341-345, 1996.

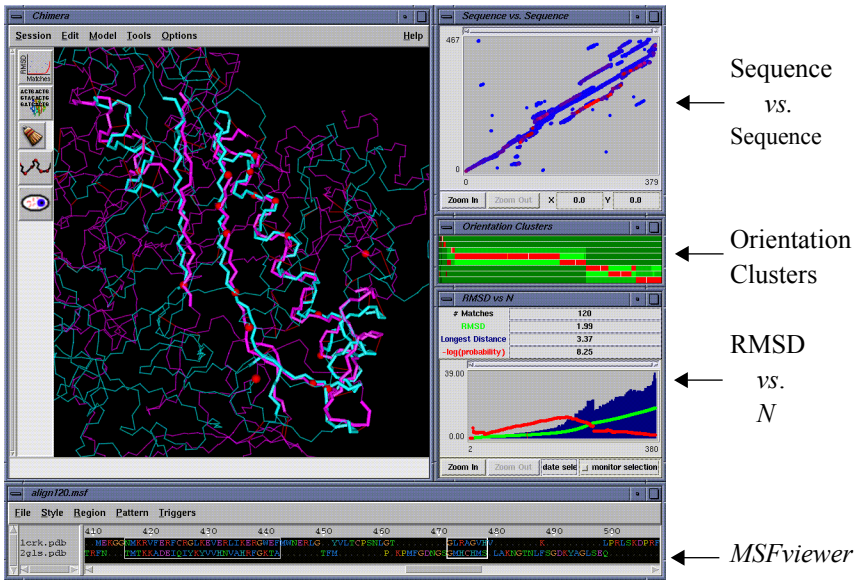


Figure 1: Screen display of *AlignPlot, MSFviewer* and *Chimera.* Glutamine synthetase is in magenta and creatine kinase is in cyan. Matched residue pairs are highlighted by red spheres and lines. See *sections 2.3 & 2.4* for detail descriptions.

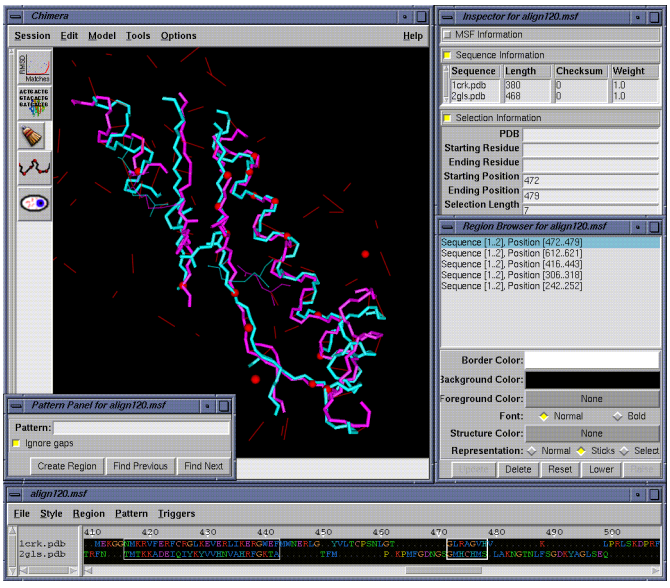
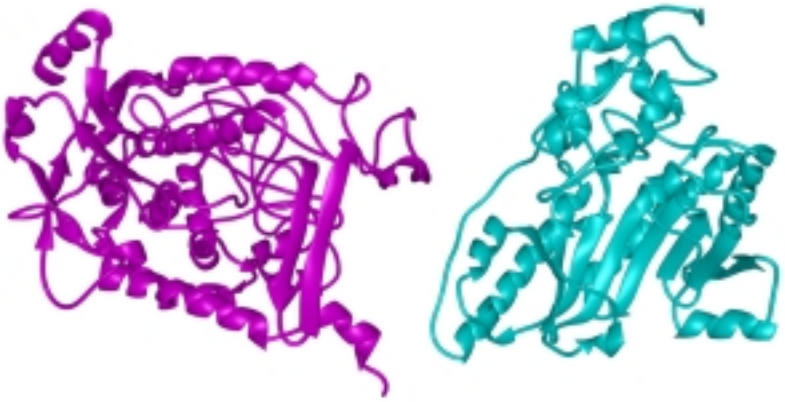


Figure 2: Graphical user interface elements of *MSFviewe*r are displayed.



.
Figure 3: Ribbon representations of glutamine synthetase (magenta) and creatine kinase (cyan) prior to alignment with *MinRMS*.

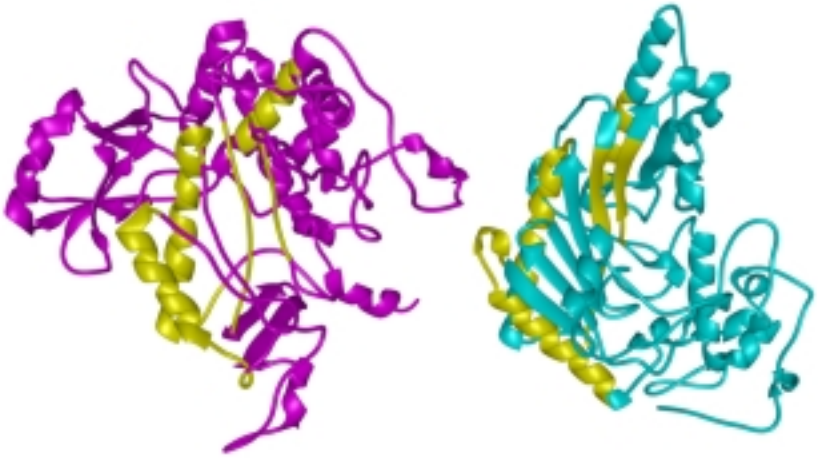


Figure 4: Ribbon representations of glutamine synthetase (magenta) and creatine kinase (cyan) post alignment with *MinRMS*. Matched regions are highlighted (yellow). The associated sequence alignment is seen in Figure 5.

```
Chimera minimal RMSD structural alignment with 120 equivalences.
RMSD = 1.988821
-----
Transform Matrix to apply to molecule: 2gls.pdb
0.580381 -0.537281 -0.611953 -9.842606
-0.744292 -0.654900 -0.130905 57.874092
-0.330435 0.531447 -0.779985 15.317198

 Name: 1crk.pdb         Len:   380  Check:    0  Weight:  1.00
 Name: 2gls.pdb         Len:   468  Check:    0  Weight:  1.00

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

1crk.pdb  HPFIKTVGMV AGDEESYEVF AEIFDPVIKA RHNGYDPRTM KHHTDL.... ..........
2gls.pdb  .......... .......... .......... .......... ......SAEH VLTMLNEHEV

1crk.pdb  .......... ..DAS..... .......... .......... .......... ..........
2gls.pdb  KFVDLRFTDT KGK..EQHVT IPAHQVNAEF FEEGKMFDGS SIGGWKGINE SDMVLMPDAS

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

1crk.pdb  .SRVRTGRSI R......... .........G. LSL....... .......... ....PPACSR
2gls.pdb  LFGPEPEFFL FDDIRFGASI SGSHVAIDDI EG.AWNSSTK YEGGNKGHRP GVKGG.....

1crk.pdb  .....AERRE VENVVVTAL. AGL..KG.DL SGKYYSLTNM SERDQQQLID DHFLFDKPVS
2gls.pdb  YFPVPPVD.S AQDIRSE.MC L.VMEQ.MGL .......... .......... ..........

1crk.pdb  PLLTCAGMAR DWPDARGIW. HNNDKTFLV. WINEED.... ..HTRVIS.. MEKGGNMKRV
2gls.pdb  .......... .........V V......E.A HHH..EVATA GQNE.VA.TR FN...TMTKK

1crk.pdb  FERFCRGLKE VERLIKERGW EFMWNERLG. .YVLTCPSNL GT........ .GLRAGVHV.
2gls.pdb  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

1crk.pdb  .......... .......... .QKRGTGGVD .TAAVADVY. .....DI.SN LD.RMGRS..
2gls.pdb  LVPGYEAPVM LAYSARNRSA SI.RIPV... VA.......S PKARRI.EV. ..RF....PD

1crk.pdb  ..EVEL...V .QIVIDGVNY .LVDCEKKLE KGQDIKVPPP LP........ ..........
2gls.pdb  PAAN..PYLC FAALLMAGLD GI..K..... ....N..... ..KIHPGEPM DKNLYDLPPE

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

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

Figure 5: MSF output from MinRMS of the sequence alignment for glutamine synthetase and creatine kinase. This structure alignment corresponding to this sequence alignment is displayed in Figure 4.