Title
Computational approaches for the study of protein structure

Author
Hearst, David Paul

Publication Date
1995

Peer reviewed|Thesis/dissertation

COMPUTATIONAL APPROACHES FOR THE STUDY OF PROTEIN STRUCTURE: GRAFTING, MODELING AND LIGAND IDENTIFICATION

by

DAVID PAUL HEARST

DISSERTATION

Submitted in partial satisfaction of the requirements for the degree of

DOCTOR OF PHILOSOPHY

in

PHARMACEUTICAL CHEMISTRY

in the

GRADUATE DIVISION

of the

UNIVERSITY OF CALIFORNIA

San Francisco

This thesis is dedicated to my family: John, Jean, Leslie, Lily, Corwin, Glacier, Denali, Taos and Nancy.

ACKNOWLEDGMENTS

I am grateful to my parents, John and Jean, for all of the love and guidance they have given me. They taught to dream and to aim high, and without their support my goals would have seemed far out of reach. I am fortunate to share their deep friendship and I have learned from them just how important good friends are.

I want to thank my sister, Leslie, for all the times that we have had together. We watched each other excel and struggle over the years, and I am incredibly fortunate to see us both achieve so much.

My grandmother, Lily, is one of life's utmost inspirations. I have learned from her that every day must be confronted with enthusiasm and vigor. If I can face life with even half her drive and determination I will do fine.

Thanks go out to Steve Isaacs and Kathy Macbride and the opportunities they gave me with H.R.I. and Steritech. The adventure that began with flashlights and empty warehouses continues to this day, and the skills that I developed with them continue to serve me well.

I am indebted to my friend and mentor, Scott Presnell, who helped me evolve from a naive first-year graduate student into a confident computer programmer. There is no way that I would be where I am today without that help. I look forward to many years of backpacking, skiing and conversation ahead!

The work described in this thesis was guided by Fred Cohen who gave me both guidance and the freedom to think on my own. Many times, when the challenges seemed unconquerable, it was Fred's patience and support that helped me push on. I also owe a great deal to Tack Kuntz for his advice and reassurance. I wish to thank

Charly Craik for his enthusiasm and willingness to go out on a limb. Computational chemists are a frightening sight in an experimental lab, and I'm grateful that he was willing to let me try...

When I joined the Cohen group, I was attracted by the people that made up the group. I am fortunate to have all of them as friends. To Chris, my guardian angel and great friend, thanks for showing the way! Without Bruce I would not have gained such a deep understanding for following my dreams and for the great history of baseball. I owe a great debt to John for the late night company, the frivolous distractions and the drive to make something right at all costs.

UCSF would not have been the same without the many friendships that I gained while there. For woodworking, music, anarchy and more baseball, Neil was always a refreshing change of pace. Without Randy I would not have remembered to stop and have a little fun. I am glad that when there was no one left to ask those tough questions, Don always seemed to know the answer. To Dennis I owe one giant stack of magazines and all the knowledge gained from them. Scott Willett was always there to share a little of his never-ending enthusiasm for the good things in life. David Corey taught me everything I know about mutagenesis and expression, and without him Chapter 3 would not exist.

I must thank Ginger for all of the conversations we shared about the twists and turns that are part of life. I know far more about glass and gardens as a result, and look forward to spring!

My final debt is to Nancy, my wife, who endured more than anyone. I look forward to many years ahead when we can share all of life's joys and challenges. We made it!

Computational Approaches for the Study of Protein Structure: Grafting, Modeling and Ligand Identification

David Paul Hearst

Abstract

Many approaches have been developed over the years for the study of macromolecular structure. The evolution of computer technology has spurred the development of computational tools for analysis of protein and nucleic acid structure. The work described here focuses on computational approaches applied toward the study of protein structure. A series of geometric search algorithms (DIST, GRAFTER and the GRAFTER suite) are summarized herein. Each of these is designed as a method for identifying potential graft sites in protein structure. Such sites are collections of residues in a protein scaffold that have suitable geometry for replacement with a functional (binding or catalytic) motif. Numerous verification examples and novel graft designs are outlined in the following chapters. The three graft-oriented algorithms represent first, second and third generation attempts at a computational solution to the grafting problem. Also described is an experimental attempt at evaluating a graft of the trypsin catalytic triad onto the Staphylococcal Nuclease scaffold. The later chapters explore another area of protein structure computations: homology modeling and inhibitor design. A model of dihydrofolate reductase from *Cryptosporidium Parvum* is predicted using homology techniques. A number of model refinement and evaluation tools that measure dihedral likelihoods, steric conflicts, and residue packing are applied during the model building process. The model is then used as the basis for ligand identification using an existing computational docking tool (DOCK) and newly developed techniques for visual screening. A large library of small molecules is scanned by computer to identify any ligands that appear to fit the active site of the model.

Table of Contents

List of Tables

List of Figures

Chapter 1: Introduction

EVOLUTION

The evolutionary process is responsible for both the diversity and refinement of the organisms on this planet. Certain aspects of an organism's genetic makeup are highly optimized due to selective pressure exerted by external factors. Without such optimization, the organism would be at a disadvantage for survival. In contrast, other aspects of a creature's characteristics are non-essential, and therefore do not experience selective pressures. These features are likely to diversify over time, leading to individuals with different physical and mental capabilities. As time passes such diversification can lead to separate sub-species and species.

We see the combination of diversification and conservation clearly at the level of DNA, RNA and protein. As we study protein sequence and structure, it is obvious that certain areas of a protein with a particular function tend to remain constant (e.g. the active or binding sites) while other areas vary tremendously (e.g. the protein surface). If we study a family of enzymes such as the serine proteases, we will find that the catalytic machinery (a catalytic Ser, His, Asp triad and an oxyanion hole) is similar in all of its members. The scaffolds that support this machinery, however, vary between the different family members. From the view of a protein engineer, one may conclude that only a subset of the enzyme's residues are crucial to the function of that enzyme. The remaining residues may function simply as a structural scaffold that supports the active components, maintaining their functional geometry. In essence, the active residues in the protein or enzyme family are grafted onto differing scaffolds, each of which maintains the functional geometry of the grafted residues.

There are two variations within the evolutionary model that have significance for protein engineering: divergent evolution and convergent evolution. Divergent evolution is the process by which a single protein can evolve into two or more distinct proteins. Over time, residues are mutated without destroying the protein's usefulness to its organism. Eventually, a protein results that is different at a certain number of residues,

and is clearly distinguishable from its ancestor. Divergent evolution often leads to a number of proteins or enzymes that retain similar function (e.g. serine proteases) but differ in their substrate specificity. In some cases (e.g. dihydrofolate reductase) the function is identical between different organisms, but the protein scaffold varies dramatically. Under other circumstances, the opposite pattern of conservation and mutation occurs, leading to two proteins with similar scaffold structures that perform different enzymatic functions.

Convergent evolution is the process by which two proteins evolve toward the same function from different starting points. Trypsin and subtilisin are examples of this phenomenon. Evolutionary studies indicate that these two enzymes evolved from distinct starting points (Graf, Hegyi et al. 1988), yet they share both function (proteolysis) and catalytic machinery (Ser, His, Asp catalytic triad, and an oxyanion hole). This is perhaps the most exciting example in nature of a functional motif "grafted" onto two unrelated scaffolds that retains its function in both cases.

GRAFTING

Our growing understanding of the basic principles governing protein structure has led to a series of increasingly ambitious experiments to redesign proteins. However, we are not yet able to design and build a functional protein with a unique folded structure from the ground up (DeGrado, Wasserman et al. 1989; Hecht, Richardson et al. 1990; Handel, Williams et al. 1993; Kamtekar, Schiffer et al. 1993).Today, protein engineering remains an exploratory field, emphasizing the modification of existing protein structures and analysis of the results. The knowledge gleaned from such studies has established the foundation for functionality transfer experiments, where a particular binding specificity or catalytic motif from one protein may be grafted onto another protein scaffold (Hedstrom, Szilagyi et al. 1992). In these applications, a motif is a collection of not necessarily contiguous residues that constitute a ligand binding or

catalytic site on a protein of known structure. Ideally such grafts would be possible between structures without known homology. This restriction raises the level of difficulty markedly for grafting experiments. Without homology as a guide for graft placement, a researcher must graphically search one or more structures to identify sites that could act as potential scaffolds. This process is not only tedious, but it is also likely to be biased by the experiences of the researcher. The result is that suitable sites may be overlooked. Molecular diversity approaches that rely on recently developed phage strategies could be adapted to this problem, but this approach is unlikely to succeed unless a relatively short linear epitope is sought (Scott and Smith 1990).

Although experiments have shown that simple functional grafts are feasible, most work to date has been based on manually designed grafts involving one or a few amino acid substitutions (Cronin, Malcolm et al. 1987; Wilks, Hart et al. 1988). Even in cases where a larger number of residues were changed, the design stage was simplified by sequence homology. Ptashne and co-workers (Wharton and Ptashne 1985) exploited repressor sequence and structural homology to graft the 434 repressor's binding specificity onto the P22 repressor. However, there is some precedence for novel graft design without the aid of homology, particularly in the area of metal-binding site grafts. Several groups have successfully used metal-binding sites as functional switches that control enzyme activity (Corey and Schultz 1989; Higaki, Haymore et al. 1990). Although homology was not an aid in these designs, the need for a binding site in proximity to the catalytic residues helped limit the extent of the search.

Computational Approaches

Several computational tools have been developed to aid protein engineering and drug design (Kuntz 1992). Each relies on a search of proteins of known structure for a subset that meets a series of constraints. These algorithms examine structure databases in search of geometric matches, and in some cases perform additional steps to modify

these structures to satisfy certain constraints. Nussinov and Wolfson (Nussinov and Wolfson 1991) have developed an effective strategy for identifying structural similarities within a number of macromolecular structures. Their approach makes use of a geometric hashing paradigm designed originally for computer vision applications. A somewhat different approach has been taken in the DEZYMER program (Hellinga and Richards 1991). This grafting tool performs an initial geometric search of a scaffold to identify optimal sites, and then explores combinations of rotamers to build the best representation of a desired functionality. Energy minimization is used to relax the structure into a favorable conformation, at which point the structure may be compared to the original motif. A related algorithm, DOCK, has been developed as an aid for drug and inhibitor design (DesJarlais, Sheridan et al. 1988). DOCK estimates the shape of clefts and pockets in a protein structure by filling these cavities with spheres. Each cluster of spheres represents an idealized ligand shape, and is used to search a database of small molecules for lead compounds. CAVEAT, a vector based tool with comparable applications to DOCK has also been developed (Bartlett, Shea et al. 1989). All of these methods limit the combinatorial complexity of an explicit conformational search by emphasizing geometric constraints at the outset.

For protein structure, the Protein Data Bank (PDB) (Bernstein, Koetzle et al. 1977) provides a vast resource for protein engineering experiments. Combined with current molecular biology techniques, this amounts to a breeding ground for novel protein structures. However, the steady increase in size of the data bank makes the aforementioned computational tools absolutely necessary for exhaustive searches.

We are particularly interested in developing tools for protein grafting. Having identified a function of interest, the greatest hurdle in the design process is the search for an ideal target site for the graft. We have developed a suite of programs (the GRAFTER suite) that can aid in the identification of the best scaffolds for placement of a given functional motif. Our goal is to automate both the search step and the primary

evaluation step. In a sense, this combines the applications of the computer vision (Nussinov and Wolfson 1991) and DEZYMER (Hellinga and Richards 1991) programs into a single tool. Our methodology allows for effective searching of large scaffold databases while providing detailed analysis of any resultant grafts. The algorithms used are designed to allow extensive databse searches while still providing a ranking scheme that narrows the search to a few sites for detailed consideration by the investigator. This approach both simplifies the search, and provides consistency to the results. The programs build the graft onto every potential site and evaluate the results in terms of overall geometry, side chain orientation, steric conflicts and accessibility. Recognizing that computational algorithms are a supplement to a researcher's intuition, we do not present a unique "best" graft. Instead, the GRAFTER suite generates a complete rank-ordered list of all matches found. A number of high scoring structures can be evaluated visually before selecting the most suitable graft.

HOMOLOGY MODELING

Evolutionary models are also important for structural modeling of proteins. A collection of proteins sharing the same function will often have remarkably similar structures while possessing less than 30% identity in amino acid sequence. If we already know the sequences and structures of a group of proteins with similar function (e.g. the NADPH-dependent reduction of dihydrofolate or folate to tetrahydrofolate) we may be able to predict the structure of a protein knowing only its sequence and the fact that it has similar function (i.e. dihydrofolate reductase). To do so, we align the sequence of the unknown protein with the structurally aligned sequences of our known proteins. Special attention is paid toward alignment of functionally significant residues or residues that tend to be conserved within the family. With this alignment in place, we borrow the structure from one or more of our known proteins and apply it in sequence to our unknown protein. Regions of the structure that cannot easily be adopted from a know

structure typically fall into loops. Loops can be predicted de novo or a dictionary lookup may be applied using loop structures from a database.

After all structural elements have been predicted, one is left with a structure describing all backbone atoms for the model. Side chains must be built onto this scaffold. Typically, model residues with identical or similar side chains to the known crystal structure are placed in identical orientations to that structure. Side chains that have no direct analogy to the known structure are placed based on side chain rotamer propensities (McGregor, Islam et al. 1987; Ponder and Richards 1987; Dunbrack and Karplus 1993). Cycles of structural evaluation and redesign typically follow before all side chains are successfully built. Evaluation criteria include residue packing, rotamer likelihoods, and steric constraints.

Constrained energy minimization may also be applied in the later stages of modeling to refine bond lengths and angles. In practice, one must be careful not to over minimize a model structure because of minimization's tendency to over-compact structures. Structure collapse is a serious side effect of energy minimization performed in the absence of solvent. Without solvent there are not enough contributions to the energy function to counterbalance the attraction of the molecule to itself. Structural collapse notwithstanding, mild minimization is useful for relaxing a model structure and revealing regions of questionable design.

The extent of model refinement varies with the application of the model. A model that is designed for the purpose of drug or inhibitor design usually must be refined only in areas near the active or binding site in question. In contrast, a model that will be compared closely to a number of refined crystal structures must be subjected to more intense scrutiny, because in this case even surface loops and remote 2° structure elements are significant.

INITIAL WORK: MODELING OF THERMOLYSIN

My fascination with the division between scaffold and catalytic or binding residues prompted the work described below. If one could demonstrate a successful graft of a catalytic or binding motif onto an unrelated scaffold, it would demonstrate the division between scaffold and motif. In other words, if the division between scaffold and motif is well defined and complete, we should be able to interchange motifs and scaffolds while retaining the activity of the motif. However, even if the line is clear between motif and scaffold, one would not expect just any scaffold to suffice for a given motif. Clearly, the geometry of the residues in the motif will dictate certain limitations on the geometry of an acceptable scaffold. Much of my work has focused on the search for appropriate new scaffolds for a variety of motifs.

In my earliest efforts, I attempted to identify new scaffolds for the functional machinery of thermolysin. These searches were performed visually, using MIDAS and a Silicon Graphics Iris workstation. I selected a set of key functional residues (Table 1-1) and noted that the catalytic motif in thermolysin is mounted on the periphery of a 4-helical bundle domain (Presnell and Cohen 1989) in the enzyme (Holmes and Matthews 1981). I proceeded to visually analyze a subset of the Protein Data Bank (Bernstein, Koetzle et al. 1977) that consisted of 4-helix bundle proteins. Although there were some interesting possibilities, I soon recognized that such an undertaking was extremely inefficient, and that some form of automated computational search tool was warranted. These early experiences formed the foundation for my efforts at UCSF which have been aimed primarily at the development and application of an automated geometric search strategy that will aid in attempts to graft motifs from protein structure. The computational tools that I have developed have taken advantage of distance matrix comparison as a means for identifying similar clusters of atoms in protein structures.

TABLE 1-1: **key functional residues from thermolysin**

Function	Residues
Catalysis	His 231, Asp 226, Glu 143
Zn binding	His 142, His 146, Glu 166
Substrate binding	Tyr 157

CHAPTER 2: DIST GRAFTING ALGORITHM

In the early stages of my work, I developed a simple search tool (DIST) that performed an a non-exhaustive, time-efficient search of a potential scaffold for sites with geometry suitable for a particular motif graft. DIST was tested on a number of known matches (trypsin catalytic triad vs. subtilisin catalytic triad, egg white lysozyme vs. T4 lysozyme and chymotrypsin inhibitor vs. BPTI) to verify its effectiveness. Subsequently, DIST was applied to the identification of suitable scaffolds for the "engraftment" of the catalytic triad of trypsin. Two strong candidates were identified, namely, Ribonuclease T1 and Staphylococcal Nuclease.

CHAPTER 3: EXPERIMENTAL GRAFT EVALUATION

Good computational tools are rarely developed without experimental insight. To evaluate our designs, it was essential that the grafted protein be prepared and assayed. Although the Ribonuclease T1 design appeared more promising from a pure modeling perspective, we were unable to acquire the necessary reagents for mutagenesis and expression of this enzyme. Fortunately, the Staphylococcal Nuclease sequences and vectors were accessible, and I was able to prepare the grafted protein. A number of major limitations observed in the assay systems available left a somewhat incomplete evaluation of the SNase/trypsin graft. As a result I realized that experimental feedback is much more readily obtained for grafts of binding motifs rather than catalytic motifs. Therefore, I shifted my computational and modeling efforts toward grafting of binding

motifs, while maintaining the belief that over the long term, these approaches would be applicable to catalytic grafts.

CHAPTER 4: GRAFTER ALGORITHM

Further analysis of the DIST algorithm revealed that it does not explore all potential matches for a particular motif in a given scaffold. In addition, the DIST algorithm is prone to spend extensive periods of time searching areas of the scaffold that are very unlikely to produce matches. These realizations prompted me to write a second-generation search tool, GRAFTER, that shares certain aspects of the DIST algorithm. GRAFTER performs more complete searches than DIST and focuses on areas of the scaffold that are most likely to generate matches. GRAFTER, like DIST, was tested using certain known matches, and was then applied to the design of a series of novel grafts. Both DIST and GRAFTER eliminate the drudgery of the initial search for potential graft sites. However, once a list of possible sites has been identified, it is still necessary to evaluate the sites visually using a graphical display program.

CHAPTER 5: THE GRAFTER SUITE - EVALUATION MODULES

To reduce the human effort involved in the evaluation of potential grafts, as well as to eliminate human bias, I developed a series of post-processing programs that combine to provide a systematic ranking scheme for grafts identified by GRAFTER. These post-processing modules are a steric check (STERIC), an accessibility weighted root-mean square deviation (ARMS), a global orientation evaluator (ORIENT) and a statistical score combiner (GSTAT). These new evaluation tools were tested on a series of known matches, and were also applied to a novel graft design.

CHAPTER 6: MODELING OF DHFR

Residual sequence homology between proteins that have grown disparate in sequence through evolution can be taken advantage of when modeling a protein of unknown structure. As two proteins diverge from a single ancestor, sequence is altered much more rapidly than structure. In fact, proteins with extremely similar structures can retain very little sequence homology (Ploegman, Drent et al. 1978).

My final project has been the identification of potential inhibitors for an enzyme (DHFR from *Cryptosporidium Parvum*) with unknown structure. Dihydrofolate reductase (DHFR) is a target for the antineoplastic agent methotrexate and the antibacterial agent trimethoprim (Oefner, D'Arcy et al. 1988). The clinical success of these agents has made DHFR a common target in so-called rational drug design. However, these compounds are not effective agents against DHFR from *C. Parvum*. DHFR was selected as the target in this study in an attempt to develop clinically significant agents against *Cryptosporidium Parvum*.

C. Parvum is a small protozoan that had minimal biomedical significance before 1980 (Fayer and Ungar 1986). Beginning around 1982 this parasite was increasingly found to be a cause of diarrheal illness in humans and some domesticated animals. In immunocompetent individuals *C. Parvum* typically causes diarrheal illness that resolves spontaneously within a month. However, in immunocompromized patients the diarrhea is usually prolonged and life-threatening. Currently, there are no effective therapies for cryptosporidiosis.

There are a number of known DHFR structures (human, chicken, *L. Caseii*) which are effective starting templates for modeling of the unknown DHFR (Bolin, Filman et al. 1982; Oefner, D'Arcy et al. 1988; McTigue, Davies et al. 1992). These proteins are in the range of 20-35% identical to the unknown DHFR. By aligning the model sequence to the structurally aligned sequences of the three known DHFRs, I was able to use the known structures to build a model for DHFR from *C. Parvum*.

Conserved secondary structure elements (i.e. helices, strands, turns) amongst the known DHFR structures were used as guides for building the corresponding regions in the model. Loops were constructed using the BLoop algorithm (Ring and Cohen 1994) which can generated a collection of loop structures that satisfy sequence, residue and distance constraints. Side chain orientations were extracted from known structures at positions of residue similarity. Otherwise, side chains were built based on rotamer preferences (McGregor, Islam et al. 1987; Ponder and Richards 1987; Dunbrack and Karplus 1993). Once all side chains were built, the model was evaluated for packing (Gregoret and Cohen 1990), bad contacts (BIOSYM Technologies, San Diego, CA) and rotamer propensities. Any questionable residues were studied and in many cases were remodeled. I repeated the evaluation/remodeling cycle until an acceptable model was generated. Finally, constrained energy minimization using AMBER (Weiner and Kollman 1981) was applied to refine any bad bond lengths or angles that may have resulted from joining the separate structural building blocks. The finished model is suitable for ligand docking studies, since the active site is its most refined area. Some additional effort would be required to generate an acceptable model for more general structural studies.

CHAPTER 7: DOCKING LIGANDS TO THE DHFR MODEL

The model for cryptosporidium DHFR was used to perform ligand docking trials. The DOCK algorithm (DesJarlais, Sheridan et al. 1988) was used to extract potential ligands from a large database. The Available Chemicals Directory (ACD) (Molecular Design, Ltd., San Leandro, CA) was used in this study. DOCK probes the small molecule database using a negative image of a binding or catalytic site. This negative image is prepared by clustering spheres inside the site or assembling spheres based on atom centers from a known ligand. In this study, three sphere sets were used: a site-based set, a methotrexate-based set and a site/methotrexate hybrid set. Two scoring schemes were applied: contact-only scoring and force-field scoring. The results from the three

sphere sets and two scoring schemes generated 15,000 potentially redundant ligands. These ligands were evaluated visually using the MIDAS graphical display program (Ferrin, Huang et al. 1988) and the list was trimmed down to 119 ligands that appeared most promising.

FINAL THOUGHTS

My journey through graduate school has been an exploration of computational approaches for the study of protein structure. My travels have taken me through highly fertile ground that in many cases is just now being sown. This is a source of great excitement for me, as well as a source of extreme frustration. It is obvious that these tools over time will have immense bearing on the future of biology, chemistry and medicine. However, currently many have yet to bear fruit. Only a small number of experiments have been performed to evaluate the existing computationally tools. Without such experiments, optimization and improvement of these tools is severely limited. I look forward to the next decade in anticipation of significant advances, not simply in the development of computationally tools for protein study, but more importantly, to the application of such tools.

REFERENCES

Bartlett, P. A., G. T. Shea, et al. (1989). CAVEAT: A Program to Facilitate the Structure-derived Design of Biological Active Molecules. Molecular Recognition: Chemical and Biochemical Problems. Exeter, Royal Society of Chemistry: 182-196.

Bernstein, F. C., T. F. Koetzle, et al. (1977). "The Protein Data Bank: A Computer-Based Archival File for Macromolecular Structure." J. Mol. Biol. **112**(3): 535-542.

Bolin, J. T., D. J. Filman, et al. (1982). "Crystal structures of Escherichia coli and Lactobacillus casei dihydrofolate reductase refined at 1.7 A resolution. I. General features and binding of methotrexate." J Biol Chem **257**(22): 13650-62.

Corey, D. R. and P. G. Schultz (1989). "Introduction of a Metal-dependent Regulatory Switch into an Enzyme." J. Biol. Chem. **264**(7): 3666-3669.

Cronin, C. N., B. A. Malcolm, et al. (1987). "Reversal of Substrate Charge Specificity by Site-Directed Mutagenesis of Aspartate Aminotransferase." J. Am. Chem. Soc. **109**: 2222-2223.

DeGrado, W. F., Z. R. Wasserman, et al. (1989). "Protein Design, a minimalist approach." Science **243**(4891): 622-628.

DesJarlais, R. L., R. P. Sheridan, et al. (1988). "Using Shape Complementarity as an Initial Screen in Designing Ligands for a Receptor Binding Site of Known Three-Dimensional Structure." J. Med. Chem. **31**(4): 722-729.

Dunbrack, R. L. J. and M. Karplus (1993). "Backbone-dependent rotamer library for proteins. Application to side-chain prediction." J. Mol. Biol. **230**(2): 543-574.

Fayer, R. and B. L. Ungar (1986). "Cryptosporidium spp. and cryptosporidiosis." Microbiol Rev **50**(4): 458-83.

Ferrin, T., C. Huang, et al. (1988). "The MIDAS Display System." J. Mol. Graph. **6**: 13-37.

Graf, L., G. Hegyi, et al. (1988). "Structural and functional integrity of specificity and catalytic sites of trypsin." Int J Pept Protein Res **32**(6): 512-8.

Gregoret, L. M. and F. E. Cohen (1990). "Novel method for the rapid evaluation of packing in protein structures." J Mol Biol **211**(4): 959-74.

Handel, T. M., S. A. Williams, et al. (1993). "Metal ion-dependent modulation of the dynamics of a designed protein." Science **261**(5123): 879-885.

Hecht, M. H., J. S. Richardson, et al. (1990). "De novo design, expression, and characterization of Felix: a four-helix bundle protein of native-like sequence (published erratum appears in Science 1990 Aug 31;249(4972):973)." Science **249**(4971): 884-891.

Hedstrom, L., L. Szilagyi, et al. (1992). "Converting trypsin to chymotrypsin: the role of surface loops." Science **255**(5049): 1249-1253.

Hellinga, H. W. and F. M. Richards (1991). "Construction of New Ligand Binding Sites in Proteins of Known Structure: I. Computer-aided Modeling of Sites with Pre-defined Geometry." J. Mol. Biol. **222**: 763-785.

Higaki, J. N., B. L. Haymore, et al. (1990). "Regulation of Serine Protease Activity by an Engineered Metal Switch." Biochem. **29**: 8582-8586.

Holmes, M. A. and B. W. Matthews (1981). "Binding of hydroxamic acid inhibitors to crystalline thermolysin suggests a pentacoordinate zinc intermediate in catalysis." Biochemistry **20**(24): 6912-20.

Kamtekar, S., J. M. Schiffer, et al. (1993). "Protein design by binary patterning of polar and nonpolar amino acids." Science **262**(5140): 1680-1685.

Kuntz, I. D. (1992). "Structure-based strategies for drug design and discovery." Science **257**(5073): 1078-1082.

McGregor, M. J., S. A. Islam, et al. (1987). "Analysis of the relationship between side-chain conformation and secondary structure in globular proteins." J Mol Biol **198**(2): 295-310.

McTigue, M. A., J. F. d. Davies, et al. (1992). "Crystal structure of chicken liver dihydrofolate reductase complexed with NADP+ and biopterin." Biochemistry **31**(32): 7264-73.

Nussinov, R. and H. J. Wolfson (1991). "Efficient detection of three-dimensional structural motifs in biological macromolecules by computer vision techniques." Proc. Natl. Acad. Sci. USA **88**: 10495-10499.

Oefner, C., A. D'Arcy, et al. (1988). "Crystal structure of human dihydrofolate reductase complexed with folate." Eur J Biochem **174**(2): 377-85.

Ploegman, J. H., G. Drent, et al. (1978). "The covalent and tertiary structure of bovine liver rhodanese." Nature **273**(5658): 124-9.

Ponder, J. W. and F. M. Richards (1987). "Tertiary Templates for Proteins: Use of Packing Criteria in the Enumeration of Allowed Sequences for Different Structural Classes." J. Mol. Biol. **193**: 775-791.

Presnell, S. R. and F. E. Cohen (1989). "Topological distribution of four-alpha-helix bundles." PNAS **86**(17): 6592-6596.

Ring, C. S. and F. E. Cohen (1994). "Conformational Sampling Of Loop Structures Using Genetic Algorithms." Israel Journal Of Chemistry **34**(2): 245-252.

Scott, J. K. and G. P. Smith (1990). "Searching for peptide ligands with an epitope library." Science **249**(4967): 386-390.

Weiner, P. K. and P. A. Kollman (1981). "AMBER: Assisted model building with energy refinement. A general program for modeling molecules and their interactions." J. Comp. Chem. **2**: 287-303.

Wharton, R. P. and M. Ptashne (1985). "Changing the binding specificity of a repressor by redesigning an α-helix." Nature **316**: 601-605.

Wilks, H. M., K. W. Hart, et al. (1988). "A Specific, Highly Active Malate Dehydrogenase by Redesign of a Lactate Dehydrogenase Framework." Science **242**: 1541-1544.

Chapter 2:
The DIST Algorithm

INTRODUCTION

Engineering is a field that relies on a strict set of rules for the design and implementation of new devices, structures and applications. The rules involved are based on sound principles. Protein engineering, in contrast, is an experimental field that explores the relationship between changes in protein structure and protein function. For protein engineering to evolve into a "true" engineering field, rules must be discovered that allow systematic design of novel proteins. In my experience, computers are excellent tools for systematic analysis and design. By applying computational tools to the study of protein structure and to the design of new proteins, I hope to take a step toward realizing protein engineering.

My efforts are aimed at functional grafting between proteins of known structure. If one could select a functional group of residues from one protein and build them onto a second protein while retaining their relative geometry, one might retain their function.

Identification of structural similarity between molecules is a subtle and difficult problem. It is influenced by how we define similarity, and how we perform our search. Manual comparisons using graphics terminals and the researcher's personal judgment are prone to bias, inefficiency and incompleteness. A computational approach to the search can eliminate much of the bias, can increase efficiency and in the ideal case, can perform a complete search. I consider these three factors to be the goals of a computational algorithm for geometric search applications.

Most known protein structures have been obtained through X-ray diffraction studies of crystals. Each such structure is represented in a particular reference frame, normally dictated by crystal axes. Unfortunately, this means that even two structures of identical proteins may not be in the same coordinate frame. A geometric search tool for grafting must somehow perform its comparison in a coordinate-frame independent manner. Distance matrices provide a simple, frame-independent representation of points

in space. Because each entry in a distance matrix is a distance, the geometries represented are completely relative, and therefore are frame-independent. A number of protein structure comparison algorithms have been developed that make use of distance matrices (Phillips 1970; Nishikawa and Ooi 1974; Sippl 1982).

I developed the DIST program in an attempt to satisfy the computational goals set forth above, i.e. to be an unbiased, efficient and complete means for comparing molecular geometries. The algorithm underlying DIST is based on distance matrix comparison. The process is accelerated by organizing the comparison steps so that only the most likely pairs are compared. The program compares a motif containing a few residues to a scaffold containing many residues. All structural groups in the scaffold that spatially align with the motif are reported.

The completed program was applied to a number of test cases to confirm its vigor. Enzyme and protein pairs that possess similar functions were analyzed by DIST: trypsin and subtilisin (Alden, Birktoft et al. 1971; Walter, Steigemann et al. 1982; Bryan, Pantoliano et al. 1986; Graf, Hegyi et al. 1988), egg white and T4 lysozymes (Matthews, Grütter et al. 1981; Matthews, Remington et al. 1981; Grutter, Weaver et al. 1983; Weaver, Grutter et al. 1984), and BPTI and chymotrypsin inhibitor (Bolognesi, Gatti et al. 1982; Eguchi and Yamamoto 1988; McPhalen and James 1988). The DIST results were examined to determine whether the known functional alignments (e.g. trypsin catalytic triad to subtilisin catalytic triad) could be detected by DIST. Subsequently, I used the program to design novel proteins, by identifying sites in proteins that had suitable geometries for grafting a known motif.

METHODS

How do we determine whether two distances are matched? The most stringent possibility is to require that the distances be exactly the same. For practical applications,

such a definition is unrealistic. A degree of tolerance is required to loosen the matching constraints to allow for real-world geometric differences. There are a number of ways that tolerance might be defined for the purposes of geometry comparison. Initially a fixed tolerance was implemented in DIST:

$$d_{ij} \text{ matches } d_{kl} \text{ if } d_{kl} - tol \leq d_{ij} \leq d_{kl} + tol$$

However, it became obvious that this tolerance scheme was too constraining for large distances and too relaxed for small distances. Instead, a fractional tolerance was implemented:

$$d_{ij} \text{ matches } d_{kl} \text{ if } d_{kl} - (d_{ij} \cdot tol_{\%}) \leq d_{ij} \leq d_{kl} + (d_{ij} \cdot tol_{\%})$$

This matching criterion scales the tolerance to the distances being compared.

A distance matrix may be used to describe a set of points in space. The matrix contains relative distances between every pair of points. This produces a description that is independent of absolute coordinates, but lacks handedness information. Hence, an object and its exact mirror image will have identical distance matrices (Figure 2-1).

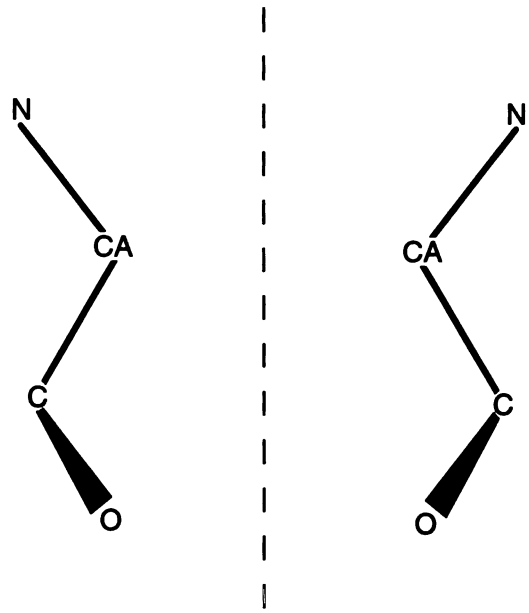
Distance matrices provide a simple means for comparing molecular structures. With small molecules their use is complicated by the occurrence of exact mirror images, i.e. structures that differ only by chirality. However, this is not a problem with macromolecules, which may be related as pseudo-mirror images, but are differentiated by local handedness (e.g. chirality of individual residues).

Comparison of distance matrices is simple when a point-to-point correspondence has already been identified. Without such an alignment, one must perform a combinatorial search to find the correct or best alignment. Simple permutation techniques will suffice when the two matrices are small, but the computational complexity increases rapidly with the number of points in each matrix. For large sets of points it is necessary to simplify the combinatorial search. It is useful to describe a distance matrix using an index, herein defined as a single row from the matrix. The index is identified by its reference atom, which is the atom label for the

FIGURE 2-1: mirror images have identical distance matrices

Two simple peptide sub-structures are shown that are related by a mirror plane. The figure shows that these structures have different coordinate sets, but identical distance matrices.

FIGURE 2-1: mirror images have identical distance matrices



COORDINATES

N	-11.714	7.720	15.167
CA	-12.021	6.630	14.259
C	-11.505	5.285	14.769
O	-11.200	4.401	13.936

N	11.714	7.720	15.167
CA	12.021	6.630	14.259
C	11.505	5.285	14.769
O	11.200	4.401	13.936

DISTANCE MATRICES

0	1.45	2.48	3.58
1.45	0	1.53	2.40
2.48	1.53	0	1.25
3.58	2.40	1.25	0

=

0	1.45	2.48	3.58
1.45	0	1.53	2.40
2.48	1.53	0	1.25
3.58	2.40	1.25	0

corresponding row in the matrix. Two indices are compared as a prelude to full distance comparison.

A pair of entries from two distance matrices are compared using the matching criterion (i.e. the fractional tolerance described previously). If the two entries are within the limits of the specified tolerance, they can be accepted as a matched pair.

Indices are sorted in order of increasing distance from the corresponding reference atoms. This facilitates rapid comparison of the two indices by a simple pair-wise approach. The initial pair containing the two reference atoms is accepted by default. The next atoms in the two indices are compared and accepted or rejected. If the pair is rejected, the point that is closer to its reference atom is eliminated from its index (Figure 2-2). This generates a new pair containing the point from the previously rejected pair that is farther from its reference along with a new point. This pair is evaluated and either accepted or rejected. This process is continued until one of the indices has no more points to compare. If the resulting trimmed indices contain at least a minimum number of pairs, the pairwise-aligned indices are used to assemble distance sub-matrices which are then subjected to a complete term-by-term comparison.

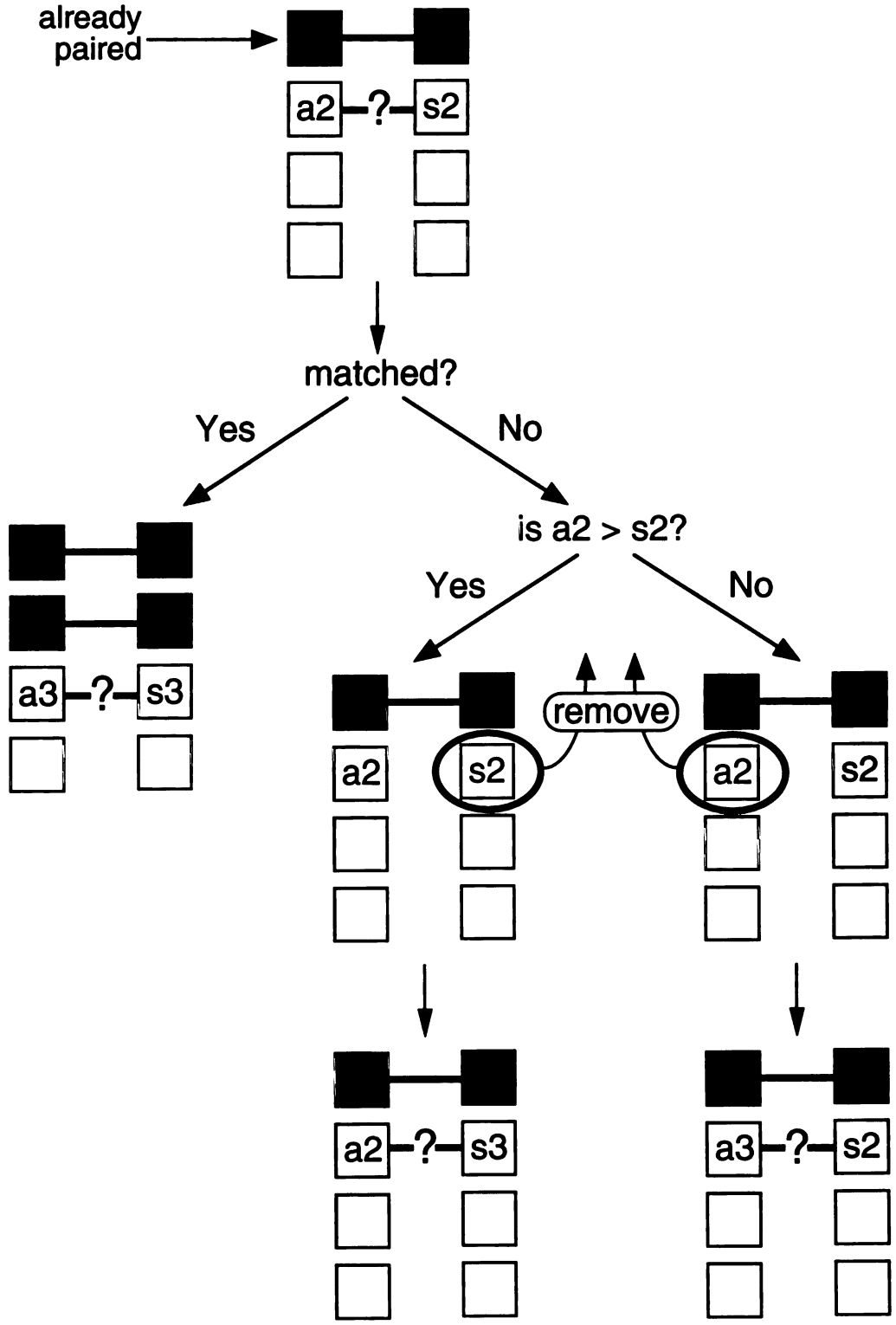
Although this simple approach to index comparison is rapid, it is flawed. It is guaranteed to be successful only in the case that there is a single exact match for every point in the first matrix and a zero tolerance is used. Otherwise, the process will not try all combinations, and suffers from being biased toward the first available path at any given pair. To overcome this, a forced 4-atom permutation has been introduced. Four atoms clearly define a geometry in three dimensions, and also describe handedness. Therefore, index comparison is modified so that every set of four atoms that 1) contain the reference atom and 2) are ordered as specified by the sorted index are used as the roots of indices to be compared.

A subset of atoms may be used to describe an individual residue. In most cases, for discontinuous motifs a Cα–only representation will not suffice. It is necessary to

FIGURE 2-2: index comparison

A schematic representation of index comparison in the DIST algorithm. Darkened boxes indicate accepted atom pairs. "a2" and "s2" are the second entries in the active site and scaffold indices respectively. Pairs that are being considered are separated by a "?".

FIGURE 2-2: index comparison scheme



include both Cα and Cβ to describe the position for side chain departure from the backbone. In applications of DIST, Cα and Cβ have been used to describe the side chain position for each residue.

POCKET DETERMINATION

Active sites typically occur in pockets or clefts in proteins. Searches for binding or catalytic motifs can be simplified by limiting the analysis to a subset of the scaffold's atoms. Usually it is most appropriate to search in areas of the scaffold that are pockets or clefts. I have developed an approach for the quick identification of pocket residues in protein structures. The approach uses differences in residue accessibility (Lee and Richards 1971; Presnell 1991) to identify pocket residues. There are two sub-strategies possible that are distinguished by whether or not the protein structure is associated with a substrate, inhibitor and/or cofactor.

When a structure contains a substrate, inhibitor and/or cofactor, accessibility calculations are performed twice, once for the protein alone, and once for the protein plus any associated molecule(s) (Figure 2-3a). Total accessibilities are compared for each residue in the presence and absence of the substrate, inhibitor and/or cofactor. Residues whose accessibilities are different for the two cases are considered to be pocket residues.

When no associated molecules are available to identify pockets, a second approach is used (Figure 2-3b). Accessibilities are calculated for the protein using a small probe and also a large probe. If a residue is accessible to a small probe, but not a large probe, it is likely to be in a pocket. Residues that are not accessible to either probe are rejected as core residues. Residues that are accessible to both probes are rejected as surface residues. The remaining residues are accepted as pocket/cleft residues.

FIGURE 2-3A: pocket determination with cofactor

A 1.4Å radius probe (dark circle) is rolled over the surface of a protein with and without the cofactor. Accessibilities are tabulated for all residues. Those residues whose accessibilities are different in the two cases are considered pocket residues.

FIGURE 2-3A: pocket identification with cofactor

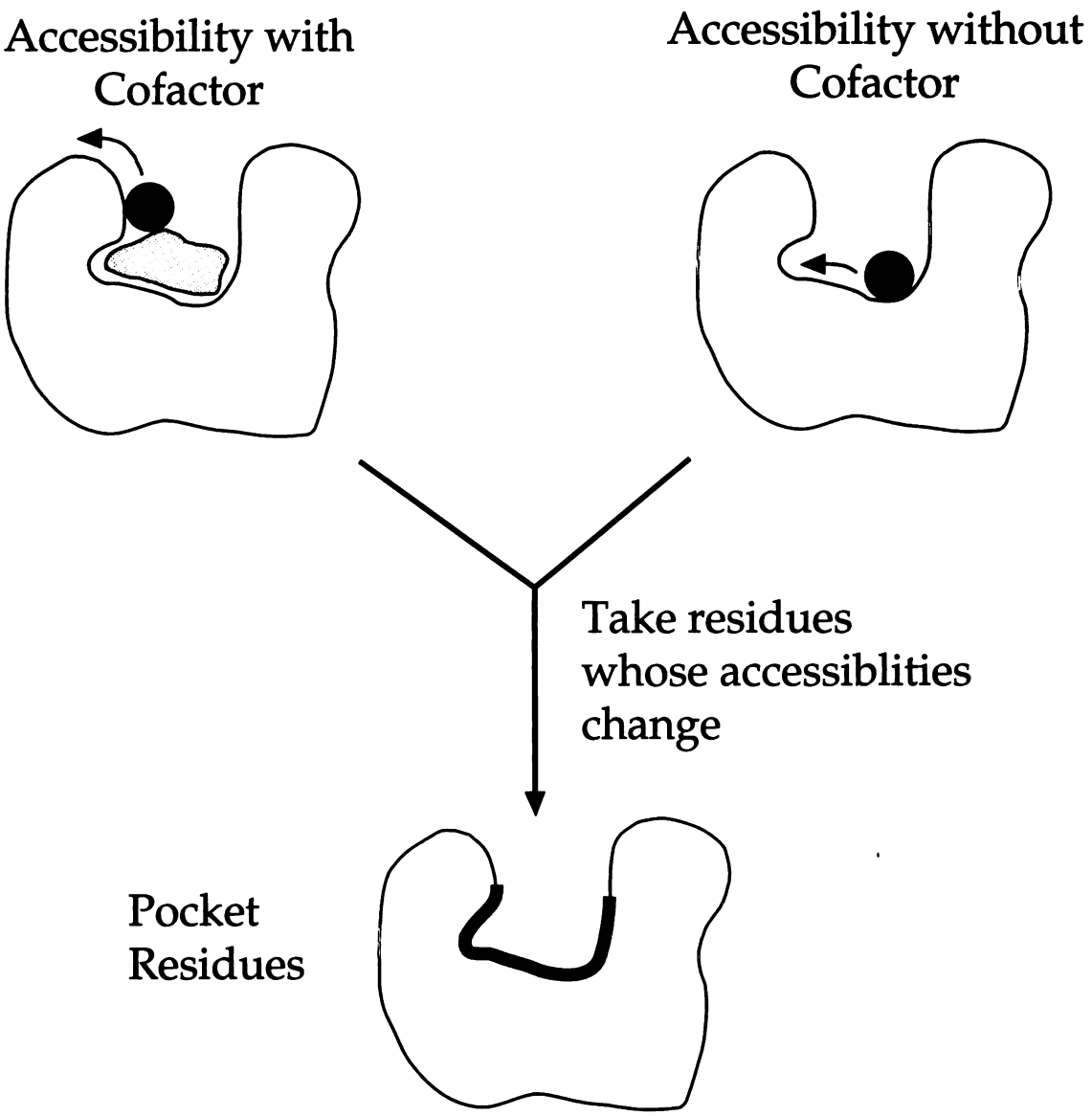
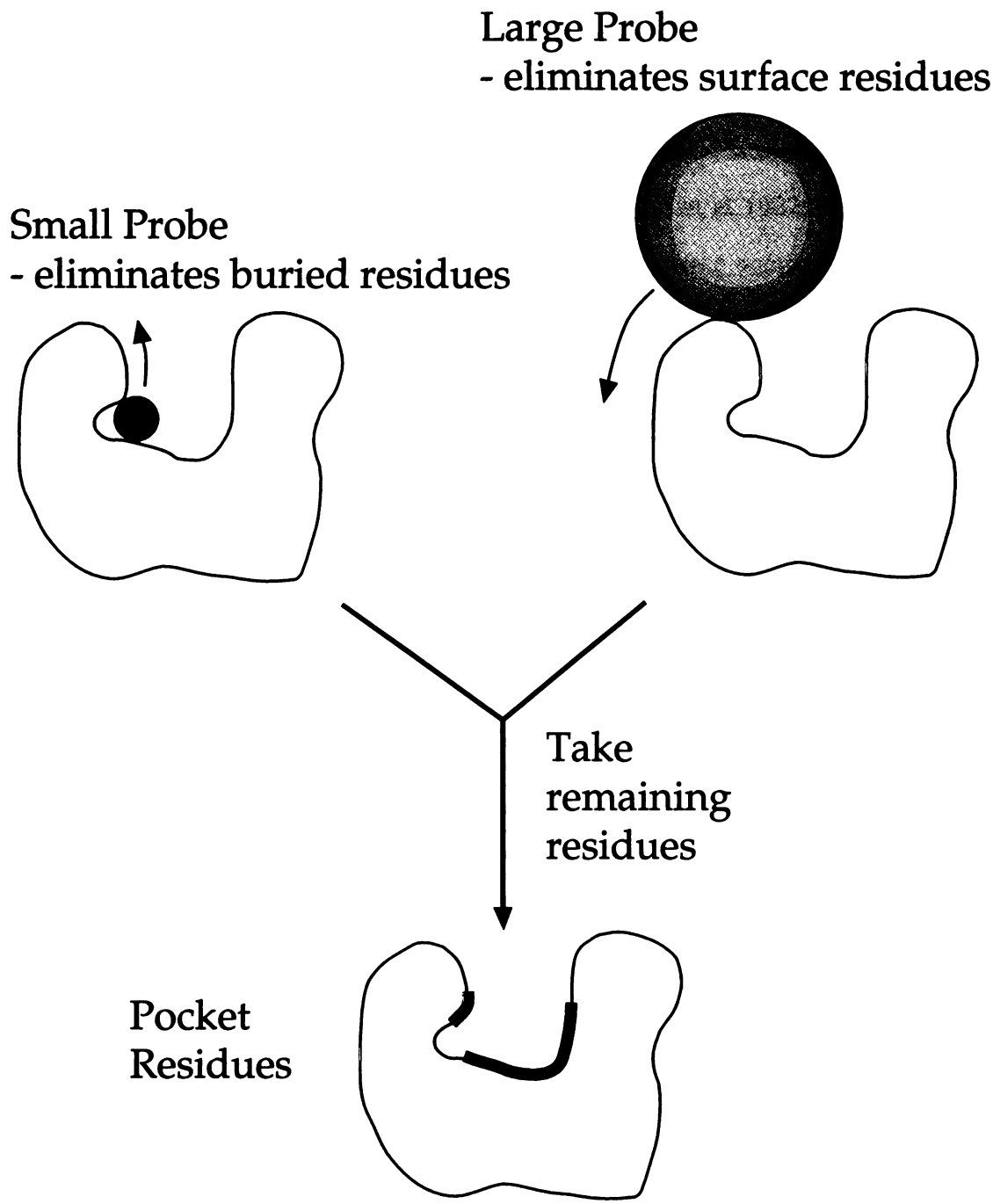


FIGURE 2-3B: pocket determination without cofactor

A small probe (1.4Å radius, dark circle) and a large probe (≥10Å radius, light circle) are rolled over a protein surface. Accessibilities are calculated in both cases for all residues. Those residues that are accessible to the small probe, but inaccessible to the large probe are pocket residues.

FIGURE 2-3B: pocket identification without cofactor



Test Comparisons

The motif from one protein was compared to the pocket region from a protein with similar function (e.g. trypsin catalytic triad vs. the pocket from subtilisin)using DIST. Any matches that aligned the functional motifs of the two proteins were recorded. Based on previously observed similarities, the trypsin/subtilisin (Alden, Birktoft et al. 1971; Walter, Steigemann et al. 1982; Bryan, Pantoliano et al. 1986; Graf, Hegyi et al. 1988), egg white/T4 lysozyme (Matthews, Grütter et al. 1981; Matthews, Remington et al. 1981) and chymotrypsin inhibitor/BPTI (Bolognesi, Gatti et al. 1982; Eguchi and Yamamoto 1988; McPhalen and James 1988) protein pairs were compared.

Novel Searches

A database of scaffolds was prepared by extracting pockets from the structures of proteins in the preliminary database. The desired motif (e.g. catalytic triad from trypsin (Walter, Steigemann et al. 1982)) was compared to each of the scaffold pockets using the DIST program. The highest scoring (i.e. lowest RMS) matches were then examined by eye using the MIDAS graphical display program (Ferrin, Huang et al. 1988).

Results

Test Cases

The DIST algorithm has been tested on a number of comparisons between motifs with known similarity. The catalytic triad from subtilisin was successfully identified using DIST and the catalytic triad of trypsin. In the reverse experiment, where the trypsin molecule was scanned using the subtilisin triad, DIST also successfully identified the trypsin catalytic triad. Parameters for all DIST comparisons are summarized in Table 2-1. Similar comparisons between egg white lysozyme and T4 lysozyme, and between

TABLE 2-1: summary of parameters used in DIST comparisons

Active Site (AS)	AS File	# Res.[1]	Scaffold (Scaf.)	Scaf. File	# Res.[2]	Tol.[3]	Min. Match[4]	Rank[5]	Residue Atoms
trypsin	1ntp	6	subtilisin	2sec	60	0.25	6	3	CA, CB
subtilisin	2sec	6	trypsin	1ntp	26	0.25	6	1	CA, CB
egg white lysozyme	1lyz	12	T4 lysozyme	2lzm	38	0.25	7	1	CA, CB
T4 lysozyme	2lzm	9	egg white lysozyme	1lyz	31	0.35	6	1-4, 5-7	CA, CB
chymotrypsin inhibitor	2sni	32	BPTI/trypsinogen	2tgp	29	0.10	7	1–6	CA, CB
trypsin (rotamer) triad	2ptn	9	scaffold database	—	—	0.40	8	—	CA, CB, N
trypsin (rotamer) diad	2ptn	6	scaffold database	—	—	0.30	6	—	CA, CB, N

[1] number of residues in active site file

[2] number of residues in scaffold file

[3] tolerance used for DIST comparison

[4] minimum number of atoms for a match to be accepted

[5] rank of the known correct match with respect to all matches found

chymotrypsin inhibitor and bovine pancreatic trypsin inhibitor resulted in expected matches. The accepted residue pairings for each case are shown in Table 2-2. In many cases, the pairing identified by DIST is a subset of the complete pairing. (idiosyncrasies of DIST and certain motif/scaffold sets prevent all atoms from being found in certain cases).

TRYPSIN SEARCH

DIST was used to probe a database of protein scaffolds for sites similar to the catalytic triad of trypsin. The database was prepared with 2 constraints: 1) there must be a structure for each protein and 2) each protein must have a well-documented expression system. The resulting database contains 48 single chains from a total of 10 proteins as shown in Table 2-3. Pockets were identified for each structure as described above. The resulting subsets of the protein structures make up the scaffold pocket database. The reader will note that in some cases, a number of structures are included for a given protein. This follows from the desire to identify any matches for a particular motif. There is enough structural variation within a single protein that matches are found in only one structure out of a group of structures for the same protein. Inclusion of all structures for a scaffold enhances the chances of finding any matches for a given motif.

To maximize the chance that the grafted side chains will conform to standard (i.e. stable) rotamers, while still occupying the same relative orientations as those in active trypsin, a modified search geometry was used. The standard rotamers of Ser, His and Asp were superimposed onto the catalytic triad of trypsin, using only the side chain atoms for alignment. This produces an active site with side chains in identical positions to the native trypsin active site, but the main chain positions are modified so that each residue conforms to standard rotamer geometry.

This modified active site was used to probe the scaffold pocket database twice. The first time, the full triad was used as a probe. The second time, only the

TABLE 2-2: known correct alignments

a)

Trypsin	Subtilisin
Ser195	Ser221
His57	His64
Asp102	Asp32

b)

Hen Egg White Lysozyme	T4 Lysozyme
Glu35	Glu11
Asp52	Asp20
Gln57	Gly30
Ile58	His31
Asn59	Leu32
Ala107	Phe104
Trp108	Glu105

c)

Chymotrypsin Inhibitor	BPTI
Val57	Pro13
Thr58	Cys14
Met59	Lys15
Glu60	Ala16
Tyr61	Arg17

TABLE 2-3: scaffold database

Proteins were chosen that have well described mutagenesis and expression systems. All available structures for a given protein were used.

PDB File	Chain(s)	Protein	Reference
8atc	A,B,C,D	Aspartate Transcarbamoylase	(Ke, Lipscomb et al. 1988)
1at1	A,B,C,D	Aspartate Transcarbamoylase	(Gouaux and Lipscomb 1990)
2at1	A,B,C,D	Aspartate Transcarbamoylase	(Gouaux and Lipscomb 1990)
3at1	A,B,C,D	Aspartate Transcarbamoylase	(Gouaux and Lipscomb 1990)
4at1	A,B,C,D	Aspartate Transcarbamoylase	(Stevens, Gouaux et al. 1990)
5at1	A,B,C,D	Aspartate Transcarbamoylase	(Stevens, Gouaux et al. 1990)
7at1	A,B,C,D	Aspartate Transcarbamoylase	(Gouaux, Stevens et al. 1990)
8at1	A,B,C,D	Aspartate Transcarbamoylase	(Gouaux, Stevens et al. 1990)
2gap	A,B	Catabolite Gene Activator Protein	(Weber and Steitz 1984)
3cla	–	Chloramphenicol Acetyltransferase	(Leslie 1990)
1r1e	E	EcoR1 Endonuclease	(Kim, Grable et al. 1990)
2rnt	–	Ribonuclease T1	(Koepke, Maslowska et al. 1989)
1snc	–	Staphylococcal Nuclease	(Loll and Lattman 1989)
2sns	–	Staphylococcal Nuclease	(Cotton, Hazen et al. 1979)
1s01	–	Subtilisin BPN (mutant)	(Pantoliano, Whitlow et al. 1989)
2ypi	A,B	Triose Phosphate Isomerase	(Lolis and Petsko 1990)
4ts1	A,B	Tyrosyl tRNA Synthetase	(Brick and Blow 1987)
4xia	A,B	Xylose Isomerase	(Henrick, Collyer et al. 1989)
5xia	A,B	Xylose Isomerase	(Henrick, Collyer et al. 1989)

Ser195/His57 dyad was used as a probe. The resulting matches from each run were ordered separately by RMS deviation[1]. The top 50 matches from each search were analyzed on computer graphics using MIDAS (Ferrin, Huang et al. 1988) to identify those that had minimal steric clashes. Matches were eliminated based on orientation of the active site side chains with respect to the scaffold pocket and their availability to a potential substrate. In addition, the matches were examined for effective positioning of the active site side chains, while maintaining reasonable side chain orientations as compared to known standard rotamers. The best match from the triad search involves the Staphylococcal Nuclease (SNase) enzyme as a scaffold. The best match in the dyad search involves the Ribonuclease (RNase) T1 enzyme as a scaffold. The substitutions involved in each of these hypothetical grafts are summarized in Table 2-4 and are displayed in the context of their scaffolds in Figures 2-4 and 2-5. These figures display the scaffold backbones with the grafted residues displayed explicitly. In each figure a peptide is shown to indicate the orientation of a substrate with respect to the catalytic triad. Figures 2-4 and 2-5a show the substrate oriented based on natural substrate placement in known trypsin/inhibitor structures (Bolognesi, Gatti et al. 1982; Eguchi and Yamamoto 1988; McPhalen and James 1988). It is clear from Figure 2-5a that a substrate will have difficulty approaching the triad from the natural direction because of steric conflicts with backbone scaffold structure. Figure 2-5b illustrates an alternate substrate approach that is the result of reflection about the catalytic triad.

[1]Within the DIST program itself, a Cartesian RMS deviation (after least-squares fit superpositioning) is used. Unlike an RMS deviation by differences in interatomic distances, an RMS deviation by least-squares superposition takes into account any handedness in the structures. The two methods perform similarly except that when there is a loose fit, the distance difference method does not penalize local inversions adequately (Cohen and Sternberg 1980)

TABLE 2-4: catalytic triad substitutions in RNase T1 and SNase

Catalytic Residue	Substitution in RNase T1	Substitution in SNase
Serine	P73S	R35S
Histidine	Y38H	D83H
Aspartate	V33D	Y85D

FIGURE 2-4: triad graft on RNase T1

The RNase T1 scaffold with a catalytic triad graft. The triad residues are highlighted and a hypothetical substrate is displayed. The substrate is oriented as it would be in known serine proteases.

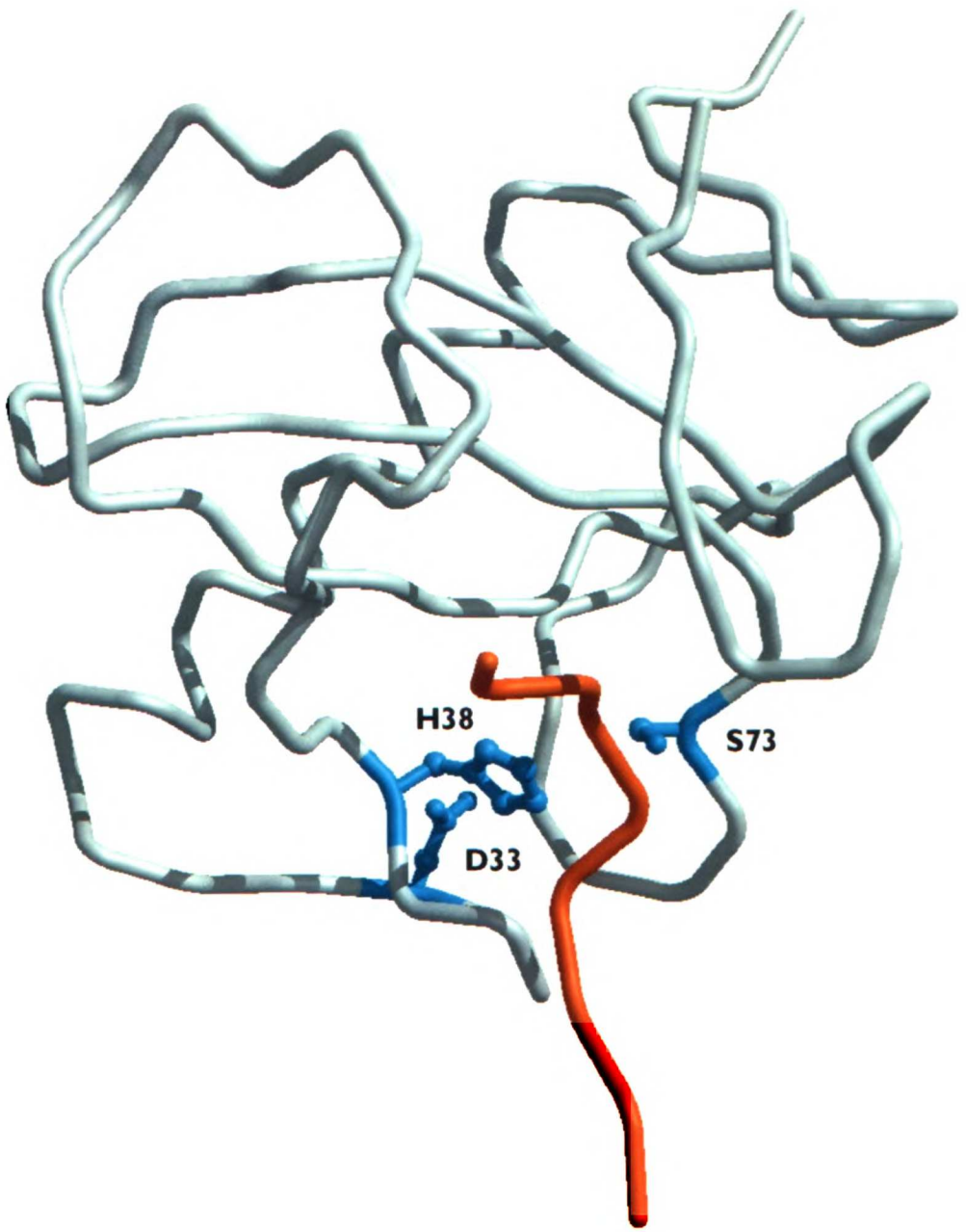


FIGURE 2-4

FIGURE 2-5A: triad graft on SNase

The SNase scaffold with a catalytic triad graft. The triad residues are highlighted and a hypothetical substrate is displayed. The substrate is oriented as it would be in known serine proteases.

FIGURE 2-5B: triad graft on SNase - reversed substrate

The SNase scaffold with a catalytic triad graft. The triad residues are highlighted and a hypothetical substrate is displayed. The substrate is in a position that is an approximate mirror image of known protease substrates with respect to the plane of the catalytic triad.

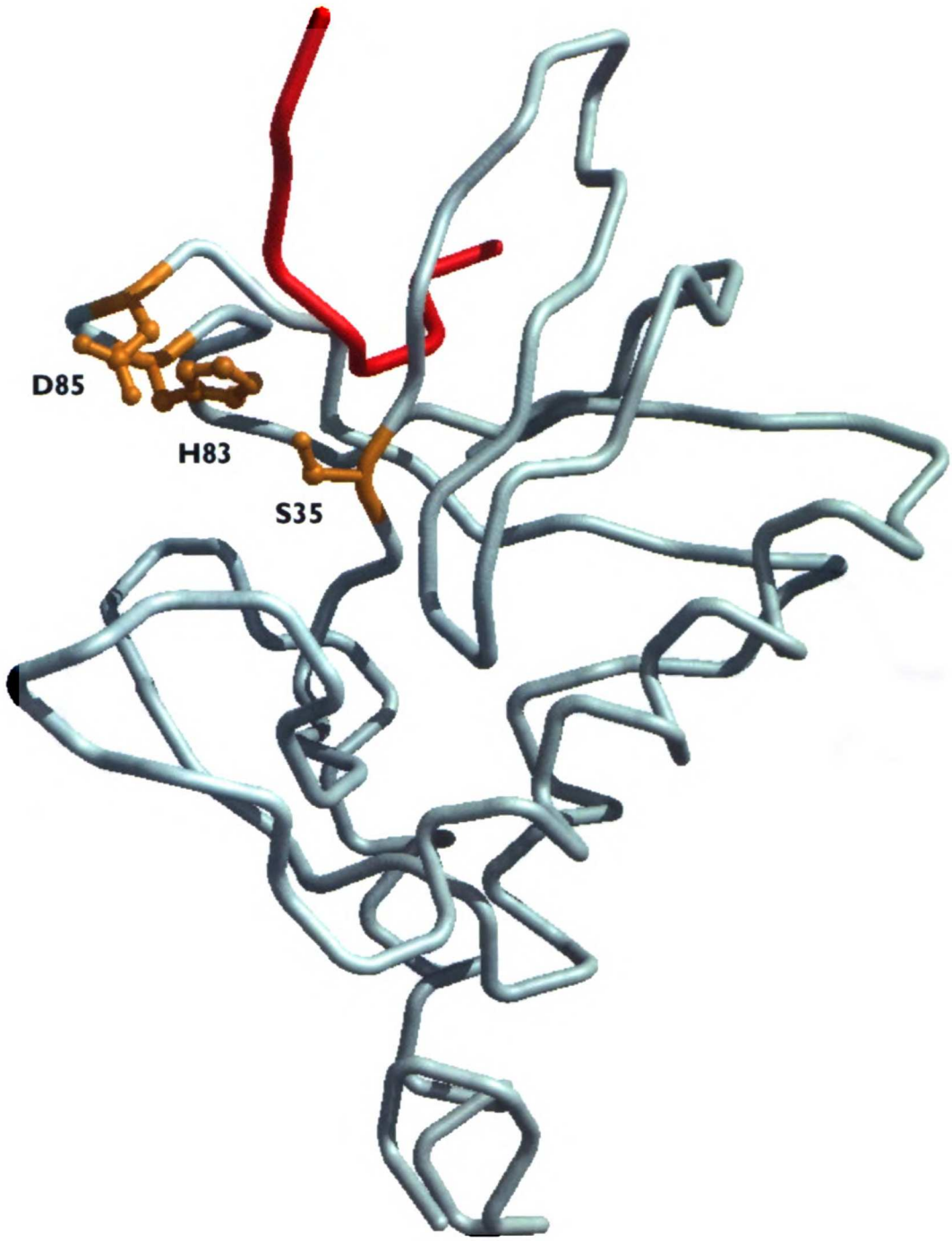


Figure 2-5a

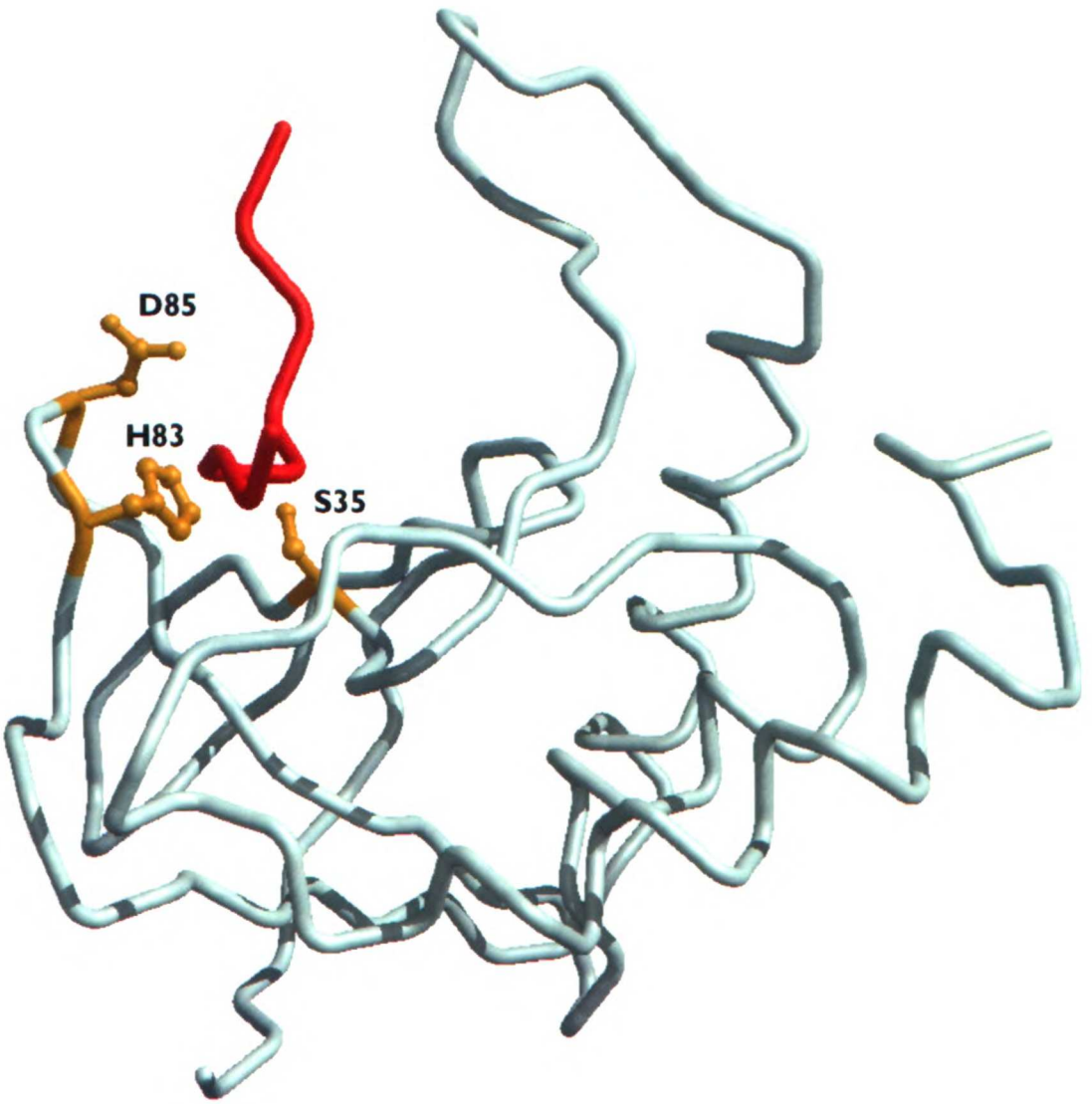


FIGURE 2-5B

DISCUSSION

The development of a computational algorithm requires extensive testing to verify its results. This validation involves two discrete steps: 1) verification that the algorithm performs as expected when the results are already known and 2) demonstration that any novel results generated are interesting. The simplest approach toward validation in Step 1 is comparison of a motif to the scaffold from the same protein. DIST consistently succeeds at identifying a motif under these circumstances. Three pairs of proteins have been used for additional verification of Step 1:

1) trypsin and subtilisin (catalytic triads)

2) egg white and T4 lysozyme (catalytic residues)

3) BPTI and chymotrypsin inhibitor (inhibitory residues)

In each of the three test sets above, DIST succeeded at identifying matches between the functionally/structurally similar proteins. However, DIST is not always able to identify all matched pairs between the proteins. This inability points to a flaw in the DIST algorithm that will be addressed below.

Step 2 of the validation process involves the identification of novel matches between a motif and a different protein scaffold. I focused my searches on the catalytic triad of trypsin. After computational and visual analysis, two protein scaffolds were identified that contain good sites for grafting of the trypsin triad, namely ribonuclease T1 (RNase T1) and staphylococcal nuclease (SNase). It is not surprising that in each graft, the relative orientations of the triad residues are quite similar to those found in trypsin. These two graft sites also stood out because of the orientation of the collective triad. In each case, one could envision a substrate or inhibitor approaching the triad without obstruction.

The RNase T1 graft is the better of the two structurally. Its triad is ideally oriented for access by a substrate. The three residues (Ser, His, Asp) have extremely similar relative geometry to those in trypsin. Although the relative triad geometry in the SNase graft is not quite as good as that in RNase T1, this graft still stands out as a close representation of a trypsin catalytic triad.

The testing with known matches revealed a flaw in the DIST algorithm. In some cases, certain atom pairs from a known match were not recognized by DIST. After further analysis of the algorithm, it became obvious that the path followed by DIST through the matrices being compared could potentially skip pairs. If more than one atom in matrix 2 could be matched to an atom in matrix 1, only the first would be tried. This would lead to the second one being skipped. This is illustrated in Figure 2-6.

CONCLUSION

DIST is a vast improvement over visual analysis as far as speed is concerned. A human being might be able to compare a motif to a scaffold in a few hours. This search would be biased and incomplete. In contrast, DIST can perform a comparable search in a few seconds. Although a DIST search is biased and may not be complete, it is at least as good, and probably much more thorough than a human search.

There is no question that DIST is a useful tool. Clearly, improvements must be made to reduce bias and enhance completeness, but it is a solid foundation. Hopefully, these improvements will not reduce DIST's speed drastically.

The two best matches for the trypsin catalytic triad (SNase, RNase T1) were considered for experimental analysis. We were unable to obtain the necessary materials to perform the mutagenesis and expression for RNAse T1, and chose to move forward with the SNase scaffold. The next chapter describes the experimental evaluation that was performed for the graft of the trypsin catalytic triad onto the SNase scaffold.

FIGURE 2-6: flaw in DIST index comparison algorithm

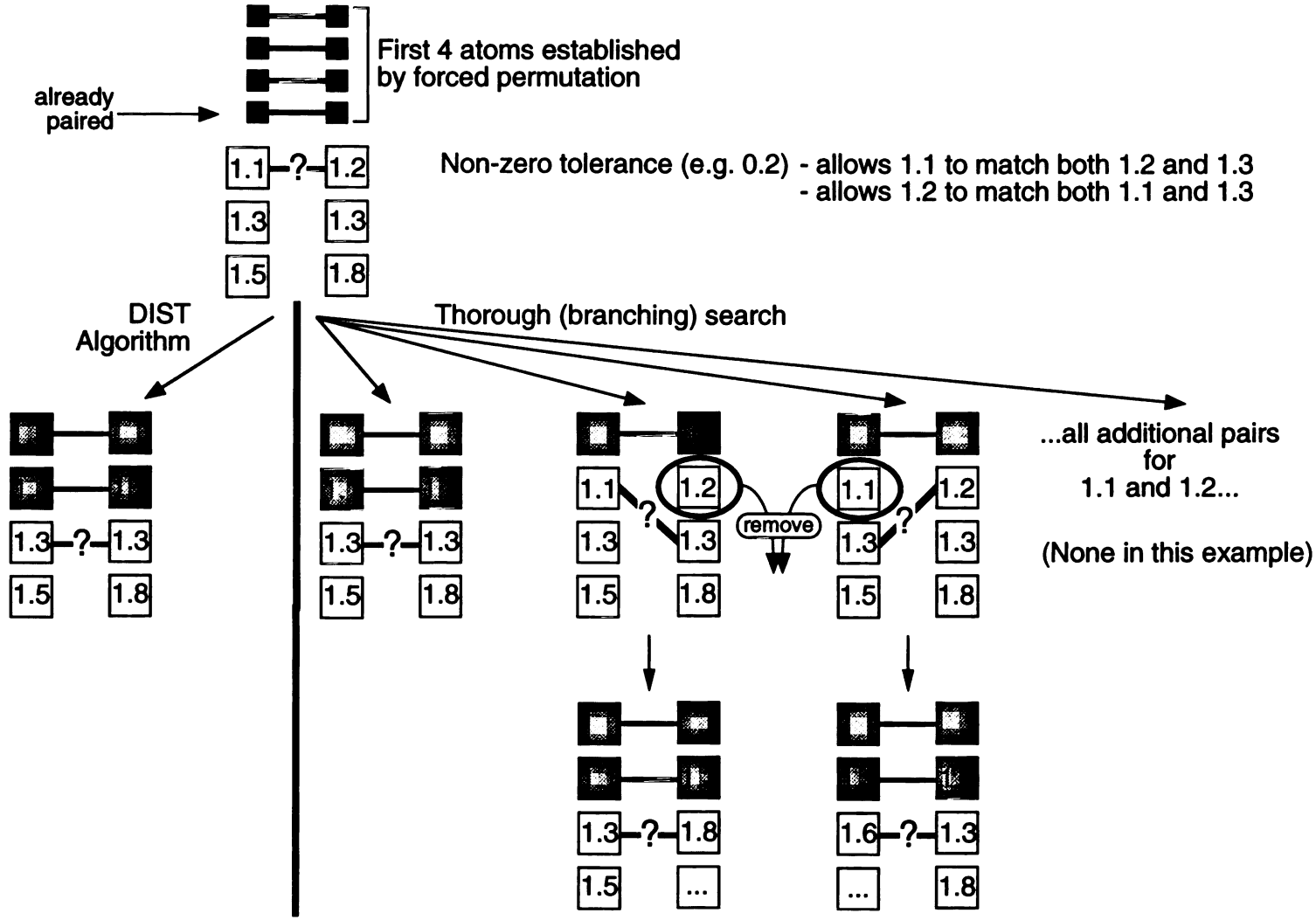
This schematic shows that the DIST index comparison scheme can overlook the correct match. If an atom in one index can be matched with more than one atom in the other index, the algorithm will select the atom with the smallest distance value. If an atom with a large distance value is the correct matched this will lead to overlooking the correct match.

Solid boxes indicate accepted pairs.

Numbers are distances.

"?" identifies the pair being compared.

FIGURE 2-6: flaw in DIST index comparison algorithm



REFERENCES

Alden, R. A., J. J. Birktoft, et al. (1971). "Atomic coordinates for subtilisin BPN' (or Novo)." Biochem Biophys Res Commun **45**(2): 337-44.

Bolognesi, M., G. Gatti, et al. (1982). "Three-dimensional structure of the complex between pancreatic secretory trypsin inhibitor (Kazal type) and trypsinogen at 1.8 A resolution. Structure solution, crystallographic refinement and preliminary structural interpretation." J Mol Biol **162**(4): 839-68.

Brick, P. and D. M. Blow (1987). "Crystal structure of a deletion mutant of a tyrosyl-tRNA synthetase complexed with tyrosine." J Mol Biol **194**(2): 287-97.

Bryan, P., M. W. Pantoliano, et al. (1986). "Site-directed mutagenesis and the role of the oxyanion hole in subtilisin." Proc Natl Acad Sci U S A **83**(11): 3743-5.

Cohen, F. E. and M. J. Sternberg (1980). "On the prediction of protein structure: The significance of the root-mean-square deviation." J. Mol. Biol. **138**(2): 321-333.

Cotton, F. A., E. E. Hazen Jr., et al. (1979). "Staphylococcal nuclease: proposed mechanism of action based on structure of enzyme-thymidine 3',5'-bisphosphate-calcium ion complex at 1.5-A resolution." Proc Natl Acad Sci U S A **76**(6): 2551-5.

Eguchi, M. and Y. Yamamoto (1988). "Comparison of serum protein inhibitors from various mammals, chicken and silkworms against four proteases." Comp Biochem Physiol [B] **91**(4): 625-30.

Ferrin, T., C. Huang, et al. (1988). "The MIDAS Display System." J. Mol. Graph. **6**: 13-37.

Gouaux, J. E. and W. N. Lipscomb (1990). "Crystal structures of phosphonoacetamide ligated T and phosphonoacetamide and malonate ligated R states of aspartate carbamoyltransferase at 2.8-A resolution and neutral pH." Biochemistry **29**(2): 389-402.

Gouaux, J. E., R. C. Stevens, et al. (1990). "Crystal structures of aspartate carbamoyltransferase ligated with phosphonoacetamide, malonate, and CTP or ATP at 2.8-A resolution and neutral pH." Biochemistry **29**(33): 7702-15.

Graf, L., G. Hegyi, et al. (1988). "Structural and functional integrity of specificity and catalytic sites of trypsin." Int J Pept Protein Res **32**(6): 512-8.

Grutter, M. G., L. H. Weaver, et al. (1983). "Goose lysozyme structure: an evolutionary link between hen and bacteriophage lysozymes?" Nature **303**(5920): 828-31.

Henrick, K., C. A. Collyer, et al. (1989). "Structures of D-xylose isomerase from Arthrobacter strain B3728 containing the inhibitors xylitol and D-sorbitol at 2.5 A and 2.3 A resolution, respectively." J Mol Biol **208**(1): 129-57.

Ke, H. M., W. N. Lipscomb, et al. (1988). "Complex of N-phosphonacetyl-L-aspartate with aspartate carbamoyltransferase. X-ray refinement, analysis of conformational changes and catalytic and allosteric mechanisms." J Mol Biol **204**(3): 725-47.

Kim, Y. C., J. C. Grable, et al. (1990). "Refinement of Eco RI endonuclease crystal structure: a revised protein chain tracing." Science **249**(4974): 1307-9.

Koepke, J., M. Maslowska, et al. (1989). "Three-dimensional structure of ribonuclease T1 complexed with guanylyl-2',5'-guanosine at 1.8 A resolution." J Mol Biol **206**(3): 475-88.

Lee, B. and F. M. Richards (1971). "The Interpretation of Protein Structures: Estimation of Static Accessibility." J. Mol. Biol. **55**: 379-400.

Leslie, A. G. (1990). "Refined crystal structure of type III chloramphenicol acetyltransferase at 1.75 A resolution." J Mol Biol **213**(1): 167-86.

Lolis, E. and G. A. Petsko (1990). "Crystallographic analysis of the complex between triosephosphate isomerase and 2-phosphoglycolate at 2.5-A resolution: implications for catalysis." Biochemistry **29**(28): 6619-25.

Loll, P. J. and E. E. Lattman (1989). "The crystal structure of the ternary complex of staphylococcal nuclease, Ca2+, and the inhibitor pdTp, refined at 1.65 A." Proteins **5**(3): 183-201.

Matthews, B. W., M. G. Grütter, et al. (1981). "Common precursor of lysozymes of hen egg-white and bacteriophage T4." Nature **290**: 334-335.

Matthews, B. W., S. J. Remington, et al. (1981). "Relation Between Hen Egg White Lysozyme and Bacteriophage T4 Lysozyme: Evolutionary Implications." J. Mol. Biol. **147**: 545-558.

McPhalen, C. A. and M. N. James (1988). "Structural comparison of two serine proteinase-protein inhibitor complexes: eglin-c-subtilisin Carlsberg and CI-2-subtilisin Novo." Biochemistry **27**(17): 6582-98.

Nishikawa, K. and T. Ooi (1974). "Comparison of homologous tertiary structures of proteins." J Theor Biol **43**(2): 351-74.

Pantoliano, M. W., M. Whitlow, et al. (1989). "Large increases in general stability for subtilisin BPN' through incremental changes in the free energy of unfolding." Biochemistry **28**(18): 7205-13.

Phillips, D. C. (1970). "The development of crystallographic enzymology." Biochem Soc Symp **30**: 11-28.

Sippl, M. J. (1982). "On the problem of comparing protein structures. Development and applications of a new method for the assessment of structural similarities of polypeptide conformations." J Mol Biol **156**(2): 359-88.

Stevens, R. C., J. E. Gouaux, et al. (1990). "Structural consequences of effector binding to the T state of aspartate carbamoyltransferase: crystal structures of the unligated and ATP- and CTP-complexed enzymes at 2.6-A resolution [published erratum appears in Biochemistry 1990 Dec 18;29(50):11146]." Biochemistry **29**(33): 7691-701.

Walter, J., W. Steigemann, et al. (1982). "On the disordered activation domain in trypsinogen. Chemical labelling and low-temperature crystallography." Acta Crystallogr. Sect. B **38**: 1462.

Weaver, L. H., M. G. Grutter, et al. (1984). "Comparison of goose-type, chicken-type, and phage-type lysozymes illustrates the changes that occur in both amino acid sequence and three-dimensional structure during evolution." J Mol Evol **21**(2): 97-111.

Weber, I. T. and T. A. Steitz (1984). "Model of specific complex between catabolite gene activator protein and B-DNA suggested by electrostatic complementarity." Proc Natl Acad Sci U S A **81**(13): 3973-7.

Chapter 3: Experimental Analysis of a Trypsin Graft

INTRODUCTION

Computational tools make the systematic design of novel proteins much simpler. If designed well, they remove much of the bias inherent in hand-built designs. However, analysis of computationally generated structures is vulnerable to extreme bias if similar tools are used for their analysis. Instead, a test case from the first DIST search was chosen for experimental evaluation. Of the two best designs for a trypsin catalytic triad graft (the RNase T1 and SNase scaffolds), the most enticing was RNase T1. Unfortunately, the necessary materials for mutagenesis and expression of RNase T1 were not available. Because a suitable system was available for SNase (courtesy of David Shortle) we chose to begin experimental evaluation using SNase.

The goal of this analysis was to generate a SNase mutant containing the triad residues from trypsin (Ser, His, Asp) in a conformation that generated measurable activity. With these residues in place, assays could be performed to identify any activity for the triad. Although proteolysis assays are obvious choices for comparison of trypsin and the SNase mutant, these assays are too specific for the purpose. The SNase mutant design contains only the catalytic triad from trypsin and does not incorporate any specificity pockets or auxiliary binding residues. With this in mind, a DIFP labeling assay was selected. DIFP is known to bind covalently to activated serines in a wide range of proteases. After the mutated SNase molecule was prepared containing the catalytic residues from trypsin, the DIFP assay was used to compare its activity to various wild type and mutant trypsins.

METHODS

The Staphylococcal Nuclease gene packaged in the phage vector M13mp9 was obtained from the laboratory of David Shortle. The mutagenesis primers shown below were synthesized using a PCR mate synthesizer (Applied Biosystems).

FIGURE 3-1: mutagenesis primers

```
5'-CA-ATG-ACA-TTC-AGC-CTA-TTA-3'     20 bases
      MET-THR-PHE-SER-LEU-LEU
                  (ARG)

5'-CAA-AGA-ACT-CAT-AAA-GAT-GGA-CGT-G-3'   25 bases
   GLU-ARG-THR-HIS-LYS-ASP-GLY-ARG
               (ASP)    (TYR)
```

In general the preparation of SNase mutants and their purification was performed as described by Shortle (Shortle and Meeker 1989; Shortle, Stites et al. 1990). Oligos were purified using a NENSORB (New England Nuclear) purification column. The method of Kunkel (Kunkel 1985) was used to introduce the desired mutations into the Staphylococcal Nuclease gene. The mutated gene sequence was confirmed using sequencing gels. The mutated gene was extracted from the mutagenesis vector by cleavage with SpeI and SphI[1] (New England Biolabs), and subsequently ligated into the pL9 expression vector. AR120 competent cells were transformed with the expression vector. Appropriate mutants were identified by Xho I[1] (New England Biolabs) digestion and then confirmed by sequencing. Protein prep cultures were grown up in SB/ampicillin media and purified by Urea extraction and Fast Flow S-Sepharose column (Pharmacia).

[1] all DNA modifying enzymes were used according to manufacturers' recommendations

TABLE 3-1: Assay #1 - DIFP Labeling

For each eppendorf tube, solutions were prepared TN1000 (1x) to obtain a final volume of 20µl. The mutants assayed and their final concentrations are listed below. Concentrations varied based on enzyme availability.

Enzyme	Final Concentration(µg/µl)
wild type bovine cationic trypsin	0.5
rat anionic trypsin - D102N mutant[2]	0.5
rat anionic trypsin - S195K mutant[2]	0.2
rat anionic trypsin - H57K mutant[2]	0.3
rat anionic trypsin - H57K, D102N mutant[2]	0.1
rat anionic trypsin - H57A, D102N mutant[2]	0.5
rat anionic trypsin - D189G mutant[2]	0.3
SNase - R35S, D83H, Y85D mutant (clone 1)	0.5

36 ng DIFP (2µCi) were added in 2 aliquots at room temperature over a 2 hour period. The reactions were loaded onto a protein gel flanked by molecular weight markers, and run at 250 volts for ~45 minutes. The gel was soaked in "Enhance " solution for 30 min. and then rinsed with ddH_2O for 10 min. After drying on filter paper, the gel and film were sandwiched between two enhancement screens and left at -70°C for 2 days.

TABLE 3-2: Assay #2 - DIFP Labeling (temperature/pH variation)

Wild type bovine cationic trypsin was compared to the D102N, S195A and H57K mutants of rat anionic trypsin under differ temperature, pH and reaction time conditions. Otherwise, conditions were similar to those used for Assay #1.

Condition	Temp	pH	Rxn. Time
1	RT	8.0	20 hr.
2	37°C	8.0	20 hr.
3	RT	10.0	20 hr.
4	RT	10.0	2 hr.
5	37°C	10.0	20 hr.
6	RT	8.0	20 hr.

[2]all mutants of rat anionic trypsin provided by the laboratory of Charles Craik.

TABLE 3-3: Assay #3 - wild type comparison

6 samples of wild type bovine cationic trypsin and 6 samples of rat anionic trypsin were run through an assay similar to Assay #1. The final concentrations of the trypsin are shown in below:

Sample	Enzyme	Final Conc.
1	wild type bovine cationic trypsin	2 μM
2	" "	0.4 μM
3	" "	0.08 μM
4	" "	16 nM
5	" "	3.2 nM
6	" "	0.64 nM
7	wild type rat anionic trypsin	0.9 μM
8	" "	0.3 μM
9	" "	0.1 μM
10	" "	33 nM
11	" "	11 nM
12	" "	3.8 nM

The film was exposed for 3 days.

TABLE 3-4: Assay #4 - inhibition with DIFP vs. PMSF using fluorogenic substrate

Wild type bovine cationic trypsin and wild type rat anionic trypsin were assayed for inhibition by DIFP and PMSF using a fluorogenic substrate:

Z-Gly-Pro-Arg-AMC where AMC = aminomethyl coumarin

A 1-2 mM stock of substrate was prepared in DMF and standardized by absorption. A trypsin stock was prepared, and a known amount of inhibitor was added. At regular time points after inhibitor/trypsin combination, aliquots were sampled and assayed as follows: Trypsin/inhibitor reaction mix (10 μl) was added to a cuvette containing 980 μl reaction buffer and 10 μl substrate stock. Initial rate was measured on an LS-5B fluorimeter (Perkin-Elmer). Initial rate was plotted against inhibition time.

RESULTS

2-2.5 mg of SNase mutant were obtained from each of the clones (1 & 3). The resulting protein was reasonably pure based on protein gel analysis. The material isolated was used in all activity assays without further purification.

The results of Assay #1 are shown in Figure 3-2. The gel shows no discernible DIFP labeling for the SNase mutant. In addition, only the G189 mutant and wild type bovine trypsin were labeled. All other mutants of rat anionic trypsin displayed no labeling by DIFP even though they are known to be active proteases.

The results of Assay #2 are shown in Figure 3-3. The data confirms the results of Assay #1 for the D102N, S195A and H57K mutants of rat trypsin. Under a number of pH values, temperatures and reaction times, bovine trypsin was effectively labeled by DIFP, but none of the rat trypsin mutants were labeled.

Assay #3 compared wild type bovine and wild type rat trypsin at a number of conditions. Consistently, the bovine trypsin was labeled by DIFP to a greater extent than the rat trypsin. In addition, assay sensitivity falls off very rapidly over a factor of 9 (0.9 μM / 0.1 μM) for rat trypsin, while a similar change in intensity is seen over a 125x (0.4 μM / 3.2 nM) dilution of bovine trypsin. It is clear that DIFP labeling is much more sensitive toward bovine cationic trypsin than toward rat anionic trypsin.

Assay #4 tested wild type bovine and rat trypsins with respect to DIFP and PMSF inhibition. This fluorogenic assay indicated minimal inhibition of rat trypsin with either DIFP or PMSF. Under the same conditions, wild type bovine trypsin was significantly inhibited within 2 minutes.

FIGURE 3-2: gel results of Assay #1

Gel shows DIFP labeling results for wild type bovine trypsin and five rat trypsin mutants. The wild type trypsin is heavily labeled. D189G is the only mutant to show any significant labeling by DIFP.

FIGURE 3-2: Gel results of Assay #1

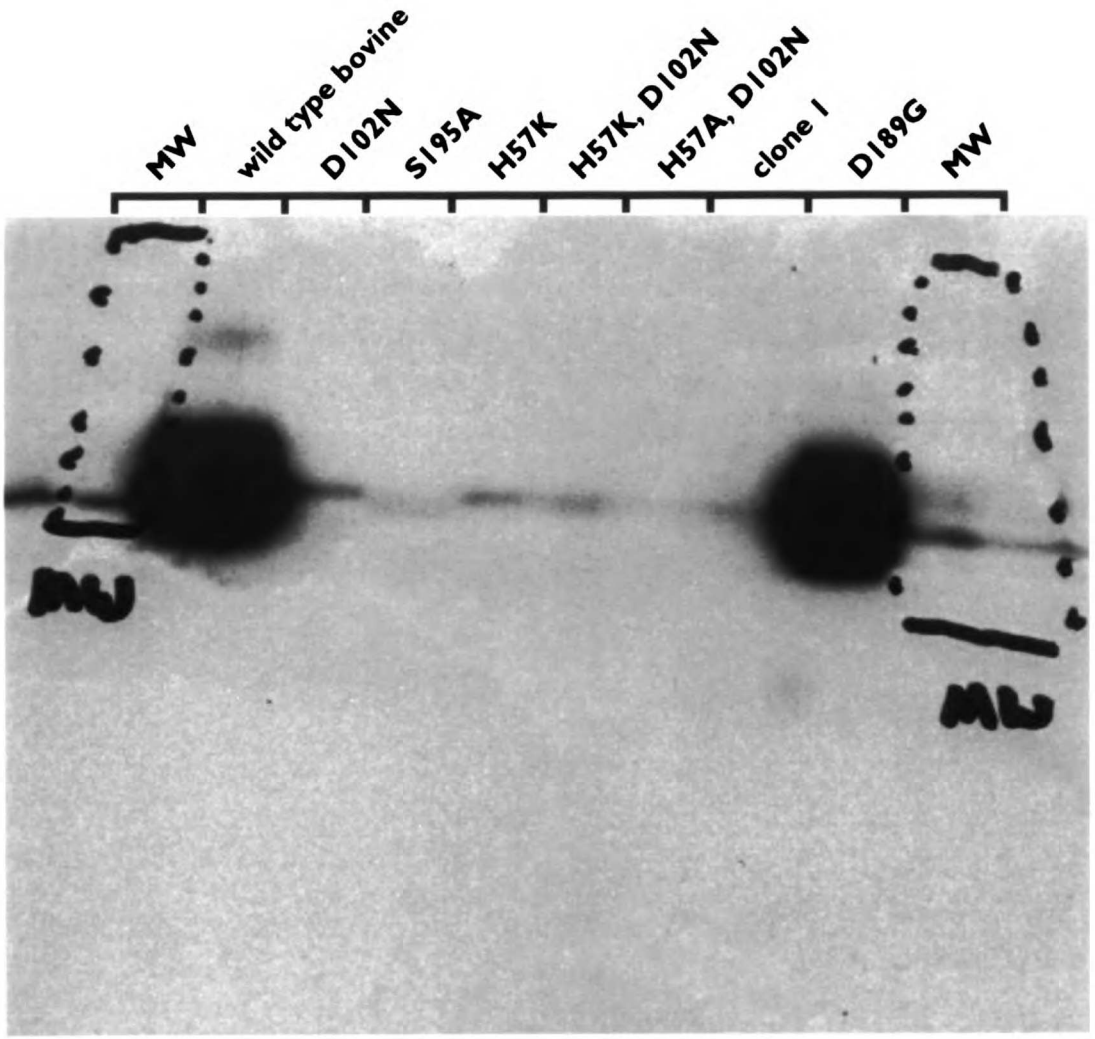


FIGURE 3-3: gel results of Assay #2

Gel shows data for DIFP labeling of wild type and mutant trypsins under varying conditions of pH, temperature and reaction time. Wild type trypsin is much more heavily labeled than the mutants.

FIGURE 3-3: Gel results of Assay #2

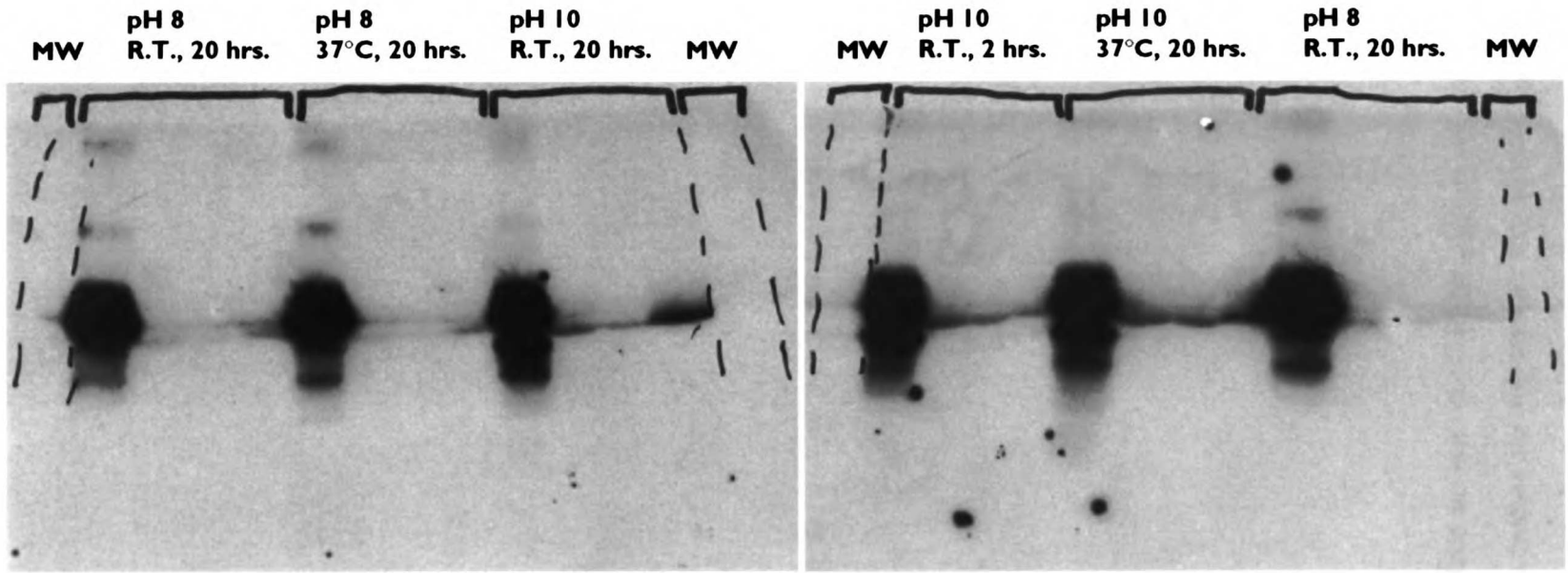


FIGURE 3-4: gel results of Assay #3

Comparison of wild type bovine cationic and rat anionic trypsins. The bovine enzyme is much more sensitive to DIFP labeling than the rat enzyme.

FIGURE 3-4: Gel results of Assay #3

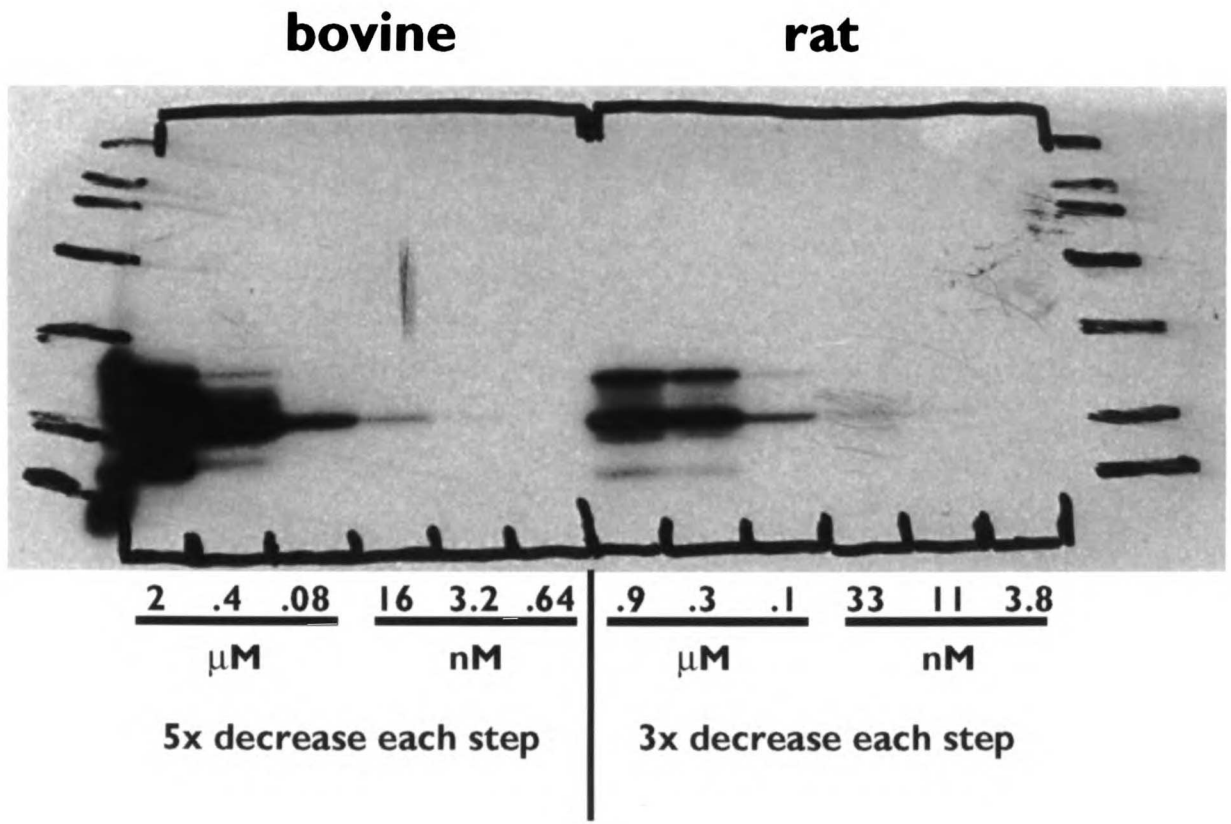
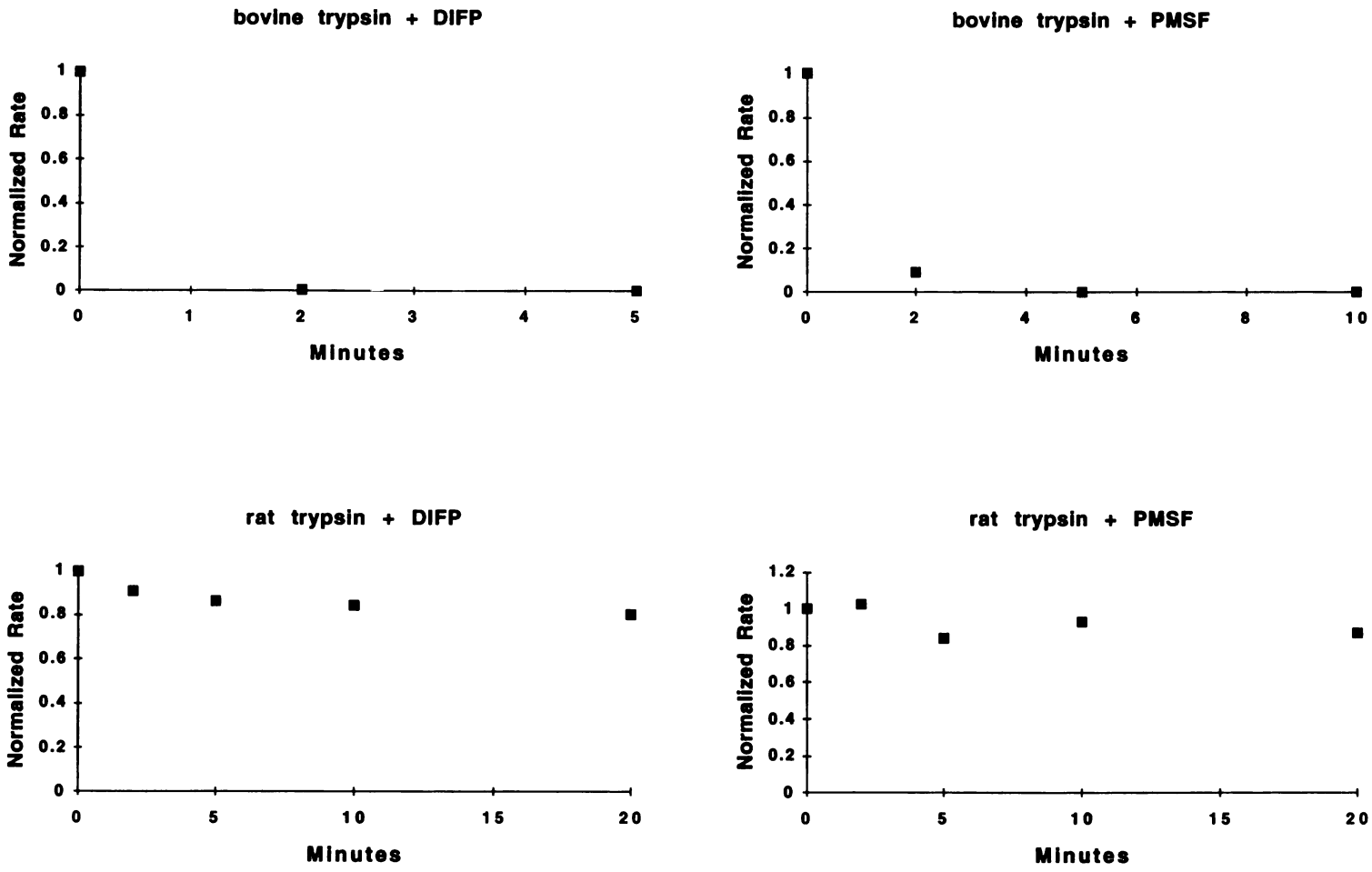


FIGURE 3-5: graph of results from Assay #4

Data from fluorometric assay of trypsin inhibition using DIFP and PMSF. Bovine cationic and rat anionic trypsin are compared. The bovine enzyme displays significant inhibition by both DIFP and PMSF. There is limited inhibition of the rat enzyme under the same conditions.

FIGURE 3-5: graph of results from Assay #4



DISCUSSION

At first glance, the SNase mutant's lack of activity toward DIFP is disappointing. Under a number of conditions where wild type trypsin is readily labeled by DIFP, the SNase mutant remains unlabeled. It would follow that the serine that was incorporated in SNase via mutagenesis is not an activated serine. This is one valid conclusion based on the data, but is complicated by the results for rat anionic trypsin. If DIFP will label any activated serine, why do all the rat anionic trypsins, mutant and wild type, show minimal activity toward DIFP? A second valid conclusion is that DIFP is a more specific reagent than previously believed. It appears to distinguish between bovine cationic and rat anionic trypsins.

The first possible conclusion bears addressing. If the SNase mutant is truly inactive, what may have been lacking in the design? The SNase mutant triad differs from trypsin's in the environment of the Asp residue. In trypsin the Asp is buried in a hydrophobic region of the enzyme. This environment is considered a requirement for proteolytic activity (Bryan, Pantoliano et al. 1986; Carter and Wells 1988; Lewendon, Murray et al. 1990). No such hydrophobic region exists in the SNase mutant.

All known serine proteases possess an oxyanion hole (Henderson 1970; Segal, Powers et al. 1971; Henderson, Wright et al. 1972; Robertus, Alden et al. 1972; Robertus, Kraut et al. 1972; Matthews, Alden et al. 1975; Poulos, Alden et al. 1976; Kraut 1977) which helps stabilize the developing negative charge on the oxyanion intermediate during proteolysis. The oxyanion hole is made up of two backbone NH groups (from Gly 193 and Ser 195) in trypsin-like proteases (Kraut 1977), while in subtilisin it is a combination of the Asn 155 side chain and the peptide NH from Ser 221 (Bryan, Pantoliano et al. 1986). A residue for this role was not included in SNase graft design.

The SNase crystal structure contains a 5-deoxythymidine bisphosphate molecule, which combined with Ca^{2+} is considered essential for structural stability (Koepke, Maslowska et al. 1989) This ligand was not included in our assays, primarily because it would block access to the triad which was introduced. The SNase mutant may not maintain a structure like the crystal structure, and this may invalidate the design. Structural stability might be obtained with addition of this nucleotide and Ca^{2+}, but this would most likely prevent DIFP from accessing the serine.

Clearly, any one of these design flaws: 1) lack of a buried Asp, 2) lack of an oxyanion hole or 3) no structural stabilization by a nucleotide could account for inactivity of the SNase mutant toward DIFP. However, independent of the lessons learned about graft design, a more fundamental lesson was learned with respect to assay choice.

To embark on a protein engineering study, one must be sure that there is a means for characterizing the resulting protein. If a chemical or biological assay is to be used, it must be general enough to recognize activity in all known proteins that match the design criteria. This is a key area of failure in the trypsin triad grafting study described here. Our assay should have been capable of recognizing any activated serine observed in nature. Unfortunately the assay has poor recognition for the activated serines from rat anionic trypsin and its mutants.

CONCLUSION

I have summarized a protein engineering study which attempted to graft the catalytic triad of trypsin onto the SNase protein. This undertaking was flawed from the outset. Future grafting attempts will need to involve more general assays, and perhaps will need to target binding function instead of catalysis. Typically, binding assays are more general than those for catalysis.

In the graft design phase, computational tools will need to evaluate aspects of protein environment other than geometry. Hydrophobic environment was a definite oversight in the SNase design. In addition, all important residues will need to be accounted for in a design. The residue corresponding to the oxyanion hole in trypsin would have to be incorporated into trypsin grafts. Anomalies of the crystalline environment must also be considered during the design process. Any compounds that are necessary for maintaining the structure identified by crystallography must be compatible with the assay conditions which will be used for design verification.

REFERENCES

Bryan, P., M. W. Pantoliano, et al. (1986). "Site-directed mutagenesis and the role of the oxyanion hole in subtilisin." Proc Natl Acad Sci U S A **83**(11): 3743-5.

Carter, P. and J. A. Wells (1988). "Dissecting the catalytic triad of a serine protease." Nature **332**(6164): 564-8.

Henderson, R. (1970). "Structure of crystalline alpha-chymotrypsin. IV. The structure of indoleacryloyl-alpha-chyotrypsin and its relevance to the hydrolytic mechanism of the enzyme." J Mol Biol **54**(2): 341-54.

Henderson, R., C. S. Wright, et al. (1972). "-Chymotrypsin: what can we learn about catalysis from x-ray diffraction?" Cold Spring Harb Symp Quant Biol **36**: 63-70.

Koepke, J., M. Maslowska, et al. (1989). "Three-dimensional structure of ribonuclease T1 complexed with guanylyl-2',5'-guanosine at 1.8 A resolution." J Mol Biol **206**(3): 475-88.

Kraut, J. (1977). "Serine proteases: structure and mechanism of catalysis." Annu Rev Biochem **46**: 331-358.

Kunkel, T. A. (1985). "Rapid and efficient site-specific mutagenesis without phenotypic selection." P.N.A.S. **82**: 488-492.

Lewendon, A., I. A. Murray, et al. (1990). "Evidence for transition-state stabilization by serine-148 in the catalytic mechanism of chloramphenicol acetyltransferase." Biochemistry **29**(8): 2075-80.

Matthews, D. A., R. A. Alden, et al. (1975). "X-ray crystallographic study of boronic acid adducts with subtilisin BPN' (Novo). A model for the catalytic transition state." J Biol Chem **250**(18): 7120-6.

Poulos, T. L., R. A. Alden, et al. (1976). "Polypeptide halomethyl ketones bind to serine proteases as analogs of the tetrahedral intermediate. X-ray crystallographic comparison of lysine- and phenylalanine-polypeptide chloromethyl ketone-inhibited subtilisin." J Biol Chem **251**(4): 1097-103.

Robertus, J. D., R. A. Alden, et al. (1972). "An x-ray crystallographic study of the binding of peptide chloromethyl ketone inhibitors to subtilisin BPN'." Biochemistry **11**(13): 2439-49.

Robertus, J. D., J. Kraut, et al. (1972). "Subtilisin; a stereochemical mechanism involving transition-state stabilization." Biochemistry **11**(23): 4293-303.

Segal, D. M., J. C. Powers, et al. (1971). "Substrate binding site in bovine chymotrypsin A-gamma. A crystallographic study using peptide chloromethyl ketones as site-specific inhibitors." Biochemistry **10**(20): 3728-38.

Shortle, D. and A. K. Meeker (1989). "Residual structure in large fragments of staphylococcal nuclease: effects of amino acid substitutions." Biochemistry **28**(3): 936-44.

Shortle, D., W. E. Stites, et al. (1990). "Contributions of the large hydrophobic amino acids to the stability of staphylococcal nuclease." Biochemistry **29**(35): 8033-41.

Chapter 4:
The GRAFTER Algorithm

INTRODUCTION

The DIST program is a simple tool for identifying some potential graft sites in a protein scaffold. Although it is highly time efficient, it is flawed by a biased search that leads to only certain potential graft sites. My attempts to alleviate these shortcomings resulted in the GRAFTER program. The goal of the GRAFTER algorithm is to identify matches in a less-biased way than in the DIST algorithm. It is also essential that this new algorithm explore a larger range of geometric space in search of matches while retaining the time efficiency of the DIST algorithm.

Current computational hardware is incapable of performing a full combinatorial comparison of two distance matrices within a reasonable calculation time. The GRAFTER algorithm needs to perform a constrained combinatorial search that examines all reasonable matrix alignments while ignoring those that have poor potential for geometric match. The concept of an index was borrowed from the DIST algorithm, but the 4-atom permutation and pairwise comparison from DIST were eliminated in favor of a more complete approach. I chose a procedure that explores every pairing combination that could be generated for two indices within the distance tolerance constraint. To avoid lengthy comparisons between poorly matches indices, a score-on-the-fly technique was developed. At intermediate steps in the comparison, scores are calculated, and only those pairings that score best are retained for expansion.

The GRAFTER algorithm was verified using two of the test cases originally applied to DIST. Trypsin was compared to subtilisin using the catalytic triad from each enzyme to search the whole structure of the other. A similar comparison was performed on hen egg white and T4 lysozyme. These applications demonstrated that GRAFTER, like DIST, can recognize known similarities between proteins.

Once the preliminary tests were completed, GRAFTER was applied to a number of novel design problems. Catalytic motifs from trypsin (Walter, Steigemann et al. 1982),

acetyl cholinesterase (Sussman, Harel et al. 1991) and pepsin (Cooper, Khan et al. 1990) were used to probe a database of protein structures as scaffolds. A binding motif from a lysozyme epitope was used to probe the same database. I present here the best match for each motif tested. Clearly, GRAFTER generates more than a single match for a given motif. However, by focusing evaluation efforts on only the best match, I hope to recognize weaknesses inherent in GRAFTER. This information will assist in the further enhancement of GRAFTER.

METHODS

GRAFTER is designed to compare a motif and a scaffold and report geometric matches between them. The algorithm (Figure 4-1) is based on the comparison of distance matrices (a matrix describing the distances between all points in a set) (Phillips 1970; Nishikawa and Ooi 1974; Sippl 1982). If two distance matrices are identical at every corresponding position, the structures that they represent are either exactly the same or exact mirror images (recall Figure 2-1). However, to make such a comparison the columns and rows in the second matrix must be aligned with the first in a fashion analogous to a sequence alignment. Therefore, geometric comparison of two structures using distance matrices requires a combinatorial approach. In the simplest case, where the two matrices contain the same number of rows, every permutation of the second matrix must be compared to one permutation of the first matrix. A matrix with n rows has n! permutations, and each permutation contains $n^2/2 - n$ significant entries. A full comparison of two matrices with n elements each requires $n!(n^2/2 - n)$ comparisons and becomes untenable even for relatively small matrices. If the sets do not contain equal numbers of atoms, or if a match can contain fewer pairs than there are rows, the combinatorial algorithm becomes even more complex.

The GRAFTER algorithm takes advantage of distance matrix comparison but incorporates a number of techniques that avoid the full combinatorial search. Since an exact geometric match between two sets of points from non-identical proteins is improbable, a tolerance value is used in determining whether matrix entries are matched. A fractional tolerance is used so that larger distances may have more uncertainty than small ones. In a single GRAFTER run, distance matrices are prepared for the motif atoms and the scaffold atoms. These matrices serve as the foundation for all subsequent comparisons. We have introduced the concept of an **index**, which is simply a single row from a distance matrix corresponding to a **reference atom.** In a complete GRAFTER execution, every index from the scaffold is compared to every index from the motif.

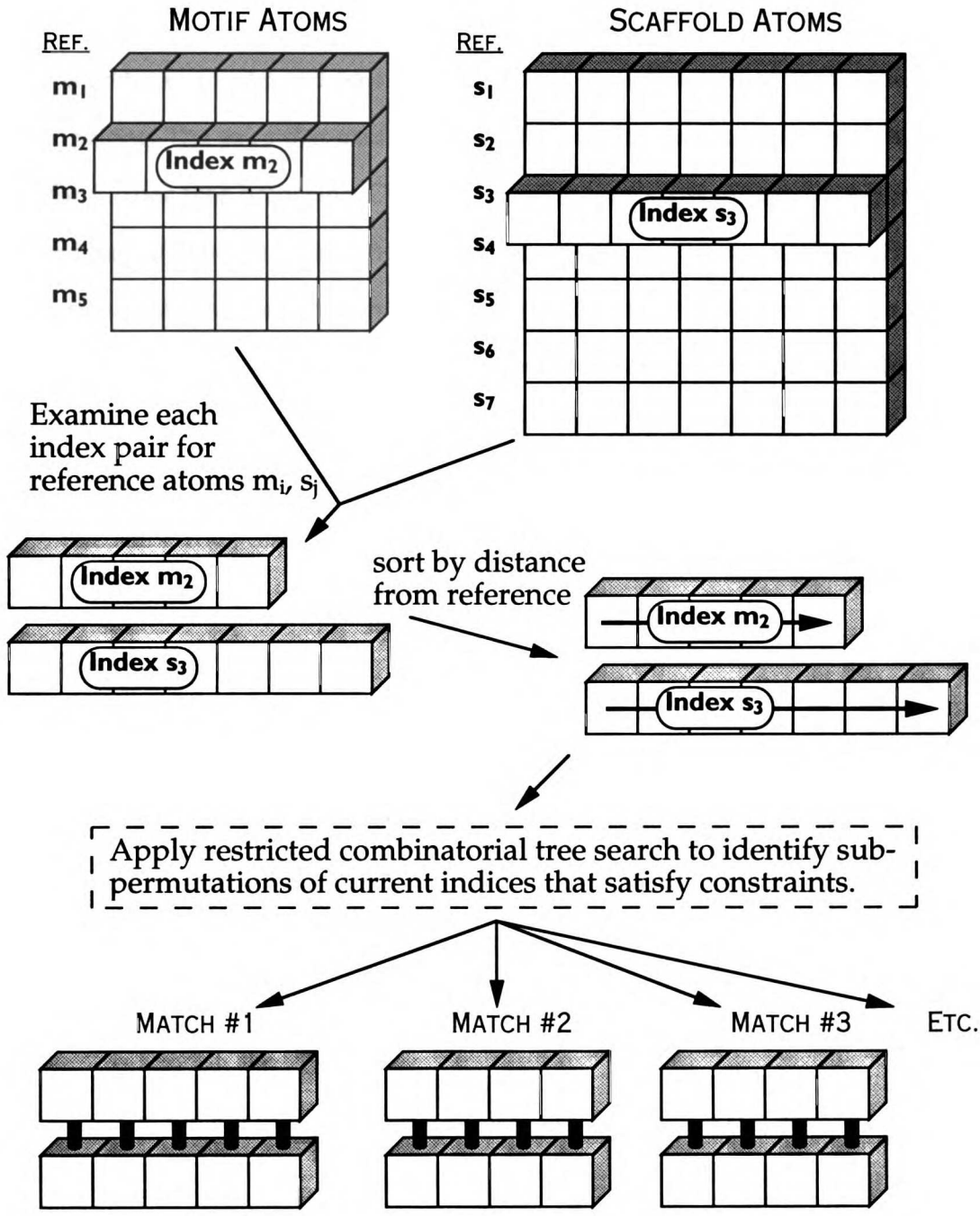
The goal in comparing two indices (index 1 and index 2) is to identify any one-to-one alignments between them containing at least a minimum number of atoms. This lower limit is specified at run time. An alignment represents a correspondence of atoms that is likely to match within the geometric constraint set by the predefined tolerance. In addition, atom types must match for atoms to be paired. Hence, a Cα atom can only be aligned with another Cα atom.

To build a list of such indices, we follow a trimmed combinatorial search. Each index is sorted with respect to increasing distance from the reference atom. This facilitates certain trimming steps during the search. The sorted nature of the indices means that all possible pairs for a given atom in index 1 are sequential, and only a subset of index 2 must be checked for each atom in index 1. Every pair of reference atoms is used as a level 1 root node. With the root node established, the next step is to identify a list of best candidates to serve as roots for level two. To accomplish this, we identify

FIGURE 4-1: GRAFTER algorithm

Simplified schematic of the GRAFTER algorithm - demonstrates the steps involved in index comparison. All indices are compared over the course of a complete GRAFTER search.

FIGURE 4-1: simplified schematic of GRAFTER algorithm



all atoms whose distance entries are within the tolerance limit of the second atom in index 1. This procedure is repeated for each level until all matches have been found.

Figure 4-2 illustrates the process of establishing all pairs for a particular level. Scores are calculated for each pair at a given level, so that the pairs may be rank ordered. GRAFTER keeps only the best scores at each level, the total number being determined by a user-defined **pair limit.** A **score cutoff** is used to limit nodes for further pairing. The pairs that pass this level of evaluation are kept, and each is used as a node for further tree branching. Once the possible atoms are exhausted, a given matched pair of indices corresponds to a matched pair of distance matrices. Each index pairing that contains more than the minimum number of pairs is reported as a match.

Our scoring function evaluates how well the atom in question is positioned based on the atoms previously paired between the indices. For each previously established atom pair in the indices, the deviation between the distances to the current pair is calculated and scaled by the active tolerance value. The score is the sum of all scaled deviations for the current pair:

$$Score = \sum_i \frac{\left|\text{dist}(M_N, M_i) - \text{dist}(S_x, S_i)\right|}{T \cdot \text{dist}(M_N, M_i)}$$

where:

dist(a,b) is the distance between atoms a and b;

M_N is the current atom in index 1 (the motif index);

M_i , S_i are the previously accepted atom pairs from the motif and scaffold;

S_x is the current atom in index 2 (the scaffold index) which is being evaluated; and

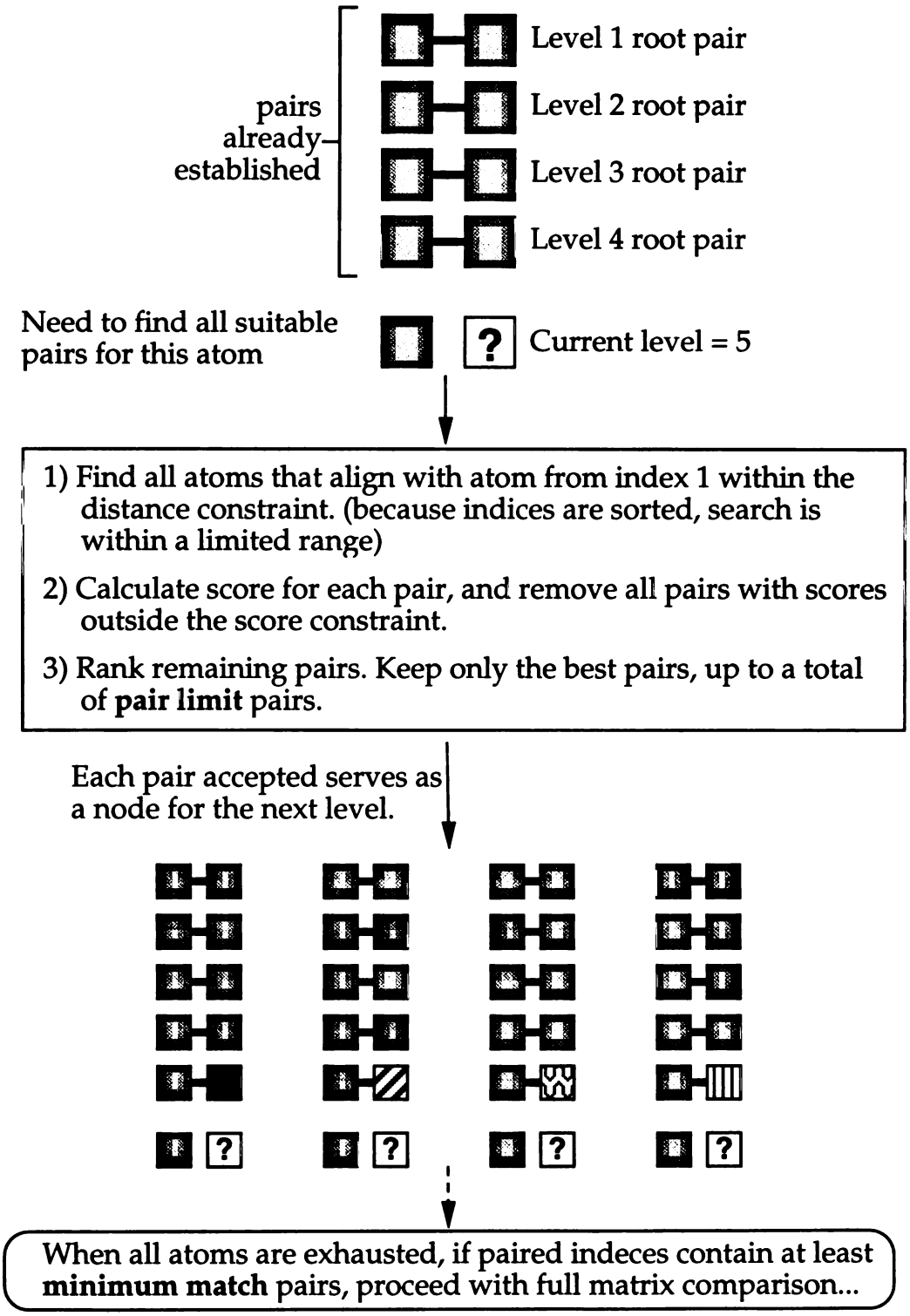
T is the fractional tolerance.

The score profile exhibits non-linear behavior; scores tend to be better for the second, third and fourth pairs than for subsequent pairs. This reflects the orientational

FIGURE 4-2: index comparison

One level in the comparison of two indices - shows two indices that have been matched through their fourth entries. Branch points are being established for level five. The diagram shows the current atom #5 in index 1 being paired with four different atoms from index 2. Each of the four resulting index pairs (containing 5 pairs) will serve as a starting node for establishing level 6 pairs.

FIGURE 4-2: establishing atom pairs



freedom available when selecting atoms 2-4. The second atom may be chosen from anywhere within a spherical shell around the reference atom. The third atom is restricted to a circular shell around the line between the reference atom and atom 2. There are two regions for the fourth atom which are differentiated only by overall handedness of the four atoms. By performing a large number of evaluations using both real structures and structures generated randomly, a score profile (S(p)) was calculated that varies with the tolerance chosen.

$$S(p) = \frac{plateau \cdot p^4}{p^4 + 45} + shift$$

where *plateau* is defined by:

$$\begin{cases} plateau = 0.0314 \cdot tolerance & 0 \le tolerance \le 0.159 \\ plateau = (0.152 \cdot tolerance) - 0.0192 & tolerance > 0.159 \end{cases}$$

Rather than using a fixed score cutoff, GRAFTER varies the cutoff so that it is highly constraining for the second pair, and becomes less of a constraint until it plateaus at about the fifth pair. The user may apply a global adjustment (*shift*) on the score profile to make it more or less constraining for a given search.

Within the GRAFTER program itself, we calculate a Cartesian RMS deviation (after least-squares fit superpositioning) for each match. We choose this approach because, unlike an RMS deviation by differences in interatomic distances, an RMS deviation by least-squares superposition takes into account any handedness in the structures. The two methods perform similarly except that when there is a loose fit, the distance difference method does not penalize local inversions adequately (Cohen and Sternberg 1980). The atom-based alignments generated by the combinatorial matrix comparison are translated into residue-based alignments and both interpretations are reported. A summary of run-time statistics is included in the report.

IMPLEMENTATION

GRAFTER is written in C. The program has been compiled and executed on a number of platforms, including Silicon Graphics IRIS, Indigo and Challenge machines, as well as Sun Sparc workstations and MIPS servers.

RESULTS AND DISCUSSION

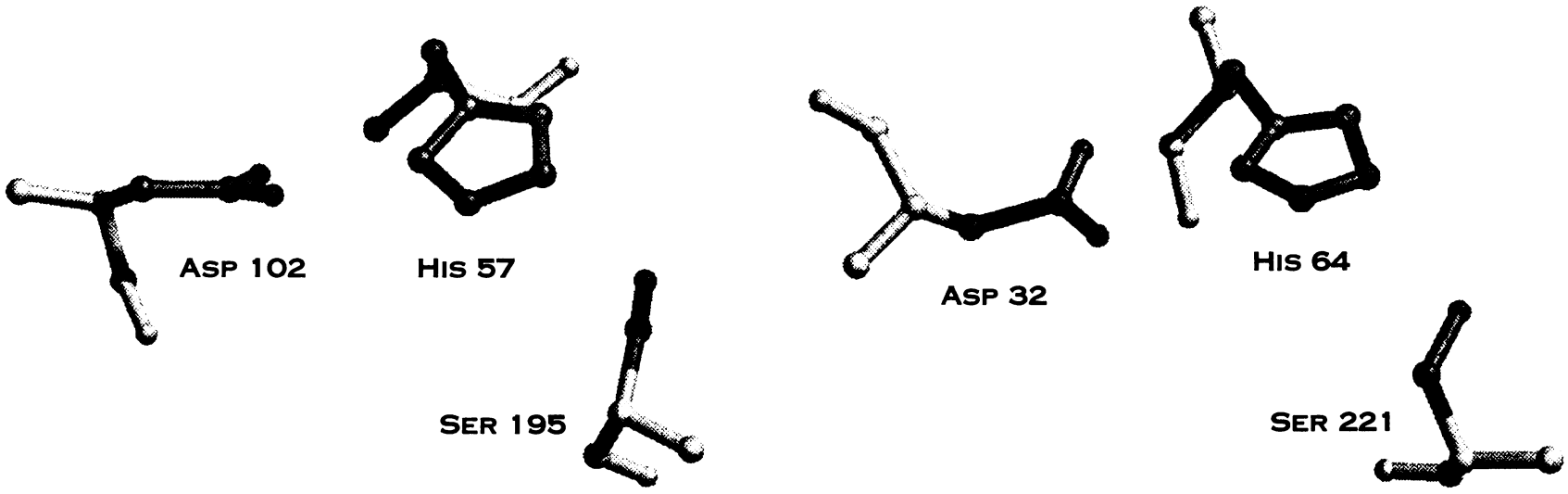
Several examples of convergent evolution have been suggested by crystallographic studies. These examples provide active site/scaffold pairs that GRAFTER should identify. The serine proteases, trypsin and subtilisin, provide an important test case. They contain a common juxtaposition of catalytic residues (Ser, His and Asp), yet lack any relationship at the level of protein sequences. With GRAFTER, the $C\alpha$, $C\beta$ and N atoms from the residues in the catalytic triad of trypsin (Ser195, His57, Asp102) were used to successfully identify the catalytic triad of subtilisin (Ser221, His64, Asp32) from a search of the entire subtilisin molecule. The reverse experiment (i.e., searching all residues in trypsin using the triad from subtilisin) was also successful.

These test comparisons are complicated by the fact that the catalytic triads of trypsin and subtilisin are not exact geometric matches (Figure 4-3). Although the catalytic R groups superimpose well and the side chain functional atoms have similar relative geometries, the backbone geometries are different. In particular, the χ_1's of Asp102 and Asp32 differ by approximately 170° with respect to each other and the side chains approach from opposite directions. The His57 and His64 side chains align well, but the main chain amide nitrogen and carbonyl are rotated about 120° from each other about the C_α-C_β bond. With a tolerance of 0.4, a limit of 20 pairs per tree level, a minimum of 8 atoms matched and a 0.0001 shift in the score profile, we obtained the

FIGURE 4-3: catalytic triads

The catalytic triads from the structures of trypsin (Walter, Steigemann et al. 1982) and subtilisin (Alden, Birktoft et al. 1971) as viewed from similar perspectives.

FIGURE 4-3: catalytic triads



TRYPSIN **SUBTILISIN**

desired subtilisin to trypsin active site match. The correct pairing was not found with a minimum of matches that required all 9 C_α, C_β and N atoms. GRAFTER did identify 9 atom matches between trypsin and subtilisin, but none of these was a match of the catalytic triads. In fact, the RMS deviations (C_α, C_β and N atoms) for these matches were in many cases lower than that for the known alignment of the catalytic triads. This reinforces our knowledge that there is more to the functional similarity between trypsin and subtilisin than just geometry. When we reduced the requirement to 8 atoms, GRAFTER was able to identify the match of the catalytic triads. With all but the His amide nitrogen matched, the resulting RMS deviation was 0.838 Å. This placed the match 105th out of 2911 matches. Once again, GRAFTER identified a substantial number of matches that are better on the basis of RMS deviation (the lowest RMS fit was 0.347 Å). This clearly coincides with our knowledge that exact geometry is not the only factor determining catalytic efficiency for an enzyme. Many of these potential grafting sites are inaccessible to the solvent (or substrate) and would make poor scaffolds. Others may provide useful scaffolds for future protein engineering experiments.

The two enzymes, hen egg white lysozyme and T4 lysozyme, provided a more complex test of the GRAFTER algorithm and serve as an example of divergent evolution (Matthews, Grütter et al. 1981; Matthews, Remington et al. 1981). We chose to compare seven residues in each lysozyme to all residues in the other lysozyme. The residues selected are shown in Table 4-1. These choices were based on the structural alignment by Matthews and coworkers. In both comparisons the appropriate match was identified. However, using a typical parameter set (tolerance = 0.25, minimum match = 15, pair limit = 15 and score shift = 0.000) we were able to obtain a match that is close to but not identical to the match proposed by Matthews. Consistently, GRAFTER matched each residue in the egg white lysozyme template with the expected residue in T4 lysozyme except for Asp52. In Matthews' alignment, Asp52 corresponds to Asp20 in T4 lysozyme. However, GRAFTER consistently matched it with Thr26. A visual analysis

TABLE 4-1: RMS summary for benchmark tests

Residue alignments for the catalytic triads of trypsin and subtilisin and the catalytic regions of egg white and T4 lysozyme. Paired residues are listed. The RMS deviations are tabulated for all atoms, side chains only and Cα, Cβ, N only. Structures used: trypsin (Walter, Steigemann et al. 1982), subtilisin (Alden, Birktoft et al. 1971), egg white lysozyme (Diamond 1974) and T4 lysozyme (Weaver and Matthews 1987)

TABLE 4-1: RMS summary for benchmark tests

		RMS for matched residues †			match ¥			
Active Site Template	Scaffold Protein	all atoms	side chains	CA, CB & N	Active Site		Scaffold	
trypsin	subtilisin	1.58 Å	0.80 Å	1.07 Å	S	195	221	S
(2ptn)	(1sbt)				H	57	64	H
					D	102	32	D
egg white lysozyme	T4 lysozyme	5.34 Å	5.87 Å	3.47 Å	E	35	11	E
(1lyz)	(2lzm)				D	52	20	D
					Q	57	30	G
					I	58	31	H
					N	59	32	L
					A	107	104	F
					W	108	105	Q

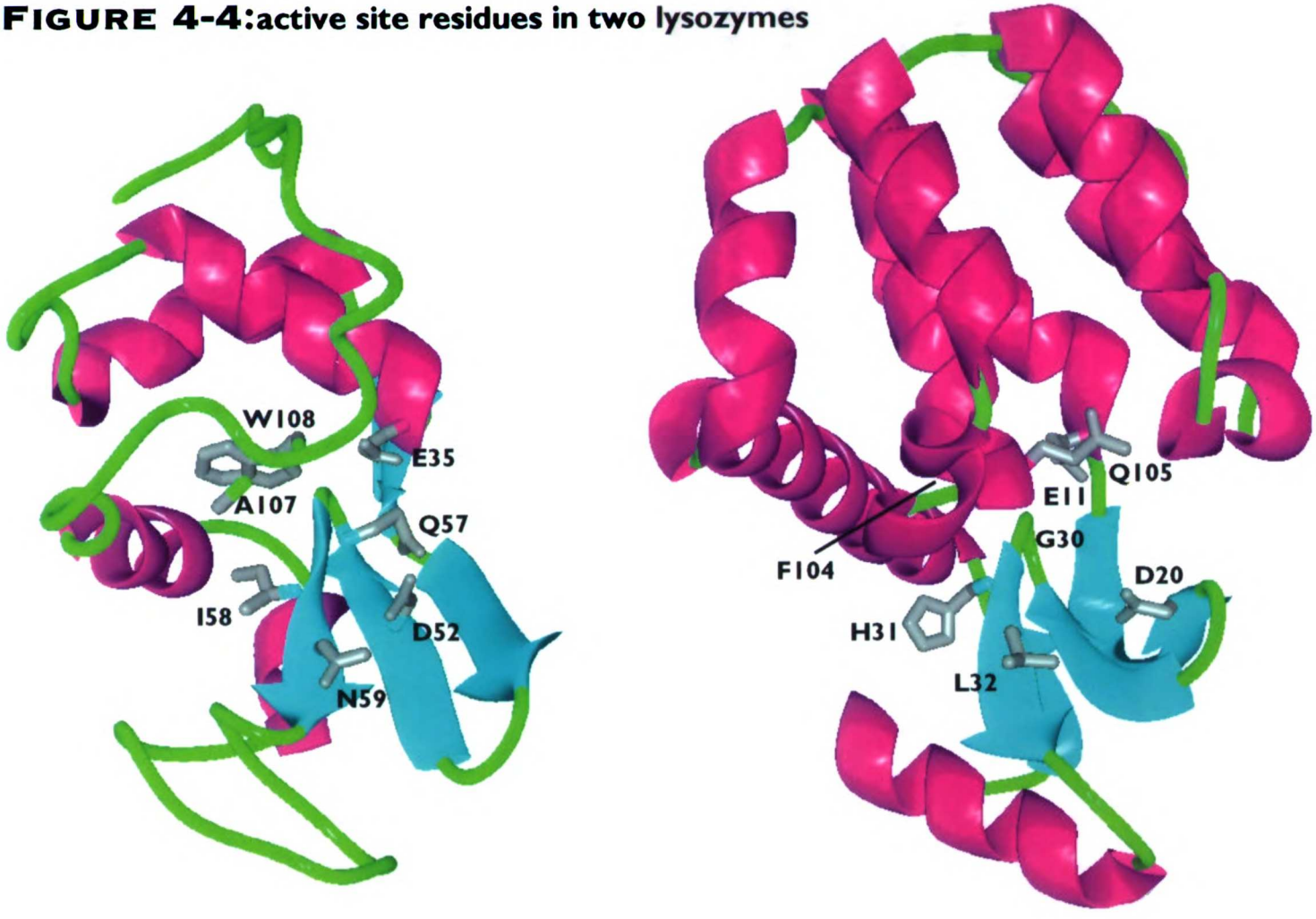
† Values of RMS deviation for the trypsin/subtilisin and egg white lysozyme/T4 lysozyme alignments

¥ Listing of corresponding residues in the two active sites

FIGURE 4-4: active site residues in two lysozymes

Egg white lysozyme and T4 lysozyme viewed from similar perspectives with respect to their catalytic residues. Corresponding catalytic residues are labeled for each structure. Structures used: egg white lysozyme (Diamond 1974) and T4 lysozyme (Weaver and Matthews 1987)

FIGURE 4-4: active site residues in two lysozymes



EGG WHITE LYSOZYME

T4 LYSOZYME

helped to explain this. As demonstrated in Figure 4-4, the carboxylate moieties of Asp52 and Asp20 are in the same vicinity, but they approach from opposite directions. The backbone atoms of Asp52 are much closer to aligning with those of Thr26 than to those of Asp20. Once again, as in the trypsin/subtilisin example, GRAFTER has indicated to us that geometric mimicry and functional mimicry are closely but not exactly correlated.

GRAFTER has also been applied to the search for novel graft sites. Using a variety of templates we have applied GRAFTER toward the examination of protein databases for geometrically useful scaffolds. Over the course of these calculations, two databases have been used. The smaller database contains a subset of proteins that are good candidates for mutagenesis and was previously described in chapter 2 (recall Table 2-3). These proteins have been successfully expressed and are readily purified. This database includes 48 structures from the PDB, representing 10 unique proteins. The larger database (Table 4-2) contains a number of structures from the PDB without consideration for expression or purification. 87 structures are included in this database, corresponding to 31 unique proteins. Multiple structures, when available, were included for proteins to allow for structural variation.

A variety of binding site and active site geometries have been used in GRAFTER searches of the scaffold databases. Table 4-3 lists the results for the active sites of trypsin, acetyl cholinesterase and pepsin as well as for an antibody binding region from lysozyme. The active site of trypsin (2ptn) was represented in two ways: 1) the catalytic triad of Ser195, His57 and Asp102 in native conformation and 2) the side chains from Ser195, His57 and Asp102 coupled to backbone atoms that are oriented according to the most favorable rotamer χ angles (Ponder and Richards 1987) The second representation was used because the side chains of the catalytic residues in trypsin are not in statistically likely orientations. Outside of the trypsin scaffold a Ser, His, and Asp may not take on side chain rotamers similar to those in trypsin. The second triad representation provides an opportunity to explore another approach to achieving the

TABLE 4-2: large scaffold database

PDB File	Chain(s)	Protein	Reference
5acn		Aconitase	(Robbins and Stout 1989)
6acn		Aconitase	(Robbins and Stout 1989)
5adh		Alcohol Dehydrogenase	(Colonna-Cesari, Perahia et al. 1986)
2aat		Aspartate Aminotransferase	(Smith, Almo et al. 1989)
1atl	A,B,C,D	Aspartate Transcarbamoylase	(Gouaux and Lipscomb 1990)
2atl	A,B,C,D	Aspartate Transcarbamoylase	(Gouaux and Lipscomb 1990)
3atl	A,B,C,D	Aspartate Transcarbamoylase	(Gouaux and Lipscomb 1990)
4atl	A,B,C,D	Aspartate Transcarbamoylase	(Stevens, Gouaux et al. 1990)
5atl	A,B,C,D	Aspartate Transcarbamoylase	(Stevens, Gouaux et al. 1990)
7atl	A,B,C,D	Aspartate Transcarbamoylase	(Gouaux, Stevens et al. 1990)
8atl	A,B,C,D	Aspartate Transcarbamoylase	(Gouaux, Stevens et al. 1990)
8atc	A,B,C,D	Aspartate Transcarbamoylase	(Ke, Lipscomb et al. 1988)
2apk		cAMP-Dependent Protein Kinase	(Weber and Steitz 1984)
1apk		cAMP-Dependent Protein Kinase	(Weber and Steitz 1984)
1bpk		cAMP-Dependent Protein Kinase	(Weber and Steitz 1984)
2bpk		cAMP-Dependent Protein Kinase	(Weber, Steitz et al. 1987)
3ca2		Carbonic Anhydrase	(Eriksson, Kylsten et al. 1988)
2gap	A,B	Catabolite Gene Activator Protein	(Weber and Steitz 1984)
1cla		Chloramphenicol Acetyltransferase	(Lewendon, Murray et al. 1990)
3cla		Chloramphenicol Acetyltransferase	(Leslie 1990)
1cts		Citrate Synthase	(Remington, Wiegand et al. 1982)
2cts		Citrate Synthase	(Remington, Wiegand et al. 1982)
3cts		Citrate Synthase	(Remington, Wiegand et al. 1982)
1ccr		Cytochrome C	(Ochi, Hata et al. 1983)
1cyc		Cytochrome C	(Tanaka, Yamane et al. 1975)
5cyt		Cytochrome C	(Takano and Dickerson 1981)
2c2c		Cytochrome C2	(Salemme, Freer et al. 1973)
3c2c		Cytochrome C2	(Salemme, Freer et al. 1973)
1cy3		Cytochrome C3	(Pierrot, Haser et al. 1982)
2cdv		Cytochrome C3	(Higuchi, Kusunoki et al. 1984)
1cc5		Cytochrome C5	(Carter, Melis et al. 1985)
155c		Cytochrome C550	(Timkovich and Dickerson 1976)

351c		Cytochrome C551	(Matsuura, Takano et al. 1982)
451c		Cytochrome C551	(Matsuura, Takano et al. 1982)
3dfr		Dihydrofolate Reductase	(Bolin, Filman et al. 1982)
6dfr		Dihydrofolate Reductase	(Bystroff, Oatley et al. 1990)
7dfr		Dihydrofolate Reductase	(Bystroff, Oatley et al. 1990)
8dfr		Dihydrofolate Reductase	(Matthews, Bolin et al. 1985)
1r1e	E	EcoR1 Endonuclease	(Kim, Grable et al. 1990)
1efm		Elongation Factor Tu	(Jurnak 1985)
1etu		Elongation Factor Tu	(la Cour, Nyborg et al. 1985)
1fd2		Ferredoxin	(Martin, Burgess et al. 1990)
1fdx		Ferredoxin	(Adman, Siefker et al. 1976)
1fxb		Ferredoxin	(Fukuyama, Matsubara et al. 1989)
3fxc		Ferredoxin	(Fukuyama, Hase et al. 1980)
3fxn		Ferredoxin	(Smith, Burnett et al. 1977)
5fd1		Ferredoxin	(Stout 1993)
1fx1		Flavodoxin	(Watenpaugh, Sieker et al. 1973)
4fxn		Flavodoxin	[(Smith et al., 1977)]
3grs		Glutathione Reductase	(Karplus and Schulz 1987)
2gbp		Glycogen Phosphorylase B	(Martin, Johnson et al. 1990)
1ecd		Hemoglobin	(Steigemann and Weber 1979)
1eco		Hemoglobin	(Steigemann and Weber 1979)
1ldm		Lactate Dehydrogenase	(Abad-Zapatero, Griffith et al. 1987)
1llc		Lactate Dehydrogenase	(Buehner, Hecht et al. 1982)
2ldb		Lactate Dehydrogenase	(Piontek, Chakrabarti et al. 1990)
3ldh		Lactate Dehydrogenase	(White, Hackert et al. 1976)
5ldh		Lactate Dehydrogenase	(Grau, Trommer et al. 1981)
8ldh		Lactate Dehydrogenase	(Abad-Zapatero, Griffith et al. 1987)
1lh1		Leghemoglobin	(Arutyunyan, Kuranova et al. 1980)
1lh2		Leghemoglobin	(Arutyunyan, Kuranova et al. 1980)
1lh3		Leghemoglobin	(Arutyunyan, Kuranova et al. 1980)
1lh4		Leghemoglobin	(Arutyunyan, Kuranova et al. 1980)
1lh5		Leghemoglobin	(Arutyunyan, Kuranova et al. 1980)
1lh6		Leghemoglobin	(Arutyunyan, Kuranova et al. 1980)
1lh7		Leghemoglobin	(Arutyunyan, Kuranova et al. 1980)
2lh1		Leghemoglobin	(Arutyunyan, Kuranova et al. 1980)
2lh2		Leghemoglobin	(Arutyunyan, Kuranova et al. 1980)
2lh3		Leghemoglobin	(Arutyunyan, Kuranova et al. 1980)
2lh4		Leghemoglobin	(Arutyunyan, Kuranova et al. 1980)
2lh5		Leghemoglobin	(Arutyunyan, Kuranova et al. 1980)
2lh6		Leghemoglobin	(Arutyunyan, Kuranova et al. 1980)
2lh7		Leghemoglobin	(Arutyunyan, Kuranova et al. 1980)

1mba		Myoglobin	(Bolognesi, Onesti et al. 1989)
1mbc		Myoglobin	(Kuriyan, Wilz et al. 1986)
1mbd		Myoglobin	(Phillips and Schoenborn 1981)
1mbn		Myoglobin	(Watson 1969)
1mbo		Myoglobin	(Phillips 1980)
1mbs		Myoglobin	(Scouloudi and Baker 1978)
3mba		Myoglobin	(Bolognesi, Onesti et al. 1989)
4mba		Myoglobin	(Bolognesi, Onesti et al. 1989)
4mbn		Myoglobin	(Takano 1977)
5mbn		Myoglobin	(Takano 1977)
1phh		p-Hydroxybenzoate Hydroxylase	(Schreuder, van der Laan et al. 1988)
4pfk		Phosphofructokinase	(Evans, Farrants et al. 1981)
3pgk		Phosphoglycerate Kinase	(Watson, Walker et al. 1982)
3pgm		Phosphoglycerate Mutase	(Winn, Watson et al. 1981)
1bp2		Phospholipase A2	(Dijkstra, Kalk et al. 1981)
3bp2		Phospholipase A2	(Dijkstra, Kalk et al. 1984)
2p21		ras P21 Protein	(Pai, Krengel et al. 1990)
1rnt		Ribonuclease T1	(Arni, Heinemann et al. 1988)
2sn3		Scorpion Neurotoxin	(Zhao, Carson et al. 1992)
1snc		Staphylococcal Nuclease	(Loll and Lattman 1989)
2sns		Staphylococcal Nuclease	(Cotton, Hazen et al. 1979)
1s01		Subtilisin Bpn (Mutant)	(Pantoliano, Whitlow et al. 1989)
2sbt		Subtilisin Novo	(Drenth, Hol et al. 1972)
4tln		Thermolysin	(Holmes and Matthews 1981)
5tln		Thermolysin	(Holmes and Matthews 1981)
7tln		Thermolysin	(Holmes, Tronrud et al. 1983)
2cln		Trimethyl Calmodulin	(Strynadka and James 1988)
2ypi	A,B	Triose Phosphate Isomerase	(Lolis and Petsko 1990)
2ptn		Trypsin	(Walter, Steigemann et al. 1982)
4ts1	A,B	Tyrosyl tRNA Transferase	(Brick and Blow 1987)
4xia	A,B	Xylose Isomerase	(Henrick, Collyer et al. 1989)
5xia		Xylose Isomerase	(Henrick, Collyer et al. 1989)
2yhx		Yeast Hexokinase B	(Anderson, Stenkamp et al. 1978)

TABLE 4-3: RMS summary for novel searches

A complete summary of RMS deviation values from each template search. The table lists the total number of matches found as well as the lowest and highest RMS deviation values from each search. The maximum RMS cutoff for visual analysis and the number of matches analyzed visually are also tabulated.

TABLE 4-3: RMS summary for novel searches

Active Site protein	database	# total matches	RMS † min. RMS	RMS † max. RMS	# analyzed visually ¥	cutoff RMS for visual *
trypsin (native)	small	146	0.335	2.969		
	large	228	0.421	3.008		
	both	374	0.335	3.008		
trypsin (std. rotamer)	small	20	0.723	2.633		
	large	24	0.290	2.621		
	both	44	0.290	2.633		
trypsin (nat. and rot.)	both	418	0.290	3.008	138	1.0 Å
acetyl cholinesterase	small	289	0.722	2.456		
	large	283	0.500	2.711		
	both	572	0.500	2.711	122	1.0 Å
pepsin	small	117	0.079	2.085		
	large	146	0.025	1.957		
	both	263	0.025	2.085	178	0.5 Å
lysozyme epitope	small	294	1.429	10.580		
	large	262	1.734	13.515		
	both	556	1.429	13.515	163	3.0 Å

† The lowest and highest RMS deviation values for each search

¥ The total number of matches analyzed visually using MIDAS

* The maximum RMS cutoff that determines which matches were analyzed visually

same side chain positions using a new scaffold and standard rotamers. The catalytic dyad of Asp32 and Asp215 was used in the pepsin searches. For acetyl cholinesterase, we selected Ser200, His440, and Glu327, another catalytic triad whose geometry is a mirror image of the triad in trypsin. In the case of the lysozyme D1.3 epitope (Fischmann, Bentley et al. 1991), a larger set of residues was used: Asp18, Asn19, Gly117, Asp119, Gln121 and Arg125.

GRAFTER has been designed as a screening tool that effectively pares down a massive geometric search. Once potential graft sites have been identified, the theoretical mutagenesis experiment must be performed to swap the active or binding site residues into their new scaffold. Residues are grafted using side chain χ angles from standard rotamers (Ponder and Richards 1987). Side chains must be rotated from these starting geometries to develop the completed site. We find that visually screening a few hundred high scoring matches from a GRAFTER run with a molecular graphics program (MIDAS) can be completed in a 1-2 day period. Although a conformational search could be automated, we rejected this because of concern that such a search would not adequately sample conformational space. I believe that GRAFTER is a supplement to and not a replacement for a protein engineer's common sense.

The best matches based on RMS deviation were evaluated visually using MIDAS. For each enzyme, from over 100 low RMS deviation matches we were able to choose a small number of outstanding matches. These results are tabulated in Table 4-4.
In the trypsin search, one scaffold contains a potential catalytic triad with a relatively low value for the RMS deviation compared to the catalytic triad from trypsin. The match involves three residues from an α-helical region in the enzyme, phosphoglycerate kinase. A model of the hypothetical hybrid protein (for convenience, named pgk_tryp) is shown in Figure 4-5 and can be more accurately described as the M237S, D252H, E302D mutant of phosphoglycerate kinase. The grafted residues have had their side chains rotated to best align with the actual trypsin catalytic triad. Although pgk_tryp shares

TABLE 4-4: RMS summary for novel matches

Values of RMS deviation are listed for the best overall matches from each GRAFTER analysis. A name has been given for each match, and the corresponding template and scaffold are listed. The residues matched and the associated RMS deviation values are tabulated.

TABLE 4-4: RMS summary for novel matches

Match Name †	Active Site Template	Match File #	Scaffold Protein	RMS for matched residues (all atoms)	RMS for matched residues (side chains)	RMS for matched residues (CA,CB & N)	match ¥ Active Site		match ¥ Scaffold	
pgk_tryp		6	phosphoglycerate kinase (3pgk)	1.37 Å	0.87 Å	0.90 Å	S	195	237	M
							H	57	252	D
							D	102	302	E
xi_ace	acetyl cholinesterase (1ace)	38	xylose isomerase (4xia)	1.46 Å	1.27 Å	1.32 Å	S	200	244A	D
							H	440	180A	E
							E	327	136A	W
thermo_pep		130	thermolysin (7tln)	1.07 Å	1.06 Å	0.52 Å	D	32	146	H
							D	215	165	N
adh_epilys	lysozyme epitope (1fdl)	47	alcohol dehydrogenase (5adh)	2.56 Å	2.60 Å	1.77 Å	D	18y	18	K
							N	19y	19	K
							G	117y	354	K
							D	119y	358	G
							Q	121y	360	D
							R	125y	364	S
							D	119y	13	L
							Q	121y	15	P
							R	125y	19	A

† A name has been given to each novel match. The prefix describes the scaffold, the suffix indicates the template.

¥ Corresponding template and scaffold atoms are listed in pairs.

Only these residues were included for the RMS deviation calculations.

similar relative side chain geometry with native trypsin, it lacks other key components. Trypsin's Asp102 is buried in the core of the protein, and while pgk_tryp's Asp302 points in towards the core, solvent accessibility calculations reveal that its C, O, Cα and Cβ atoms are solvent accessible. Also, there is some steric clash between the His252 side chain position and a scaffold loop (residues 306-310) that appears in the foreground of Figure 4-5b.

The best overall match for the acetyl cholinesterase active site positions the triad in the core of an α/β barrel in xylose isomerase. This match, xi_ace, is shown in Figure 4-6 and contains a Ser, His, Glu triad that matches the triad geometry from acetyl cholinesterase well, both by RMS fit and visual analysis. The D244S, E180H and W136E mutations required are in the A chain of the scaffold. Unfortunately, the serine is positioned so that its hydroxyl group points away from the accessible area within the barrel. There does not appear to be any way that a substrate could gain access to the serine.

Thermo_pep is perhaps the most provocative of the matches encountered. This match introduces the H146D and N165D mutations into thermolysin (Figure 4-7). The Asp dyad is grafted into the active site cleft in thermolysin between the helical and β-sheet domains in the scaffold enzyme. The geometry of the Asp pair is very close to that in native pepsin. There is a groove running between the two Asp side chains that could accommodate a peptide substrate. Substitution of an aspartate into position 146 replaces one of the metal chelating His residues involved in thermolysin's proteolytic activity.

While catalytic activity will require precise side chain geometries, the successful grafting of an antigenic conformational epitope onto a new scaffold is a more modest problem. In the top match for the lysozyme D1.3 epitope, 4 of the 6 residues: Asp18, Asn19, Gln121 and Arg125, were matched very well. The adh_epilys match is built on the alcohol dehydrogenase scaffold with the K18D, K19N, K354G, G358D, D360Q and

FIGURE 4-5: graft of trypsin onto phosphoglycerate kinase

Views of pgk_tryp aligned with the catalytic triad from trypsin (white). Grafted residues are shown (violet). In the figure, as is the case with all comparable figures, the grafted side chains have been rotated to best align with the template side chains.

a) close view of triad

b) overall view of scaffold with graft in place

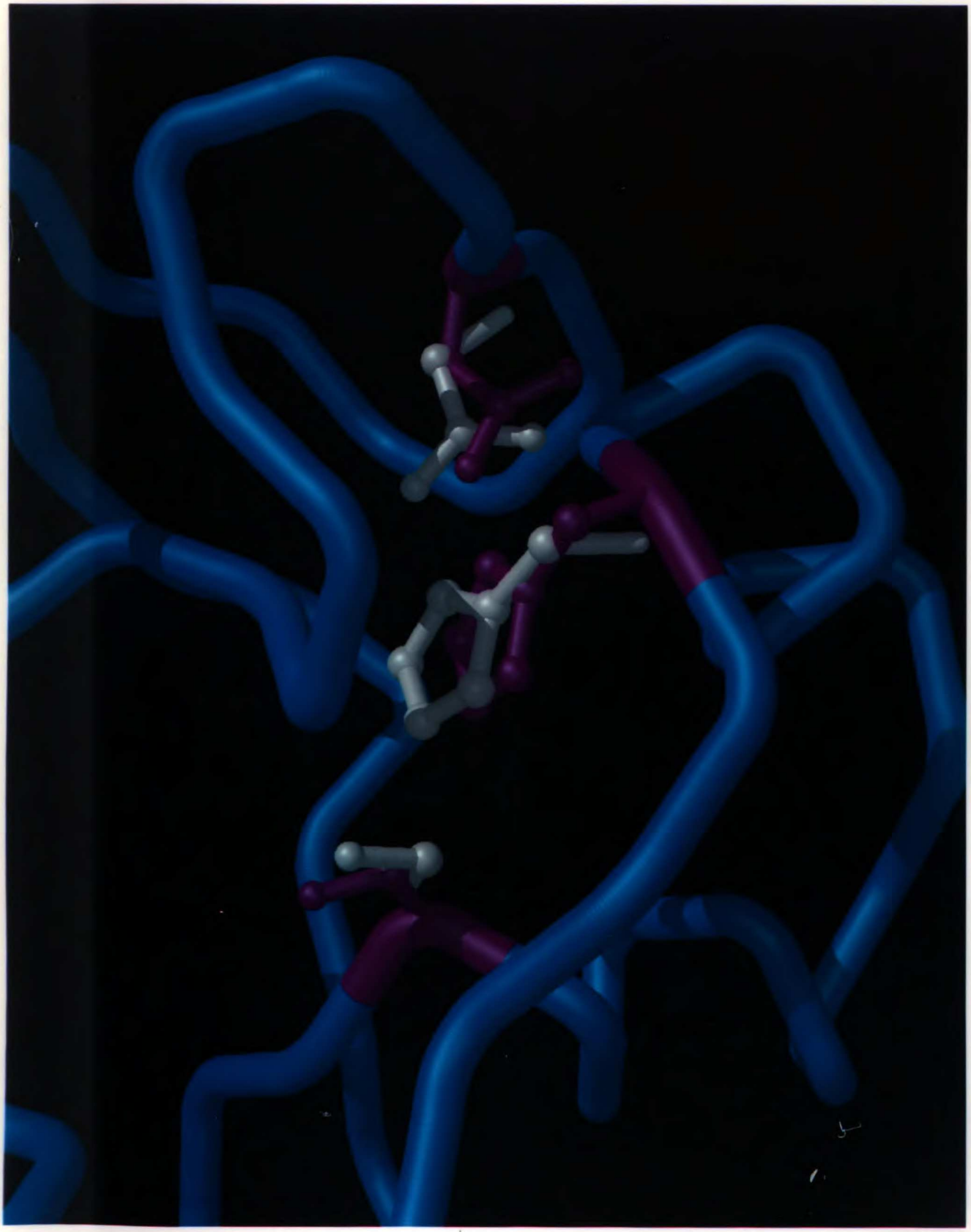
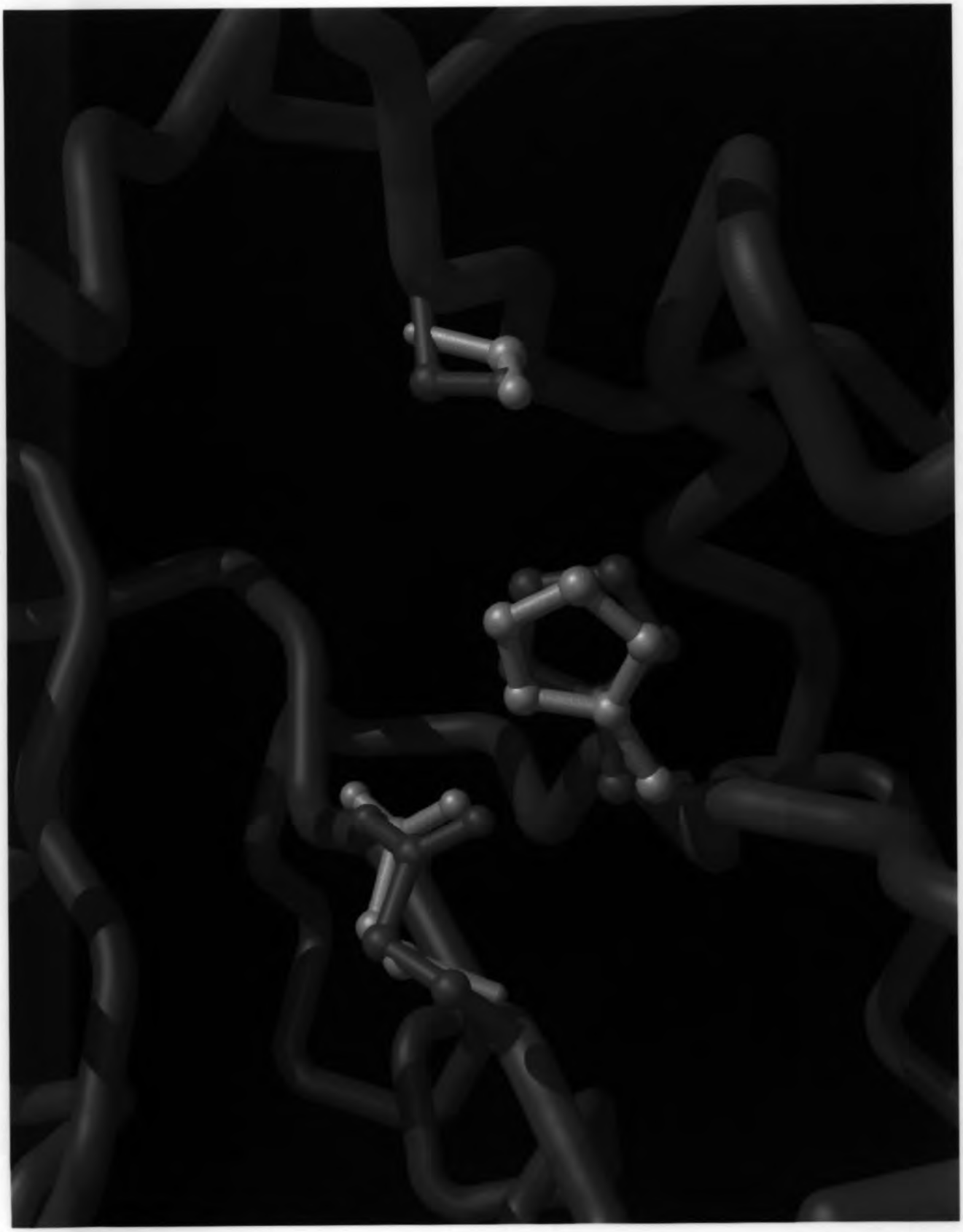


FIGURE 4-6: graft of acetylcholinesterase onto xylose isomerase

Views of xi_ace aligned with the catalytic triad from acetyl cholinesterase (white). Grafted residues are shown (blue).

a) close view of triad

b) overall view of scaffold with graft in place



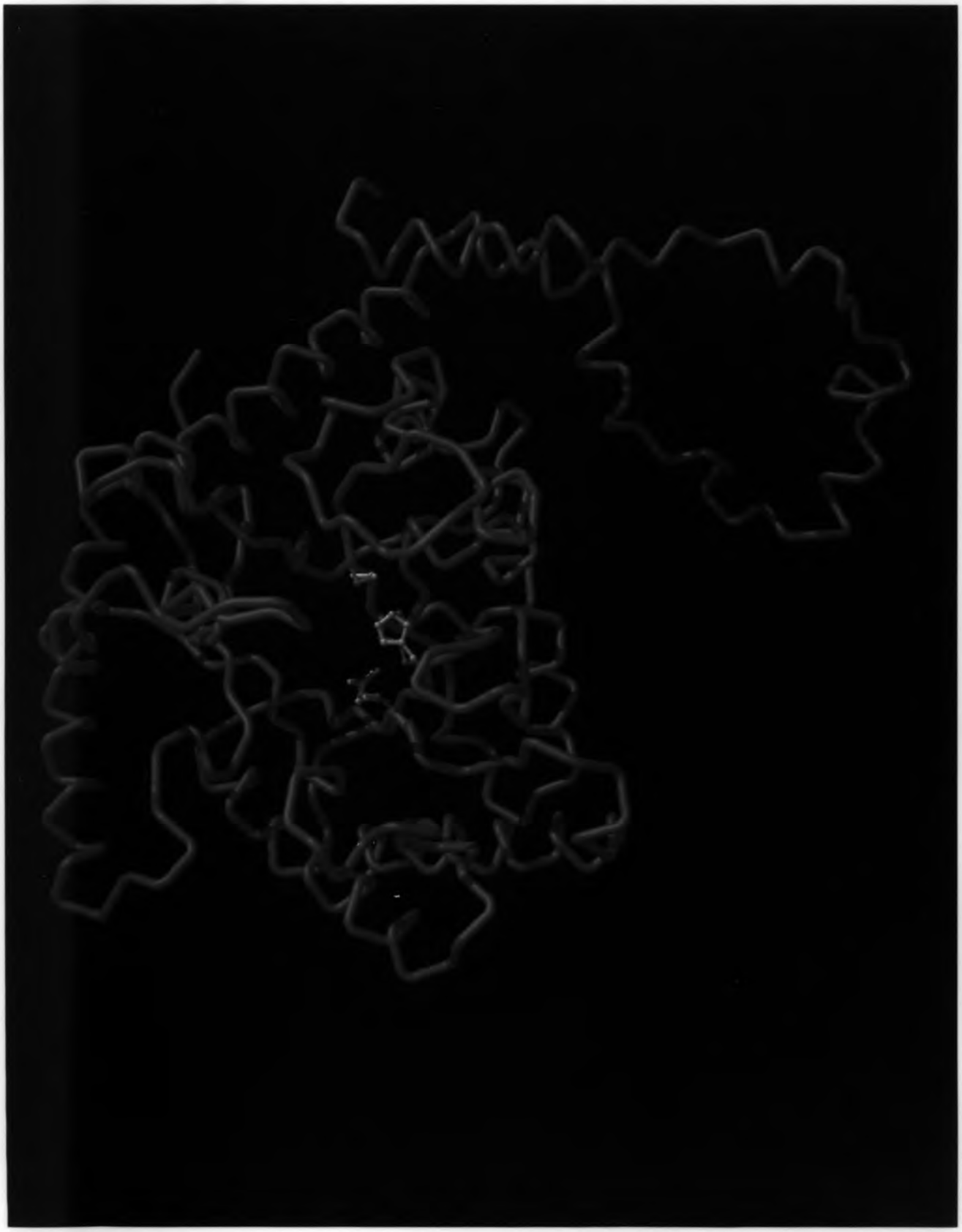
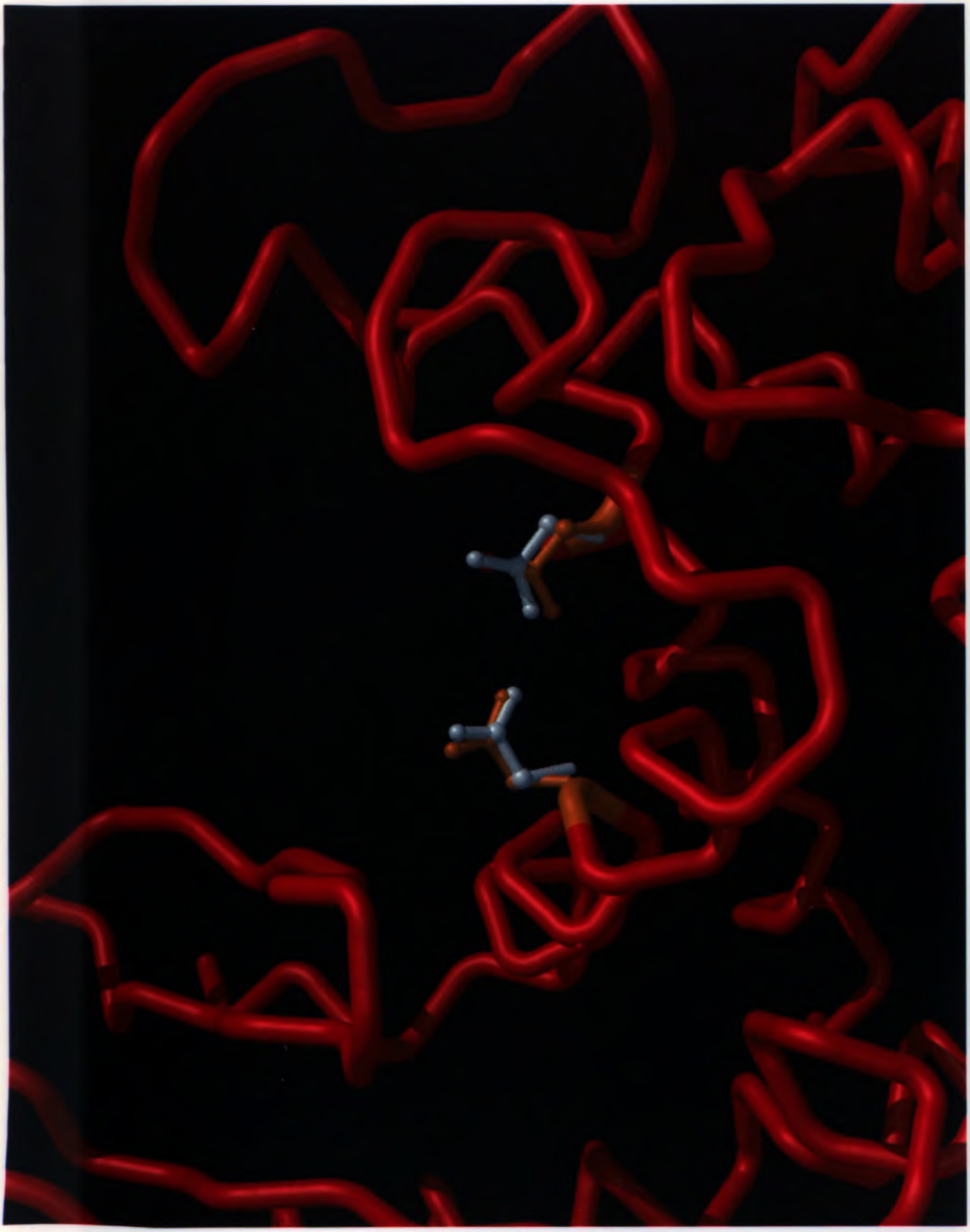


FIGURE 4-7: graft of pepsin onto thermolysin

Views of thermo_pep aligned with the catalytic dyad from pepsin (white). Grafted residues are shown (gold).

a) close view of dyad

b) overall view of scaffold with graft in place



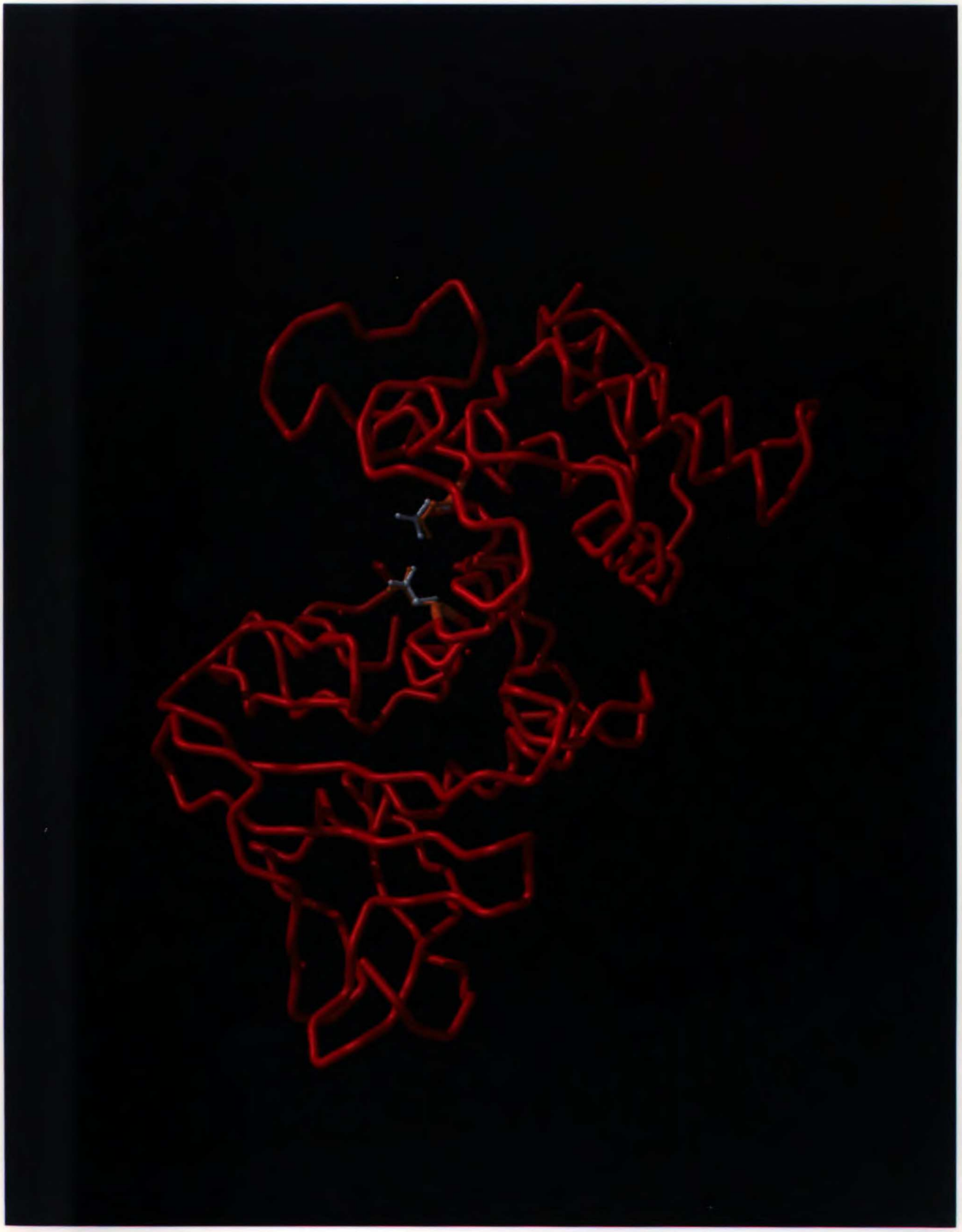
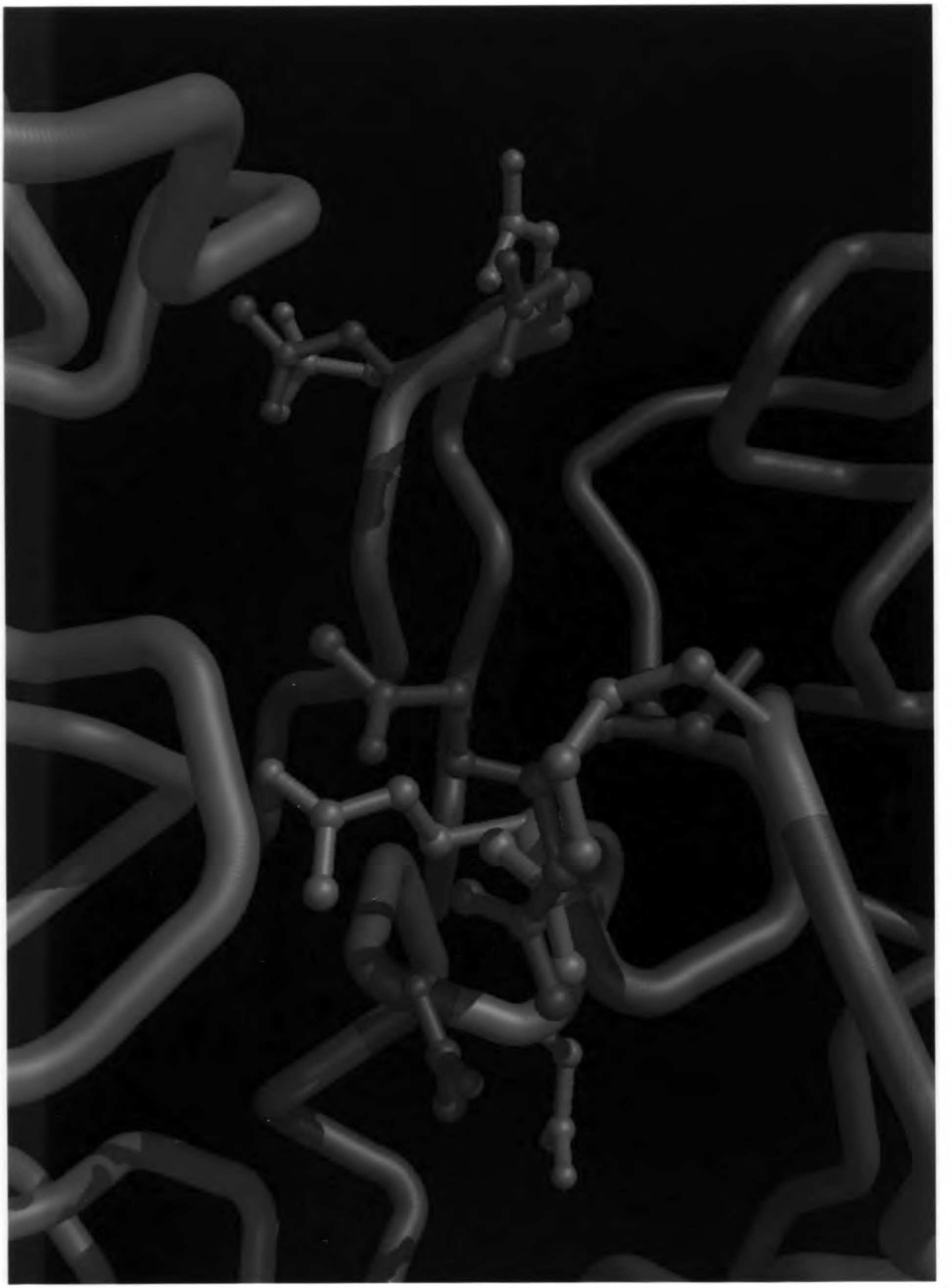


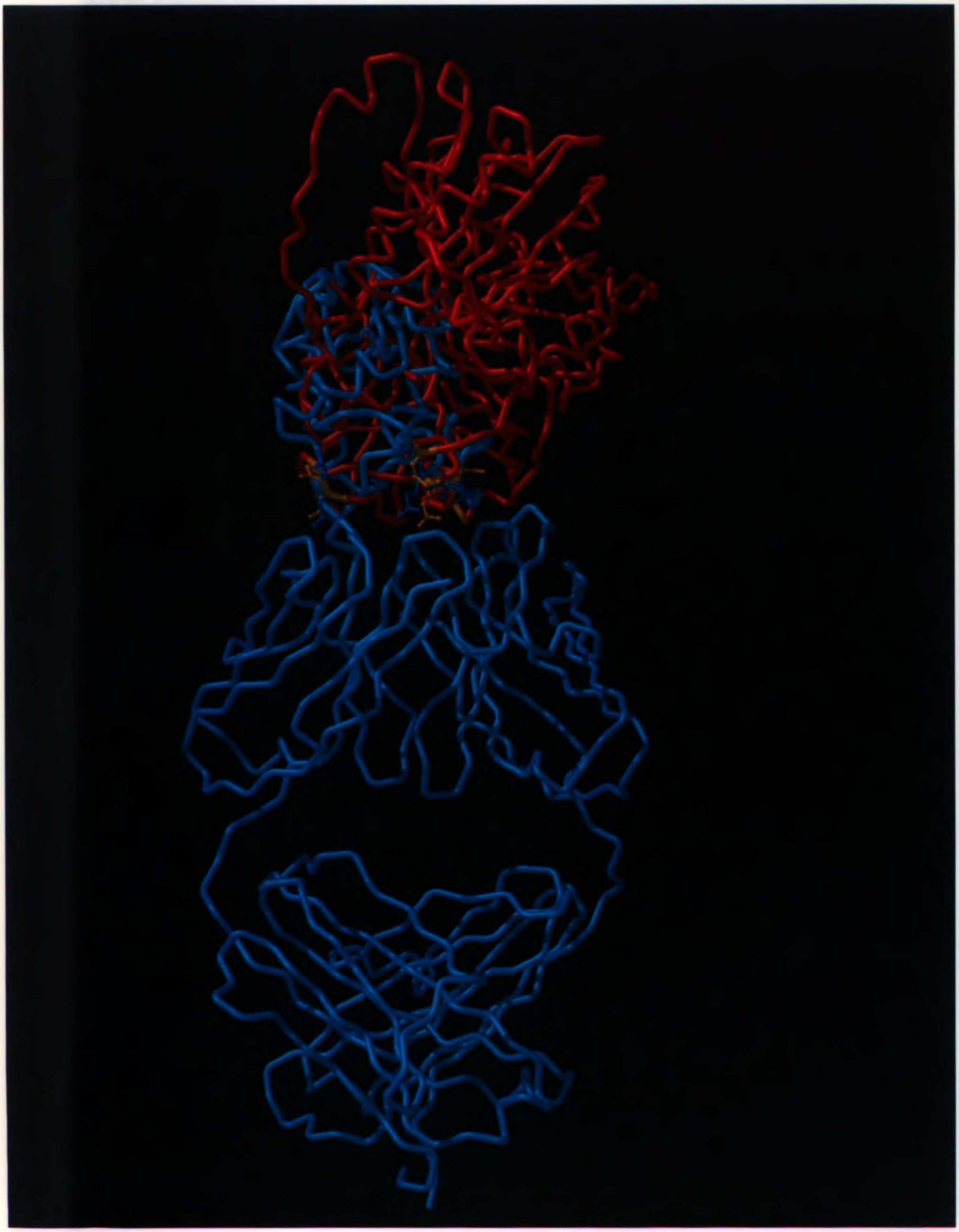
FIGURE 4-8: graft of a lysozyme epitope onto alcohol dehydrogenase

Views of adh_epilys (red) aligned to the lysozyme epitope in a structure (1fdl) of lysozyme (light blue) bound to antibody (light blue). Grafted residues are shown (gold) aligned to the epitope residues (dark blue).

a) close view of epitope and binding interface. The lysozyme molecule is not shown.

b) overall view of lysozyme bound to antibody with adh_epilys aligned to the epitope





S364R substitutions. The two residues, Gly117 and Asp119, were poorly matched in adh_epilys as seen in Figure 4-8. The four well matched residues are in upper two-thirds of the figure, while the poorly matched residues appear near the bottom. Figures 4-8b demonstrates the proposed interaction of the scaffold with the antibody as compared to the interaction of lysozyme with the antibody. The match falls on a convex surface of the scaffold, potentially allowing the exposed, grafted epitope to interact with the corresponding antibody. This graft appears to mimic the native lysozyme epitope effectively.

CONCLUSION

When successful, active site grafting will provide a means for producing new catalysts using old scaffolds. A scaffold possessing desired physical properties or molecular specificity could be altered to perform a new catalytic function. Grafting would allow us to build catalysts and binding motifs onto scaffolds that are efficiently expressed, purified and characterized structurally. Additional rounds of engineering on such a molecule would take less time both for production and analysis.

In our studies, we have attempted wherever possible to learn from examples available in nature. Has evolution developed enzymes that provide evidence that an active site grafting approach can work? The two enzymes subtilisin and trypsin demonstrate that some minimal set of structural features can function effectively when mounted onto completely different molecular scaffolds. While the Ser, His, Asp catalytic triad, an oxyanion hole and substrate specificity pockets are present in both of these serine proteases, the two proteins have completely different primary sequence and overall tertiary structures (Kraut 1977). Although this is not a result of concerted effort on the part of Nature, convergent evolution has generated two distinct scaffolds onto which the serine protease active site geometry and functionality have been grafted.

The two lysozymes, hen egg white and T4, are also examples of similar functionality with distinct scaffolds. However, in this case, the two enzymes are related by divergent evolution (Matthews, Grütter et al. 1981; Matthews, Remington et al. 1981). Their primary sequences are so disparate that it is difficult to determine a reasonable alignment. Yet, Matthews and coworkers have shown that the two enzymes can be aligned structurally, and that there is a correspondence between active site residues. We have used this pairing of active site residues as one of our test cases for the GRAFTER algorithm.

As described above, I have explored the possibility of grafting the same catalytic or binding motif onto different scaffolds while retaining functionality. Can the opposite experiment also succeed? Can the same protein serve as a scaffold for two distinct functions? Nature has demonstrated more than once that this is possible. Perhaps the most intriguing example of this is the two enzymes: mandelate racemase and muconate lactonizing enzyme (Neidhart, Kenyon et al. 1990). While sharing nearly identical overall protein structures, these enzymes contain different active site residues and perform different functions. Catalytic antibodies serve as another example of this; any number of functionalities can be mounted on very similar antibody scaffolds (Schultz 1989; Lerner, Benkovic et al. 1991).

A great deal of current research involves the use of antigenic peptides to induce antibodies toward a protein of interest. A short sequence of residues that is known to encompass an epitope for a given protein is used as the pattern for a synthetic peptide. This peptide is often coupled to protein carrier to stimulate antibody formation. Both native and synthetic polymers have been used successfully as carriers.

We propose an alternate approach to generating antibodies toward a particular epitope. The 3-dimensional structure of the epitope in question can be used as a pattern for a GRAFTER search. Any matches are potential scaffolds for mutagenesis to effectively graft the epitope structure onto a new scaffold. Scaffolds may be selected

based on stability, ease of purification and expression or other physical properties. In addition, an ideal scaffold might be one that is itself not antigenic and is a good candidate to be delivered as a therapeutic or prophylactic vaccine. Alternatively, if the scaffold were itself antigenic, a polyvalent vaccine might be the product of epitope grafting.

The trypsin/subtilisin comparison would appear to be a simple test case. Both enzymes are serine proteases that have catalytic Ser, His, Asp triads (Alden, Birktoft et al. 1971; Walter, Steigemann et al. 1982; Bryan, Pantoliano et al. 1986; Graf, Hegyi et al. 1988). In both, the His is positioned between Ser and Asp. They both possess oxyanion holes (Henderson 1970; Segal, Powers et al. 1971; Henderson, Wright et al. 1972; Robertus, Alden et al. 1972; Robertus, Kraut et al. 1972; Matthews, Alden et al. 1975; Poulos, Alden et al. 1976; Kraut 1977) and the Asp is buried amidst hydrophobic residues. Then why does GRAFTER have such difficulty matching all CA, CB and N atoms between the two triads? Figure 4-3 revealed the answer: side chains are positioned fairly similarly between the two, but this is accomplished through remarkably different backbone orientations. Does this indicate a failing in GRAFTER? In fact, it indicates that subtilisin and trypsin are poor choices for a test case. GRAFTER is based on a search algorithm that looks for similar backbone geometries. Similar side chain geometries are then extrapolated from the similar main chain geometries. Without similarity in main chain orientations, GRAFTER will not recognize geometries as matched. Therefore, GRAFTER must be applied with an understanding of the geometric assumptions made by the algorithm.

It is possible that this limitation of GRAFTER may, to some degree, be circumvented by applying additional evaluation tools. If GRAFTER were applied with looser constraints, more geometric matches would be generated. These matches would not be as exacting for main chain atom positions, but as we know from subtilisin and trypsin, that is not always necessary. The resulting matches could then be evaluated with

other tools, perhaps measuring accessibility, bad contacts or global motif orientation. The following chapter describes efforts toward that end,

REFERENCES

Abad-Zapatero, C., J. P. Griffith, et al. (1987). "Refined crystal structure of dogfish M4 apo-lactate dehydrogenase." J Mol Biol **198**(3): 445-67.

Adman, E. T., L. C. Siefker, et al. (1976). "Structure of Peptococcus aerogenes ferredoxin. Refinement at 2 A resolution." J Biol Chem **251**(12): 3801-6.

Alden, R. A., J. J. Birktoft, et al. (1971). "Atomic coordinates for subtilisin BPN' (or Novo)." Biochem Biophys Res Commun **45**(2): 337-44.

Anderson, C. M., R. E. Stenkamp, et al. (1978). "Sequencing a protein by x-ray crystallography. II. Refinement of yeast hexokinase B co-ordinates and sequence at 2.1 A resolution." J Mol Biol **123**(1): 15-33.

Arni, R., U. Heinemann, et al. (1988). "Three-dimensional structure of the ribonuclease T1 2'-GMP complex at 1.9-A resolution." J Biol Chem **263**(30): 15358-68.

Arutyunyan, E. G., I. P. Kuranova, et al. (1980). "X-ray structural investigation of leghemoglobin. VI. Structure of acetate-ferrileghemoglobin at a resolution of 2.0 angstroms (Russian)." Kristallografiya **25**: 80.

Bolin, J. T., D. J. Filman, et al. (1982). "Crystal structures of Escherichia coli and Lactobacillus casei dihydrofolate reductase refined at 1.7 A resolution. I. General features and binding of methotrexate." J Biol Chem **257**(22): 13650-62.

Bolognesi, M., S. Onesti, et al. (1989). "Aplysia limacina myoglobin. Crystallographic analysis at 1.6 A resolution." J Mol Biol **205**(3): 529-44.

Brick, P. and D. M. Blow (1987). "Crystal structure of a deletion mutant of a tyrosyl-tRNA synthetase complexed with tyrosine." J Mol Biol **194**(2): 287-97.

Bryan, P., M. W. Pantoliano, et al. (1986). "Site-directed mutagenesis and the role of the oxyanion hole in subtilisin." Proc Natl Acad Sci U S A **83**(11): 3743-5.

Buehner, M., H. J. Hecht, et al. (1982). "Crystallization and preliminary crystallographic analysis at low resolution of the allosteric L-lactate dehydrogenase from Lactobacillus casei." J Mol Biol **162**(4): 819-38.

Bystroff, C., S. J. Oatley, et al. (1990). "Crystal structures of Escherichia coli dihydrofolate reductase: the NADP+ holoenzyme and the folate.NADP+ ternary complex. Substrate binding and a model for the transition state." Biochemistry **29**(13): 3263-77.

Carter, D. C., K. A. Melis, et al. (1985). "Crystal structure of Azotobacter cytochrome c5 at 2.5 A resolution." J Mol Biol **184**(2): 279-95.

Cohen, F. E. and M. J. Sternberg (1980). "On the prediction of protein structure: The significance of the root-mean-square deviation." J. Mol. Biol. **138**(2): 321-333.

Colonna-Cesari, F., D. Perahia, et al. (1986). "Interdomain motion in liver alcohol dehydrogenase. Structural and energetic analysis of the hinge bending mode." J Biol Chem **261**(32): 15273-80.

Cooper, J. B., G. Khan, et al. (1990). "X-ray analyses of aspartic proteinases. II. Three-dimensional structure of the hexagonal crystal form of porcine pepsin at 2.3 A resolution." J Mol Biol **214**(1): 199-222.

Cotton, F. A., E. E. Hazen Jr., et al. (1979). "Staphylococcal nuclease: proposed mechanism of action based on structure of enzyme-thymidine 3',5'-bisphosphate-calcium ion complex at 1.5-A resolution." Proc Natl Acad Sci U S A **76**(6): 2551-5.

Diamond, R. (1974). "Real-space refinement of the structure of hen egg-white lysozyme." J Mol Biol **82**(3): 371-91.

Dijkstra, B. W., K. H. Kalk, et al. (1984). "Role of the N-terminus in the interaction of pancreatic phospholipase A2 with aggregated substrates. Properties and crystal structure of transaminated phospholipase A2." Biochemistry **23**(12): 2759-66.

Dijkstra, B. W., K. H. Kalk, et al. (1981). "Structure of bovine pancreatic phospholipase A2 at 1.7A resolution." J Mol Biol **147**(1): 97-123.

Drenth, J., W. G. Hol, et al. (1972). "A comparison of the three-dimensional structures of subtilisin BPN' and subtilisin novo." Cold Spring Harb Symp Quant Biol **36**: 107-16.

Eriksson, A. E., P. M. Kylsten, et al. (1988). "Crystallographic studies of inhibitor binding sites in human carbonic anhydrase II: a pentacoordinated binding of the SCN- ion to the zinc at high pH." Proteins **4**(4): 283-93.

Evans, P. R., G. W. Farrants, et al. (1981). "Phosphofructokinase: structure and control." Philos Trans R Soc Lond [Biol] **293**(1063): 53-62.

Fischmann, T. O., G. A. Bentley, et al. (1991). "Crystallographic Refinement of the Three-dimensional Structure of the FabD1.3-Lysozyme Complex at 2.5Å Resolution." J. Bio. Chem. **266**(20): 12915-12920.

Fukuyama, K., T. Hase, et al. (1980). "Structure of S. platensis (2Fe-2S) ferredoxin and evolution of chloroplast-type ferredoxins." Nature **286**: 522.

Fukuyama, K., H. Matsubara, et al. (1989). "Structure of [4Fe-4S] ferredoxin from Bacillus thermoproteolyticus refined at 2.3 A resolution. Structural comparisons of bacterial ferredoxins." J Mol Biol **210**(2): 383-98.

Gouaux, J. E. and W. N. Lipscomb (1990). "Crystal structures of phosphonoacetamide ligated T and phosphonoacetamide and malonate ligated R states of aspartate carbamoyltransferase at 2.8-A resolution and neutral pH." Biochemistry **29**(2): 389-402.

Gouaux, J. E., R. C. Stevens, et al. (1990). "Crystal structures of aspartate carbamoyltransferase ligated with phosphonoacetamide, malonate, and CTP or ATP at 2.8-A resolution and neutral pH." Biochemistry **29**(33): 7702-15.

Graf, L., G. Hegyi, et al. (1988). "Structural and functional integrity of specificity and catalytic sites of trypsin." Int J Pept Protein Res **32**(6): 512-8.

Grau, U. M., W. E. Trommer, et al. (1981). "Structure of the active ternary complex of pig heart lactate dehydrogenase with S-lac-NAD at 2.7 A resolution." J Mol Biol **151**(2): 289-307.

Henderson, R. (1970). "Structure of crystalline alpha-chymotrypsin. IV. The structure of indoleacryloyl-alpha-chyotrypsin and its relevance to the hydrolytic mechanism of the enzyme." J Mol Biol **54**(2): 341-54.

Henderson, R., C. S. Wright, et al. (1972). "-Chymotrypsin: what can we learn about catalysis from x-ray diffraction?" Cold Spring Harb Symp Quant Biol **36**: 63-70.

Henrick, K., C. A. Collyer, et al. (1989). "Structures of D-xylose isomerase from Arthrobacter strain B3728 containing the inhibitors xylitol and D-sorbitol at 2.5 A and 2.3 A resolution, respectively." J Mol Biol **208**(1): 129-57.

Higuchi, Y., M. Kusunoki, et al. (1984). "Refined structure of cytochrome c3 at 1.8 A resolution." J Mol Biol **172**(1): 109-39.

Holmes, M. A. and B. W. Matthews (1981). "Binding of hydroxamic acid inhibitors to crystalline thermolysin suggests a pentacoordinate zinc intermediate in catalysis." Biochemistry **20**(24): 6912-20.

Holmes, M. A., D. E. Tronrud, et al. (1983). "Structural analysis of the inhibition of thermolysin by an active-site-directed irreversible inhibitor." Biochemistry **22**(1): 236-40.

Jurnak, F. (1985). "Structure of the GDP domain of EF-Tu and location of the amino acids homologous to ras oncogene proteins." Science **230**(4721): 32-6.

Karplus, P. A. and G. E. Schulz (1987). "Refined structure of glutathione reductase at 1.54 A resolution." J Mol Biol **195**(3): 701-29.

Ke, H. M., W. N. Lipscomb, et al. (1988). "Complex of N-phosphonacetyl-L-aspartate with aspartate carbamoyltransferase. X-ray refinement, analysis of conformational changes and catalytic and allosteric mechanisms." J Mol Biol **204**(3): 725-47.

Kim, Y. C., J. C. Grable, et al. (1990). "Refinement of Eco RI endonuclease crystal structure: a revised protein chain tracing." Science **249**(4974): 1307-9.

Kraut, J. (1977). "Serine proteases: structure and mechanism of catalysis." Annu Rev Biochem **46**: 331-358.

Kuriyan, J., S. Wilz, et al. (1986). "X-ray structure and refinement of carbon-monoxy (Fe II)-myoglobin at 1.5 A resolution." J Mol Biol **192**(1): 133-54.

la Cour, T. F., J. Nyborg, et al. (1985). "Structural details of the binding of guanosine diphosphate to elongation factor Tu from E. coli as studied by X-ray crystallography." Embo J **4**(9): 2385-8.

Lerner, R. A., S. J. Benkovic, et al. (1991). "At the Crossroads of Chemistry and Immunology: Catalytic Antibodies." Science **252**: 659-667.

Leslie, A. G. (1990). "Refined crystal structure of type III chloramphenicol acetyltransferase at 1.75 A resolution." J Mol Biol **213**(1): 167-86.

Lewendon, A., I. A. Murray, et al. (1990). "Evidence for transition-state stabilization by serine-148 in the catalytic mechanism of chloramphenicol acetyltransferase." Biochemistry **29**(8): 2075-80.

Lolis, E. and G. A. Petsko (1990). "Crystallographic analysis of the complex between triosephosphate isomerase and 2-phosphoglycolate at 2.5-A resolution: implications for catalysis." Biochemistry **29**(28): 6619-25.

Loll, P. J. and E. E. Lattman (1989). "The crystal structure of the ternary complex of staphylococcal nuclease, Ca2+, and the inhibitor pdTp, refined at 1.65 A." Proteins **5**(3): 183-201.

Martin, A. E., B. K. Burgess, et al. (1990). "Site-directed mutagenesis of Azotobacter vinelandii ferredoxin I: [Fe-S] cluster-driven protein rearrangement." Proc Natl Acad Sci U S A **87**(2): 598-602.

Martin, J. L., L. N. Johnson, et al. (1990). "Comparison of the binding of glucose and glucose 1-phosphate derivatives to T-state glycogen phosphorylase b." Biochemistry **29**(48): 10745-57.

Matsuura, Y., T. Takano, et al. (1982). "Structure of cytochrome c551 from Pseudomonas aeruginosa refined at 1.6 A resolution and comparison of the two redox forms." J Mol Biol **156**(2): 389-409.

Matthews, B. W., M. G. Grütter, et al. (1981). "Common precursor of lysozymes of hen egg-white and bacteriophage T4." Nature **290**: 334-335.

Matthews, B. W., S. J. Remington, et al. (1981). "Relation Between Hen Egg White Lysozyme and Bacteriophage T4 Lysozyme: Evolutionary Implications." J. Mol. Biol. **147**: 545-558.

Matthews, D. A., R. A. Alden, et al. (1975). "X-ray crystallographic study of boronic acid adducts with subtilisin BPN' (Novo). A model for the catalytic transition state." J Biol Chem **250**(18): 7120-6.

Matthews, D. A., J. T. Bolin, et al. (1985). "Refined crystal structures of Escherichia coli and chicken liver dihydrofolate reductase containing bound trimethoprim." J Biol Chem **260**(1): 381-91.

Neidhart, D. J., G. L. Kenyon, et al. (1990). "Mandelate Racemase and Muconate Lactonizing Enzyme are Mechanistically Distinct and Structurally Homologous." Nature **347**(6294): 692-694.

Nishikawa, K. and T. Ooi (1974). "Comparison of homologous tertiary structures of proteins." J Theor Biol **43**(2): 351-74.

Ochi, H., Y. Hata, et al. (1983). "Structure of rice ferricytochrome c at 2.0 A resolution." J Mol Biol **166**(3): 407-18.

Pai, E. F., U. Krengel, et al. (1990). "Refined crystal structure of the triphosphate conformation of H-ras p21 at 1.35 A resolution: implications for the mechanism of GTP hydrolysis." Embo J **9**(8): 2351-9.

Pantoliano, M. W., M. Whitlow, et al. (1989). "Large increases in general stability for subtilisin BPN' through incremental changes in the free energy of unfolding." Biochemistry **28**(18): 7205-13.

Phillips, D. C. (1970). "The development of crystallographic enzymology." Biochem Soc Symp **30**: 11-28.

Phillips, S. E. (1980). "Structure and refinement of oxymyoglobin at 1.6 A resolution." J Mol Biol **142**(4): 531-54.

Phillips, S. E. and B. P. Schoenborn (1981). "Neutron diffraction reveals oxygen-histidine hydrogen bond in oxymyoglobin." Nature **292**(5818): 81-2.

Pierrot, M., R. Haser, et al. (1982). "Crystal structure and electron transfer properties of cytochrome c3." J Biol Chem **257**(23): 14341-8.

Piontek, K., P. Chakrabarti, et al. (1990). "Structure determination and refinement of Bacillus stearothermophilus lactate dehydrogenase." Proteins **7**(1): 74-92.

Ponder, J. W. and F. M. Richards (1987). "Tertiary Templates for Proteins: Use of Packing Criteria in the Enumeration of Allowed Sequences for Different Structural Classes." J. Mol. Biol. **193**: 775-791.

Poulos, T. L., R. A. Alden, et al. (1976). "Polypeptide halomethyl ketones bind to serine proteases as analogs of the tetrahedral intermediate. X-ray crystallographic comparison of lysine- and phenylalanine-polypeptide chloromethyl ketone-inhibited subtilisin." J Biol Chem **251**(4): 1097-103.

Remington, S., G. Wiegand, et al. (1982). "Crystallographic refinement and atomic models of two different forms of citrate synthase at 2.7 and 1.7 A resolution." J Mol Biol **158**(1): 111-52.

Robbins, A. H. and C. D. Stout (1989). "Structure of activated aconitase: formation of the [4Fe-4S] cluster in the crystal." Proc Natl Acad Sci U S A **86**(10): 3639-43.

Robertus, J. D., R. A. Alden, et al. (1972). "An x-ray crystallographic study of the binding of peptide chloromethyl ketone inhibitors to subtilisin BPN'." Biochemistry **11**(13): 2439-49.

Robertus, J. D., J. Kraut, et al. (1972). "Subtilisin; a stereochemical mechanism involving transition-state stabilization." Biochemistry **11**(23): 4293-303.

Salemme, F. R., S. T. Freer, et al. (1973). "The structure of oxidized cytochrome c 2 of Rhodospirillum rubrum." J Biol Chem **248**(11): 3910-21.

Schreuder, H. A., J. M. van der Laan, et al. (1988). "Crystal structure of p-hydroxybenzoate hydroxylase complexed with its reaction product 3,4-dihydroxybenzoate." J Mol Biol **199**(4): 637-48.

Schultz, P. G. (1989). "Catalytic Antibodies." Acc. Chem. Res. **22**(8): 287-294.

Scouloudi, H. and E. N. Baker (1978). "X-ray crystallographic studies of seal myoglobin. The molecule at 2.5 A resolution." J Mol Biol **126**(4): 637-60.

Segal, D. M., J. C. Powers, et al. (1971). "Substrate binding site in bovine chymotrypsin A-gamma. A crystallographic study using peptide chloromethyl ketones as site-specific inhibitors." Biochemistry **10**(20): 3728-38.

Sippl, M. J. (1982). "On the problem of comparing protein structures. Development and applications of a new method for the assessment of structural similarities of polypeptide conformations." J Mol Biol **156**(2): 359-88.

Smith, D. L., S. C. Almo, et al. (1989). "2.8-A-resolution crystal structure of an active-site mutant of aspartate aminotransferase from Escherichia coli." Biochemistry **28**(20): 8161-7.

Smith, W. W., R. M. Burnett, et al. (1977). "Structure of the semiquinone form of flavodoxin from Clostridum MP. Extension of 1.8 A resolution and some comparisons with the oxidized state." J Mol Biol **117**(1): 195-225.

Steigemann, W. and E. Weber (1979). "Structure of erythrocruorin in different ligand states refined at 1.4 A resolution." J Mol Biol **127**(3): 309-38.

Stevens, R. C., J. E. Gouaux, et al. (1990). "Structural consequences of effector binding to the T state of aspartate carbamoyltransferase: crystal structures of the unligated and ATP- and CTP-complexed enzymes at 2.6-A resolution [published erratum appears in Biochemistry 1990 Dec 18;29(50):11146]." Biochemistry **29**(33): 7691-701.

Stout, C. D. (1993). "Crystal structures of oxidized and reduced Azotobacter vinelandii ferredoxin at pH 8 and 6." J Biol Chem **268**(34): 25920-7.

Strynadka, N. C. and M. N. James (1988). "Two trifluoperazine-binding sites on calmodulin predicted from comparative molecular modeling with troponin-C." Proteins **3**(1): 1-17.

Sussman, J. L., M. Harel, et al. (1991). "Atomic structure of acetylcholinesterase from Torpedo californica: a prototypic acetylcholine-binding protein." Science **253**(5022): 872-9.

Takano, T. (1977). "Structure of myoglobin refined at 2-0 A resolution. II. Structure of deoxymyoglobin from sperm whale." J Mol Biol **110**(3): 569-84.

Takano, T. and R. E. Dickerson (1981). "Conformation change of cytochrome c. I. Ferrocytochrome c structure refined at 1.5 A resolution." J Mol Biol **153**(1): 79-94.

Tanaka, N., T. Yamane, et al. (1975). "The crystal structure of bonito (katsuo) ferrocytochrome c at 2.3 A resolution. II. Structure and function." J Biochem (Tokyo) **77**(1?): 147-62.

Timkovich, R. and R. E. Dickerson (1976). "The structure of Paracoccus denitrificans cytochrome c550." J Biol Chem **251**(13): 4033-46.

Walter, J., W. Steigemann, et al. (1982). "On the disordered activation domain in trypsinogen. Chemical labelling and low-temperature crystallography." Acta Crystallogr. Sect. B **38**: 1462.

Watenpaugh, K. D., L. C. Sieker, et al. (1973). "The binding of riboflavin-5'-phosphate in a flavoprotein: flavodoxin at 2.0-Angstrom resolution." Proc Natl Acad Sci U S A **70**(12): 3857-60.

Watson, H. C. (1969). "The stereochemistry of the protein myoglobin." Prog. Stereochem. **4**: 299.

Watson, H. C., N. P. Walker, et al. (1982). "Sequence and structure of yeast phosphoglycerate kinase." Embo J **1**(12): 1635-40.

Weaver, L. H. and B. W. Matthews (1987). "Structure of bacteriophage T4 lysozyme refined at 1.7 A resolution." J Mol Biol **193**(1): 189-99.

Weber, I. T. and T. A. Steitz (1984). "Model of specific complex between catabolite gene activator protein and B-DNA suggested by electrostatic complementarity." Proc Natl Acad Sci U S A **81**(13): 3973-7.

Weber, I. T., T. A. Steitz, et al. (1987). "Predicted structures of cAMP binding domains of type I and II regulatory subunits of cAMP-dependent protein kinase." Biochemistry **26**(2): 343-51.

White, J. L., M. L. Hackert, et al. (1976). "A comparison of the structures of apo dogfish M4 lactate dehydrogenase and its ternary complexes." J Mol Biol **102**(4): 759-79.

Winn, S. I., H. C. Watson, et al. (1981). "Structure and activity of phosphoglycerate mutase." Philos Trans R Soc Lond [Biol] **293**(1063): 121-30.

Zhao, B., M. Carson, et al. (1992). "Structure of scorpion toxin variant-3 at 1.2 A resolution." J Mol Biol **227**(1): 239-52.

Chapter 5:
The GRAFTER Scoring Modules

INTRODUCTION

GRAFTER is an efficient tool for motif identification in scaffolds. It can generate very long lists of potential graft sites for a particular motif. Although geometric similarity is a useful search criterion, it proves to be less suitable for subsequent scoring of matches. To enhance the ability to recognize the best matches for a functional motif, I have developed a series of scoring modules to evaluate graft sites. These modules (STERIC, aRMS and ORIENT) focus on steric conflicts, accessibility and motif environment. The scores generated by these modules can be combined in an evenly weighted manner so as to generate a total score for each match. This approach provides us with an additional level of screening before visual analysis on a graphics terminal is necessary. In addition, these scoring modules allow us to generate more matches with GRAFTER, using less stringent geometric constraints. Even though GRAFTER will generate more matches, the scoring modules allow the list to be trimmed back down to a manageable size.

As with GRAFTER itself, the new suite of programs (the GRAFTER suite) made up of GRAFTER and the additional scoring modules had to be evaluated on known test systems before being applied to novel design efforts. To verify and confirm the operation of GRAFTER in recognizing structural features, the λ repressor and the 434 repressor were compared. Our goal was to identify a design analogous to Ptashne's specificity swap between the 434 and P22 repressors (Wharton and Ptashne 1985). Relatively loose constraints were applied in the GRAFTER geometric search. Therefore, the scoring functions were mainly responsible for the resulting list of top matches.

To evaluate the scoring modules from the GRAFTER suite, loops from the complementarity-determining regions (CDRs) of 26 antibody and antigen-binding fragment (FAB) structures were compared. As with the repressor comparisons, minimal constraints were applied during the geometric search. The resulting data was used to

group the CDR into families of similar structure. These comparisons are significant from a protein engineering perspective. The residues from one CDR can be swapped into corresponding positions on a second CDR using standard mutagenesis techniques. The resulting antibodies can be analyzed in binding assays to determine whether they adopt the binding characteristics from the new residues. This approach has been used previously to place desired binding properties from one antibody onto an antibody scaffold from another organism (Winter and Harris 1993).

Finally, a known epitope in human growth hormone (hGH) was used to probe the structure of interleukin-4 (IL-4) to identify sites for epitope grafting. Unlike the two previous examples, the hGH epitope contains 5 non-contiguous residues (Jin and Wells 1992). This test case is an example of a graft which is not simply a replacement of a linear sequence of residues. Grafts of this sort are the most difficult to identify visually, because often there are no secondary structure elements to serve as clues.

By following these steps, I have evaluated the GRAFTER suite in both stringent verification tests and real world applications. Final assessment and optimization for real world use requires experimental testing and feedback. I hope that shortly such experimental data will be available so that GRAFTER suited can be further refined.

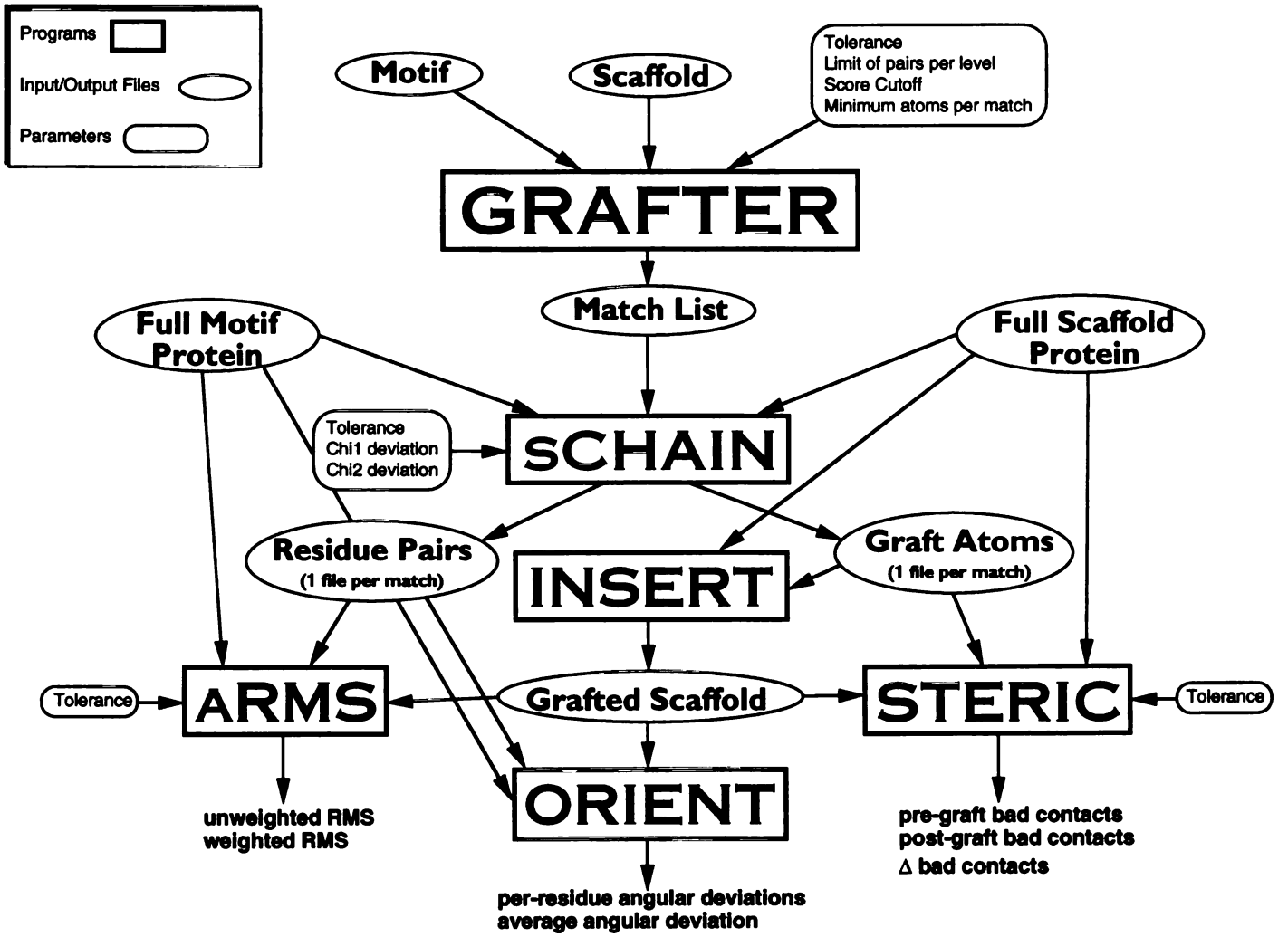
METHODS

To augment the geometric criteria that govern a GRAFTER search, we have developed a number of post-processing modules that help evaluate matches based on general structural properties. The overall scheme relating GRAFTER and these post-processing modules is depicted in Figure 5-1. SCHAIN helps eliminate redundant matches from GRAFTER's output, and builds side chains onto the graft site following the backbone conformation dependent rotamer library of Dunbrack and Karplus (Dunbrack and Karplus 1993). INSERT prepares the grafted structures by inserting the graft

FIGURE 5-1: GRAFTER suite overview

Overview of the Grafter Suite - indicates all programs, parameters and input/output.

FIGURE 5-1: overview of GRAFTER suite



residues into the full scaffold structure. STERIC, ARMS and ORIENT evaluate each graft in terms of bad contacts, accessibility-weighted RMS and overall orientation, respectively. The suite may be used in a mode where the geometric constraints applied in GRAFTER are not very restrictive. This produces a large number of proposed matches as output from GRAFTER. In this mode, any narrowing of the graft search is a result of post-processor scoring. This mode also serves as a test of the effectiveness of the post-GRAFTER scoring modules. The following provides greater detail on each module:

SCHAIN

GRAFTER's atom-based combinatorics can potentially lead to duplicate residue-based matches, as well as numerous sub-matches that are all simply part of a larger match. The bookkeeping necessary to eliminate such duplication within GRAFTER can become an overwhelming burden on the combinatorial algorithm itself. It is more efficient to allow these duplicates to be generated by GRAFTER, and filter them in subsequent step. The SCHAIN module eliminates all sub-matches that are part of a larger match that is also present, and removes any duplicate matches. The SCHAIN module also builds the side chains from the motif residues onto the corresponding residues from the scaffold. Side chains are built using side-chain dihedral (χ) angles chosen from Dunbrack's database (Dunbrack and Karplus 1993), which correlates preferred χ angles with a residue's main-chain dihedral values. This method incorporates information from a residue's local environment (its backbone conformation) in selecting a side chain orientation, unlike the rotamer library (Ponder and Richards 1987) used by DEZYMER. We allow for some adjustment of the χ values, within user defined constraints (the default is $\pm 20^{\circ}$ for each dihedral). The angles are adjusted toward the angles found in the motif side-chains.

INSERT

The GRAFTER combinatorial search often generates large numbers (>100) of possible graft sites for a particular motif. If we assemble a complete grafted protein for every graft simultaneously, we would build a database with size comparable to the PDB. Disk storage restrictions prevent such wholesale assembly. Instead, only the graft residues are prepared and stored by SCHAIN. INSERT takes a collection of newly built residues and inserts them into a scaffold structure, replacing the corresponding residues in the structure. INSERT allows us to store a large number of grafts without duplicating the scaffold structure for each graft.

STERIC

Given the success of hard sphere models in describing many of the features in proteins, we have incorporated a steric evaluation (based on the hard sphere model) into the GRAFTER suite. The STERIC module evaluates how well a given set of atoms fits into a grafted structure. Only those atoms that are grafted are scored. We speed the calculation by limiting the atoms included to those that are within the sphere that is centered on the graft site and has radius large enough to enclose all atoms in the graft, plus a user-defined tolerance. The number of bad contacts is defined by the sigmoidal switching function

$$badContacts = \sum_i \sum_j \frac{1}{1 + 50 \cdot \left(\frac{\mathrm{dist}(i,j)}{r_i + r_j} \right)^{50}}$$

where

r_n is the radius for atom n; and

i and j are chosen so that the two atoms are not within 3 bonds of each other.

The steric score is determined for all target residues prior to the graft, and for the corresponding residues after the graft. We report the change in the number of bad contacts as well.

ARMS

Although appropriate geometry is essential for a potential graft site, there are many sites in proteins that score well by RMS deviation with respect to a motif, but are either too buried or too exposed to mimic the native environment of the motif. We have designed a modified RMS deviation routine that incorporates accessibility as well as geometry into the evaluation. ARMS is this accessibility-weighted RMS deviation module. It is designed to compare a graft geometry and environment to those of the original motif. ARMS evaluates how similar each residue's accessibility and orientation are to the motif. Accessibility is calculated using the ACCESS program (Lee and Richards 1971) The RMS deviation (by rotation) is calculated with uniform weights, and also with weights based on accessibility differences between the corresponding atoms in the motif and the graft (Kabsch 1976). In the weighted calculation, the minimum weight is 1.0 (corresponding to the uniform weight value). The greater the difference between a given atom's accessibility in the motif and its corresponding value in the graft, the more heavily it is weighted in the RMS calculation. Thus, a grafted structure is penalized if its atoms do not retain their accessibility values from the motif. The accessibility calculation takes a significant amount of time, so we limit the atoms included to a tolerance adjusted sphere around the grafted residues. This module (unlike the RMS calculation within GRAFTER) generates an RMS value for the full graft, including all main-chain and side-chain atoms. This full graft-motif RMS is essential for evaluating the overall orientation of all of the residues in the motif and is possible at this stage because the side chains have been placed.

ORIENT

In some searches, matches are identified that score well in terms of weighted RMS deviation and steric score, but have an overall orientation different from the desired motif. One such case is a collection of residues on two helices in a four-helix bundle (Presnell and Cohen 1989). In searches of other four-helix bundles, we observe numerous matches that score well by STERIC and ARMS evaluations, but face into the center of the bundle. The residues on the inside of the bundle are sufficiently accessible to the solvent that ARMS scores them well compared to the motif. To recognize such cases, the ORIENT module was developed. For each residue in the potential graft site, the Cα-Cβ-(center of mass) angle is calculated, and the difference between it and the corresponding angle in the motif is recorded. The overall score is the average of all angular differences for the graft. We find that the ORIENT module clearly differentiates matches on the inside of a bundle from those on the outside.

GSTAT

We wish to combine individual scores into a composite score. By representing scores in terms of standard deviations from the mean, we can sum them to generate a composite score. However, if the distribution for each individual score is not well balanced, the scores will not contribute equally to the composite. Therefore, outlying scores are eliminated from the distributions. Due to the nature of the scores, outliers only occur on the high (worse scoring) side of the distribution. The GSTAT module performs all necessary statistical processing on scores from the GRAFTER suite. It accepts multiple scores and determines the mean and standard deviation for each score. Outliers are removed from the distributions based on Chauvenet's Criterion (a rule for eliminating data that appears to be in error with respect to the overall distribution) (Taylor 1982). When an outlying score is removed, the row containing it is removed from the multiple column listing. Hence, that particular match is completely eliminated

from the summary. This process is repeated until there are no outliers and typically requires that 1-2% of the matches be eliminated. For all of our post-processing modules, negative displacements represent good scores and positive displacements represent poor scores.

IMPLEMENTATION

All post-processing modules in the GRAFTER suite are written in C++. These additional programs are all built around a library of C++ classes developed to handle collections of residues and atoms based on the standard PDB format. The class designed to handle atoms makes use of a PDB library routine implemented locally (Pettersen, Couch et al. submitted 1994). The authors note that we observed substantial benefits from switching to C++ for development. Once the core classes were developed, all of the additional modules were easily programmed. Both programming time and additional lines of code were minimized because of the underlying class hierarchy.

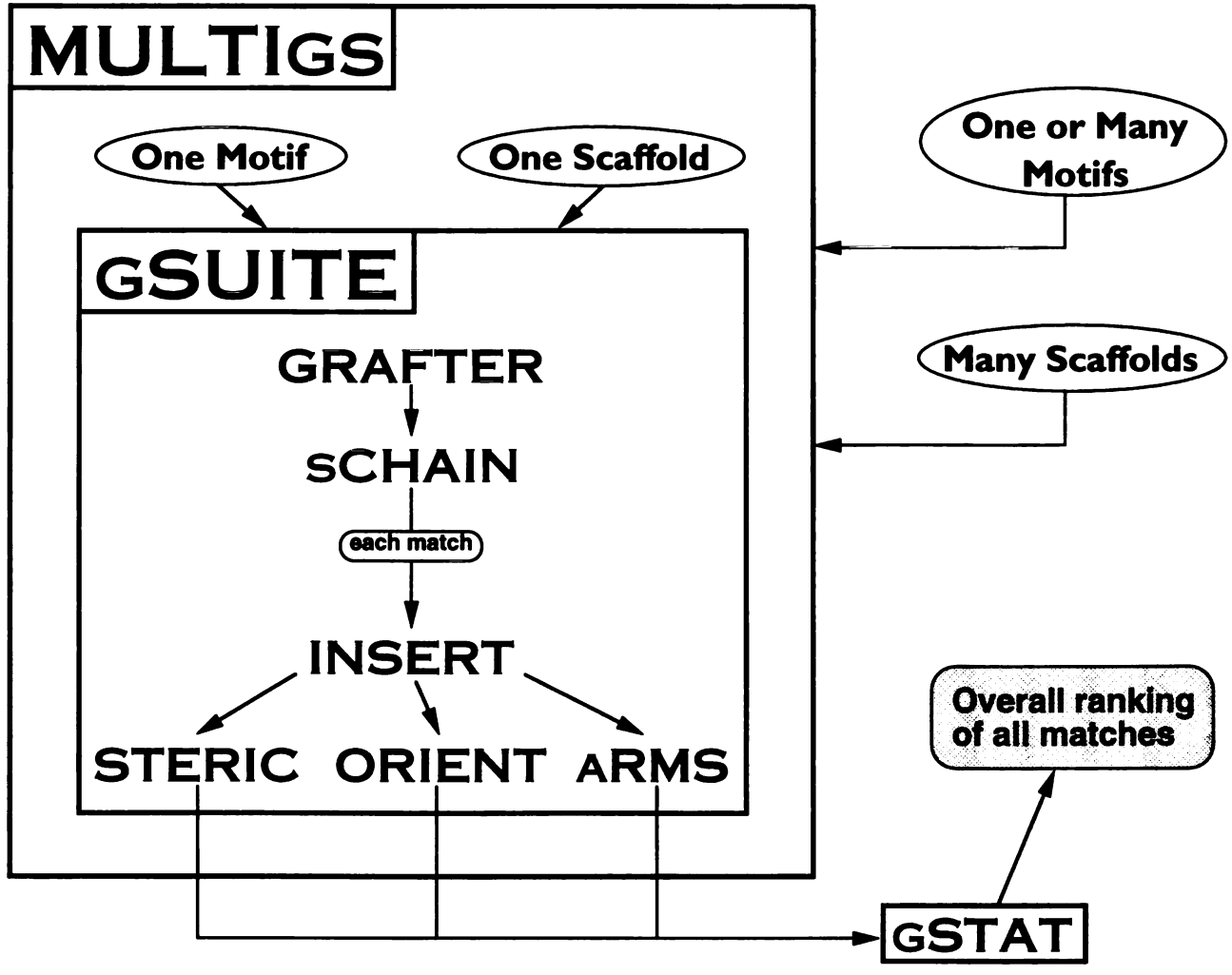
The suite of programs is managed by a Perl script (GSUITE), which performs all manipulations to complete a single comparison (one site vs. one scaffold). In turn, another Perl script (MULTIGS) handles multiple comparisons, using as input a simple text file describing the necessary input to GSUITE. The relationship between MULTIGS and GSUITE is depicted in Figure 5-2.

Data collection was performed on Silicon Graphics Iris, Indigo-2 and Challenge workstations, although GRAFTER has been successfully ported to and executed on a variety of machines including MIPS and Sun workstations. The code for the GRAFTER suite is available on request.

FIGURE 5-2: GRAFTER suite management scripts

Management scripts for the GRAFTER suite - GSUITE is responsible for managing the comparison of one motif and one scaffold, including the GRAFTER search and subsequent post-processor scoring. MULTIGS manages multiple GSUITE executions. Typically, all scores produced by a MULTIGS run are evaluated using GSTAT, which produces a ranked list of all matches.

FIGURE 5-2: GRAFTER suite management scripts



RESULTS & DISCUSSION

REPRESSORS

For the comparison of the λ and 434 repressors, we used relatively loose constraints (tolerance = 0.40, pair limit = 30, score shift = 0.005, minimum atoms = 12). More stringent parameter values result in only a single match between the repressors. Although this match is the trivial match which corresponds to an exact alignment of the specificity helices, we were anxious to probe a more substantial list of plausible matches. When GRAFTER generates more than one match, rank ordering is possible and is determined by the scores from our post-processing modules.

Two chains from the λ repressor structure were used, and a single chain from the 434 structure was used. Six comparisons were performed, i.e., all possible comparisons were explored except those involving the same structure as the motif and the scaffold. The comparisons were based on Cα-only representations of the motif and scaffold. Cα atoms were sufficient because the helices contain enough residues to describe a geometry clearly using only Cα atoms. The specificity helices themselves contain 10 residues. In addition, two extra residues were included at the N-terminus of each specificity helix. These residues appear to interact with the DNA in the bound structure, and help eliminate the ambiguity inherent with the symmetrical helix geometry.

Table 5-1 displays summary data for the GRAFTER searches between repressor structures. The post-processing scores from all 6 comparisons were collected and GSTAT analysis was performed on the group as a whole. The structures from the resulting ranking were analyzed visually. The highest ranked match for each search was the expected alignment of the specificity helices. The top 20 matches were clustered based on geometry and are summarized in Table 5-2. 12 of these (matches AA, AB, BA,

TABLE 5-1: summary of repressor comparisons

Summary information from GRAFTER repressor comparisons. Times are shown for a Silicon Graphics Indigo 2 and represent real times (not processor times).

Structures used: 1lmb (λ repressor) (Beamer and Pabo 1992), 1pra (434 repressor) (Neri, Billeter et al. 1992)

motif	scaffold file	#motif atoms	#scaf atoms	total matches	tolerance	time (sec.)	RMSD[1] mean	sdom[2]
1lmb3	1pra1	12	69	13	0.4	562	2.594	0.551
1lmb3	1lmb4	12	92	51	0.4	1020	2.48	0.184
1lmb4	1pra1	12	69	13	0.4	558	1.649	0.249
1lmb4	1lmb3	12	87	51	0.4	982	2.188	0.146
1pra1	1lmb3	12	87	55	0.4	993	2.322	0.166
1pra1	1lmb4	12	92	53	0.4	1024	2.209	0.157

[1] root mean square deviation

[2] standard deviation of the mean

TABLE 5-2: top 20 repressor matches

Summary of top 20 matches in the repressor comparisons. Low STERIC, aRMS and ORIENT scores are better. More negative composite scores are better. Matches are grouped by motif and scaffold.

Structures used: 1lmb (λ repressor) (Beamer and Pabo 1992), 1pra (434 repressor) (Neri, Billeter et al. 1992)

TABLE 5-2: summary of top repressor matches

Motif	Scaffold	Match	Class[1]	Sequence Numbers	Score STERIC	Score aRMS	Score ORIENT	Composite Score[2]
1lmb, chain 3				**42–53**				
	1pra, structure 1	AA	1	26–37	22.80	1.86	7.66	**-5.08**
	1pra, structure 1	AB	1	21,27–37	29.15	2.24	16.85	-3.72
	1pra, structure 1	AC	2	38–27	14.61	3.67	32.45	-1.99
	1lmb, chain 4	BA	1	42-53	45.98	1.51	8.71	**-4.13**
	1lmb, chain 4	BB	4	10–21	13.11	2.99	24.62	-3.29
	1lmb, chain 4	BC	3	73–74,78–87	11.53	2.78	29.14	-3.26
	1lmb, chain 4	BD	3	76–87	12.80	3.31	26.45	-2.86
	1lmb, chain 4	BE	1	37,43–53	51.57	2.07	15.86	-2.78
	1lmb, chain 4	BF	3	76,74,78–87	12.07	3.45	26.56	-2.75
	1lmb, chain 4	BG	2	54-43	24.27	3.34	20.61	-2.63
1lmb, chain 4				**42–53**				
	1pra, structure 1	CA	1	26–37	21.51	1.80	12.61	**-4.86**
	1pra, structure 1	CB	1	21,27–37	27.79	2.24	22.65	-3.39
	1lmb, chain 3	DA	1	42–53	42.01	1.47	5.35	**-4.61**
	1lmb, chain 3	DB	4	10–21	15.12	3.25	25.40	-2.87
	1lmb, chain 3	DC	1	37,43–53	54.61	2.06	15.94	-2.63
1pra, structure 1				**26–37**				
	1lmb, chain 3	EA	1	42–53	40.78	1.53	9.13	**-4.36**
	1lmb, chain 3	EB	1	37,43–53	51.90	2.02	15.52	-2.84
	1lmb, chain 3	EC	5	77,14–24	46.69	2.60	21.17	-2.15
	1lmb, chain 4	FA	1	42–53	39.73	2.12	14.55	**-3.45**
	1lmb, chain 4	FB	1	37,43–53	48.15	2.54	19.99	-2.21

[1] Matches are classified into groups based on the residues matched. Matches from different motif/scaffold comparisons fall into the same class if they involve similar alignments.

[2] A sum of the STERIC, aRMS and ORIENT scores after representing them in terms of standard deviations from the mean. The top score within each motif/scaffold comparison is shown in bold.

BE, CA, CB, DA, DC, EA, EB, FA and FB) correspond to an exact helix alignment, or a similar alignment with a single residue mispairing. Three other groups (made up of 1 match [EC], 2 matches [BB, DB], and 3 matches [BC, BD, BF], respectively) contain a total of 6 matches that align the motif elsewhere in the scaffold. Finally, there are two matches [AC, BG] that align the motif helix to the specificity helix in the scaffold, but reverse the chain direction and offset the helices by one residue.

Visual analysis of the grafts identified for the two repressors, λ and 434, reveals certain general characteristics (Figure 5-3): 1) Backbone alignment between the specificity helices of λ and 434 repressors is very good, and 2) as built by the GRAFTER suite, side-chains on the grafted helices deviate somewhat from those in the motif. Considering that the Cα trace for one helix was used as the motif in probing the other repressor, it is reasonable that the greatest similarity between proposed graft and motif is in the backbone trace. Side-chain discrepancies may be accounted for by the fact that all side-chain dihedrals are currently chosen from a database, and that only ±20° deviations from the database value are allowed. Often, the side-chains of residues in the target motifs do not abide by the values summarized in the database. Therefore, we cannot expect the matching side-chain in the scaffold to adopt a preferred dihedral value.

The one-to-one alignment of the specificity helices from λ repressor (residues 42-53) and 434 repressor (residues 26-37) consistently achieves the best score from our analysis. This helps to affirm the effectiveness of the GRAFTER suite's searching and scoring criteria. Ptashne's experiments have already demonstrated that the analogous graft between 434 and P22 repressors possesses specificity based on the grafted residues and not the underlying scaffold (Wharton and Ptashne 1985). Consequently, we would expect the GRAFTER suite to recognize the 434/P22 graft and the analogous λ/434 graft as potential grafts with high scores. We are unable to perform the 434/P22 repressor comparison because we do not have access to a structure for the P22

FIGURE 5-3: repressor grafts - views of structural alignments

a) alignment of the specificity helix from chain 3 of the λ-repressor (black) with the proposed graft onto 434-repressor (white). Side chains in the grafted structure are positioned by the GRAFTER suite.

b) alignment of the specificity helix from 434-repressor (black) with the proposed graft onto chain 3 of the λ-repressor (white). Side chains in the grafted structure are positioned by the GRAFTER suite.

repressor, but GRAFTER clearly picks out the corresponding graft in the λ/434 repressor comparison.

The two proposed grafts in the top-20 scores that involve reversed chain traces are intriguing. This result suggests a possible pseudo-twofold axis for the specificity helix in a repressor. Perhaps reversing the sequence of residues along the helix will produce a new repressor that has some binding affinity for the original operator. Otherwise, perhaps each repressor has one specificity based on the forward sequence of its helix, and another (possibly unused by nature) that is the result of its reverse helix sequence.

CDR COMPARISON

We wished to investigate GRAFTER's ability to select sensible antibody scaffolds for particular CDR loops. This question is relevant to groups interested in the humanization of mouse monoclonal antibodies (Hakimi, Ha et al. 1993; Roguska, Pedersen et al. 1994). CDR regions were selected from 26 antibody structures from the PDB for evaluation using the GRAFTER suite. The loops range from 11 residues to 17 residues. Each of the 26 loops was used to search every one of the 26 full structures. As for the repressor comparisons, Cα-only representations were used for the motifs and scaffolds. After a few preliminary comparisons of the loops from the 3hfm and 1dfb structures, GRAFTER parameters were selected that were slightly less constrained than the default values. The values for tolerance (0.20), pair limit (15) and score shift (0.001) were used consistently throughout the searches. The minimum atoms parameter was adjusted so that its value was always equal to 2 less than number of atoms in the motif loop.

A full summary of the CDR comparisons is shown in Figure 5-4. Three distinct families of CDR structures are evident in the matrix. In Figure 5-5 we show all 26 loops grouped based on the families from Figure 5-4. The Ω, ζ and "Compound" designations

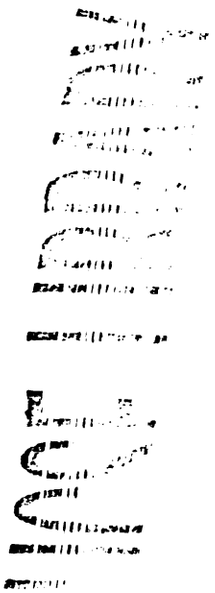


FIGURE 5-4: CDR comparison matrix

Grid positions are colored according to the best score obtained for that particular search. Darker squares correspond to better scores. White squares indicate comparisons where no acceptable match was identified. Matches were acceptable as long as the target site was contiguous in sequence, so that the graft would not be broken into multiple segments.

Structures used: 1bbd (FAB fragment from 8F5 antibody against HRV2) (Tormo, Stadler et al. 1992), 1dfb (3D6 FAB fragment) (He, Ruker et al. 1992), 1fdl (IgG1 FAB fragment - lysozyme complex) (Fischmann, Bentley et al. 1991), 1fvc (FV fragment of humanized antibody 4D5) (Eigenbrot, Randal et al. 1993), 1hil (IgG2a FAB fragment) (Rini, Schulze-Gahmen et al. 1992), 1igi (IgG1 FAB fragment) (Jeffrey, Strong et al. 1993), 1igm (IgM FV fragment) (Fan, Shan et al. 1992), 1mam (IgG2b FAB fragment) (Rose, Przybylska et al. 1993), 1mcw (Ig heterologous light chain) (Ely, Herron et al. 1990), 1ncd (N9 neuroamidase-NC41 FAB complex) (Tulip, Varghese et al. 1992), 1rei (Bence-Jones Ig REI variable portion) (Epp, Lattman et al. 1975), 2fb4 (FAB fragment) ((Marquart and Huber 1989), 2fbj (IgA FAB fragment (J539) Galactan-binding) (Bhat, Padlan et al. 1989), 2hfl (IgG1 FAB fragment HyHEL-5 and lysozyme complex) (Sheriff, Silverton et al. 1987), 2igf (IgG1 FAB' fragment (B1312) complex with peptide) (Stanfield, Fieser et al. 1990), 2mcp (Ig MC PC603 FAB-phosphocholine complex) (Padlan, Cohen et al. 1985), 2rhe (Bence-Jones protein (λ, variable domain)) (Furey , Wang et al. 1983), 3hfm (IgG1 FAB fragment (HyHEL-10) and lysozyme) (Padlan, Silverton et al. 1989), 3mcg (Ig λ light chain dimer (MCG)) (Ely, Herron et al. 1989), 4fab (4-4-20 (IgG2A κ) FAB fragment - fluoroscein (dianion) complex) (Herron , He et al. 1989), 6fab (FAB 36-71) (Strong, Campbell et al. 1991), 7fab (λ Ig FAB' - NEW) (Saul and Poljak 1992), 8fab (FAB fragment from human IgG1 (λ, HIL)) (Saul and Poljak 1993)

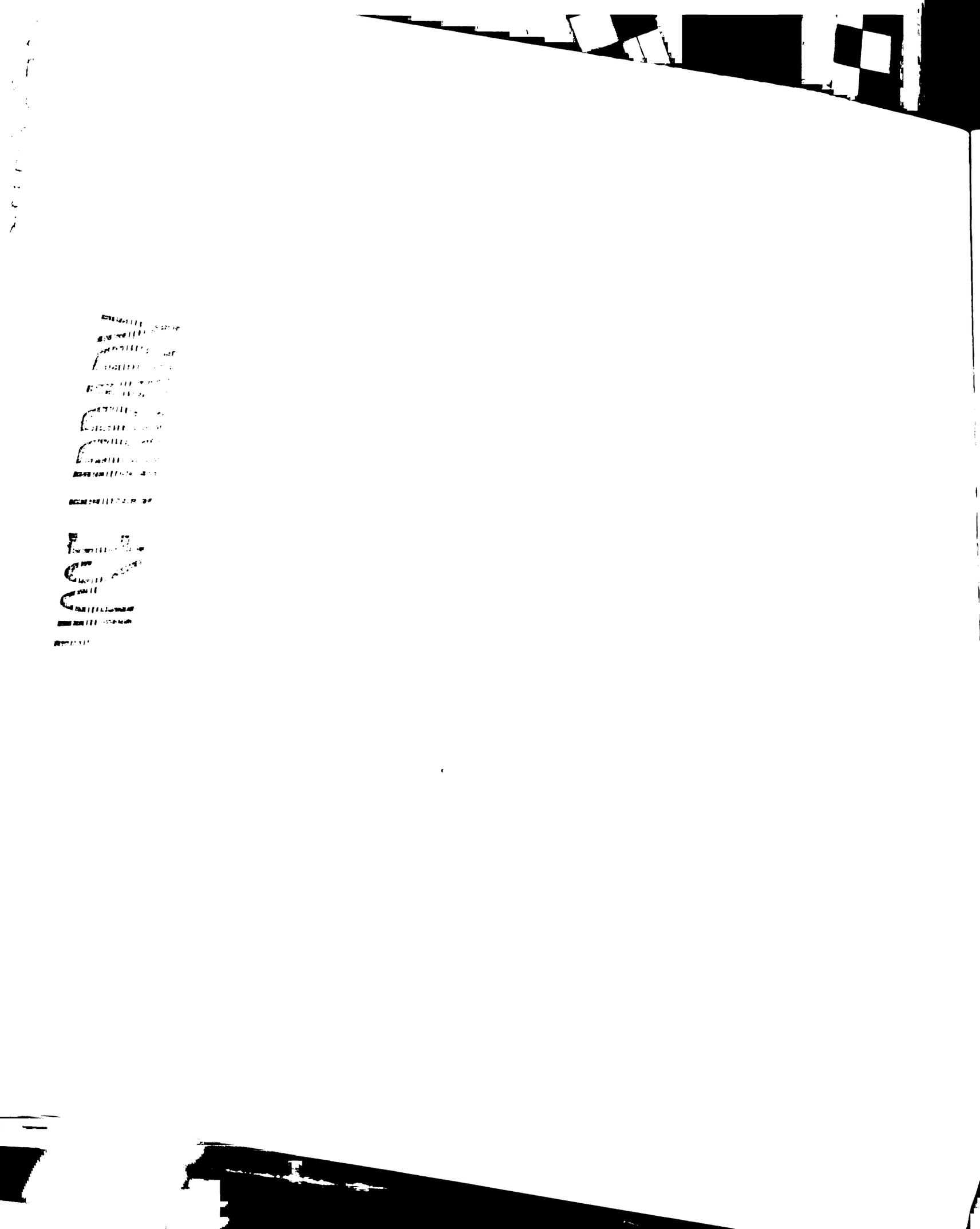


FIGURE 5-4: CDR comparison matrix

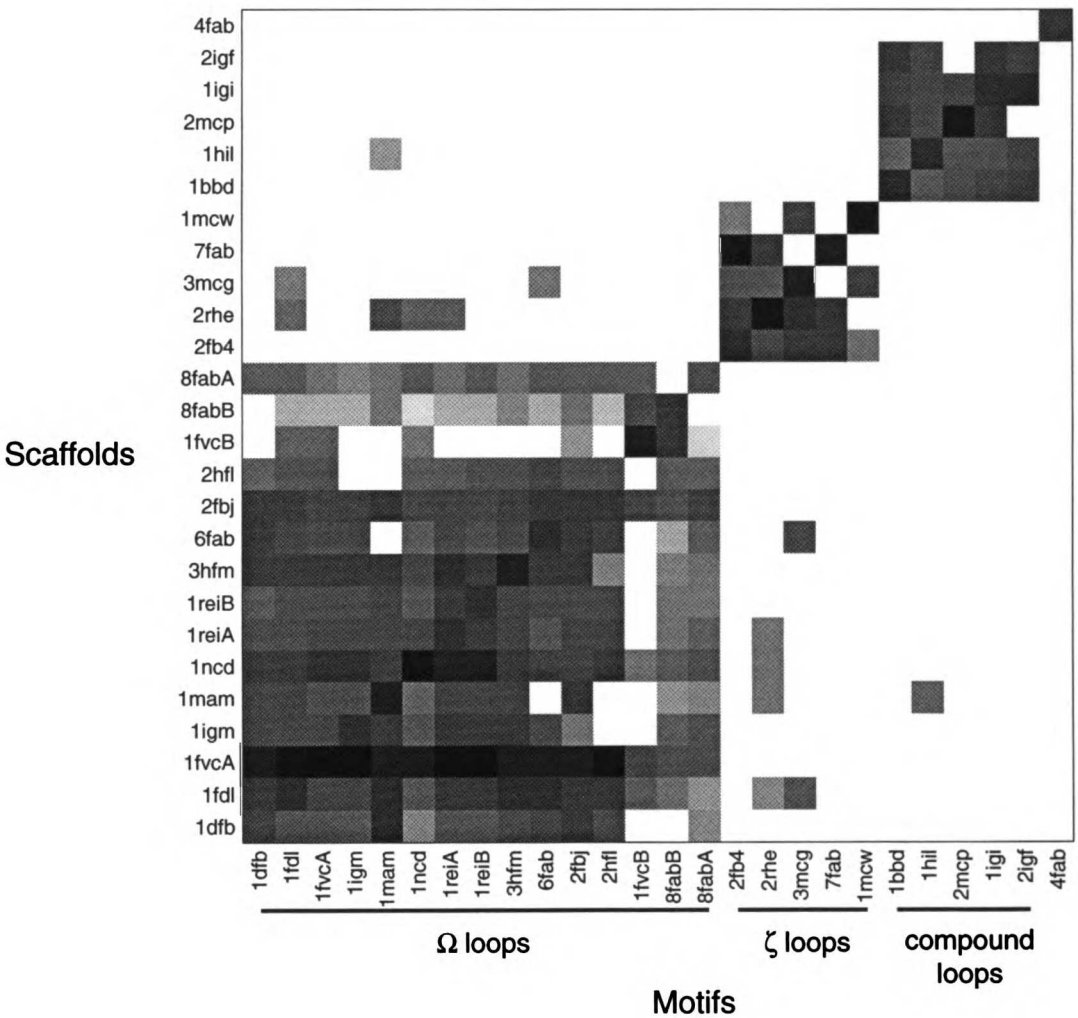
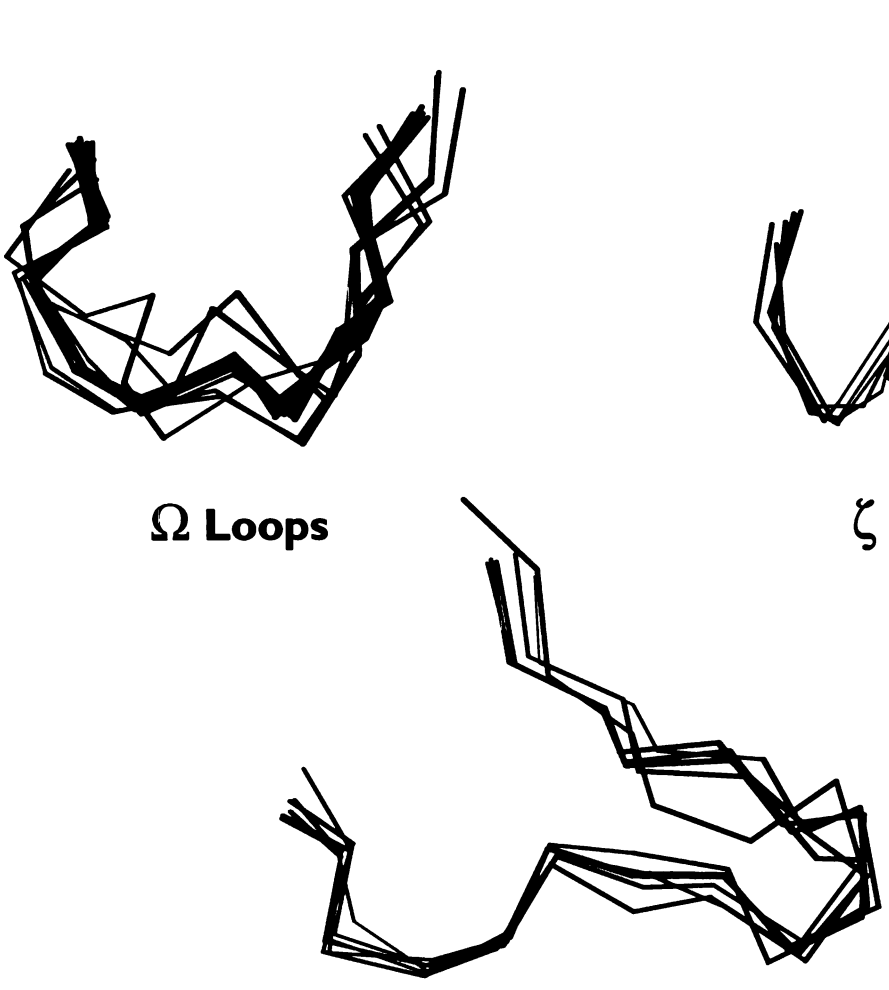
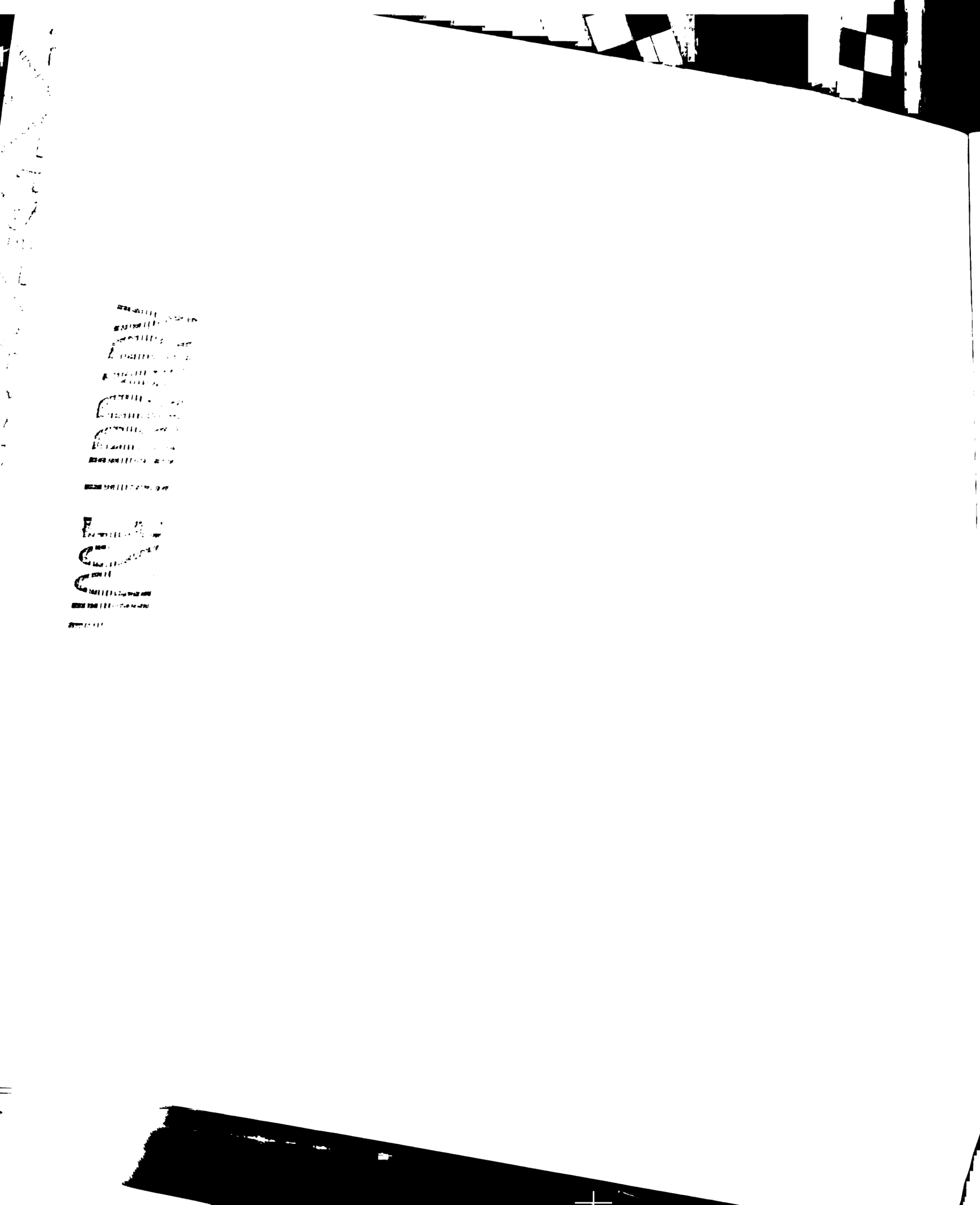


FIGURE 5-5: antibody loop classes

The 26 CDR regions used in our comparison are shown. The loops are classified according to the families identified in Figure 5.

FIGURE 5-5: Antibody Loop Classes





have been adopted based on previous studies (Leszczynski and Rose 1986; Ring, Kneller et al. 1992). Of all the loops, 4fab is singled out as distinct from all others. It bears the most resemblance to the compound loop family, and is shown in that family in Figure 5-5. Careful scrutiny of the data and the loop structures suggests that there are sub-families within the three main structure groups. We could define a sub-family as a group of loops that are all similar based on the data in Figure 5-4. For example, the loops from 1dfb, 1fdl, 1fvcA, 1igm, 1mam, 1ncd, 1reiA, 1reiB, 3hfm, 2fbj and 8fabA could be grouped into a sub-family of the omega loop family. The 1fvcB and 8fabB loops would make up a second omega loop sub-family.

The CDR loop comparisons were performed primarily to evaluate the effectiveness of GRAFTER and the post-processing modules. The results confirm the GRAFTER suite's ability to identify similar structural motifs within protein scaffolds. Those CDR loops that are most similar based on our classification are the best candidates for loop grafts. The GRAFTER suite provides a simple tool for identifying the best scaffolds for a particular CDR loop.

GROWTH HORMONE

Parameters for the comparison of the hGH epitope and IL-4 were selected after a series of computational controls were run. The values were chosen based on a trial comparison of the epitope to a mutant structure of hGH. We selected parameters that were just loose enough to find the correct five residues in the mutant (tolerance = 0.265, pair limit = 35, score shift = 0.005, minimum atoms = 13). Because $C\alpha$ atoms alone do not capture the overall geometry of this cluster of 5 residues, we opted for a $C\alpha$, $C\beta$ and N atom representation in this search. The minimum atoms value of 13 allows GRAFTER to overlook 2 of the 15 total atoms in the selected motif.

The summary information for the hGH/IL-4 GRAFTER search is shown in Table 5-3. The post-processing scores were analyzed by GSTAT, and the results appear in

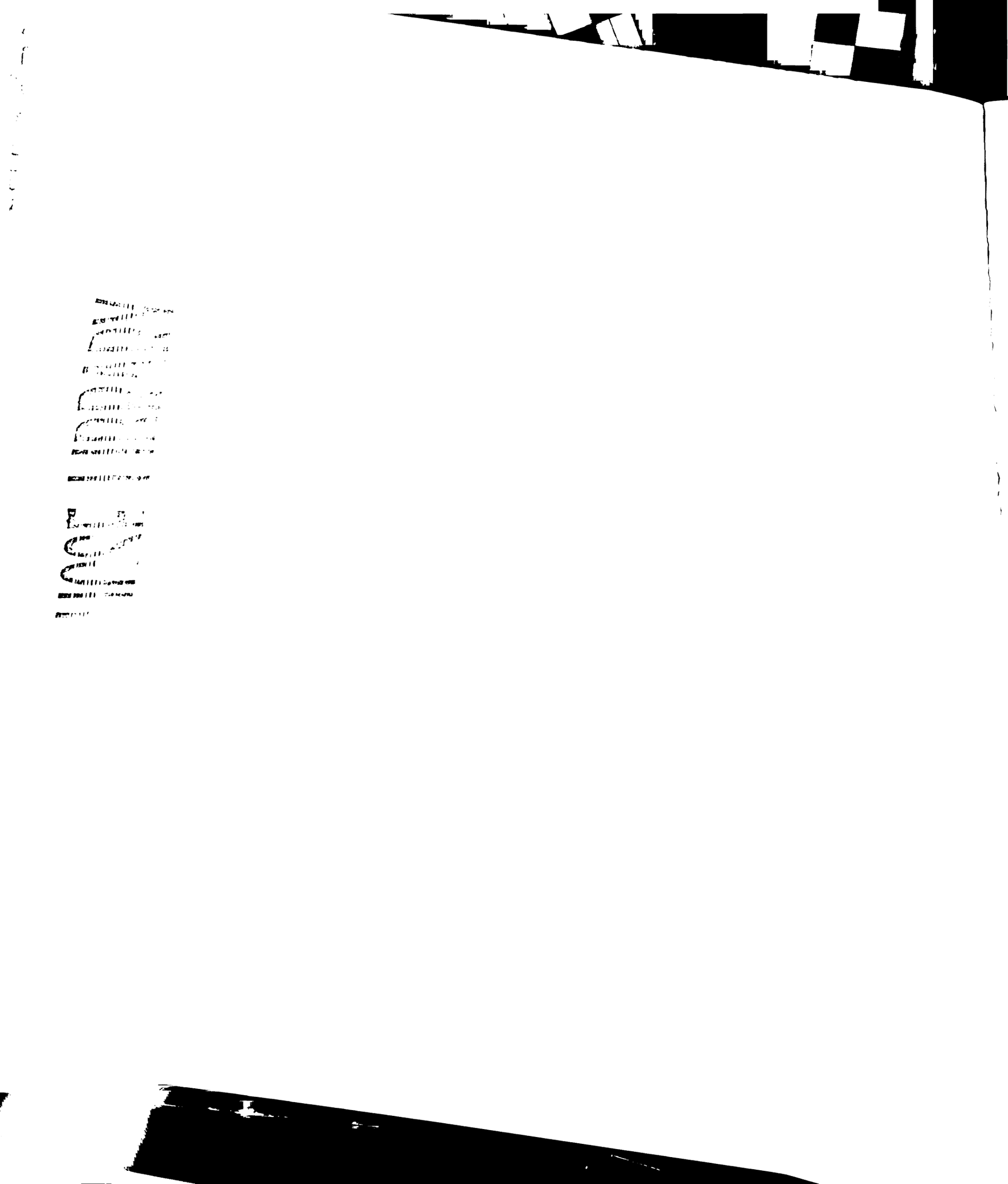


TABLE 5-3: summary from hGH epitope search

Summary information from the GRAFTER hGH epitope search. Times are shown for a Silicon Graphics Indigo 2 and represent real times (not processor times).

Structures used: (Ultsch, de Vos et al. 1991; de Vos, Ultsch et al. 1992), IL-4 (interleukin-4) (Smith, Redfield et al. 1992).

	Scaffold	# Motif	# Scaf.	# atoms per match			Total		time	RMSD[3]	
Motif	File	atoms	atoms	13	14	15	matches	Tol.[4]	(sec)	mean	SDOM[5]
hGH	IL-4	15	387	4581	486	23	5090	0.265	1.83×10^5	4.832	0.046

[3]root mean square deviation

[4]tolerance used in GRAFTER comparison

[5]standard deviation of the mean

Table 5-4. In the top 20 matches, 7 clusters were identified. Two of these clusters (classes 1 and 2, containing 4 and 6 matches, respectively and encompassing matches A-J) are closely related, and differ by a simple rotation of the scaffold. The remaining 9 matches are grouped into clusters containing 1 match (1 cluster, [T]), 2 matches (3 clusters, [K, L], [P, Q] and [R, S]) and 3 matches (1 cluster, [M, O]).

The hGH/IL-4 comparison demonstrates that structural matches for a 5 residue motif can be identified between these two structures. These matches are conceptually more difficult than the repressor and CDR comparisons, in that the residues are not contiguous in sequence. It is interesting to note that all of the high scoring matches graft the 5 residues onto two neighboring helices, just as they are mounted in their native structure. GRAFTER has no knowledge of the helical placement of the original motif during the search, so we can conclude that the helical character is clearly described by the geometries of the 5 residues. It has been shown by Lei Jin and Jim Wells that these 5 residues are the only residues that contribute significantly to the energetics of hGH epitope/antibody binding (Jin, Fendly et al. 1992). It is interesting that there are 15-20 residues that can be considered to be part of the hormone/antibody structural interface. However, homolog- and alanine-scanning mutagenesis experiments reveal that only these five contribute energetically to the interaction. This evidence encouraged us to limit our search motif to these 5 residues.

Because the scaffold (IL-4) is a 4-helix bundle protein, the GRAFTER suite identifies a number of high scoring matches that position the graft on different pairs of helices (Figure 5-6). Hence, there are a number of potential graft sites for the hGH epitope within the IL-4 structure. Under some circumstances this may allow for multiple grafts with the same functional/binding properties onto a single scaffold. It is interesting to note that this implicit pseudo twofold nature could help to explain why hGH and several 4-helix bundle cytokines can support a 1:2 ligand to receptor stoichiometry (de Vos, Ultsch et al. 1992).

TABLE 5-4: top 20 hGH/IL-4 matches

Summary of top 20 matches from the hGH/IL-4 comparison - Low STERIC, aRMS and ORIENT scores are better. More negative composite scores are better. Matches are grouped by family (i.e., similar matches).

Structures used: hGH (human growth hormone) (Ultsch, de Vos et al. 1991; de Vos, Ultsch et al. 1992), IL-4 (interleukin-4) (Smith, Redfield et al. 1992)

TABLE 5-4: summary of top matches from hGH/IL-4 comparison

	hGH epitope motif								
Match	R 8	N 12	R 16	D 112	D 116	STERIC	aRMS	OTN	Composite Score
			Family #1						
A	F 73	Q 78	F 82	K 12	S 16	1.26	2.81	13.46	**-5.60**
B	E 60	Q 78	F 82	K 12	S 16	1.08	3.89	15.28	-4.08
C	L 66	Q 78	F 82	K 12	S 16	16.76	3.01	22.63	-3.89
D	F 73	Q 78	F 82	K 12	N 15	-0.26	3.68	22.74	-3.89
			Family #2						
E	H 74	Q 78	F 82	K 12	S 16	1.91	2.04	9.15	**-6.87**
F	H 74	Q 78	F 82	K 12	N 15	0.39	2.67	18.41	-5.47
G	H 74	Q 78	F 82	I 11	L 14	-1.85	3.06	15.15	-5.31
H	H 74	Q 78	F 82	E 9	T 13	5.73	2.64	22.04	-4.97
I	H 74	Q 78	F 82	I 10	T 13	0.00	3.05	24.96	-4.53
J	H 74	Q 78	F 82	I 10	L 14	-1.26	3.36	21.59	-4.43
			Family #3						
K	I 5	E 9	T 13	R 81	R 85	19.51	1.65	18.32	**-5.82**
L	K 2	E 9	T 13	R 81	R 85	7.10	2.42	21.84	-5.20
			Family #4						
M	K 12	S 16	Q 20	Q 71	R 75	-1.87	3.12	17.63	**-5.06**
N	K 12	S 16	Q 20	Q 72	R 75	-2.00	3.54	17.96	-4.51
O	N 15	Q 20	T 22	Q 71	R 75	7.87	3.30	16.26	-4.43
			Family #5						
P	Q 8	K 12	S 16	Q 78	F 82	13.85	2.74	14.63	**-4.97**
Q	Q 8	K 12	S 16	I 80	F 82	13.17	3.47	12.28	-4.22
			Family #6						
R	I 80	R 85	N 89	I 5	E 9	21.14	2.94	19.79	**-3.96**
S	Q 78	R 85	N 89	I 5	E 9	3.23	3.61	20.86	-3.93
			Family #7						
T	R 75	L 79	L 83	V 51	F 55	8.64	1.74	43.70	**-4.43**

FIGURE 5-6: views of top-scoring hGH grafts

Top-scoring grafts for the hGH epitope onto IL-4 - the epitope is shown in black. Scaffold chains are shown in white.

a) hGH

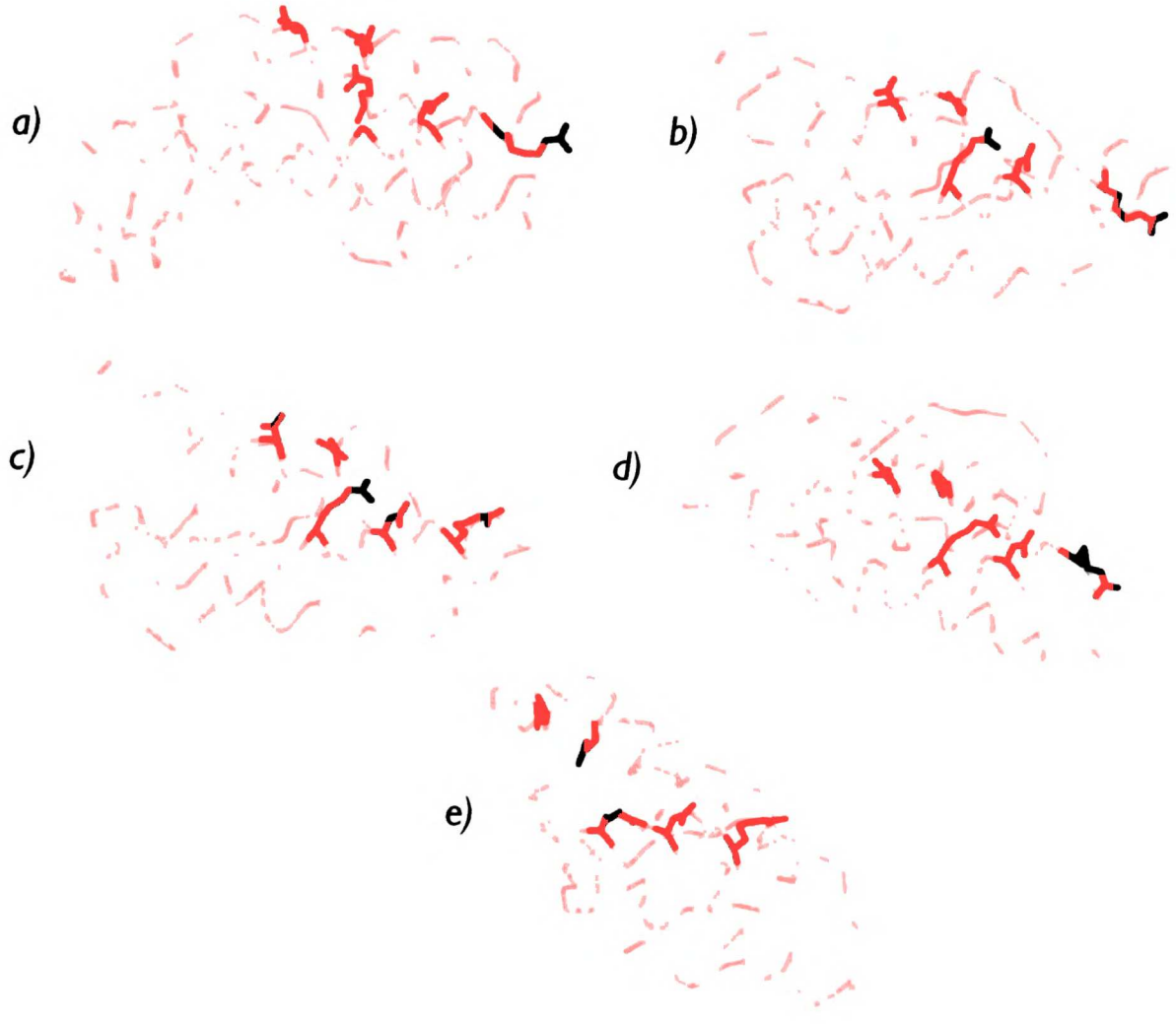
b) Match A

c) Match E

d) Match K

b) Match M

FIGURE 5-6: hGH epitope grafts on IL-4 scaffold



Using antibody binding experiments, a graft of an hGH epitope onto the IL-4 scaffold should be readily testable. Any number of the grafts identified in our searches could be constructed via site-directed mutagenesis. The resulting hybrid proteins could be analyzed for binding to the antibody corresponding to the relevant epitope. Structural analysis of the grafted proteins would supplement the binding data, and help us to understand any observed binding properties of the grafts.

CONCLUSION

As we strive to understand and manipulate protein structure, computational tools become more and more essential for their efficiency and thoroughness. There are many goals in the field of protein engineering, but all revolve around the desire to manipulate a protein's function by altering its structure. One area that has been targeted by protein engineers is that of grafting: the transfer of a functional collection of residues from one protein onto another.

Grafting protein motifs provides a means for producing new functionalities from current protein scaffolds. A scaffold possessing desired physical properties or molecular specificity could be altered to take on a new function, or to simply place a desired motif onto a scaffold that is more efficiently expressed, purified and characterized than its native scaffold. In our experience, the GRAFTER suite achieves a series of goals as an aid for protein engineering. It automates the search for protein scaffolds that contain graft sites for a desired motif. The suite's systematic nature eliminates the bias that had been inherent in our previous manual searches. The efficiency of the algorithm allows us to perform large scale searches without sacrificing thoroughness.

As grafting techniques develop, it is possible to imagine the creation of multifunctional proteins. In its simplest form, grafting could add catalytic or binding functionality onto a scaffold that retains its native function. These polyfunctional proteins

would be a first step towards macromolecular assembly lines where a sequence of enzymatic steps are performed on the same scaffold. Also, by combining two binding regions on the same scaffold, we may generate molecules capable of bridging two distinct species. This could reproduce the characteristics of superantigens, which are known to cross-link T-cell receptors and class II histocompatibility molecules. Superantigens produce a much more general immune response than typical antigens and generate exaggerated T-cell proliferation and increased cytokine release (Swaminathan, Furey et al. 1992). Such stimulation of immune response presents a variety of therapeutic applications.

REFERENCES

Beamer, L. J. and C. O. Pabo (1992). "Refined 1.8 A crystal structure of the lambda repressor-operator complex." J. Mol. Biol. **227**(1): 177-196.

Bhat, T. N., E. A. Padlan, et al. (1989). Refined crystal structure of the Galactan-binding immunoglobulin FAB J539 at 1.95 Å resolution., to be published.

de Vos, A. M., M. Ultsch, et al. (1992). "Human growth hormone and extracellular domain of its receptor: crystal structure of the complex." Science **255**(5042): 306-312.

Dunbrack, R. L. J. and M. Karplus (1993). "Backbone-dependent rotamer library for proteins. Application to side-chain prediction." J. Mol. Biol. **230**(2): 543-574.

Eigenbrot, C., M. Randal, et al. (1993). "X-ray structures of the antigen-binding domains from three variants of humanized anti-p185HER2 antibody 4D5 and comparison with molecular modeling." J. Mol. Biol. **229**(4): 969-995.

Ely, K. R., J. N. Herron, et al. (1990). "Three-dimensional structure of a hybrid light chain dimer: protein engineering of a binding cavity." Mol. Immunol. **27**(2): 101-114.

Ely, K. R., J. N. Herron, et al. (1989). "Three-dimensional structure of a light chain dimer crystallized in water. Conformational flexibility of a molecule in two crystal forms." J. Mol. Biol. **210**(3): 601-615.

Epp, O., E. E. Lattman, et al. (1975). "The molecular structure of a dimer composed of the variableportions of the Bence-Jones protein REI refined at 2.0-A resolution." Biochem. **14**(22): 4943-4952.

Fan, Z. C., L. Shan, et al. (1992). "Three-dimensional structure of an Fv from a human IgM immunoglobulin." J. Mol. Biol. **228**(1): 188-207.

Furey , W. J., B. C. Wang, et al. (1983). "Structure of a novel Bence-Jones protein (Rhe) fragment at 1.6 A resolution." J. Mol. Biol. **167**(3): 661-692.

Hakimi, J., V. C. Ha, et al. (1993). "Humanized Mik beta 1, a humanized antibody to the IL-2 receptor beta-chain that acts synergistically with humanized anti-TAC." J. Immunol. **151**(2): 1075-1085.

He, X. M., F. Ruker, et al. (1992). "Structure of a human monoclonal antibody Fab fragment against gp41 of human immunodeficiency virus type 1." PNAS **89**(15): 7154-7158.

Herron , J. N., X. M. He, et al. (1989). "Three-dimensional structure of a fluorescein-Fab complexcrystallized in 2-methyl-2,4-pentanediol." Proteins **5**(4): 271-280.

Jeffrey, P. D., R. K. Strong, et al. (1993). "26-10 Fab-digoxin complex: affinity and specificity due to surface complementarity." PNAS **90**(21): 10310-10314.

Jin, L., B. M. Fendly, et al. (1992). "High resolution functional analysis of antibody-antigen interactions." J Mol Biol **226**(3): 851-65.

Jin, L. and J. Wells (1992). personal communication.

Kabsch, W. (1976). "A solution for the best rotation to relate two sets of vectors." Acta Cryst. **A32**: 922-923.

Lee, B. and F. M. Richards (1971). "The Interpretation of Protein Structures: Estimation of Static Accessibility." J. Mol. Biol. **55**: 379-400.

Leszczynski, J. F. and G. D. Rose (1986). "Loops in globular proteins: a novel category of secondary structure." Science **234**(4778): 849-855.

Marquart, M. and R. Huber (1989). , personal communication.

Neri, D., M. Billeter, et al. (1992). "NMR determination of residual structure in a urea-denaturedprotein, the 434-repressor." Science **257**(5076): 1559-1563.

Padlan, E. A., G. H. Cohen, et al. (1985). "On the specificity of antibody/antigen interactions: phosphocholine binding to McPC603 and the correlation of three-dimensional structure and sequence data." Ann. Inst. Pasteur Immunol. **136C**(2): 271-276.

Padlan, E. A., E. W. Silverton, et al. (1989). "Structure of an antibody-antigen complex: crystal structure of the HyHEL-10 Fab-lysozyme complex." PNAS **86**(15): 5938-5942.

Pettersen, E. F., G. S. Couch, et al. (submitted 1994). "The Object Technology Framework: An Object-Oriented Interface to Molecular Data." Information Systems.

Ponder, J. W. and F. M. Richards (1987). "Tertiary Templates for Proteins: Use of Packing Criteria in the Enumeration of Allowed Sequences for Different Structural Classes." J. Mol. Biol. **193**: 775-791.

Presnell, S. R. and F. E. Cohen (1989). "Topological distribution of four-alpha-helix bundles." PNAS **86**(17): 6592-6596.

Ring, C. S., D. G. Kneller, et al. (1992). "Taxonomy and conformational analysis of loops in proteins. (published erratum appears in J. Mol. Biol. 227(3):977)." J. Mol. Biol. **224**(3): 685-699.

Rini, J. M., U. Schulze-Gahmen, et al. (1992). "Structural evidence for induced fit as a mechanism for antibody-antigen recognition." Science **255**(5047): 959-965.

Roguska, M. A., J. T. Pedersen, et al. (1994). "Humanization of murine monoclonal antibodies through variable domain resurfacing." PNAS **91**(3): 969-973.

Rose, D. R., M. Przybylska, et al. (1993). "Crystal structure to 2.45 A resolution of a monoclonal Fab specific for the Brucella A cell wall polysaccharide antigen." Protein Sci. **2**(7): 1106-1113.

Saul, F. A. and R. J. Poljak (1992). "Crystal structure of human immunoglobulin fragment Fab New refined at 2.0 A resolution." Proteins **14**(3): 363-371.

Saul, F. A. and R. J. Poljak (1993). Crystal structure of the FAB fragment from the human myeloma immunoglobulin IgG HIL at 1.8 Å resolution., to be published.

Sheriff, S., E. W. Silverton, et al. (1987). "Three-dimensional structure of an antibody-antigen complex." PNAS **1987**(22): 8075-8079.

Smith, L. J., C. Redfield, et al. (1992). "Human interleukin 4. The solution structure of a four-helix bundle protein." J. Mol. Biol. **224**(4): 899-904.

Stanfield, R. L., T. M. Fieser, et al. (1990). "Crystal structures of an antibody to a peptide and its complex with peptide antigen at 2.8 A." Science **248**(4956): 712-719.

Strong, R. K., R. Campbell, et al. (1991). "Three-dimensional structure of murine anti-p-azophenylarsonate Fab 36-71. 1. X-ray crystallography, site-directed mutagenesis, and modeling of the complex with hapten." Biochem. **30**(15): 3739-3748.

Swaminathan, S., W. Furey, et al. (1992). "Crystal structure of staphylococcal enterotoxin B, a superantigen." Nature **359**(6398): 801-806.

Taylor, J. R. (1982). An introduction to error analysis: the study of uncertainties in physical measurements. Mill Valley, University Science Books.

Tormo, J., E. Stadler, et al. (1992). "Three-dimensional structure of the Fab fragment of a neutralizing antibody to human rhinovirus serotype 2." Protein Sci. 1(9): 1154-1161.

Tulip, W. R., J. N. Varghese, et al. (1992). "Crystal structures of two mutant neuraminidase-antibody complexes with amino acid substitutions in the interface." J. Mol. Biol. **227**(1): 149-159.

Ultsch, M., A. M. de Vos, et al. (1991). "Crystals of the complex between human growth hormone and the extracellular domain of its receptor." J. Mol. Biol. **222**(4): 865-868.

Wharton, R. P. and M. Ptashne (1985). "Changing the binding specificity of a repressor by redesigning an α-helix." Nature **316**: 601-605.

Winter, G. and W. J. Harris (1993). "Humanized Antibodies." Immunology Today **14**(6): 243-246.

Chapter 6:
Modeling of DHFR from Cryptosporidium Parvum

INTRODUCTION

Present day approaches to the study of protein function are heavily dependent on structural data. Techniques for structural determination such as X-ray crystallography and NMR have evolved dramatically over the last decade (Gronenborn and Clore 1990). Unfortunately even with present day structural evaluation the rate of sequence determination exceeds the rate of structure determination by at least an order of magnitude. Fortunately, not every new sequence is completely unique. Many sequences are related to the sequences of proteins whose structures have already been determined. The interrelation of sequences and the need for protein structures has driven the development of homology modeling.

To model the structure of an unknown protein, homology modeling takes advantage of any sequence similarity there is to proteins of known structure. Sequence alignments, either manual or automatic, and secondary structure prediction help the researcher to align the sequences of a family of proteins, some with known structure, some without. Typically, the evolutionary process leads to greater differences in loop regions than in α-helices and β-sheets. Any regions with clear sequence similarity can be modeled based on the structures in those regions from the protein with known structure. Typically, after this phase of modeling, the model structure consists of a number of helices and/or strands, as well as a few short turns. Longer loop regions are usually absent.

Prediction of loop regions falls into two major categories: 1) analogy to known structures and 2) *de novo* design. Loop building by analogy is performed by dictionary lookup, where a collection of loops from known structures is searched for sequence and/or length similarity. In *de novo* design, loops are built based on residue preferences, side chain dihedral propensities and steric constraints. In either approach, most often a list of potential loops is generated for each loop region. A researcher will typically

evaluate these loops visually using a molecular graphics program. Eventually, a loop will be selected for each unknown loop.

Once all loops and secondary structure regions have been predicted, a preliminary structure can be assembled from these units. The resulting structure contains only backbone (i.e. N, CA, C, O) atoms. Side chains must be built as determined by the sequence of the protein being modeled. When the residue type is the same or similar to that residue in one of the known structures, the known side chain orientation is often used. When no similar residue is available, side chain orientation can be predicted based on dihedral propensity databases (Dunbrack and Karplus 1993).

After all side chains have been placed, it is useful to apply a number structure evaluation tools to the model. This will often reveal areas of unlikely or conflicting structure. Tools are readily available for evaluation of packing (Gregoret and Cohen 1990), steric conflicts (BIOSYM Technologies, San Diego, CA), solvent accessibility (Lee and Richards 1971; Presnell 1991), and side chain orientation (McGregor, Islam et al. 1987; Ponder and Richards 1987; Dunbrack and Karplus 1993). Based on this feedback, the researcher will often need to redesign loops, reorient side chains and adjust bond lengths and angles. By cycling between redesign and evaluation, the model can be refined to a point that satisfies the requirements of the modeling project. Often, only certain portions of the structure need to be highly refined (e.g. active site, binding site).

Once a model structure is complete with all main chain and side chain atoms, energy minimization and/or dynamics may be used to clean up bond lengths and angles. Usually constraints are used to prevent positional shifts in areas of well defined structure. Hot spots for minimization are typically loops and 2° structure junctions that may have been poorly built during modeling.

It is valuable to perform a final evaluation using steric, accessibility, packing and rotamer evaluation tools to verify that no major shifts have been introduced by minimization. Final assessment can also involve comparison to the known structures in

regions of functional importance. Generally, residues involved in binding and/or catalysis are structurally conserved amongst a family of related proteins.

Dihydrofolate reductase (DHFR) from *Cryptosporidium Parvum* was selected for my enzyme modeling efforts. Often in drug design projects, a target is selected that has already proven to be invaluable for clinical applications. Both clinically significant drugs, methotrexate and trimethoprim (Oefner, D'Arcy et al. 1988) target DHFR. However, neither of these agents is effective against DHFR from *C. Parvum*, a fact that has fueled our attempts to identify a clinical agent for *C. Parvum* DHFR. With a structural model of the enzyme we could move forward toward the identification of potential drugs that target *C. Parvum* DHFR.

Prior to 1980, *C. Parvum* was of little biomedical significance ((Fayer and Ungar 1986). This protozoan began to apppear more often in clinical logs beginning around 1982 when it was identified as a cause of diarrheal illness in humans and some domesticated animals. *C. Parvum* typically causes diarrheal illness in immunocompetent individuals that lasts no longer than a month. Immunocompromized patients are at a much greater risk with respect to *C. Parvum*. For these patients, typically the diarrhea is long-lasting and often fatal. Due to the current lack of available therapies for cryptosporidiosis, DHFR from *C. Parvum* was selected as a drug target in this study.

In *C. Parvum*, DHFR is structurally contiguous with thymidylate synthetase (TS) (Nelson 1994) forming a DHFR•TS combined enzyme. Sequence data alone has not revealed the exact dividing line between DHFR and TS in the structure, but it was inferred from sequence alignment to known DHFR structures. The sequence of DHFR•TS from *C. Parvum* was combined with the DHFR structures from human, chicken and lactobacillus caseii to build a homology model. This model was refined using the structural analysis tool previously mentioned and energy minimization.

The model structure bears a great deal of resemblance to the three crystal structures used to build it. It most closely resembles the structure of human DHFR, as a

result of the modeling approach. The active site appears to accommodate the inhibitors that are contained in the three structures (methotrexate, folate and biopterin).

METHODS

The sequence of DHFR•TS from *Cryptosporidium Parvum* was obtained from the laboratory of Richard Nelson (Nelson 1994). Structures for DHFR from human (1drf) (Oefner, D'Arcy et al. 1988), chicken (1dr1) (McTigue, Davies et al. 1992) and *Lactobacillus Caseii* (3dfr) (Bolin, Filman et al. 1982) were obtained from the Protein Data Bank.

The three DHFR structures (1drf, 1dr1 and 3dfr) were aligned using the graphical display program MIDAS (Ferrin, Huang et al. 1988). Only structurally conserved regions (SCRs) were included in the RMS fit. This approach left disparate loop regions out of the structural alignment. The resulting structurally-based sequence alignment was extracted for use in the alignment of the *C. Parvum* DHFR sequence.

Preliminary attempts with automated alignment tools (PIMA, fasta, blast) failed to produce suitable sequence alignments of DHFR from a number of species including *Cryptosporidium Parvum*. Long, unreasonable gaps were introduced by the automated tools. In addition, the tools failed to correctly align human, chicken and *L. Caseii* DHFR as compared to the known structural alignment. As a result it was necessary to prepare a manual alignment. This alignment takes into account regions of conserved sequence, as well as functionally significant residues that have been shown to be conserved (Figure 6-1). Note that the C-terminal (TS) section of the DHFR•TS sequence has been truncated based on the alignment.

Selection of manual alignment techniques was made only after automatic methods had been exhausted. Automatic alignment techniques have been invaluable in aligning large families of closely related sequences. Also, these automatic approaches have monumentally aided in sequence searches. However, regardless of their

effectiveness in such cases, the automatic tools can perform poorly for a set of distantly related sequences. This poor performance was observed for the alignment of DHFRs required by this modeling effort. I found that the automatic techniques were overlooking certain obvious sequence similarities and conserved residues. As a result I chose to perform the alignment manually.

The human structure (1drf) was selected as the best basis structure for modeling the *C. Parvum* DHFR. This decision was made based on alignment scores calculated using a variety of matrices. In all cases the scores for the chicken and human DHFRs were higher than those for *L. Caseii.* In 3 of 4 cases the human alignment scored higher than the chicken. The alignment scores are summarized below:

TABLE 6-1: alignment scores

Scores for a number of applied to comparison of the *C. Parvum* DHFR manual alignment to the three DHFR structures (human, chicken and *L. Caseii*). In all cases high scores are better. Scores in bold are the highest for a given matrix.

Matrix	human	chicken	*L. Caseii*
codon	176	**178**	125
% identity	**31%**	30%	23%
PAM 120	**130**	125	51
PAM 250	**177**	175	95

The residue types from the *C. Parvum* sequence were substituted onto corresponding positions in the human structure (1drf). Side chain orientations were preserved as shown in Table 6-2.

FIGURE 6-1: DHFR alignment

Structural alignment of the three crystal structures for human, chicken and *L. Caseii* DHFR along with the aligned sequence for *C. Parvum*. Structurally significant and conserved positions are indicated with alphabetic codes as listed in the following table. Secondary structure types are shown. Residue positions where the *C. Parvum* sequence is identical to at least one of the structures are indicated with black bars.

Code	Feature[1]	Same?[2]	Model	1drf	1dr1	3dfr	Notes
a	cis	√	G114 -	G116 -	G116 -	G98 -	cis Gly-Gly bond
	Gly-Gly		G115	G117	G117	G99	
b	Asp	√	D32	E30	E30	D26	MTX pteridine binding
	or Glu						
c	Thr	√	T134	T136	T136	T116	MTX 2-amino group binding
d	Ala	√	A11	A9	A9	A6	non-polar interactions bet. enzyme & pteridine
e	Phe	√	F36	F34	F34	F30	non-polar interactions bet. enzyme & pteridine
f	Trp	√	W27	W24	W24	W21	H-bonded to Asp26 & water
g	Leu	√	129	L27	L27	L23	H-bonded to Asp26 & water
	or Ile						
h	hϕ	√	L33	F31	Y31	L27	extend from Helix B, interact w. MTX
j	Arg	√	R70	R70	R70	R57	H-bonded to MTX glutamyl-COOH
k	Thr	√	T40	T38	T38	T34	conserved, helps orient Arg57
l	Asn		172	N72	N72	N59	conserved in 6 of 7 species
m	cis	√	R65 -	R65 -	R65 -	R52 -	cis Arg-Pro
	Arg-Pro		P66	P66	P66	P53	
o	Leu	√	L67	L67	L67	L54	conserved in known structures
p	Pro	√	G36	P61	P61	P50	conserved in known structures
q	Ile	√	162	160	160	149	conserved in known structures

[1] consensus based on crystal structures

[2] does residue in the model match the consensus found in the known structures?

[redacted]

FIGURE 6-1: DHFR alignment

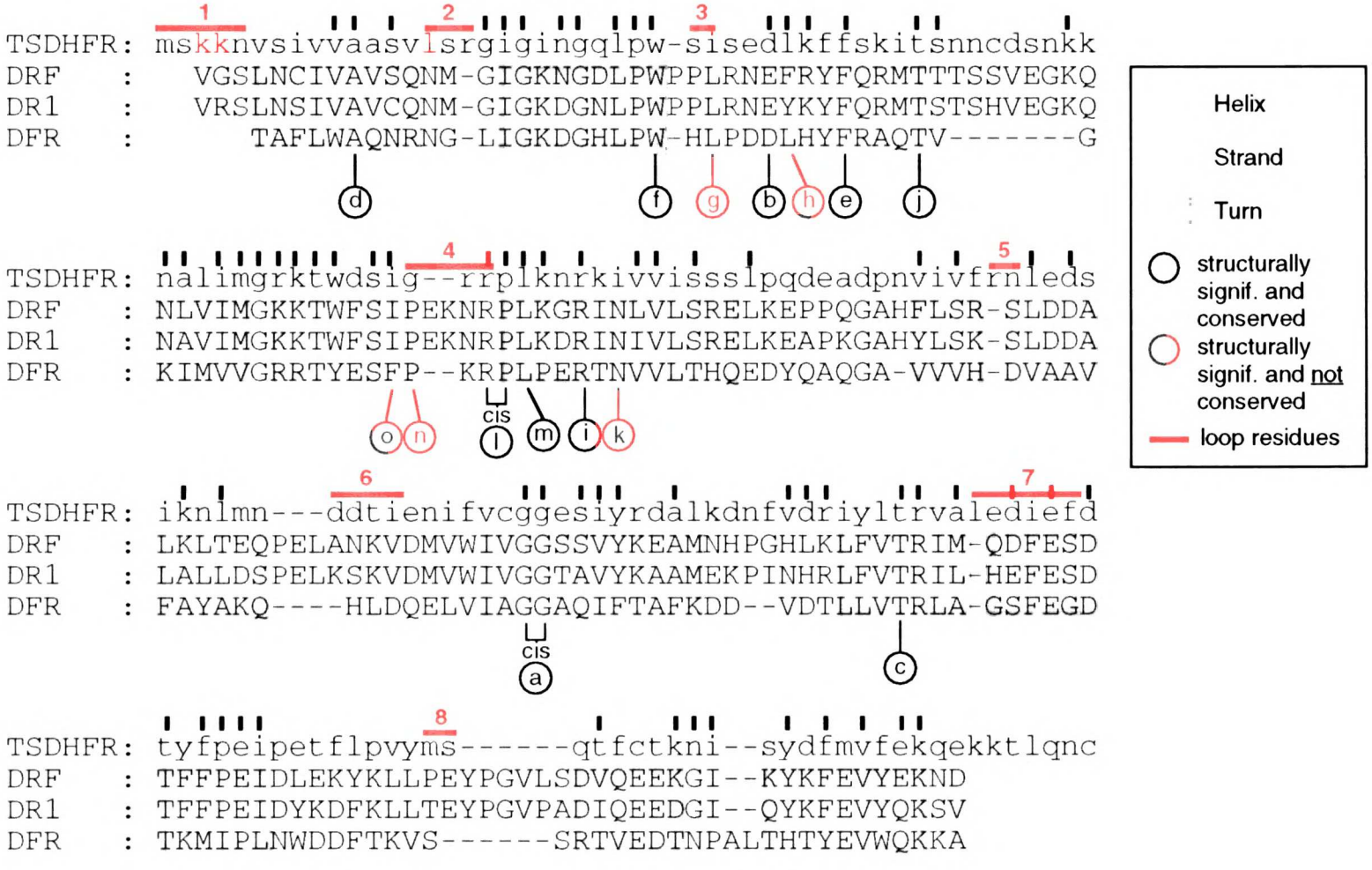


TABLE 6-2: side chain substitution chart

Residue substitutions where side chain orientation was retained

Original Residue	New Residue Type
Asp	Asp, Asn
Glu	Glu, Gln, Asp, Asn
Phe	Phe, Tyr
Ile	Ile, Leu, Met, Val
Lys	Lys, Arg
Leu	Leu, Ile, Met, Val
Met	Met, Ile, Leu, Val
Asn	Asn, Asp
Gln	Gln, Glu, Asp, Asn
Arg	Arg, Lys
Ser	Ser, Thr
Thr	Thr, Ser
Val	Val, Ile, Leu
Tyr	Tyr, Phe

Otherwise, the new side chain was placed using the most common (i.e. primary) rotamer (Ponder and Richards 1987). Insertions and gaps, which occurred only in loop regions, were ignored until loop placement. The positions of loops and 2° structure regions are shown in Figure 6-1.

Loops were generated using the BLoop algorithm (Ring and Cohen 1994). The algorithm accepts the new loop sequence along with the anchor residue(s) as input. I used the default mode, which generates 50 loops for each run. These loops were visualized in the presence of the model structure (with loops removed). For each loop region, the best loop structure was selected based on avoidance of steric clashes and

structural comparison with the known loops from 1drf, 1dr1 and 3dfr. In one case (loop #6, the loops generated by BLoop were not suitable, but the corresponding loop in 3dfr contained the same number of residues. As a result, this loop (#6) was modeled based on the corresponding loop in 3dfr. Side chains were placed on the loops using the most common rotamers.

The resulting model structure was evaluated for bad contacts, and reasonable side chain dihedrals. The number of bad contacts was measured using InsightII (BIOSYM Technologies, San Diego, CA) with a bump criterion of 0.6Å. Pairs of residues that clash were adjusted by selecting less common rotamers until the clash was eliminated.

Once all bumps had been dealt with, side chain dihedrals were evaluated in two ways: 1) side chain propensities based on 2° structure class (McGregor, Islam et al. 1987) 2) side chain propensities based on main-chain dihedrals (Dunbrack and Karplus 1993). Those positions that were flagged as unlikely were compared to known rotamers to determine why they were flagged. In many cases, particularly when evaluated with CHICHECK (Dunbrack and Karplus 1993). the side chains were in known rotamers. This typically indicates one of two things: 1) the rotamer is statistically common, but not for that particular phi/psi pair or 2) that particular phi/psi bin in the CHICHECK dihedral propensity library is empty, and hence there is no data available on side chain propensity. These two cases can be distinguished using output from CHICHECK. Side chains that fell into the first category were replaced with the most likely orientation based on phi/psi. Side chains in the second category were replaced with orientations derived from nearby phi/psi bins in Dunbrack's rotamer database.

Bumps were re-evaluated, and any problems were eliminated by replacing side chains with other known rotamers, biased if possible by the phi/psi pair for that residue. Side chain orientations were re-evaluated, and then it was observed that all residues passed the side chain evaluation. CHICHECK still flags residues that fall into empty phi/psi bins, but its clear that those residues have already been handled manually.

The program, QPACK (Gregoret and Cohen 1990), which measures the packing of residues within a protein structure was use to evaluate the model at this stage. Residues with packing values more than 20% away from the corresponding value for human DHFR were flagged for further inspection. Twelve of these 49 residues were then replaced using Dunbrack's rotamer database. One of the three possible rotamers from the database was selected based on visual inspection, with the goal of improving the packing without introducing new steric conflicts. The structure was re-evaluated using QPACK and I observed a net improvement of 7 residues. This process was repeated 2 more times resulting in net improvements of 7 and 8 residues respectively. The residues changed and those affected are listed in Table 6-3. The 27 residues that remain flagged after the final cycle are either on the surface, or in loops that were modeled by BLoop.

The structure at this stage was minimized using the AMBER force-field (Weiner and Kollman 1981) to eliminated any unusual bond lengths or angles that may have been introduced during the structural cutting and pasting. The minimization was highly constrained for all non-loop backbone atoms (N, CA, C) and was run for a total of 200 cycles without dynamics. The minimized structure was re-evaluated using the chi-checking tools and QPACK to ensure that minimization was not detrimental to the model.

RESULTS & DISCUSSION

The homology model built for DHFR from *Cryptosporidium* most closely resembles human DHFR. The parallel backbone traces for the crystal structures of human, chicken and *L. Caseii* DHFR along with the model are shown in Figure 6-2. It is clear that all four structures are similar except in certain loop regions. The key active

TABLE 6-3: residue changes and effect on packing

The residues changed in each of three passes are listed. Residues that improved and worsened in each pass are shown. In addition, for each residue changed, the type of change is listed. Most changes involved swapping in one of three preferred rotamers. A few changes involved manual adjustment of chi3, chi4 or both.

TABLE 6-3: rotamer replacement and effect on packing

Pass #1			
Changed	Improved†	Worse†	Rotamer
	Gly23		
Gln24			1
	Phe35		
Ser37	Ser37		3
		Ser41	
Asn42			2
Asn43			3
Ser46			1
	Asn50		
Ser78			2
	Gln81		
Asp82	Asp82		1
	Pro86		
Asn100	Asn100		1
Asp105			2
		Glu108	
	Pro151		
Thr153			3
Phe154			1
Ser169			1

Pass #2			
Changed	Improved†	Worse†	Rotamer
Asn43	Asn43		1
	Lys48		
Asn50			*
Ser78			3
Met102			*
	Glu108		
Glu116	Glu116		3
Lys124	Lys124		2
Asn126	Asn126		1
		Gly129	
Arg135			1
	Ala137		
		Glu139	
Asp140			3
Ile141			3
Glu142	Glu142		3
Glu152			3
Met159			1
Ser160			3
Thr162	Thr162		3
Asn167			3
Phe172			1

Pass #3			
Changed	Improved†	Worse†	Rotamer
Arg64	Arg64		3
Arg65	Arg65		*
	Pro66		
Lys68	Lys68		3
	Asn69		
	Ala84		
Asp85	Asp85		*
	Met173		

† a range of acceptable packing was set at 1.0±0.2 neighbors. Any time a residue began within this range and an adjustment forced it out of the range it is reported as "Worse". Similarly, any time a residue was originally outside the range and moved into the range after adjustment it was listed as "Improved".

* Adjustment was not a simple rotamer swap. These entries represent adjustments in either chi3, chi4 or both. The rotamer library specifies preferences only for chi1 and chi2.

site residues previously described in Figure 6-1 are in similar orientations in the human and model DHFR structures (Figure 6-3). The final model described here contained no steric clashes at 0.6 Å. Final dihedral and packing data are shown in Tables 6-4 and 6-5 Although there are still regions where packing and side chains are questionable, these are concentrated in surface loops. Because the primary goal of this modeling effort is active site/ligand docking, these exterior surface discrepancies were ignored. If this model is to be used for more general structural studies, additional work will be necessary to adjust these regions.

The active site of the model bears strong resemblance to human DHFR, except at in the loop regions. It is easy to dock folate, biopterin or methotrexate into the model's active site based on their orientations in known structures (1drf, 3dfr, 1dr1) (Bolin, Filman et al. 1982; Oefner, D'Arcy et al. 1988; McTigue, Davies et al. 1992). The charge distribution for the active site of the model was calculated and is shown in Figure 6-4.

CONCLUSION

In this study, homology modeling has provided a valuable tool for structure prediction. The model obtained for DHFR from *Cryptosporidium Parvum* provides us with an excellent starting point for ligand docking. The application of the DOCK algorithm (DesJarlais, Sheridan et al. 1988) to the modeled active site is described in the next chapter.

FIGURE 6-2: model compared to 3 crystal structures

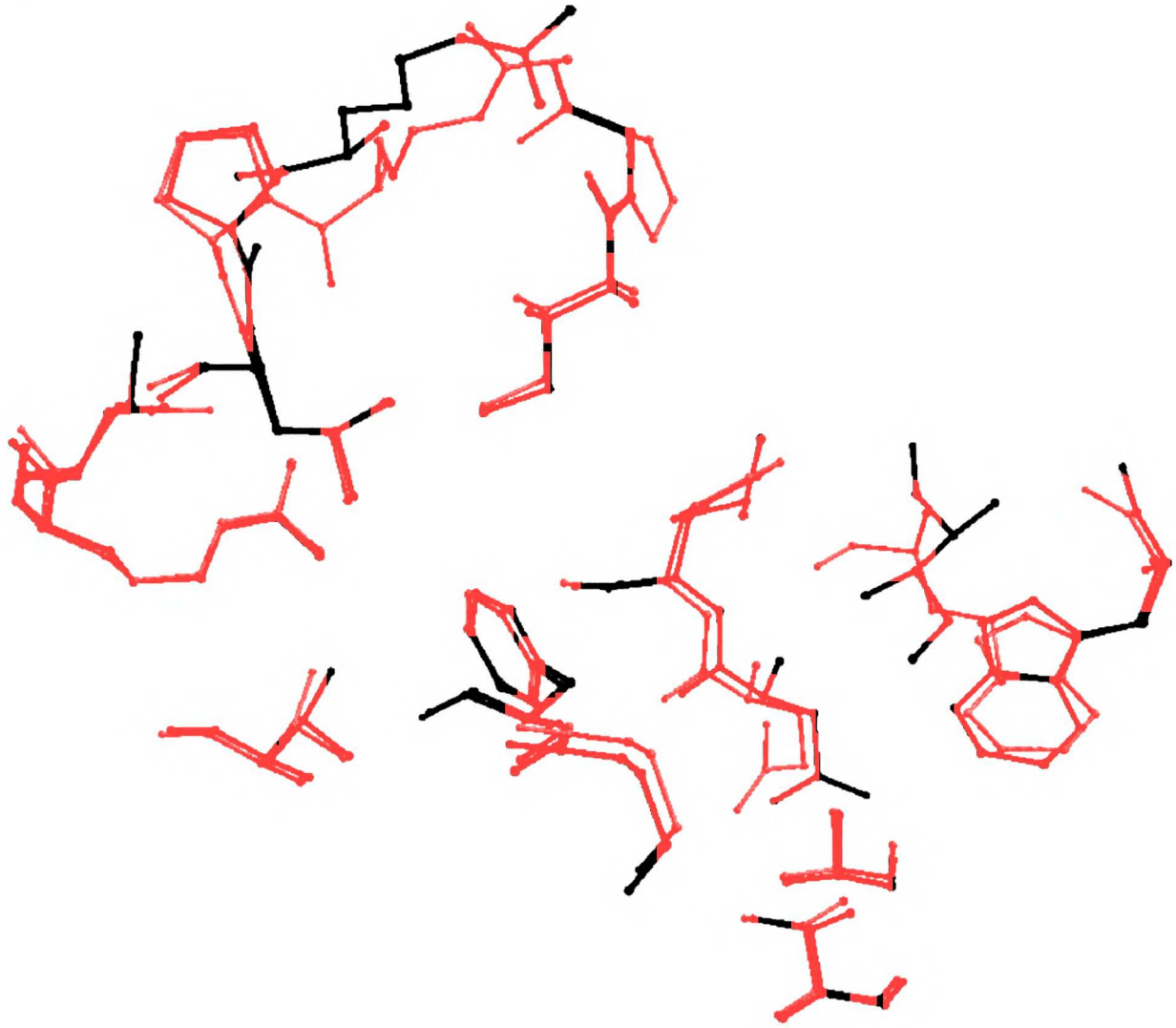
The structural alignment of the three crystal structures (human, chicken and *L. Caseii)* are shown aligned to the *C. Parvum* model. Only backbone chain traces are shown. The traces are colored with a rainbow color cycle to clarify the chain progression.

FIGURE 6-2: backbones of model and 3 crystal structures

FIGURE 6-3: similarity of active site residues

The active site residues (as listed in Figure 6-1) are shown for the model DHFR and the human structure. The human structure is shown in red.

FIGURE 6-3: active site residues from model and human DHFR



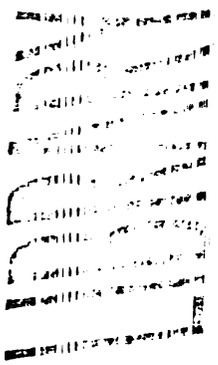
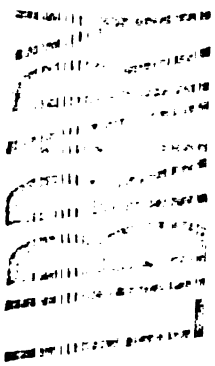


TABLE 6-4: final packing data for the model

Packing data are shown for the model and human DHFR as calculated using QPACK. The difference in packing for corresponding residues is also tabulated. The data is based on an ideal value of 1.0.



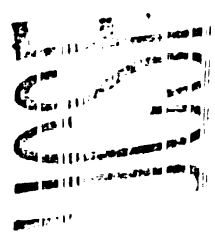


TABLE 6-4: final packing data for the model

human DHFR			crypto DHFR				
Residue	Residue #	Packing	Residue	Residue #	Packing	Δ Packing	
			MET	1	1.78		
			SER	2	0.91		
VAL	1	1.00	LYS	3	0.91	-0.09	
GLY	2	0.93	LYS	4	0.75	-0.18	
SER	3	0.93	ASN	5	0.82	-0.11	
LEU	4	0.98	VAL	6	1.06	0.08	
ASN	5	0.88	SER	7	0.99	0.11	
CYS	6	0.91	ILE	8	0.89	-0.02	
ILE	7	0.90	VAL	9	0.95	0.05	
VAL	8	0.98	VAL	10	0.96	-0.02	
ALA	9	0.91	ALA	11	0.91	0.00	
VAL	10	1.05	ALA	12	0.71	-0.34	<----
SER	11	0.84	SER	13	0.79	-0.05	
GLN	12	0.84	VAL	14	0.93	0.09	
ASN	13	0.80	LEU	15	0.93	0.13	
MET	14	1.10	SER	16	1.15	0.05	
			ARG	17	0.59		
GLY	15	0.93	GLY	18	0.62	-0.31	<----
ILE	16	1.00	ILE	19	1.03	0.03	
GLY	17	1.09	GLY	20	1.11	0.02	
LYS	18	0.99	ILE	21	0.87	-0.12	
ASN	19	1.04	ASN	22	1.17	0.13	
GLY	20	1.46	GLY	23	1.40	-0.06	
ASP	21	0.93	GLN	24	1.23	0.30	<----
LEU	22	0.93	LEU	25	0.85	-0.08	
PRO	23	0.92	PRO	26	0.87	-0.05	
TRP	24	0.80	TRP	27	0.91	0.11	
PRO	25	1.05					
PRO	26	1.10	SER	28	0.91	-0.19	
LEU	27	0.94	ILE	29	0.95	0.01	
ARG	28	1.22	SER	30	1.15	-0.07	
ASN	29	0.98	GLU	31	1.05	0.07	
GLU	30	0.99	ASP	32	0.95	-0.04	
PHE	31	1.20	LEU	33	1.05	-0.15	
ARG	32	0.98	LYS	34	1.05	0.07	
TYR	33	1.02	PHE	35	1.19	0.17	
PHE	34	1.20	PHE	36	1.19	-0.01	
GLN	35	1.34	SER	37	1.20	-0.14	

human DHFR			crypto DHFR				
Residue	Residue #	Packing	Residue	Residue #	Packing	Δ Packing	
ARG	36	0.99	LYS	38	1.16	0.17	
MET	37	0.98	ILE	39	0.96	-0.02	
THR	38	0.99	THR	40	1.02	0.03	
THR	39	0.85	SER	41	1.09	0.24	<----
THR	40	1.47	ASN	42	1.09	-0.38	<----
SER	41	0.95	ASN	43	1.00	0.05	
SER	42	0.98	CYS	44	1.03	0.05	
VAL	43	0.98	ASP	45	1.06	0.08	
GLU	44	1.07	SER	46	0.75	-0.32	<----
GLY	45	0.97	ASN	47	0.88	-0.09	
LYS	46	0.97	LYS	48	1.04	0.07	
GLN	47	1.15	LYS	49	1.01	-0.14	
ASN	48	0.94	ASN	50	1.11	0.17	
LEU	49	0.96	ALA	51	0.99	0.03	
VAL	50	1.03	LEU	52	0.96	-0.07	
ILE	51	1.00	ILE	53	1.18	0.18	
MET	52	1.10	MET	54	0.93	-0.17	
GLY	53	1.14	GLY	55	1.07	-0.07	
LYS	54	1.05	ARG	56	0.98	-0.07	
LYS	55	1.05	LYS	57	0.98	-0.07	
THR	56	0.85	THR	58	0.85	0.00	
TRP	57	0.99	TRP	59	0.93	-0.06	
PHE	58	1.31	ASP	60	1.17	-0.14	
SER	59	1.03	SER	61	1.03	0.00	
ILE	60	1.03	ILE	62	1.03	0.00	
PRO	61	1.07	GLY	63	1.23	0.16	
GLU	62	0.95					
LYS	63	0.90					
ASN	64	0.90	ARG	64	1.02	0.12	
ARG	65	0.95	ARG	65	0.87	-0.08	
PRO	66	1.07	PRO	66	0.89	-0.18	
LEU	67	0.89	LEU	67	0.87	-0.02	
LYS	68	1.18	LYS	68	1.09	-0.09	
GLY	69	1.18	ASN	69	1.09	-0.09	
ARG	70	0.85	ARG	70	0.87	0.02	
ILE	71	0.96	LYS	71	1.01	0.05	
ASN	72	1.01	ILE	72	1.01	0.00	
LEU	73	0.89	VAL	73	0.84	-0.05	
VAL	74	0.99	VAL	74	1.04	0.05	
LEU	75	1.14	ILE	75	1.07	-0.07	
SER	76	0.91	SER	76	0.91	0.00	

human DHFR			crypto DHFR				
Residue	Residue #	Packing	Residue	Residue #	Packing	Δ Packing	
ARG	77	1.23	SER	77	1.09	-0.14	
GLU	78	1.44	SER	78	1.09	-0.35	<----
LEU	79	0.91	LEU	79	0.91	0.00	
LYS	80	1.39	PRO	80	1.29	-0.10	
GLU	81	0.99	GLN	81	1.11	0.12	
PRO	82	0.92	ASP	82	0.90	-0.02	
PRO	83	1.04	GLU	83	0.85	-0.19	
GLN	84	1.11	ALA	84	1.08	-0.03	
GLY	85	1.06	ASP	85	0.89	-0.17	
ALA	86	0.94	PRO	86	0.90	-0.04	
HIS	87	1.00	ASN	87	1.01	0.01	
PHE	88	0.89	VAL	88	0.84	-0.05	
LEU	89	0.92	ILE	89	0.94	0.02	
SER	90	0.94	VAL	90	0.92	-0.02	
ARG	91	1.23	PHE	91	1.04	-0.19	
			ARG	92	1.17		
SER	92	0.89	ASN	93	0.85	-0.04	
LEU	93	0.95	LEU	94	1.03	0.08	
ASP	94	1.26	GLU	95	1.26	0.00	
ASP	95	0.89	ASP	96	0.85	-0.04	
ALA	96	0.94	SER	97	0.92	-0.02	
LEU	97	0.95	ILE	98	1.03	0.08	
LYS	98	1.26	LYS	99	1.26	0.00	
LEU	99	1.11	ASN	100	1.19	0.08	
THR	100	1.00	LEU	101	1.04	0.04	
GLU	101	1.23	MET	102	1.53	0.30	<----
GLN	102	1.08	ASN	103	0.99	-0.09	
PRO	103	1.08					
GLU	104	1.09					
LEU	105	1.04					
ALA	106	0.90	ASP	104	0.99	0.09	
ASN	107	0.90	ASP	105	1.32	0.42	<----
LYS	108	1.07	THR	106	0.95	-0.12	
VAL	109	1.04	ILE	107	0.88	-0.16	
ASP	110	0.89	GLU	108	0.84	-0.05	
MET	111	0.89	ASN	109	0.82	-0.07	
VAL	112	0.98	ILE	110	0.99	0.01	
TRP	113	1.03	PHE	111	0.96	-0.07	
ILE	114	0.91	VAL	112	0.89	-0.02	
VAL	115	1.03	CYS	113	1.13	0.10	
GLY	116	0.85	GLY	114	0.85	0.00	

human DHFR			crypto DHFR				
Residue	Residue #	Packing	Residue	Residue #	Packing	Δ Packing	
GLY	117	0.99	GLY	115	0.98	-0.01	
SER	118	1.06	GLU	116	0.99	-0.07	
SER	119	1.12	SER	117	0.99	-0.13	
VAL	120	1.05	ILE	118	1.05	0.00	
TYR	121	1.12	TYR	119	1.10	-0.02	
LYS	122	0.94	ARG	120	1.10	0.16	
GLU	123	1.05	ASP	121	0.92	-0.13	
ALA	124	1.02	ALA	122	0.90	-0.12	
MET	125	1.19	LEU	123	1.16	-0.03	
ASN	126	1.26	LYS	124	1.28	0.02	
HIS	127	1.02	ASP	125	0.90	-0.12	
PRO	128	1.03	ASN	126	1.01	-0.02	
GLY	129	1.25	PHE	127	1.04	-0.21	<----
HIS	130	0.95	VAL	128	0.93	-0.02	
LEU	131	1.07	ASP	129	1.06	-0.01	
LYS	132	0.93	ARG	130	0.98	0.05	
LEU	133	1.06	ILE	131	0.99	-0.07	
PHE	134	0.90	TYR	132	0.95	0.05	
VAL	135	1.04	LEU	133	0.96	-0.08	
THR	136	0.91	THR	134	0.91	0.00	
ARG	137	0.91	ARG	135	1.15	0.24	<----
ILE	138	0.93	VAL	136	0.93	0.00	
MET	139	1.02	ALA	137	1.15	0.13	
			LEU	138	0.93		
GLN	140	1.01	GLU	139	1.22	0.21	<----
ASP	141	0.80	ASP	140	1.22	0.42	<----
PHE	142	1.07	ILE	141	0.79	-0.28	<----
GLU	143	0.99	GLU	142	0.90	-0.09	
SER	144	0.93	PHE	143	0.80	-0.13	
ASP	145	1.04	ASP	144	1.06	0.02	
THR	146	1.04	THR	145	1.03	-0.01	
PHE	147	1.23	TYR	146	1.15	-0.08	
PHE	148	1.00	PHE	147	0.59	-0.41	<----
PRO	149	0.94	PRO	148	0.99	0.05	
GLU	150	1.69	GLU	149	1.74	0.05	
ILE	151	1.07	ILE	150	1.09	0.02	
ASP	152	0.83	PRO	151	0.96	0.13	
LEU	153	1.07	GLU	152	1.45	0.38	<----
GLU	154	1.41	THR	153	1.11	-0.30	<----
LYS	155	0.83	PHE	154	1.11	0.28	<----
TYR	156	0.88	LEU	155	0.79	-0.09	

human DHFR			crypto DHFR				
Residue	Residue #	Packing	Residue	Residue #	Packing	Δ Packing	
LYS	157	1.16	PRO	156	1.44	0.28	<----
LEU	158	0.90	VAL	157	1.01	0.11	
LEU	159	1.00	TYR	158	0.89	-0.11	
PRO	160	1.19	MET	159	0.89	-0.30	<----
GLU	161	1.19	SER	160	0.76	-0.43	<----
TYR	162	0.95					
PRO	163	1.10					
GLY	164	1.16					
VAL	165	0.95					
LEU	166	1.07					
SER	167	1.19					
ASP	168	1.09	GLN	161	1.20	0.11	
VAL	169	1.09	THR	162	0.94	-0.15	
GLN	170	1.07	PHE	163	0.83	-0.24	<----
GLU	171	0.72	CYS	164	0.93	0.21	<----
GLU	172	0.91	THR	165	1.08	0.17	
LYS	173	0.91	LYS	166	0.94	0.03	
GLY	174	1.19	ASN	167	0.99	-0.20	
ILE	175	1.01	ILE	168	0.97	-0.04	
LYS	176	0.72	SER	169	0.93	0.21	<----
TYR	177	0.93	TYR	170	1.01	0.08	
LYS	178	1.02	ASP	171	0.94	-0.08	
PHE	179	1.12	PHE	172	1.20	0.08	
GLU	180	0.91	MET	173	0.76	-0.15	
VAL	181	0.97	VAL	174	1.10	0.13	
TYR	182	0.88	PHE	175	1.00	0.12	
GLU	183	0.93	GLU	176	1.09	0.16	
LYS	184	0.92	LYS	177	0.72	-0.20	
ASN	185	0.95	GLN	178	0.93	-0.02	
ASP	186	1.02	GLU	179	0.72	-0.30	<----

TABLE 6-5: final dihedral data for model

The main chain (phi, psi) and side chain (chi1, chi2) dihedral values are shown for residues 2-178. The "# in bin" and "% in bin" columns refer to the phi/psi bin from Dunbrack's database (Dunbrack and Karplus 1993). The "Rotamer #" column refers to the primary, secondary or tertiary rotamer as indicated by the database.

TABLE 6-5: final dihedral data for model

Residue	Res. #	phi	psi	# in bin	chi 1	chi 2	Rotamer #	% in bin
SER	2	-90	166	40	57	0	1	88
LYS	3	-144	154	58	-83	90	3	60
LYS	4	-99	-67	2	-98	66	3	50
ASN	5	-165	148	4	-176	12	2	50
VAL	6	-139	127	94	177	0	2	72
SER	7	-142	158	102	53	0	1	61
ILE	8	-124	140	138	-64	169	3	76
VAL	9	-128	144	195	-177	0	2	70
VAL	10	-164	152	13	51	0	1	85
SER	13	-84	-180	21	56	0	1	95
VAL	14	-60	-105	0	-172	0	2	0
LEU	15	-7	-75	1	-69	178	3	100
SER	16	-116	-23	7	52	0	1	57
ARG	17	109	59	0	-162	-154	2	0
ILE	19	-146	-16	1	64	164	1	0
ILE	21	-139	112	16	-68	178	3	88
ASN	22	52	47	66	-81	-44	3	65
GLN	24	-156	177	3	60	172	1	100
LEU	25	-71	134	88	-72	-76	3	43
PRO	26	-65	-8	97	-28	47	3	13
TRP	27	-140	-152	0	52	58	1	0
SER	28	-139	152	116	63	0	1	44
ILE	29	-154	161	20	-71	179	3	5
SER	30	-132	-63	2	-60	0	3	50
GLU	31	-61	-32	322	-66	-175	3	63
ASP	32	-66	-26	255	-133	-51	2	12
LEU	33	-71	-35	622	-56	170	3	64
LYS	34	-65	-40	520	-64	-162	3	49
PHE	35	-63	-50	235	173	-102	2	80
PHE	36	-52	-49	141	168	83	2	79
SER	37	-60	-58	33	71	0	1	21
LYS	38	-57	-46	366	178	179	2	55
ILE	39	-55	-56	45	-67	161	3	96
THR	40	-69	-26	170	57	0	1	51
SER	41	-91	-42	27	64	0	1	30
ASN	42	-65	134	26	-74	-27	3	23
ASN	43	-131	176	9	-77	-78	3	0
CYS	44	-173	-138	0	-68	0	3	0

Residue	Res. #	phi	psi	# in bin	chi 1	chi 2	Rotamer #	% in bin
ASP	45	-47	142	15	-67	-29	3	60
SER	46	94	-10	0	-56	0	3	0
ASN	47	-174	-122	0	167	-116	2	0
LYS	48	-160	-37	1	-68	-171	3	100
LYS	49	-130	151	64	-71	-175	3	73
ASN	50	-89	151	24	-75	-41	3	75
LEU	52	-109	119	121	-67	-168	3	37
ILE	53	-106	126	202	-71	-176	3	90
MET	54	-150	162	18	65	-169	1	50
ARG	56	-58	-52	267	-170	-167	2	57
LYS	57	-65	-36	418	-86	-141	3	53
THR	58	-62	-46	230	-58	0	3	92
TRP	59	-49	-55	16	171	-108	2	75
ASP	60	-61	-22	83	-68	-35	3	71
SER	61	-77	-18	131	86	0	1	73
ILE	62	-90	137	62	-52	152	3	84
ARG	64	-81	138	39	-74	-171	3	72
ARG	65	-98	26	1	-66	-173	3	0
PRO	66	-142	127	0	31	-19	1	0
LEU	67	-72	105	5	-60	-178	3	20
LYS	68	-52	138	24	90	-116	1	17
ASN	69	78	6	11	-59	-40	3	82
ARG	70	-130	151	44	-65	151	3	70
LYS	71	-84	118	33	-67	-179	3	36
ILE	72	-96	123	136	-58	159	3	96
VAL	73	-124	123	237	-178	0	2	92
VAL	74	-100	139	91	-180	0	2	78
ILE	75	-106	125	176	-69	-164	3	94
SER	76	-169	152	14	-178	0	2	50
SER	77	-109	-19	12	51	0	1	67
SER	78	-82	-45	82	55	0	1	41
LEU	79	-66	128	88	-104	14	3	43
PRO	80	-83	-27	21	31	-35	1	76
GLN	81	-149	151	26	-169	-145	2	27
ASP	82	-67	152	58	-140	-10	2	22
GLU	83	-62	-174	0	-70	-164	3	0
ASP	85	100	-23	1	-80	-40	3	0
PRO	86	-78	171	60	31	-23	1	90
ASN	87	-106	-45	3	-63	-39	3	100
VAL	88	-137	156	94	-58	0	3	82
ILE	89	-125	154	78	-58	165	3	31
VAL	90	-146	151	55	176	0	2	20

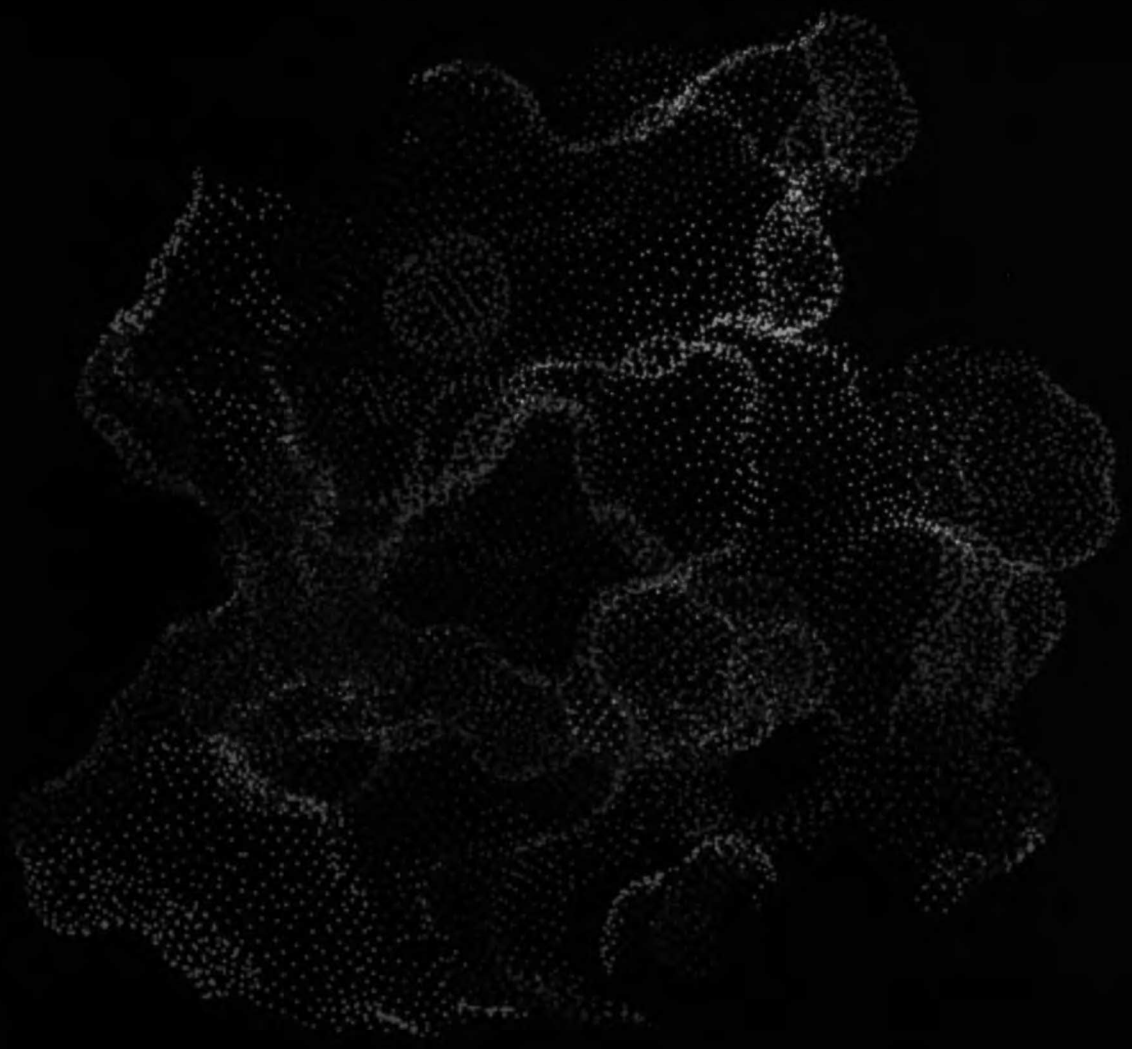
Residue	Res. #	phi	psi	# in bin	chi 1	chi 2	Rotamer #	% in bin
PHE	91	-69	31	0	-75	-77	3	0
ARG	92	55	27	7	-157	-176	2	14
ASN	93	-89	-29	10	-65	-35	3	70
LEU	94	66	-46	1	-169	88	2	0
GLU	95	-56	-32	322	-65	-173	3	63
ASP	96	-74	-40	317	-73	-167	3	82
SER	97	-67	-44	289	67	0	1	35
ILE	98	-67	-42	324	-62	166	3	90
LYS	99	-58	-41	520	-81	169	3	49
ASN	100	-59	-34	118	-63	-34	3	79
LEU	101	-67	-9	51	-60	176	3	86
MET	102	-66	-7	21	-68	-68	3	71
ASN	103	-66	85	3	-101	82	3	100
ASP	104	57	-112	1	-58	-61	3	100
ASP	105	-123	69	6	-63	-41	3	50
THR	106	-94	-28	42	69	0	1	67
ILE	107	-146	134	19	179	180	2	63
GLU	108	-92	-99	0	-178	-156	2	0
ASN	109	-85	150	24	-46	-50	3	75
ILE	110	-114	119	176	-55	154	3	94
PHE	111	-113	125	38	-61	-105	3	74
VAL	112	-90	123	105	-72	0	3	4
CYS	113	-129	13	1	68	0	1	100
GLU	116	-60	-39	593	-64	175	3	56
SER	117	-66	-36	289	32	0	1	35
ILE	118	-70	-42	324	-97	50	3	90
TYR	119	-62	-53	182	-69	137	3	18
ARG	120	-51	-52	127	-173	76	2	61
ASP	121	-70	-45	162	75	-45	1	7
LEU	123	-57	-32	348	-68	174	3	78
LYS	124	-78	-23	95	-70	175	3	69
ASP	125	-53	124	10	-73	-14	3	10
ASN	126	-65	151	22	-68	-43	3	55
PHE	127	113	-165	0	-59	112	3	0
VAL	128	-82	126	88	178	0	2	88
ASP	129	-151	153	13	-176	24	2	85
ARG	130	-134	140	48	-74	-160	3	56
ILE	131	-111	120	176	-43	-50	3	94
TYR	132	-101	105	7	-72	76	3	43
LEU	133	-110	127	145	-62	-166	3	50
THR	134	-100	116	93	-57	0	3	94
ARG	135	-92	105	8	-59	-70	3	75

Residue	Res. #	phi	psi	# in bin	chi 1	chi 2	Rotamer #	% in bin
VAL	136	-83	109	36	-57	0	3	6
LEU	138	-68	165	53	-55	155	3	98
GLU	139	-93	-68	1	-65	178	3	100
ASP	140	-89	89	37	-69	-27	3	14
ILE	141	-121	141	138	-34	164	3	76
GLU	142	-114	157	21	-63	-174	3	76
PHE	143	-157	152	30	-64	89	3	3
ASP	144	-111	-7	19	64	19	1	42
THR	145	-142	133	48	-150	0	2	17
TYR	146	-130	157	55	-51	84	3	65
PHE	147	-88	141	36	-176	95	2	31
PRO	148	-62	150	259	19	-30	1	40
GLU	149	-61	137	50	-81	-159	3	58
ILE	150	-90	121	107	-165	163	2	2
PRO	151	-79	99	3	27	-33	1	100
GLU	152	-63	-8	42	-66	-175	3	43
THR	153	-76	31	1	59	0	1	100
PHE	154	178	-35	0	-71	103	3	0
LEU	155	-88	107	59	-70	176	3	46
PRO	156	-90	173	24	27	-39	1	100
VAL	157	-108	142	156	169	0	2	77
TYR	158	-102	-79	0	-73	98	3	0
MET	159	-110	147	17	-67	-61	3	82
SER	160	-143	136	75	81	0	1	27
GLN	161	-147	113	4	-54	-57	3	0
THR	162	-60	170	8	-44	0	3	13
PHE	163	-176	172	1	-72	95	3	0
CYS	164	-154	129	10	-178	0	2	70
THR	165	-147	139	24	-167	0	2	58
LYS	166	53	32	9	-47	-116	3	100
ASN	167	75	10	11	-56	-46	3	82
ILE	168	-108	122	176	-53	-177	3	94
SER	169	-91	133	53	-177	0	2	58
TYR	170	-140	158	48	57	87	1	71
ASP	171	-126	146	11	-56	-33	3	36
PHE	172	-99	141	59	-61	132	3	78
MET	173	-141	143	30	-66	-64	3	37
VAL	174	-130	126	223	-63	0	3	7
PHE	175	-115	147	46	-55	69	3	87
GLU	176	-130	152	52	-175	173	2	13
LYS	177	-151	153	44	95	100	1	34
GLN	178	-142	101	4	-58	-170	3	25

FIGURE 6-4: charge distribution in active site

Surface representation of charge in active site of the model for DHFR from *C. Parvum*. Colors at the yellow end of the spectrum are more positive and at the blue end are more negative

FIGURE 6-4: electrostatic surface of active site



REFERENCES

Bolin, J. T., D. J. Filman, et al. (1982). "Crystal structures of Escherichia coli and Lactobacillus casei dihydrofolate reductase refined at 1.7 A resolution. I. General features and binding of methotrexate." J Biol Chem **257**(22): 13650-62.

DesJarlais, R. L., R. P. Sheridan, et al. (1988). "Using Shape Complementarity as an Initial Screen in Designing Ligands for a Receptor Binding Site of Known Three-Dimensional Structure." J. Med. Chem. **31**(4): 722-729.

Dunbrack, R. L. J. and M. Karplus (1993). "Backbone-dependent rotamer library for proteins. Application to side-chain prediction." J. Mol. Biol. **230**(2): 543-574.

Fayer, R. and B. L. Ungar (1986). "Cryptosporidium spp. and cryptosporidiosis." Microbiol Rev **50**(4): 458-83.

Ferrin, T., C. Huang, et al. (1988). "The MIDAS Display System." J. Mol. Graph. **6**: 13-37.

Gregoret, L. M. and F. E. Cohen (1990). "Novel method for the rapid evaluation of packing in protein structures." J Mol Biol **211**(4): 959-74.

Gronenborn, A. M. and G. M. Clore (1990). "Protein structure determination in solution by two-dimensional and three-dimensional nuclear magnetic resonance spectroscopy." Anal Chem **62**(1): 2-15.

Lee, B. and F. M. Richards (1971). "The Interpretation of Protein Structures: Estimation of Static Accessibility." J. Mol. Biol. **55**: 379-400.

McGregor, M. J., S. A. Islam, et al. (1987). "Analysis of the relationship between side-chain conformation and secondary structure in globular proteins." J Mol Biol **198**(2): 295-310.

McTigue, M. A., J. F. d. Davies, et al. (1992). "Crystal structure of chicken liver dihydrofolate reductase complexed with NADP+ and biopterin." Biochemistry **31**(32): 7264-73.

Nelson, R. (1994). .

Oefner, C., A. D'Arcy, et al. (1988). "Crystal structure of human dihydrofolate reductase complexed with folate." Eur J Biochem **174**(2): 377-85.

Ponder, J. W. and F. M. Richards (1987). "Tertiary Templates for Proteins: Use of Packing Criteria in the Enumeration of Allowed Sequences for Different Structural Classes." J. Mol. Biol. **193**: 775-791.

Ring, C. S. and F. E. Cohen (1994). "Conformational Sampling Of Loop Structures Using Genetic Algorithms." Israel Journal Of Chemistry **34**(2): 245-252.

Weiner, P. K. and P. A. Kollman (1981). "AMBER: Assisted model building with energy refinement. A general program for modeling molecules and their interactions." J. Comp. Chem. **2**: 287-303.

Chapter 7:
Docking of Ligands to a DHFR Model

INTRODUCTION

Traditionally, drug discovery has been the result of large scale screening trials or serendipitous compound identification. It has been discovered that many such drugs have biologically significant macromolecules (proteins, nucleic acids, etc.) as their targets. In recent years, a new, computationally-based approach has developed that makes use of macromolecular structure (Sun and Cohen 1993; Whittle and Blundell 1994). If we know the structure of the target molecule, this should help to suggest ligands that fit the target well enough to be inhibitors. This is often considered to be a "lock and key" model for drug/macromolecule interaction.

Simple visual analysis of a protein structure in an attempt to identify ligands is a laborious process. Without a previously identified drug to serve as a starting point, it is difficult to even know what family of compounds to consider let alone which exact ligand to select. Because protein structure is easy to represent mathematically in terms of atom types, bonding patterns and coordinates, a number of approaches have evolved for computational selection of potential ligands. CAVEAT (Bartlett, Shea et al. 1989), is a vector-based algorithm that can identify ligands that match the profile of an active or binding site. DOCK (DesJarlais, Sheridan et al. 1988) is a well accepted algorithm for screening a large database of small molecules in search of possible ligands for a macromolecular target. The site on the target, which is almost always a pocket or invagination, is described using spheres that fill the site. These spheres are clustered to create a negative image of the target site. The small molecule database can then be screened for ligands that fit this negative image. A number of scoring methods have been developed including shape-based and force-field scoring. Ligands may be ranked based on such scores and the best scoring ligands can then be selected for further assessment. Ligands may be further evaluated by using of graphical display programs where a researcher can view the ligand in the target site in the orientation suggested by DOCK.

This allows one to narrow the list of ligands to a small enough collection that can easily be screened for binding using experimental assays.

In this work, a previously prepared model of DHFR from *C. Parvum* was used as the target for DOCK ligand screening. The Available Chemicals Directory (ACD) (Molecular Designs, San Leandro, CA, USA) was screened for ligands that might fit the pocket of this DHFR. Three different representations of the pocket were screened: 1) a sphere cluster generated by filling the active site in DHFR, 2) a cluster made up of spheres positioned at the atom centers from the known DHFR inhibitor (methotrexate) and 3) a hybrid cluster created by merging and re-clustering the spheres from the DHFR pocket and the spheres from methotrexate.

The spheres produced by filling the active site serve as a negative image of the site. If one assumes that the enzyme does not reorient significantly on binding a ligand, then this collection of spheres resembles an average of all possible ligands for the site. This assumption is not completely valid. Many enzymes do reorient in major ways upon binding a ligand. By generating a collection of spheres based on a known ligand, we make no assumptions about reorientation of the enzyme. We know that methotrexate itself does not inhibit DHFR from *C. Parvum*. However, its shape may fit the pocket, and minor side chain and charge modifications might produce an effective inhibitor. The methotrexate-based sphere set could identify compounds during the DOCK search that are closer to a true ligand in shape than those resulting from negative-image spheres. The final, hybrid sphere set combines these two views. In this case, we have a shape that is an average of all available space in the static binding site plus any additional space represented by the shape of methotrexate. This volume is a superset of cases one and two, but could identify matches that would not score highly enough to be retained during the first two searches.

A total of 5000 high scoring ligands were selected in each search, where 2500 were generated using contact scoring and 2500 resulted from force-field scoring.

Because the final list was a combination of three independent searches, there was significant redundancy within the 15000 compounds. Multiple rounds of visual analysis were undertaken to trim out unique compounds that were judged to be of potential interest. A list of 119 compounds was extracted as a result of these efforts.

METHODS

The model of dihydrofolate reductase (DHFR) described in the previous chapter was used as a basis for ligand docking studies. The program, DOCK (DesJarlais, Sheridan et al. 1988) was used to identify possible ligands that might bind in the active site of the model. Such compounds could then be evaluated experimentally and those that exhibited binding to the site could be used as lead compounds in drug development.

The DOCK program uses spheres clustered in the active or binding site of a macromolecule as a pattern that can be efficiently screened against a database of small molecule ligands. A number of scoring schemes have been developed, but this study focused on 2 scoring options: contact-only scoring and force-field scoring.

Three different DOCK runs were performed on the model:

1) Site-based search: Spheres were generated automatically using the SPHGEN utility from the DOCK program suite. The collection of spheres was reduced manually using the MIDAS graphics package (Ferrin, Huang et al. 1988) so that redundant spheres were eliminated. (22 spheres)

2) Inhibitor-based search: Spheres were assigned to every atom position in methotrexate, based on its orientation in the *Lactobacillus Caseii* structure (Bolin, Filman et al. 1982) while aligned to the model. (33 spheres)

3) Site- and ligand-based hybrid search: The spheres from runs 1 & 2 were combined. Redundant spheres were trimmed out using MIDAS. (36 spheres)

The Available Chemicals Directory (ACD) (Molecular Designs, San Leandro, CA, USA) was used as a source of ligand structures for these searches. The database was divided into five sub-databases that could be searched independently, facilitating more efficient use of CPU time on multiple workstations. From each DOCK run (one sphere site compared to 1/5 of the ACD, one scoring option), the 500 top-scoring ligands were retained. Hence for the complete database and all three sphere sets, 15,000 (i.e. 500 x 2 x 3 x 5) ligands were accepted.

These 15,000 possible ligands were then subjected to multiple rounds of visual screening to identify a small number for experimental assessment. The first trim was intended to quickly and effectively reduce the collection of ligands to a few thousand. MIDAS was used to view the site and the ligands, and a delegate program was prepared that allowed a single button to step through the ligands. As shown in Figure 7-1, I looked for ligands that filled zone C, plus occupying either zone A or zone B, or both. In addition, I trimmed out any ligands that clearly exceeded the limits of the site by a large amount. This initial trim reduced the number of ligands from 15,000 to 2,267.

The second trim step involved a different goal than the first. It was still important to eliminate uninteresting ligands, but in addition I began clustering the ligands into families of similar structures. Because the results from three separated DOCK searches were combined to generate the 15,000 ligands, there was a significant likelihood for duplication amongst the ligands. Some examples of templates for these ligand families are shown in Figure 7-2. All ligands from a given family were extracted and placed in a separate database file. In this step, 13 families were extracted in addition to a "Miscellaneous" category containing 611 ligands.

The third trim step repeated the goals of the second trim, but was concentrated on the Miscellaneous category. Ten more ligand families were extracted as well as the "Miscellaneous 2" category which contained 347 ligands.

FIGURE 7-1: ligand selection zones in active site

The electrostatic surface of DHFR model active site is shown with a cartoon of the ligand selection zones.

FIGURE 7-1: ligand sub-zones in electrostatic surface

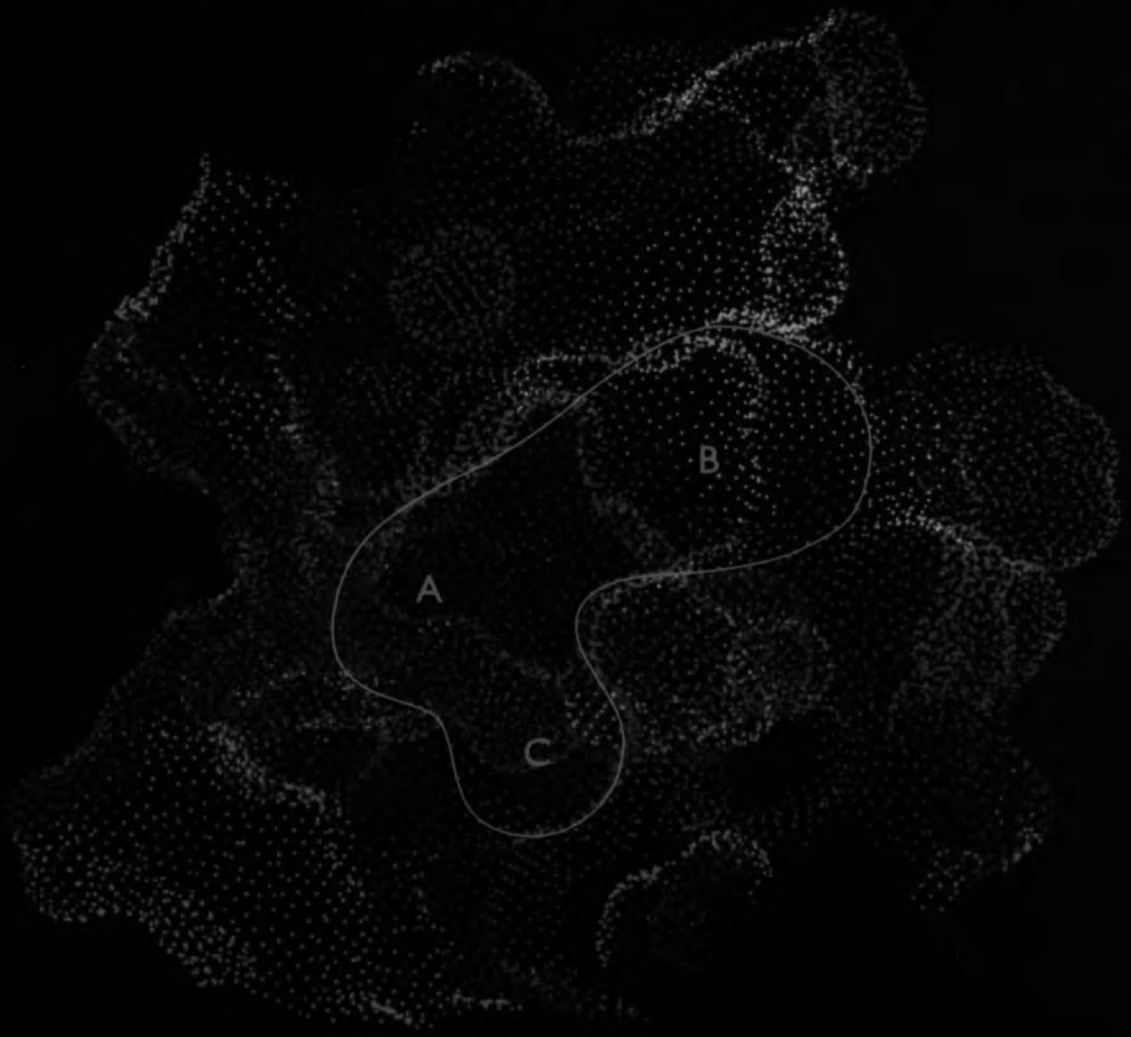
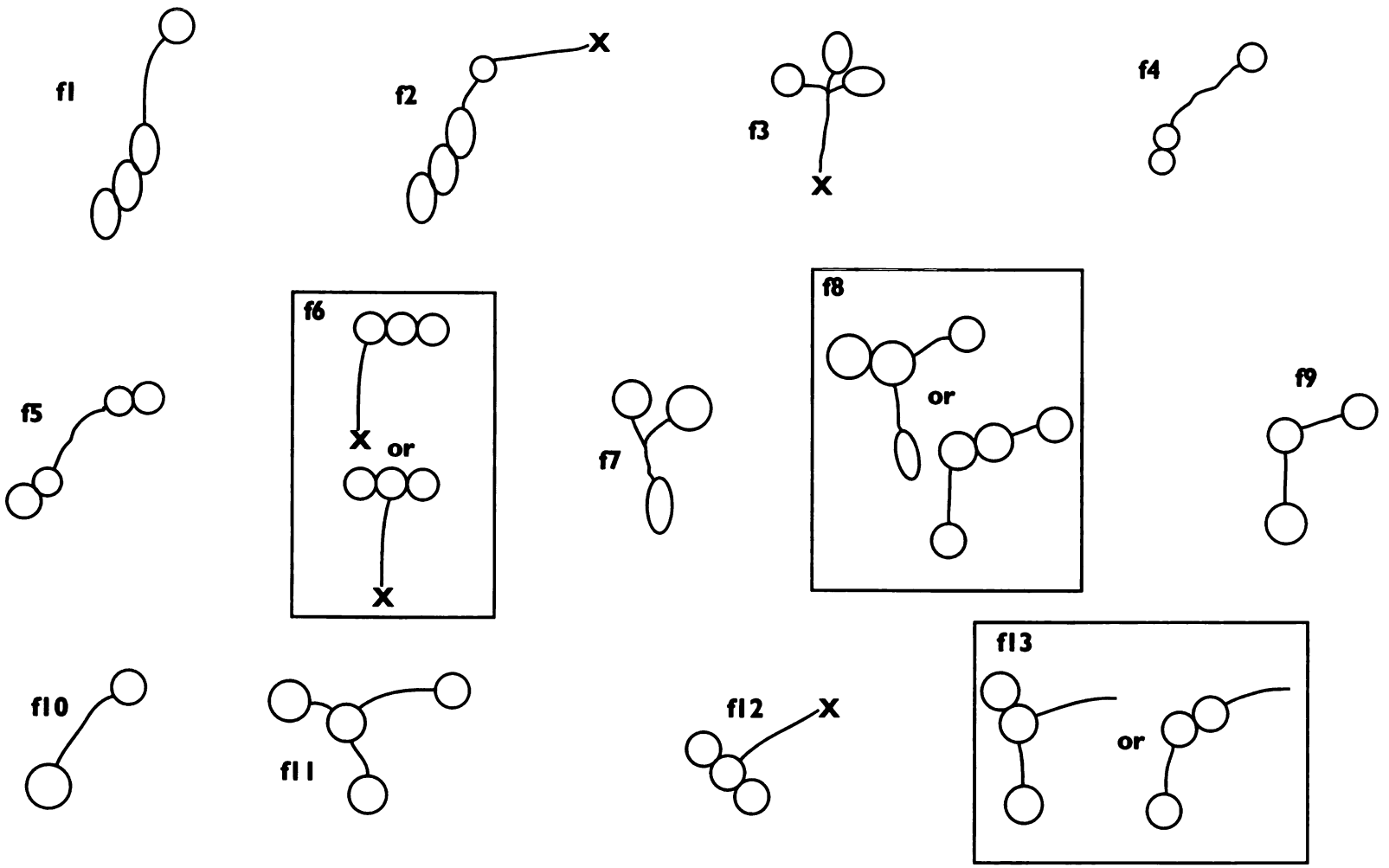


FIGURE 7-2: ligand shape families

Cartoon representations of some of the ligand families used to group the small molecules from the DOCK run. Circles and ovals represent rings containing varying numbers of atoms. Straight and curved lines represent varying numbers of linear bonds.

FIGURE 7-2: ligand shape families



A fourth trim, comparable to trims 2 and 3 was applied to the Miscellaneous 2 category. I extracted 14 more ligand families, and left 260 ligands in the "Miscellaneous 3" category.

The resulting 37 families plus the Miscellaneous 3 category were then screened visually to produce a final list of structures. This final visual screen was more detailed, focusing more on the identification of good candidates instead of screening out bad ones. An electrostatic surface representation for the model's active site was prepared using MS (Connolly 1992). This surface was displayed in MIDAS and the ligands were screened against it. Based on the electrostatic representation, zones A and C appear to be more negatively charged, while zone B is more positively charged (Figure 7-1). Ligands were chosen that complemented these electrostatic characteristics well. Whole families could quickly be eliminated, which proved to be a worthwhile benefit of ligand clustering. Large families that did appear interesting were reduced to a few top candidates. The final list after this screen contained 131 ligands, but 12 of these were duplications, resulting in a final list of 119 unique ligands.

RESULTS & DISCUSSION

Table 7-1 lists the 119 ligands selected by my docking efforts. Certain trends were observed during the trimming steps:

- Compounds that matched the pocket well typically had 2-3 rings joined by linkers.
- Numerous nucleotide mono-, di- and triphosphates were identified by the DOCK algorithm.
- Numerous blocked amino acids were identified by the DOCK algorithm.

During the visual trim step, I attempted to narrow the list to a broad variety of compounds with little repetition. Hopefully this will lead to a better range of possible

TABLE 7-1: ligand finalists

119 ligands selected from the DOCK runs for the *C. Parvum* DHFR model.

ACD Registry #	Compound Name
MFCD00003889	ACID ALIZARIN VIOLET N
MFCD00003934	ERIOCHROME BLUE BLACK B
MFCD00005095	6,7-DIHYDRO-5,8 -DIMETHYLDIBENZO[B,J][1,10]PHENANTHROLINE
MFCD00005649	5-FLUORO-DL-TRYPTOPHAN
MFCD00005714	4-(2-KETO-1-BENZIMIDAZOLINYL)PIPERIDINE
MFCD00005756	2',3'-ISOPROPYLIDENEADENOSINE
MFCD00006708	TRIAMTERENE
MFCD00009731	9,10-DIHYDRO-2-METHYL-4-(4-METHYL-1-PIPERAZINYL)-4H-BENZO[5.6]CYCLOHEPT[1,2-D]-
MFCD00012297	2,4,6-TRIPHENYLPHENOL
MFCD00012656	TRIFLUOPERAZINE DIHYDROCHLORIDE
MFCD00013261	9-FLUORENYLMETHYL PENTAFLUOROPHENYL CARBONATE
MFCD00029275	1-(2-FLUOROPHENYLIMINOMETHYL)-2-NAPHTHOL
MFCD00029296	2-(1-NAPHTHYL)-N-(3-SULFAMOYLPHENYL)ACETAMIDE
MFCD00030572	2-(2-FURYL)-4-(3-NITROBENZYLIDENE)-2-OXAZOLIN-5-ONE
MFCD00030744	3,4-DIHYDRO-3,3-DIPHENYL-1H-2-BENZOPYRAN-1-ONE
MFCD00031539	P-NITROPHENACYLTRIPHENYLPHOSPHONIUM BROMIDE
MFCD00033447	4B,11-DIHYDRO-4B-PHENYLISOINDOLO[1,2-B]BENZOTHIAZOL-11-ONE
MFCD00033608	4-(2-CHLOROBENZYLIDENE)-2-(3,4,5-TRIMETHOXYPHENYL)-2-OXAZOLIN-5-ONE
MFCD00033610	4-(2-HYDROXYBENZYLIDENE)-2-(3,4,5-TRIMETHOXYPHENYL)-2-OXAZOLIN-5-ONE
MFCD00034682	6-METHYL-N-EPSILON-(P-TOSYL)LYSINE HYDROCHLORIDE
MFCD00037020	SINEFUNGIN ACETONE COMPLEX
MFCD00038261	N-ALPHA-BOC-N-EPSILON-TOSYL-L-LYSINE
MFCD00038532	N-ALPHA-FMOC-L-PROLINE P-NITROPHENYLESTER
MFCD00038852	ADENYLYL-(3'-5')-URIDINE
MFCD00065672	N-(9-FLUORENYLMETHOXYCARBONYL)-L-PROLINE PENTAFLUOROPHENYL ESTER
MFCD00065673	FMOC-PRO-OSU
MFCD00065681	FMOC-TRP-OPFP
MFCD00066260	(TRITYLOXYMETHYL)-GAMMA-BUTYROLACTONE
MFCD00067001	1,2-BENZISOXAZOL-3-YL-DIPHENYLPHOSPHATE
MFCD00069713	ADENOSINE 5'-DIPHOSPHOMANNOSE SODIUM SALT
MFCD00069727	S-ADENOSYL-D-HOMOCYSTEINE
MFCD00070108	6-MERCAPTOPURINE RIBOSIDE 5'-PHOSPHATE SODIUM SALT
MFCD00070300	DANSYL AMINO ACIDS DANSYLTRYPTAMINE FREE ACID
MFCD00071932	KAYANOL RED NB
MFCD00078937	BASIC BLUE 47
MFCD00079227	3-PYRIDINEALDEHYDE ADENINE DINUCLEOTIDE
MFCD00079609	THIONICOTINAMIDE ADENINE DINUCLEOTIDE
MFCD00080143	[GLU1]-TRH
MFCD00102173	MAYBRIDGE BTB 10396
MFCD00102479	MAYBRIDGE NRB 02672
MFCD00102819	MAYBRIDGE NRB 03086
MFCD00102854	MAYBRIDGE NRB 03248
MFCD00102924	MAYBRIDGE NRB 03153
MFCD00102926	MAYBRIDGE NRB 03155
MFCD00103389	MAYBRIDGE SPB 01749
MFCD00103532	MAYBRIDGE KM 02027
MFCD00103566	MAYBRIDGE S 11716
MFCD00103777	MAYBRIDGE S 11834

MFCD00103833	MAYBRIDGE S 11860
MFCD00104317	MAYBRIDGE MWP 00355
MFCD00104319	MAYBRIDGE MWP 00357
MFCD00104830	MAYBRIDGE MWP 00412
MFCD00104832	MAYBRIDGE MWP 00414
MFCD00105454	MAYBRIDGE SE 01411
MFCD00106014	MAYBRIDGE GK 01006
MFCD00107446	MAYBRIDGE SPB 03022
MFCD00107876	MAYBRIDGE BTB 08114
MFCD00108178	MAYBRIDGE SPB 03253
MFCD00108588	MAYBRIDGE BTB 08354
MFCD00108909	MAYBRIDGE RDR 03552
MFCD00108969	MAYBRIDGE RDR 03559
MFCD00109338	MAYBRIDGE MWP 01013
MFCD00109612	MAYBRIDGE RDR 03865
MFCD00109719	MAYBRIDGE MWP 01054
MFCD00109815	MAYBRIDGE MWP 01065
MFCD00109920	MAYBRIDGE S 13370
MFCD00110125	MAYBRIDGE KM 03973
MFCD00110801	MAYBRIDGE MWP 00940
MFCD00110883	MAYBRIDGE SPB 02291
MFCD00110915	MAYBRIDGE SPB 05152
MFCD00110919	MAYBRIDGE SPB 05137
MFCD00111178	MAYBRIDGE BTB 09042
MFCD00111797	MAYBRIDGE DSHS 0613
MFCD00114924	MAYBRIDGE BTB 10571
MFCD00115076	MAYBRIDGE SEW 04037
MFCD00126289	MAYBRIDGE KM 06577
MFCD00126290	MAYBRIDGE KM 06584
MFCD00127386	2,6-DICHLOROBENZYLTHIO-4-PHENYLTRIAZOLOPYRIMIDINE
MFCD00127387	2,6-DICHLOROBENZYLTHIO-4-METHOXYPHENYLTRIAZOLOPYRIMIDINE
MFCD00129236	NAME NOT SUPPLIED
MFCD00129393	NAME NOT SUPPLIED
MFCD00129570	NAME NOT SUPPLIED
MFCD00129578	NAME NOT SUPPLIED
MFCD00129591	NAME NOT SUPPLIED
MFCD00134745	4'-ANILINOMALEANILIC ACID
MFCD00138061	NAME NOT SUPPLIED
MFCD00138281	NAME NOT SUPPLIED
MFCD00138432	NAME NOT SUPPLIED
MFCD00138610	NAME NOT SUPPLIED
MFCD00138611	NAME NOT SUPPLIED
MFCD00139872	HYDROXYMETHYL-THIOCHROMAN-4-ONE-DIBENZOATE
MFCD00140019	1-(1-(3,4-DICHLOROBENZYL))-3-(2-METHOXYCARBONYLPHENYL)UREIDO-2-PYRIDONE
MFCD00140328	NAME NOT SUPPLIED
MFCD00141072	1,3-DIPHENYL-6-BENZOYLAMIDO-7-OXO-8-OXA-1,2-DIAZA INDENE
MFCD00141109	1-(2,4-DICHLOROBENZYLOXY)2-PHENYL BENZIMIDAZOLE
MFCD00141199	METHYL 1-(4-CHLOROPHENYL)4-(2-PYRIMIDYLTHIO)6-PYRIDAZINONE-3-CARBOXYLATE
MFCD00141253	2-PHENYL-3-(2,4-DICHLOROBENZYLOXY)IMIDAZO[4,5-B]PYRIDINE
MFCD00141457	3-(4-BROMOPHENYL)-5-ANILINO-1,2,4-TRIAZOLO[4,3-C]-QUINAZOLINE
MFCD00141458	3-(4-BROMOPHENYL)-5-(4-METHOXYPHENYLAMINO)-1,2,4-TRIAZOLO[4,3-C]-QUINAZOLIN
MFCD00141477	2,4-DICHLOROPHENYL-PHENYLAMINO-1,2,4-TRIAZOLO[3,4-C]QUINAZOLINE
MFCD00141478	2,4-DICHLOROPHENYL-(4-METHOXYPHENYLAMINO)-1,2,4-TRIAZOLO[3,4-C]QUINAZOLINE
MFCD00142771	2-(2-ETHOXY-1-BUTENYL)-5-PHENYL-3-(3-SULFOBUTYL)BENZOXAZOLIUM INNER SALT
MFCD00142941	2-METHYL-3-SULFOPROPYL-5-PHENYL-BENZOXAZOLE-BETAINE

compounds. Multiple examples of the same compound type were minimized in favor of greater variety. Five examples of ligands that were selected as finalists are shown in Figure 7-3. Both similarities and differences are evident in these ligand structures. Figure 7-4 shows the active site of cryptosporidium DHFR with all of the 119 finalist compounds docked into it. The protein is shown as an electrostatic surface. This depiction gives a good idea for the extent of coverage by these compounds.

Clearly the next step in this study is the experimental assay of these 119 compounds. My hope is that a few compounds will be shown to inhibit the function of *C. Parvum* DHFR. If such compounds are found, then each may serve as a lead compound for chemical modification studies. Synthetic methods can be applied to the modification of these lead compounds in an effort to fine-tune the effectiveness of these ligands. The final list of compounds was made available to the laboratory of Richard Nelson, and I hope that we will have binding data for these compounds in the near future.

If no potential lead compounds are identified, then it will be important to review the docking approach in an effort to understand where it failed. It may be necessary at that point to try additional sets of spheres as well as more scoring approaches. As long as we are able to learn from the initial docking efforts described here, the study will have been worthwhile.

CONCLUSION

Computational approaches toward ligand docking have shown great promise in their ability to screen large sets of ligands (100,000+) and narrow a search by factors of a thousand or more. Although these methods are incapable of identifying a single, ideal ligand out of a database of small molecules, they are able to weed out compounds that are completely unreasonable. This trimming ability can dramatically reduce the number of compounds that need to be assayed in the laboratory or reviewed on a graphics

FIGURE 7-3: structures of some ligands

5 examples of ligands selected as finalists from the DOCK search

a) Maybridge #SPB03022

b) dichlorobenzyl thio-4-phenyl triazolopyrimidine

c) Maybridge #S13370

d) 2,6-triphenyl phenol

e) no supplier, ACD #38432

FIGURE 7-3: five example ligands

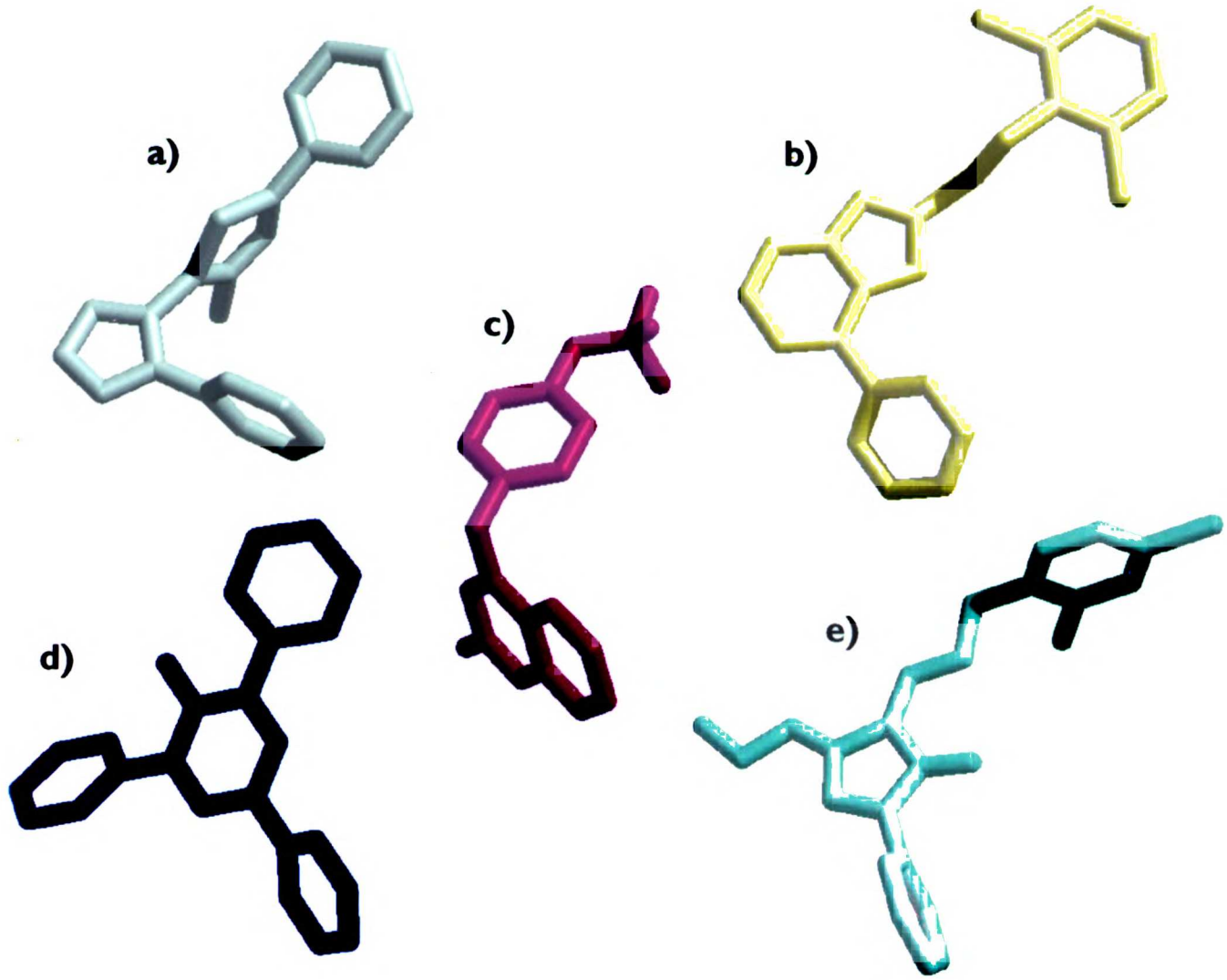
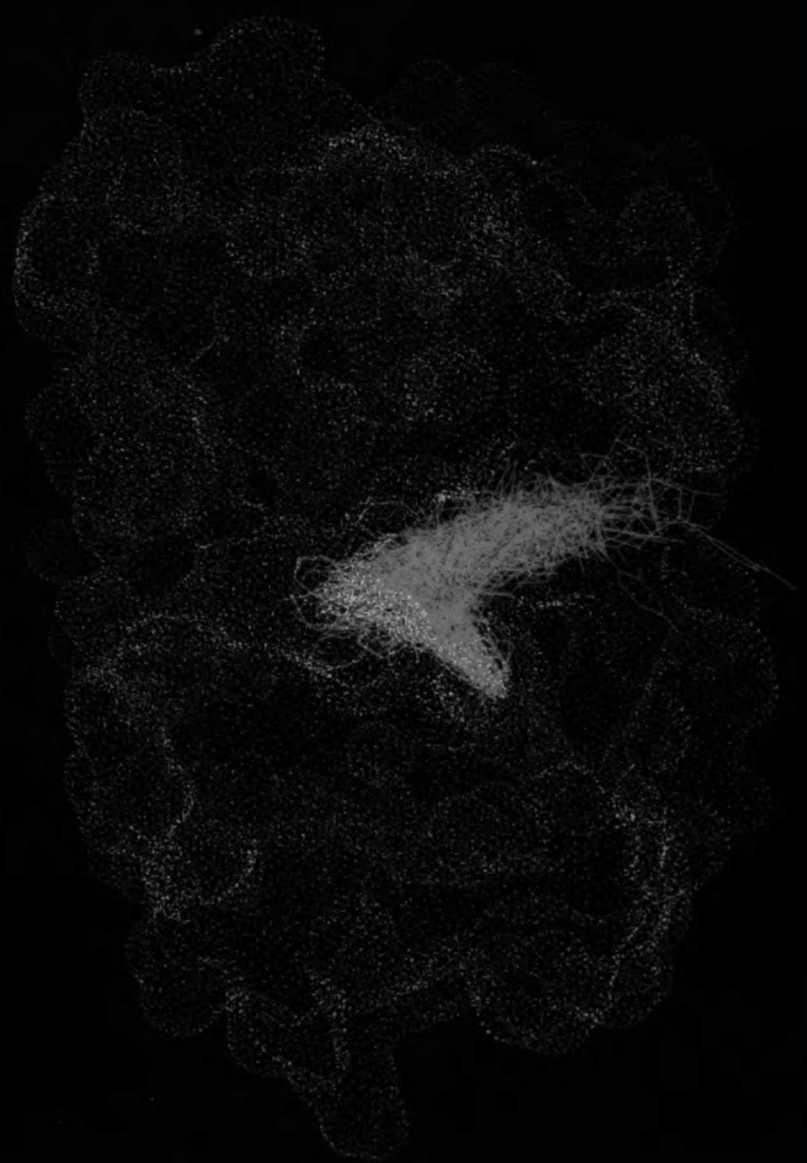


FIGURE 7-4: view of all ligands

The active site of the DHFR model with all 119 ligands docked. Presents an overall view of ligand coverage in the active site of the model.

FIGURE 7-4: DHFR model with 119 ligands



terminal. By allowing researchers to focus on the most likely ligands, computational docking tools have dramatically altered the world of drug discovery.

REFERENCES

Bartlett, P. A., G. T. Shea, et al. (1989). CAVEAT: A Program to Facilitate the Structure-derived Design of Biological Active Molecules. Molecular Recognition: Chemical and Biochemical Problems. Exeter, Royal Society of Chemistry: 182-196.

Bolin, J. T., D. J. Filman, et al. (1982). "Crystal structures of Escherichia coli and Lactobacillus casei dihydrofolate reductase refined at 1.7 A resolution. I. General features and binding of methotrexate." J Biol Chem **257**(22): 13650-62.

DesJarlais, R. L., R. P. Sheridan, et al. (1988). "Using Shape Complementarity as an Initial Screen in Designing Ligands for a Receptor Binding Site of Known Three-Dimensional Structure." J. Med. Chem. **31**(4): 722-729.

Ferrin, T., C. Huang, et al. (1988). "The MIDAS Display System." J. Mol. Graph. **6**: 13-37.

Sun, E. and F. E. Cohen (1993). "Computer-assisted drug discovery--a review." Gene **137**(1): 127-32.

Whittle, P. J. and T. L. Blundell (1994). "Protein structure--based drug design." Annu Rev Biophys Biomol Struct **23**: 349-75.