Title
Molecular docking

Author
Shoichet, Brian Kenton

Publication Date
1991

Peer reviewed|Thesis/dissertation

Molecular Docking:
Theory and Application to Recognition and Inhibitor Design

by

Brian Kenton Shoichet

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

My parents, Dorothy and Irving Shoichet

Preface

Since Newton, it has been the habit of scientists to acknowledge their predecessors and colleagues, upon whose work their own progress has depended. Never are such acknowledgements more appropriate than at the conclusion of a course of doctoral study. A successful graduate career depends on the good wishes and support of mentors and peers. At the end of a long apprenticeship, one is, moreover, conscious of how many and how varied have been one's teachers, and how patient. I have space only to acknowledge here those most closely tied to this, my last studentship.

I am deeply grateful to Irwin Kuntz, under whose wise and generous mentorship my researches have been at every turn encouraged. Tack gave me direction when I needed it, and gave me the freedom to pilot my own course, however unlikely to reach land, when the wind was in my sails.

The research in this dissertation has been supported and encouraged by the congenial community of intellects here at UCSF. I would especially like to thank the groups of Robert Stroud and Dan Santi, who have provided a critical experimental context for theory, and who have supported my efforts to test our models in the lab. Kathy Perry, who has occasionally had more of the courage of my convictions than I myself have had, deserves particular mention. My work on the theoretical aspects this dissertation has been enriched by my learned colleagues in the DOCK group, especially Elaine Meng, Dale Bodian, George Seibel and Renee DesJarlais, and more generally by the strong theory group here at UCSF. David Rossman's course on the philosophy of science helped shape my understanding of how science is best accomplished. I must also thank my readers, Professors Ken Dill, Peter Kollman and Tack Kuntz for their patience and kindness in reading and critiquing this *opus*.

In the span of time of a doctoral research program, one is supported and entertained by a group of like minded individuals, without whom the excitement of discovery would be a colourless drudgery. In San Francisco I have been lucky to find so varied a group of sharp minds and swift spirits as anyone could hope to count as his friends. For his enduring friendship and critical mind, I would like to thank John Altman. Lydia Gregoret and Chuck Wilson were daily companions at the heart of our graduate careers - our preparation for the Oral Defense. Leslie Taylor has been a constant source of delight in conversation. Jason Swedlow has been as serious and frivolous a friend as one could wish for. Inke Nathke always seemed to know when I needed a wink or a hug. Jeremy Minshul and Abby Durnberg have seen me in mid-leap and jumped. Thanks also to John Irwin, Adrian Yovanovich and Robert Hall, who are with me always.

The support of my family has always been the shield rock of my strength. Molly and Richard Shoichet remain my models for grace and courage.

This dissertation is dedicated to Dorothy and Irving Shoichet, whose love, support and interest has never failed or waned.

The dissertation is divided into four chapters, an introduction and a conclusion. Each chapter is in the form of a manuscript submitted for publication. Chapter three has been published already (Shoichet and Kuntz, 1991) and chapter two is in press (Shoichet et al., 1992). Chapter four is about to be submitted for publication. Chapter one was submitted to *Biochemistry*, but was not accepted.

While publication in scholarly journals is the appropriate way to communicate with the scientific community, collecting several such together and calling it a "Thesis" makes for choppy reading. To smooth the way, I have written glosses to each of the main chapters in which their central points are summarized and their contexts explained. Since some of the work is already a few years old, I take the opportunity to comment, in the glosses, on how the testable predictions made in the chapters have weathered with time.

Tack encouraged me to conclude the dissertation with some thoughts on "future directions". I have taken his suggestion, but I warn the reader that my acuity in these matters is weak. My own research has been lifted by such unexpected tides, that I expect the river of work in our group to promptly overleap the channel I have cut for it, and find its own course to stranger seas than I can now imagine. I have tried to limit my ruminations to techniques which are easy to implement and seem obvious to succeed, and to hypotheses which can be practically tested.

The issue of experimental falsification is one to which I return in the following chapters. Sir Peter Medawar commented that "biologists work very close to the frontier between bewilderment and understanding"(Medawar, 1969). How much greater the possibility for bewilderment and error for one who tries to predict biological systems with drastically simplified molecular models? I am convinced that to have

currency in the dynamic world of the experimental biochemist, the theoretician must commit himself not only to interesting problems, but to hypotheses which lead to practical and definitive experiments. Occasionally, he must be even be willing to descend from Parnassus and test his hypotheses himself. Chapter four is an example of the ambushes, pleasant and unpleasant, which sometimes await him in such an enterprise.

Medawar, P. B. (1969). Induction and Intuition in Scientific Thought. American Philosophical Society.

Shoichet, B., Bodian, D. L. and Kuntz, I. D., *J. Comp. Chem.* **13**, in press. (1992)

Shoichet, B. and Kuntz, I. D., *J. Mol. Biol.* **221**, 327-346 (1991)

ABSTRACT

Predicting the favourable co-positioning of two interacting molecules is a necessary first step towards predicting function or designing inhibitors based on the structures of biological receptors. In the following chapters I take up some of the theoretical and practical aspects of molecular docking, often involving the use and elaboration of the computer program DOCK.

In Chapter one, I use DOCK and molecular mechanics to predict a ternary complex of Thymidylate Synthase (TS) with its two natural substrates, dUMP and CH_2-H_4folate. My predictions are testable by site-directed mutagenesis, crystallography and enzymology.

I develop new docking algorithms in chapter two. The new algorithms improve the efficiency and accuracy of the docking method. I use molecular organization and sampling techniques to remove the exponential time dependence on molecular size in docking calculations. The new techniques allow me to study systems that were prohibitively large for the original method, including 7 complexes where the ligand is itself a protein. In all cases, the new algorithms successfully reproduces the experimentally determined configurations.

These new algorithms are used in chapter three to address "the protein docking problem." I accurately regenerate the structures of three known protein complexes, using both the bound and unbound conformations of the interacting molecules. I also find geometries that did not resemble the crystal structure. Simple complementarity methods (surface area burial, solvation free energy, packing, electrostatic interaction energy and mechanism-based filtering) can not distinguish between 'native' and 'non-

native' complexes. Energy minimization is more reliable, though the energy differences are surprisingly small. The regeneration of the crystallographic configurations using the unbound conformations suggests that DOCK will be useful in predicting the structures of unsolved complexes.

In chapter four, I screen a structural database for chemicals complementary to TS using DOCK. Besides retrieving the natural substrate and known inhibitors, I find molecules previously unknown to bind to TS. I test several of these and find two different classes of novel inhibitors. The crystallographic solution of complexes of several of these inhibitors allow us to test our models to atomic resolution. These structures suggest new ligands with improved binding.

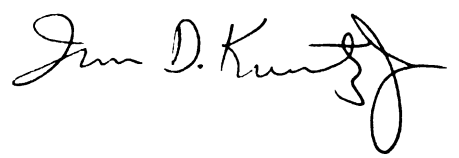


Table of Contents

List of Figures

List of Tables

Introduction

We live chemically. A muscle contracts, a hand closes, food is digested, very thought arcs across synapses because of chemical reactions. The purview of science is to ask what sort of reactions, and how?

Since the turn of the century, it has been paradigmatic that biological reactions are mediated through large molecular assemblies. Emil Fischer first proposed the theory of enzyme specificity in the 1890's (Fischer, 1894), Paul Ehrlich took the same approach with the theory of receptors for xenobiotics (Ehrlich, 1907). Since the 1920's, most enzymes and receptors have been understood to be proteins (DNA can also be a receptor, but this notion did not develop until later). The crystallization of Urease in the late 1920's followed by the X-ray diffraction experiments of Bernal and Hodgkin on pepsin in the mid-1930's showed that proteins had definite, stable structure, and were not the colloid-like substances that their name would suggest (Perutz, 1964).

The chemical bases for the association between a ligand and its cognate receptor or enzyme - what is today referred to as molecular recognition - have been broadly understood since the early 1940's. In a famous critique of Pascual Jordan's proposal that identical chromosomes would attract each other - a quantum mechanically reasoned argument that amounted to 'like attracts like'- Pauling wrote (Pauling and Delbruck, 1940):

> It is our opinion that the process of synthesis and folding of highly complex molecules in the living cell involve, in addition to covalent-bond formation, only the intermolecular interactions of van der Waals attraction and repulsion, electrostatic interactions, hydrogen-bond formation, etc., which are now rather well understood. These interaction are such as to give stability to a system of

> two molecules with *complementary* structures in juxtaposition, rather than of two molecules with necessarily identical structures; we accordingly feel that complementariness should be given primary consideration in the discussion of the specific attraction between molecules and enzymatic synthesis of molecules.
>
> A general argument regarding complementariness may be given. Attractive forces between molecules vary inversely with a power of distance, and maximum stability of a complex is achieved by bringing molecules as close together as possible, in such a way that positively charged groups are brought near to negatively charge groups....In complementary surfaces, like die and coin, and also a complementary distribution of active groups.

This 1940 communication needs little amendment today.* The complementarity argument powerfully influenced the heroic age of molecular biology: it was at the heart of Pauling's alpha-helix proposal (Pauling et al., 1951) and informed Francis Crick's "coiled coil" analysis of keratin (Crick, 1952; Crick, 1953). Proteins, of course, are described to this day in terms of Pauling's two structural elements, the alpha-helix and the beta-pleated sheet. Crick's attack on alpha-keratin structure is still used in crystallographic analysis of coiled coil motifs in proteins (O'Shea et al., 1991) and is referred to in papers on molecular docking (Connolly, 1985). The most important end to which Pauling's complementarity paradigm was put was, undoubtedly, the solution of the double helical structure of DNA by Watson and Crick (Watson and Crick, 1953). The double helix not only gave the gene a physical reality, but explained its hetero- and homo-catalytic roles, and in general revolutionized molecular biology, giving the field the form and ambit that characterizes it to this day.

Perutz and Kendrew's solution of the first protein structures, six years after Watson and Crick's first *Nature* paper, was to do for receptors what the double helix had done for the gene. The hemoglobin and myoglobin structures succeeded in giving physical

---

* Scholars today would probably also refer to the hydrophobic effect, a statistical phenomenon unrelated to complementarity.

meaning to Ehrlich's receptors, which until then had been mostly a useful mental construct. The complicated architecture that the X-ray analysis revealed, however, frustrated the sort of rich *functional* interpretation of proteins that Watson and Crick were able deduce from their DNA structure. Where the structure of DNA seduced with simplicity, the first proteins repelled with their tangled complexity, which seemed to speak to few of the questions of life. Perutz himself asked, "Could the search for ultimate truth really have revealed so hideous and visceral looking an object? Was the nugget of gold a lump of lead?" (Perutz, 1964)

Perutz's question resonates today. X-ray crystallographers have gone from strength to strength in their ability to solve protein structures, which we can agree with Perutz form the central point of attack on the molecular basis for biological function. Jane Richardson (Richardson, 1981) has shown us how to read some of the underlying structural motifs in Perutz's viscera. What remains recondite is the relevance of such structures to life. Atomic resolution structures of proteins have infrequently contributed to the understanding of a disease, and have never suggested a cure. With several notable exceptions, the structure of the HLA (Bjorkman et al., 1987) molecule being one of the best, atomic resolution protein structures have rarely, since the double helix, led to profound insight into biological function.

Molecular biology thus finds itself in an awkward position. The field that took its point of departure from the spectacular molecular reductionism of Watson, Crick *et al.*, finds that structure is only rarely informative about the question that most concerns it: what is life? While the notion that biological function can be understood by studying molecular structure remains paradigmatic, structure-function analyses infrequently guide the most active areas of research in biology today.

The essays that follow reflect, in a minor key, the some of the problems and aspirations that characterize structural, especially theoretical, biology generally. Like most of my colleagues, I still draw on the reserve of credibility for modeling that Watson, Crick and Pauling first banked 40 years ago. Likewise, I am plagued by the uncertainties of the complex structures that so dismayed Perutz and Kendrew. These uncertainties are in plainest view in my early work on modeling thymidylate synthase's reaction path (chapter one), and are still to be seen in the some of the work, and much of the style, of the essay on the protein docking problem (chapter three). In undertaking the protein docking work, I had hoped to show that molecular docking could treat recognition on the scale of protein-protein interfaces. What I found was that, though our method could regenerate the native configuration of these complex, we could not reliably distinguish between the native and non-native, even when we used some of the more sophisticated energy evaluations schemes that are current in the field. I tried one method after another, each time expecting that *this* one, for sure, will be able to distinguish amongst the complexes. Each time I was disappointed.

Under the best circumstances, continuous attention to a problem leads either to its solution, or a declaration that it is not, after all, a problem. My conclusions at the end of the protein docking work have the latter quality. What I decided was that if theoretical methods could not distinguish between the right and the wrong answer, then maybe the 'wrong' answer was not so wrong. The peculiar geometries that I suggest as plausible in chapter three (Shoichet and Kuntz, 1991) remain untested. They are, however, *testable*, and I believe, generative, even if they turn out to be incorrect. What they point to is a receptor-based logic of molecular function, a logic whose vision I had lost in the minutiae of the day-to-day modeling.

A receptor-based logic for molecular design, which I finally consciously take up in chapter four, led to unexpected results that were hard to explain because I saw them, at first, through the lens of another logic. Thirty years of drug design in thymidylate synthase has resulted in a bestiary of very similar molecules - designed after and resembling the natural substrates of the enzyme. I therefore expected to find molecules that resembled either folates or nucleotides. Instead, molecular docking in thymidylate synthase led me to some very strange beasts, not at all like the natural substrates, that do not appear to bind in either of the traditional binding pockets. These results hint that there is more in the structure of a receptor than can be suggested by its function. Perhaps this is the silver lining in Perutz's viscera, if you will excuse the confused metaphor. There is so *much* going on in a receptor, that we can expect to find ligands that it binds to that have nothing to do with the 'purpose' to which the receptor was ostensibly designed.

Slowly (but *how* slowly) then, a theme emerges from the false notes and aborted beginnings that litter the pages that follow. Looking back over five years of work, I believe that I was, all along, pursuing the logic of the receptor. If only I had known it - 'but our beginnings can never know our ends.' I hope, gentle reader, that you too will find this shadowy unity in the pages that follow.

Bjorkman, P. J., Saper, M. A., Samraoui, B., Bennett, W. S., Strominger, J. L. and Wiley, D. C., *Nature* **329**, 506-512 (1987)

Connolly, M. L., *Biopolymers* **25**, 1229-1247 (1985)

Crick, F. H. C., *Nature* **170**, 882 (1952)

Crick, F. H. C., *Acta. Crystallogr.* **6**, 689-697 (1953)

Ehrlich, P., *Chem. Berichte* **42**, 17 (1907)

Fischer, E., *Chem. Berichte* **27**, 2985-2993 (1894)

O'Shea, E. K., Klemm, J. D., Kim, P. S. and Alber, T., *Science* **254**, 539-544 (1991)

Pauling, L., Corey, R. B. and Branson, H. R., *Proc. Natl. Acad. Sci. U.S.* **37**, 205 (1951)

Pauling, L. and Delbruck, M., *Science* **92**, 77-79 (1940)

Perutz, M. F., *Scient. Am.* **211**, 64-76 (1964)

Richardson, J., *Adv. Prot. Chem.* **34**, 167-339 (1981)

Shoichet, B. and Kuntz, I. D., *J. Mol. Biol.* **221**, 327-346 (1991)

Watson, J. D. and Crick, F. H. C., *Nature* **171**, 737-738 (1953)

Gloss to Chapter I

We set out, nearly five years ago, to predict the ternary complex of thymidylate synthase (TS) with its two substrates, deoxyuridine mono-phosphate (dUMP) and methylene-tetrahydrofolate (CH_2-H_4folate). Our starting models were a 3 Å resolution crystal structure of the enzyme, several small molecule crystal structures of dUMP, and some models of CH_2-H_4folate based on NMR and crystal data. All of our starting models were of the molecules as they existed in isolation from each other. Both dUMP and CH_2-H_4folate are flexible molecules, and the enzyme was known to undergo significant conformational change on ternary complex formation, so beginning from structures of the isolated molecules was a serious handicap (see for instance the recent articles in *Science* on this subject (Jorgensen, 1991; Wuthrich et al., 1991)). Our goal was therefore quite ambitious. The theoretical tools we brought to the task, conversely, were relatively simple.

Why did we even attempt it? What could we hope to learn about TS that would be useful? And why should anyone care about TS anyway? I will briefly pause to address these questions, before continuing with the details of the work.

TS is an interesting enzyme. It makes two contributions to the life of the cell, which is a little unusual in the specialized world of enzymes, that mostly are responsible for one product only. The first chemical TS produces is thymidine monophosphate. Without thymidine, a cell eventually dies ("thymineless death"). The second chemical TS produces is dihydrofolate, an important metabolite that is used in a number of different pathways in the cell. TS makes these two chemicals by a remarkable series of reactions involving no less than six distinct steps wherein a chemical bond is formed or cleaved (Hardy et al., 1987). Studying the enzyme's

mechanism is therefore interesting and medicinally important. Sitting at the end of the only *de novo* pathway for thymidine biosynthesis, TS is a logical target for drug design (Hardy et al., 1987). If one could predict the structural aspects of how TS interacted with its substrates, then one would go a long way to understanding the molecular basis for the enzyme mechanism, and at the same time set the stage for designing molecules that could derail the enzyme from its normal path. So goes the argument.

I came into the lab wishing to do "structure-based inhibitor design." One obviously needs a structure. In mid-1986, Janet Finer-Moore, Larry Hardy, Bob Stroud and his colleagues at UCSF had solved the structure of TS, a well respected medicinal target, so this was a good place for me to begin my researches. Modeling the ternary complex seemed to Tack and me a good way for me to learn the 'in's and out's' of a complicated system.

Early on, I came to understand that modeling the catalytic pathway would be difficult and prone to error. I persevered because for all the problems I was having, I was learning a lot about modeling and about the enzyme. Also, I reasoned that if I could hew to the line of testable predictions, based on the modeling, then as complicated as things were, and as wrong as I was likely to be, the work would at least be relevant, by proposing doable experiments, and would perhaps focus debate.

As I describe below, we had some measure of both success and failure in our predictions. The details of what we did correctly and where we made our mistakes will interest, I hope, the specialist, either modelers or TS heads. If I have a general criticism of the work, however, it is not that we were wrong in such or so specific prediction - grievous though such errors were. The greatest problem with this effort was that I failed to design the proposals so that a negative result was interpretable. I

was so intent on getting things right, that I didn't think through what it would mean if I was wrong. It has taken me a long time to understand that the best science is that where any result, positive or negative, improves one's understanding of a system (see chapter four). For if it is true, as Medawar claims, that "science has been incomparably the most successful enterprise human beings have engaged upon," (Medawar and Medawar, 1983) than surely this is not because scientists are so often correct, but rather because they know what to do when they are wrong, which, if common experience in any way reflects my own, is most of the time.

Back to the modeling.

Many of the molecular properties and interactions that this paper sought to predict have since been determined experimentally by Bill Montfort, Janet Finer-Moore and colleagues, working in the lab of Bob Stroud (Finer-Moore et al., 1990; Montfort et al., 1990), as well as by the Agouron group (Matthews et al., 1990). We can therefore evaluate how well we did in this predictive paper.

Our ultimate goal was to predict, in atomic detail, the configuration of dUMP, CH_2-H_4folate and TS. The experimental structures from both X-ray groups are inconsistent with our final predictions; our predicted folate geometries were wrong. In the crystal structure ternary complexes, the folate/folate-analogues have their pterin rings above and parallel to the plane of the pyrimidine ring, with the paba-moiety bent back towards the 'top' of the site. While one of our four suggested complexes resembles this structure in the placement of the pterin relative to the pyrmidine (the C6R C5S, figure IVd. in chapter I), many of the important details of the folate interactions are incorrect, especially regarding the placement of the paba, which is wholly wrong. Mostly this owes, I believe, to a difference in conformation between the CH_2-H_4folate

found in the crystal structures and that used in our modeling. The crystal structures show CH_2-H_4folate (or in the case of the Stroud group papers, the ICI folate analogue CB3717) after it has committed to catalysis by undergoing a ring opening reaction, while our modelling was done with models of CH_2-H_4folate before its imidazolidine ring had opened. To a lessor degree, the different final states also reflect what Bill Montfort and Eric Fauman have called the "segmental accommodation" of TS to its substrates. This effect was, however, only a small impediment to the modelling. Like most enzymes, the active site of TS is largely established in the absence of substrates; conformational change acts more to 'lock' substrates in place than to form a cognate binding pocket. Probably more important was the existence of a two residue frame shift error in the last third of the enzyme (thus, Y261 was modeled into density that belonged to H259, and so on). This error existed in all of the models that the Stroud group was then using, and was only discovered in 1989 by Bill Montfort, after this work had been completed. Given this error, and the differences between our model of CH_2-H_4folate and that found crystallographically, it's slightly amazing to me that so many of the predictions in this manuscript were as correct as they were.

For, overall, our analyses were fairly accurate. The most important predictions, after that of the ternary geometry, are summarized in Tables I and II of chapter one. Comparing Table I with Table I in Finer-Moore *et al* and Tables 3, 4, 6, and 7 in Matthews *et al.*, one notes that of the 15 residues that we flagged as contributing directly to substrate recognition, only four (L144, E84, F64 and Y233) were not similarly flagged by the crystallographers. Moreover, it is unclear to me why one of these four, Y233, was left off of the list of residues defining the "alternate" folate binding site (Montfort et al., 1990). On the other hand, of the 22 residues in Table I of Finer-Moore *et al.*, eight cannot be found in our own Table I. Of the six residues in Table II, whose putative roles we felt were particularly worthy of experimental

testing, five have turned out to have at least one of the two specific interactions we proposed. Even the sixth residue, D221, interacts with the pterin moiety in a manner that resembles our predictions: rather than interacting directly with the exo-cyclic amino group of CH_2-H_4folate, it hydrogen-bonds with the vicinal pyridinal nitrogen and speaks to the exo-cyclic nitrogen through a bound water.

At least two of our predictions took issue with what was then the accepted model for TS's mechanism. The first structure paper (Hardy et al., 1987) had implicated H199 as a necessary general base/acid in the reaction pathway. Our modeling led us to predict that this was not the case, and mutagenesis experiments have since falsified this role for H199 (Frasca et al., 1988), which is no longer considered important for catalysis (Finer-Moore et al., 1990). The very ability of DOCK to model dUMP into the TS site ran counter to the crystallographer's interpretation of their structure. At the beginning of the project our colleagues in the Stroud group warned that they had not been able to find an acceptable configuration for dUMP in the site, and had hypothesized that the conformation of the unbound enzyme did not allow dUMP to bind (Hardy et al., 1987). We found, however, that we could dock dUMP into TS without much trouble. We did not emphasize this finding in the manuscript that follows because it was difficult to draw a useful conclusion from it. Just last year, however, Janet Finer-Moore discovered that the structure with which we began our modeling, solved as a substrate-free structure, in fact contained a bound dUMP molecule in its active site. The structure that supposedly could not accommodate a molecule of dUMP without conformational adjustment had had tightly bound to it just such a molecule all the time. This is an amusing example of "just because you can't see it, doesn't mean that it isn't really there." The converse is also true.

I can offer three reasons for why our predictions turned out to be as accurate as they

did, given the errors in our models. First, most of our predictions concerned residues that interacted with dUMP, which we were able to more highly constrain than the folate. Second, DOCK is pretty good at topographically mapping a region, and there are usually a small number of residues in a site that are likely to be consistent with ligand binding. Third, our commitment to testable predictions was a discipline that forced me to talk a great deal with the enzymologists and crystallographers. This gave me access to a great mine of knowledge and saved me from much foolishness.

Finer-Moore, J. S., Montfort, W. R. and Stroud, R. M., *Biochem.* **29**, 6977-6986 (1990)

Frasca, V., LapPat-Polasko, L., Maley, G. F. and Maley, F., 149 (1988)

Hardy, L. W., Finer-Moore, J. S., Montfort, W. R., Jones, M. O., Santi, D. V. and Stroud, R. M., *Science* **235**, 448-455 (1987)

Jorgensen, W. L., *Science* **254**, 954-955 (1991)

Matthews, D. A., Appelt, K., Oatley, S. J. and NG, H. X., *J. Mol. Biol.* **214**, 923-936 (1990)

Medawar, P. B. and Medawar, J. S. (1983). Aristotle to Zoos. Cambridge, Harvard University Press.

Montfort, W. R., Perry, K. M., Fauman, E. B., Finer-Moore, J. S., Maley, G. F., Hardy, L., Maley, F. and Stroud, R. M., *Biochem.* **29**, 6964-6976 (1990)

Wuthrich, K., Freyberg, B. v., Weber, C., Wider, G., Traber, R., Widmer, H. and Braun, W., *Science* **254**, 953-954 (1991)

Shape Complementarity and Molecular Recognition: Predicting the Structure of the Thymidylate SyntaseTernary Complex†

Brian K. Shoichet, Renee L. Desjarlais and Irwin D. Kuntz*

Department of Pharmaceutical Chemistry

University of California

San Francisco, California 94143-0446

Running Title: TS Ternary Complex Prediction.

† This work was supported in part by grants from the National Institutes of Health (GM19267C and GM39552) and from the Defense Advanced Research Projects Agency (N0014-86-K-0757). I.D.K was partially supported by a National Institutes of Health grant (GM31497).

[1]Abbreviations used in this paper: TS (Thymidylate Synthase), H4-folate (methylenetetrahydrofolate), dUMP (deoxyuridine monophosphate), PABA (para-amino benzoic amide), SG (gamma sulphur of cysteine).

[2] The coordinates for the native TS structure were graciously provided to us by Dr. Robert Stroud, Dept. of Biochemistry and Biophysics, University of California, San Francisco.

[3] Residues from the other monomer in the *L. casei* TS dimer are denoted with primes.

[4] Dr. Janet Finer-Moore, UCSF Dept. of Biochemistry/Biophysics, personal communication.

[5] HND is the hydrogen of the delta nitrogen of His, OP is a phosphate oxygen.

[6] Our work was done with the 3 Å crystal structure coordinates (Hardy *et al.*, 1987). Crystallographic work with this enzyme has continued; the resolution has recently improved to 2.3 Å.

ABSTRACT

Predicting the favourable co-positioning of two interacting molecules remains a difficult problem in chemistry and biochemistry. We present an approach for addressing this question in the case of predicting the ternary complex of *L. casei* Thymidylate Synthase (TS) with its two natural substrates, deoxyuridine monophosphate (dUMP) and -methylenetetrahydrofolate (CH_2-H4folate). We began with the uncomplexed structure of each molecule. Using the docking method of Kuntz (Kuntz *et al.*, 1982) we generated starting configurations of the substrates in the enzyme. These starting structures were energy-minimized using the molecular mechanics program AMBER (Weiner *et al.*, 1984). Using this process we generated in turn non-covalent binary complexes between TS and dUMP, TS-dUMP covalent binary complexes and finally TS-dUMP-CH_2-H4folate partially covalent ternary complexes involving a covalent bond between TS and dUMP and non-covalent interactions between the enzyme and the CH_2-H4folate. We have generated four possible ternary complex structures. All four structures have a covalent bond between the SG of Cys 198 of TS and the C6 of dUMP, with the phosphate of dUMP minimally tethered by Arg 218 and Arg 179'. The dihydro-pyrimidine is in a half-chair conformation, with the SG and the CH_2-H4folate in a trans-diaxial relationship. Two possible stereochemistries are predicted for the dUMP in the fully covalent ternary complex: either C6 'R' C5 'S' or C6 'S' C5 'R'. In the former, the CH_2-H4folate lies parallel to the major axis of the dUMP in either of two possible orientations, while in the latter the CH_2-H4folate lies orthogonal to the major axis of the dUMP, again in either of two possible orientations. The modeling suggests specific roles for certain TS residues in the binding of the substrates. Our predictions are testable by site-directed mutagenesis, crystallography and enzymology.

Introduction

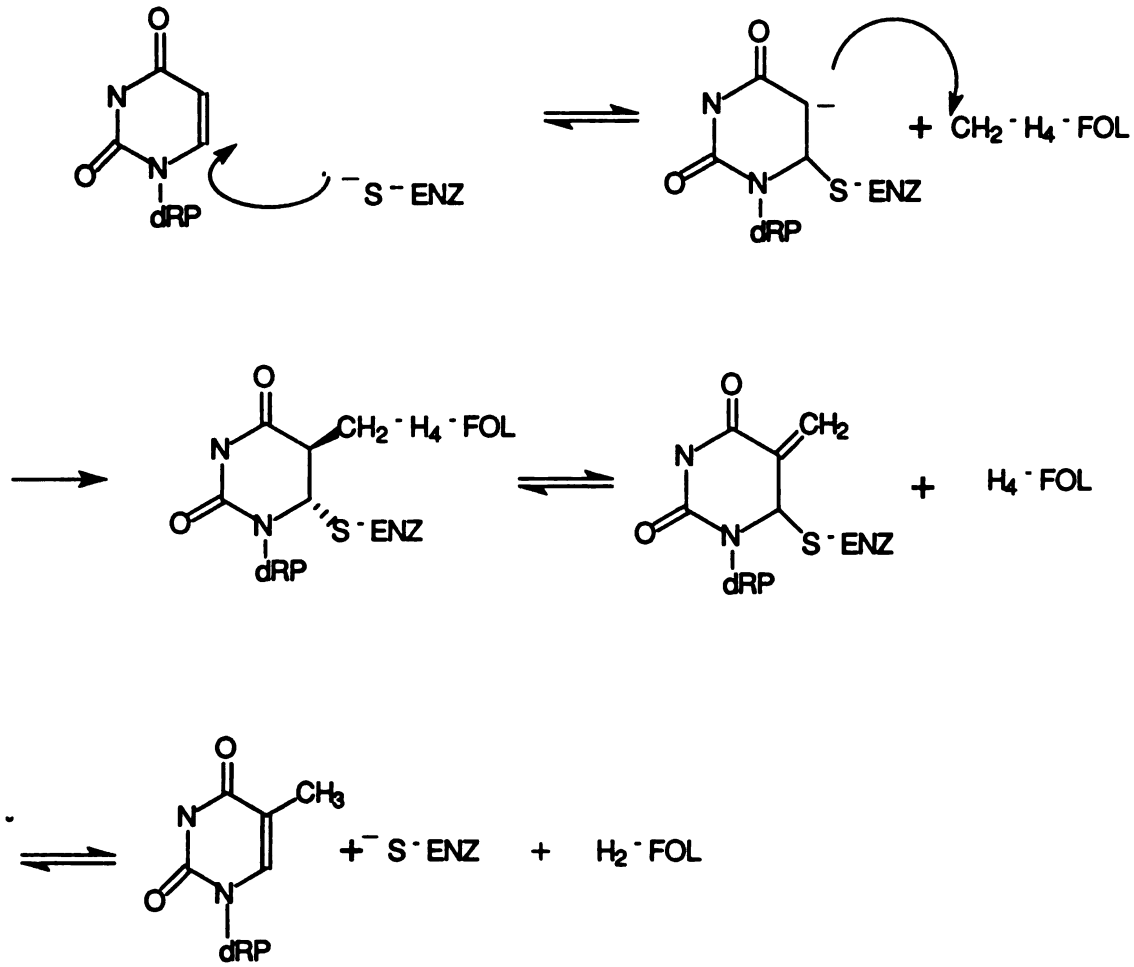
One of the more difficult issues confronting chemists and biochemists is that of molecular recognition - what determines the configuration of two molecules so that they interact most favourably with one another. The essence of the problem is that molecules of any complexity generally have large numbers of internal degrees of freedom, complicated potential energy surfaces and many conformational minima (Kuntz et. al., 1982; Paine and Scheraga, 1985; Vasquez and Scheraga, 1985; Paine and Scheraga, 1986). When the molecules in question are biological polymers the numbers of internal degrees of freedom are so large that a comprehensive search of the conformational space of a ligand-macromolecule interaction cannot be done. The problem of predicting how such polymers interact with ligands or other macromolecules has therefore resisted a general solution.

While progress in structural biochemistry has increased our understanding of the intermolecular forces responsible for organizing macromolecule-ligand systems (Bolin *et al.*, 1982; Filman *et al.*,1982; Goodford, 1984), it remains difficult to use such knowledge predictively in situations where the interaction is not already geometrically well-defined. When attempts have been made to apply this knowledge of intermolecular forces to the problem of molecular docking they have almost always occurred in the context of active sites whose relationship to the putative ligands has been largely established by crystallography (Goodford, 1985; Cody, 1986). Using techniques such as interactive graphics (Hansch *et al.*, 1982), electrostatic grid searches (Pattabiraman *et al.*, 1985) and free energy calculations (Wong & McCammon, 1986; Bash *et al.*, 1987) some progress has been made in predicting how small modifications to a ligand in an existing configuration will affect its interactions with the protein. While it is certainly possible to model these sorts of interactions, however, these efforts distract attention from the more general and admittedly less

tractable issue of predicting how a given ligand will bind to a macromolecule when a structural model for such binding is unavailable.

Some recent attempts to address the general problem of molecular recognition have drawn back from energy calculations, with their attendant sophistications and restrictions, towards a simpler statement of the problem. Both Lesk (Lesk, 1986) and Kuntz and co-workers (Kuntz *et al.*, 1982) have proposed that predicting how ligands will interact with macromolecules is, to a first approximation, a question of understanding how a small, mostly rigid, 'convex' object can be fit into a large, mostly rigid, 'concave' one. Such a formulation returns to a fairly simple geometrical interpretation of the seminal "lock and key" ideas of Fischer (Fischer, 1894). The great advantage of such an approach is that it reduces the degrees of freedom of the system from many thousand to six - the three rotational and three translational degrees of freedom that define the movements of rigid bodies - thus allowing one to investigate much more of orientation space than would otherwise be possible. Connolly has developed a similar approach for the docking of macromolecules (Connolly, 1986). The disadvantages of such a procedure lie in its potential insensitivity to subtle chemistry as well as the difficulties inherent in any static approach to modeling the results of dynamic events. Induced fit mechanisms, for instance, might well be overlooked by this methodology. These potential problems notwithstanding, progress has been reported in applying these methods to the design of new lead compounds for drug design (Desjarlais *et al.*, 1988; Sheridan & Venkataragthavan, 1987) and the reproduction of known ligand orientations and conformations in active sites from a knowledge of the structures of the independent molecules (Desjarlais *et al.*, 1986).

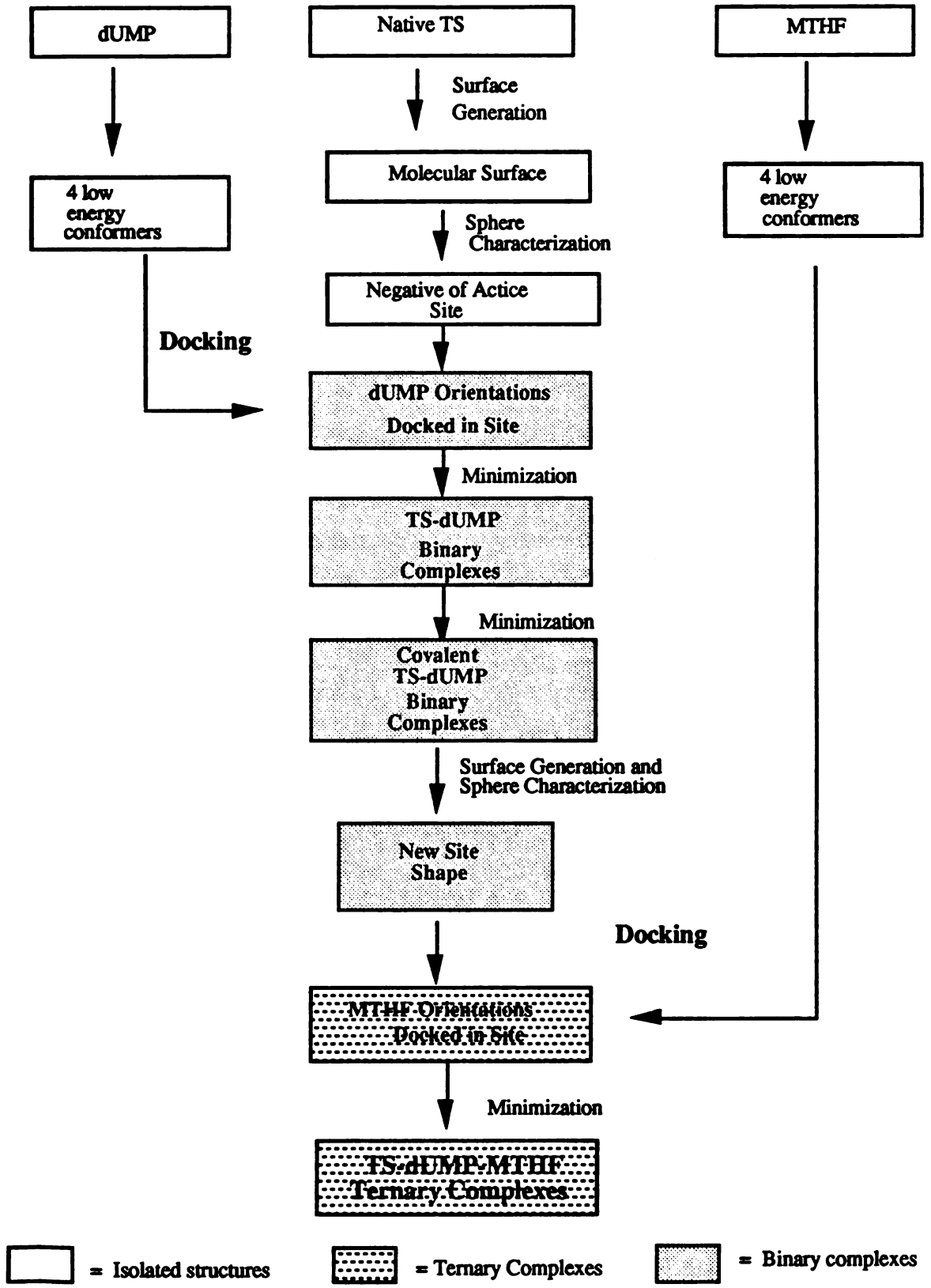
In this paper we report an extension of the docking method of Kuntz *et al.*, used in conjunction with the molecular mechanics package AMBER (Weiner and Kollman, 1981; Weiner *et al.*, 1984) and the interactive molecular graphics program MIDAS (Huang *et al.*,1983) to the prediction of a set of possible structures for the ternary complex of *L. casei* Thymidylate Synthase (TS)[1] with its two natural substrates, deoxyuridine monophosphate (dUMP) and methylenetetrahydrofolate (CH_2-H4folate). TS is a critical enzyme in the bio-synthesis of thymidylate for DNA synthesis and has been the target of drug design efforts for 30 years (Santi and Danenberg, 1984). The actual structure of the TS ternary complex was unknown to us when this work was done and remains so at the time of this writing. The predictions were made based on a knowledge of the structures of the independent ligands and of the uncomplexed 3.0 Å crystal structure of *L. casei* TS as reported by Hardy and colleagues (Hardy *et al.*, 1987).[2] As the mechanism by which TS catalyses the methylation of dUMP by CH_2-H4folate to produce dTMP and dihydrofolate is complex (see scheme I), an attempt to predict the TS ternary complex provides a difficult test of the method. Our results are formulated as hypotheses that are testable by experiments involving site-directed mutagenesis, enzymology and crystallography. These experiments will provide a direct indication of the usefulness and limitations of this extension of the docking method to the prediction of the results of dynamic chemical events on the basis of individual static structures.



Scheme I: Proposed TS mechanism used for the modeling (Hardy *et al.*, 1987).

Methods

The steps we used to arrive at the final, partially covalent ternary complexes from the uncomplexed molecules (i.e. TS, dUMP and CH_2-H4folate) followed the sequence of physical binding events as they occur in one of the proposed mechanisms for this reaction (scheme I). Starting orientations of dUMP in the active site of the protein were generated using the program DOCK (Desjarlais *et al.*, 1988). These orientations were scored on the basis of geometrical fit and basic chemical criteria. The molecular mechanics module of AMBER (Weiner and Kollman, 1981; Weiner *et al.*, 1984) was then used to refine the best of the dUMP orientations, yielding non-covalent binary complexes that were complementary to the detailed chemistry of the TS active site. The most reasonable of the non-covalent binary complexes were further energy-minimized using constraints to produce covalent binary complexes via Michael addition of Cys 198 to the dUMP. The dUMP-TS binary complexes were used to form new molecular surfaces and the entire docking process was repeated to find new possible locations for CH_2-H4folate. High scoring orientations of CH_2-H4folate were minimized in the TS active site containing the modeled Michael adduct of Cys 198 and dUMP to produce a set of possible "partially covalent" ternary complexes. The approach is shown schematically in scheme II. A more detailed explanation of each of the steps follows.



Scheme II: Modeling Strategy

DOCKING: The essence of the docking approach is the geometrical fitting of rigid bodies based on shape. To produce a geometric object with the shape of the active site requires two steps: first the molecular surface of the entire protein is generated; second, any invaginations in the surface are filled with a set of spheres which constitute a negative image of the original site. The negative image of the site is then matched against the positive image of the ligand on the basis of shape complementarity. The details of this algorithm have been described previously (Kuntz *et al.*, 1982; Desjarlais *et al.*, 1988).

The active site of the protein is usually associated with the largest cluster of spheres, which generally describe the largest or most well defined invagination in the protein. The largest cluster of spheres in TS was composed of 245 spheres that defined a shallow elongated well. This well contained all of the residues described as being in the active site in the crystal structure paper on TS by Hardy and co-workers (Hardy *et al.*, 1987). These included Cys 198, His 199, Arg 218, Arg 179',[3] Tyr 261 and Tyr 146, as well as other residues such as Lys 58 that were thought to be important to catalysis but were not included in the active site described by Hardy *et al.*

Following earlier work (Desjarlais *et al.*, 1986), we represented the shape of the substrates directly from atomic coordinates. Since both dUMP and CH_2-H4folate have bonds around which rotation is energetically feasible and since the TS-bound geometry of neither molecule is known, it was decided to represent each molecule in a number of different low energy conformations. Each conformer of the two substrates was used in independent docking runs. Thus a set of conformers of dUMP and a set of conformers of CH_2-H4folate which were energetically reasonable and yet

geometrically disimilar were used in the docking runs rather than only that geometry which corresponded to the lowest energy conformer for the particular molecule.

Four conformers of dUMP were used in the docking runs and subsequent minimizations (figure I). Both the syn and anti conformers of dUMP were examined. Moreover, each of the syn and anti conformers was represented as two rotomers about the C4'-C5' part of the phosphate tail of dUMP in either the *gauche-minus* or *gauche-plus* conformation, which gave the molecule either an extended or a compact shape, respectively. The atomic coordinates of dUMP in the anti, *gauche-minus* conformation (figure I, a.) were obtained from crystal structure data.[4] The coordinates of dUMP in the anti, *gauche-plus* conformation (figure I, b.) were obtained using the structure COBFOO (Aoki and Saenger, 1984) from the Cambridge Crystallographic data base (Allen *et al.*, 1973). The two syn conformers (II c., d.) were created from their respective *gauche-minus* and *gauche-plus* anti conformers by

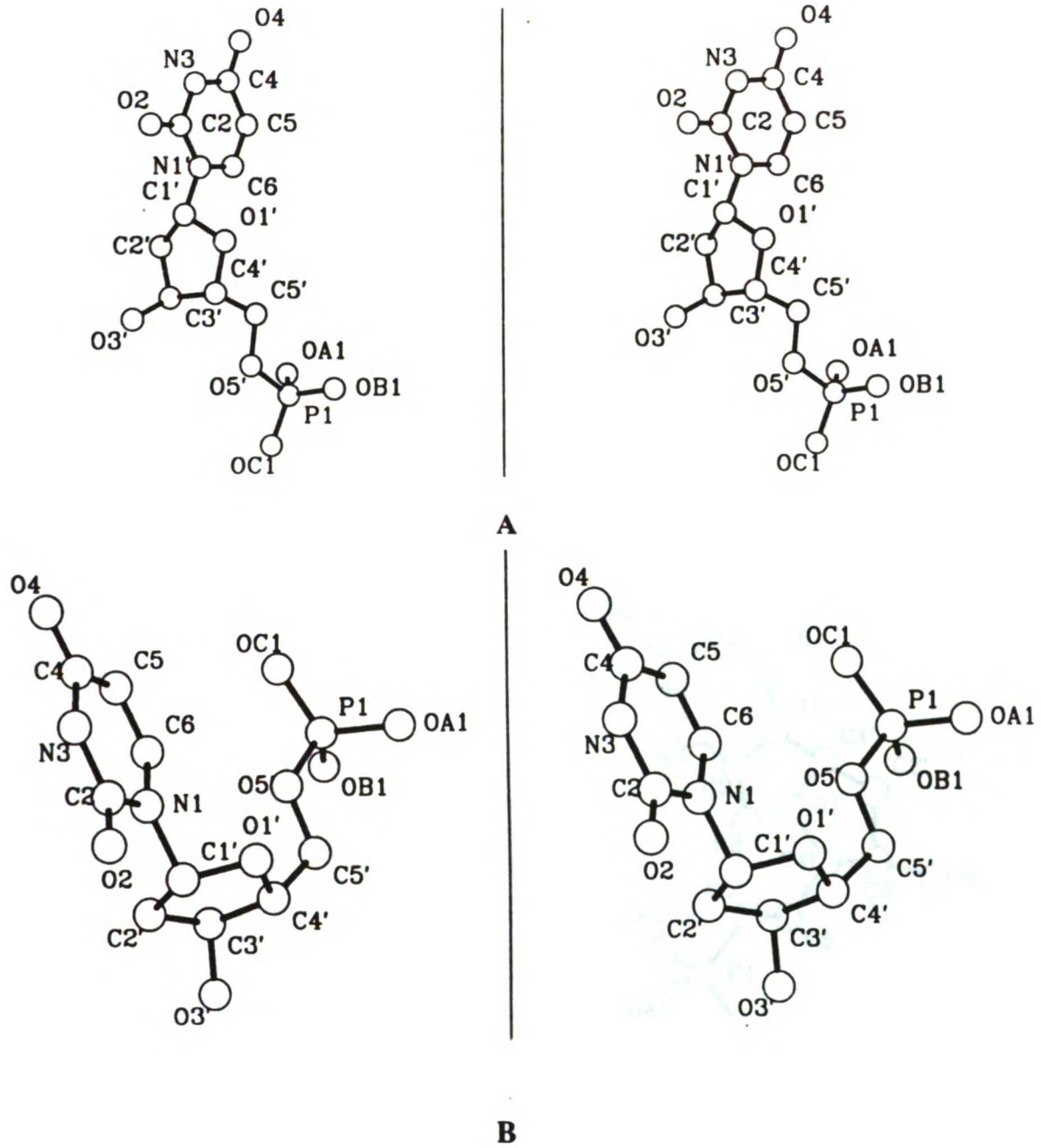
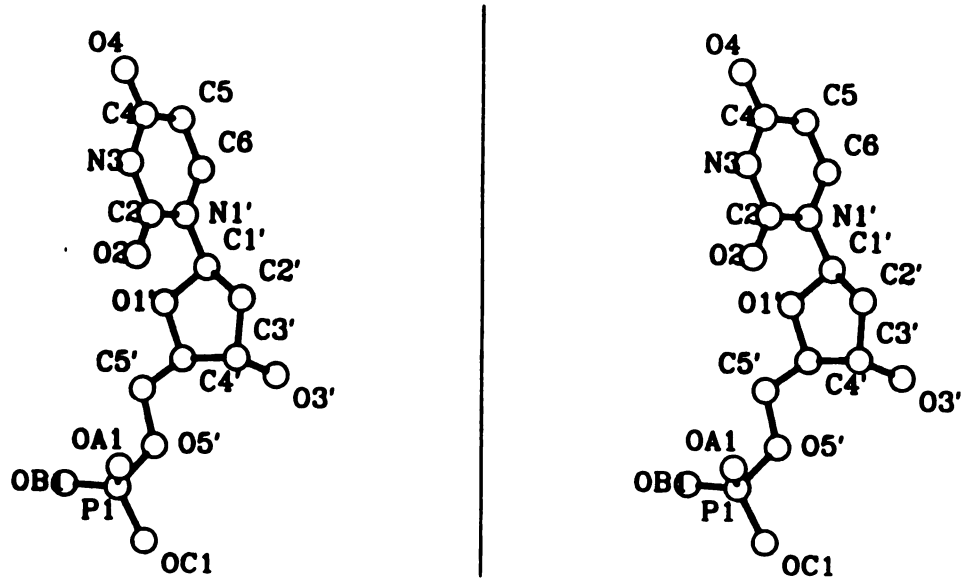
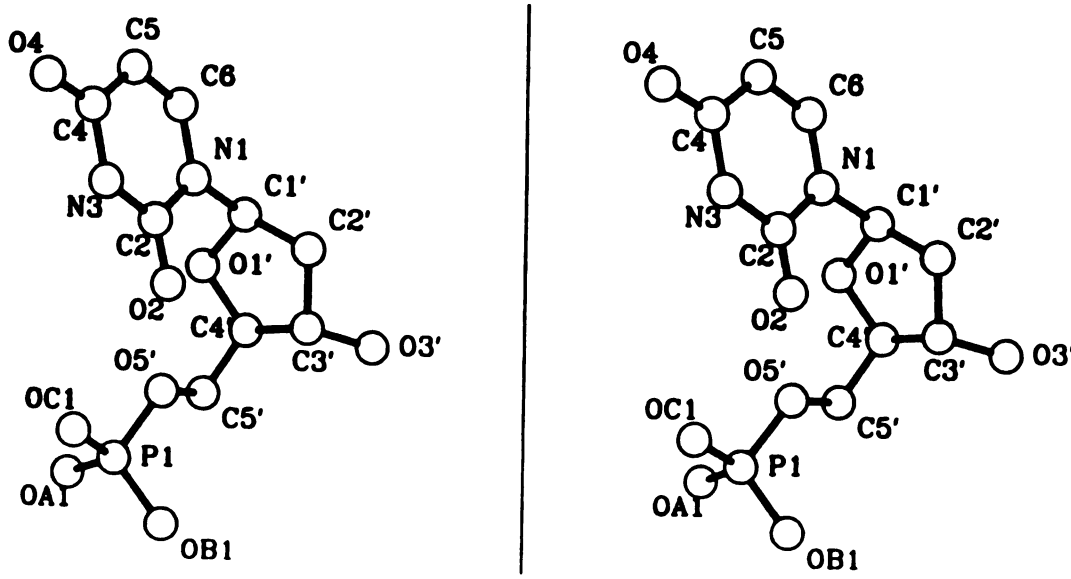


Figure I: Stereo views of the four different conformers of dUMP that were used in the docking runs: A. 'anti', *gauche-minus* conformation, B. 'anti', *gauche-plus* conformation.



C



D

Figure I: Stereo views of the four different conformers of dUMP that were used in the docking runs: C. 'syn', *gauche-minus* conformation and D. 'syn', *guache-plus* conformation.

rotating the pyrimidine by 180 degrees around the N1 - C1' bond using a graphics based editor.

Four conformations of CH_2-H4folate were used in the docking runs (figure II a., b., c. and d.). Three of these were based on the published solution structure of CH_2-H4folate as determined by NMR (Poe *et al.*, 1979; Slieker and Benkovic, 1984). These three structures differed in the conformation of the PABA phenyl ring which is attached to the rest of the molecule at N10. One conformation had the phenyl ring co-planer with the imidazolidine ring (figure II a.), another had it canted at 45 degrees to the imidazolidine (figure II b.) and the third had the two rings staggered at 90 degrees to one another (figure II c.). Coordinates of these structures were generated using the interactive graphics programs INSIGHT (Biosym Technologies) MIDAS (Huang *et al.*, 1983) and FRODO (Jones, 1982). The fourth conformation of CH_2-H4folate, (figure II d), was generated with the semi-empirical quantum mechanics program AM1 of Dewar (Dewar *et al.*, 1985). This structure closely resembled the structures of the model-built folates, except that the imidazolidine ring was puckered slightly causing the PABA moiety of CH_2-H4folate to fold slightly back towards the pteridine system. In the NMR-based structures the imidazolidine ring was puckered in the other direction causing the PABA to extend away from the pteridine ring system. Structure III d. was used only for docking purposes and was not included in the AMBER runs. The poly-glutamate tail of CH_2-H4folate, which is thought to be important for substrate binding to TS *in vivo*, was not included in any of the folates used in our modeling owing to its large conformational flexibility.

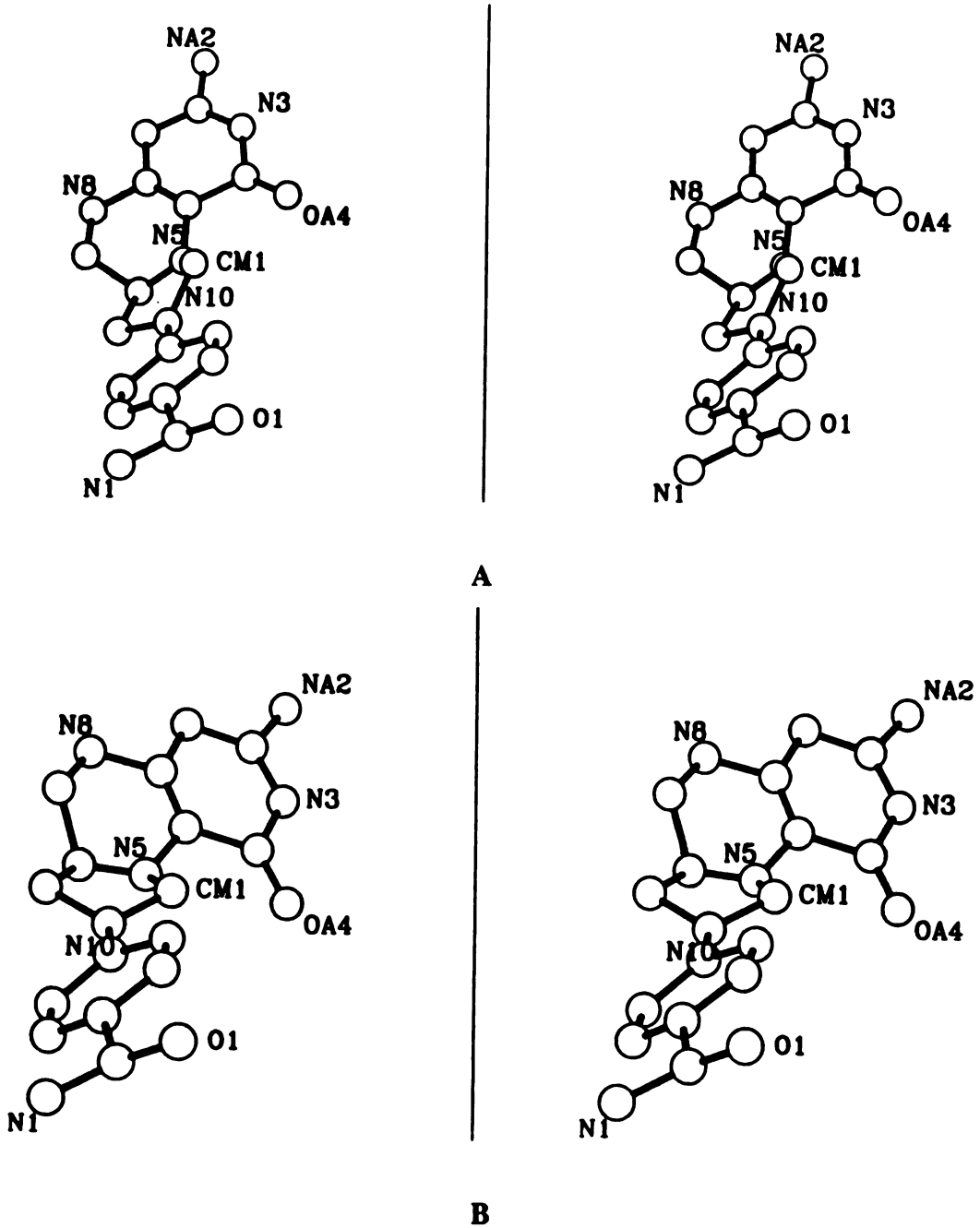


Figure II: Stereo views of the four different conformers of CH_2-H4folate that were used in the docking runs: A. plane of the PABA is at 0 degrees to the plane of the imidazolidine ring, B. plane of the PABA is at 45 degrees to the plane of the imidazolidine ring.

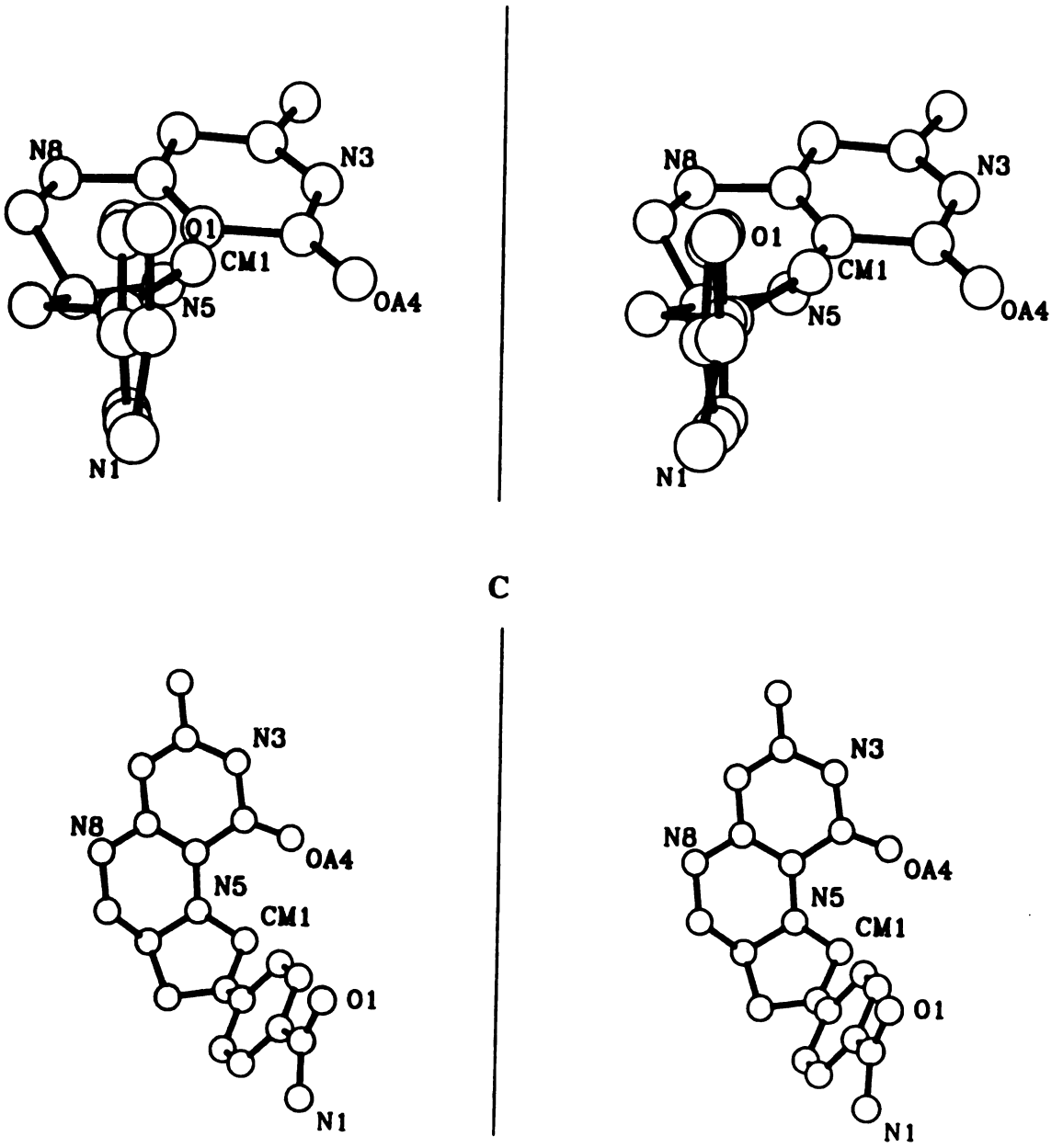


Figure II: Stereo views of the four different conformers of CH_2-H4folate that were used in the docking runs: C. plane of the PABA is at 90 degrees to the plane of the imidazolidine ring and D. conformation of CH_2-H4folate generated by the program AM1.

The conformations of dUMP were independently docked into the negative of the active site groove of the native enzyme structure (Hardy *et al.*, 1987). Several thousand orientations were generated for each conformation. Because of the size and diversity of the site defined by the spheres it was necessary to filter the orientations which emerged according to some elementary chemical criteria known on the basis of crystallography (Hardy *et al.*, 1987) and enzymology (Santi & Danenberg, 1984) In the crystal structure of the native enzyme, phosphate density from buffer was observed 5.0 Å and 4.0 Å from the guanidinium carbons Arg 218 and Arg 179', respectively. Only those orientations which put the phosphate of dUMP within 6 Å of the guanidinium carbon of Arg 218 and Arg 179' and put the C6 of dUMP within 5 Å of the catalytic thiol were accepted. The CH_2-H4folate conformers were independently docked into the negative image of the binary covalent complex of dUMP-TS (see below). Once again, several thousand orientations of CH_2-H4folate in TS resulted; these were filtered according to whether the C11 of the folate was within 5 Å of the C5 of dUMP. The filtered orientations of both dUMP and CH_2-H4folate were scored on the basis of shape complementarity to the active site of the protein.

Molecular Mechanics and Graphics: The energies of the various high-scoring orientations of dUMP in TS were minimized with the molecular mechanics module MIN of AMBER (Weiner and Kollman, 1981; Weiner *et al.*, 1984) using a combination of steepest descent and conjugate gradient methods. Minimizations were continued until an RMS gradient change of less than 0.1 between minimization steps was achieved. This took approximately 1800 cycles for the molecular coordinates of TS, using all residues which were within 5 Å of the active site and the docked dUMP - approx 2000 atoms. Energy-minimization of the binary and ternary complexes starting from the equilibrated protein structure took considerably less time: from approximately

600 cycles for non-covalent CH_2-H_4folate minimizations to 900 cycles for forming the covalent binary complex. A distance-dependent dielectric was used in all runs to approximate the effects of solvent (Weiner *et al.*, 1984). The partial charges and coordinates file for the dUMP was created by combining the uracil portion of uridine mono-phosphate with the ribose portion of thymidine monophosphate and the standard AMBER phosphate. We changed the charges on the phosphate oxygens such that the overall charge on the molecule was set to -2.

The results of the non-covalent binary complex minimization runs were judged on the basis of the energies of the final complexes and whether the Cys 198 SG - dUMP C5-C6 angles would allow bond formation based on orbital overlap requirements (Houk *et al.*, 1986). Only if the SG-C6-C5 angle and the C4-C5-C6-SG dihedral were reasonable (i.e. over 90 degrees) was the geometry of a particular complex accepted. The next step was to form the Michael adduct between Cys 198 and the dUMP across the C6-C5 double bond.

The charges and coordinates file for the Michael adduct of dUMP bonded to Cys 198 was adapted from that of methionine and the previously created dUMP. The atom types of C5 and C6 were changed from an aromatic character to an sp^3 character (from CA to CT atom types in AMBER) reflecting the reduction of the double bond. The partial atomic charges were left largely unchanged except that C5, C6 and their associated hydrogens were given charges of 0.0. It was also necessary to allow for the new torsion and angular relationships in the new molecule. These were created on the basis of small molecule crystal structures and standard AMBER energy parameters (see appendix I).

Most of the resulting binary complexes had the dihydropyrimidine in a 'half-chair' conformation, with the sulphur axial to the ring. The few complexes which did not achieve a 'half-chair' were discarded at this point. Parenthetically, the structures which were not kept were also those with the highest relative energies.

The final stage in the modeling involved predicting CH_2-H_4folate configurations in the active site of the TS-dUMP covalent binary complex. Using the coordinates of the acceptable covalent binary complexes, new molecular surfaces were generated, redefining the shape of the active site. Using this new active site shape, new docking runs were undertaken using the CH_2-H_4folate structures a., b. and d. from figure II., as described above. The highscoring CH_2-H_4folate orientations were then minimized with each of four representative binary complex structures. In order to do so, AMBER parameters had to be derived for the CH_2-H_4folate. Angular, torsional and bond distance parameters were adapted, when possible, from the nucleic and amino acid parameter set. When this was not possible, these values were taken from crystallographic studies of analogues of CH_2-H_4folate such as methotrexate (Bolin *et al.*, 1982) and methenyl tetra-hydrofolate (Fontecilla-Camps *et al.*, 1979) as well as from quantum mechanical studies of pyrimidines and folate analogues conducted by Gready (Gready, 1984; 1985; 1986). Charges were derived for this molecule using the Gaussian 80 UCSF program with ESP smoothing (Singh & Kollman, 1984). Unlike methotrexate, CH_2-H_4folate is expected to be a neutral molecule at physiological pH (Kallen and Jenks, 1966; Cody, 1986) and was so modeled. A single point calculation was undertaken using a fixed geometry and the STO-3G basis set. A full description of the additions to the AMBER parameter set used in the CH_2-H_4folate runs as well as the atom types and charges used may be found in appendix I.

Results

DOCKING: From the docking runs we chose four starting orientations of dUMP and six orientations of CH_2-H4folate in the TS active site. Of the four beginning dUMP orientations, the anti, *gauche-minus* conformation (figure I a.) scored the highest in the DOCK program, indicating that its shape corresponded most closely to the active site topography around Cys 198. Structure I a. was represented by two orientations, while the anti, *gauche-plus* conformation (I b.) and the syn, *gauche-plus* (I d.) were represented by one orientation each. The orientations of the syn, *gauche-minus* conformation (I c.) did not score well and were not included. The anti conformers scored better in all respects than the syn conformers. The I a. orientations scored better than the I b. orientation on the basis of best fit and slightly less well than I b. by bond formation criteria. All three model-built conformations of CH_2-H4folate were represented in the six highest scoring CH_2-H4folate orientations. The conformation of CH_2-H4folate from AM1 (II d.) did not score as well as the three NMR (Poe *et al.*, 1979; Slieker & Benkovic, 1984) based CH_2-H4folate structures did. Nevertheless, the high scoring orientations of II d. reproduced the same pattern of configurations as those of the model-built structures, relative to the enzyme.

The ternary complexes which emerged from the minimization runs may be placed into four groups of internally similar structures; representatives of each of these groups are briefly described in Table I. It should be noted that while the energies reported are useful guides to chemical reasonableness, they are prone to error and should be interpreted with caution. The major source of uncertainty in the energy values from AMBER derives from difficulties in getting good partial charges for the CH_2-H4folate. The charges produced by Gaussian 80 UCSF seem to be very sensitive to the conformation of the molecule - the AM1 conformation (II d.) had significant deviations in its charge profile from the model-built structures, even though they were quite

Table I
Ternary Complex Structures

COMPLEX	ENERGY[a] (kcal/mol)	IMPORTANT[b] RESIDUES	BIND[c]	COMMENTS
IV b.	-61 (-36)	R218, R179', H259 Y261 C198, H199@NH D221 W82 L144@O	PHOS 3'-OH O4 NA2 PABA NH (PABA)	C4-C5-C11 = 84 N5-C11-C5 = 86 C6 'S', C5 'R' PABA points away from lys 50, 51 CH_2-H4folate orthog. to dUMP
IV a.	-48 (-30)	R218, R179', H259 Y261 C198, H199@NH D221 E84 W82	PHOS 3'-OH O4 PABA HN8 PTERIN	C4-C5-C11 = 64 C6 'S', C5 'R' PABA points towards lys 50, 51 CH_2-H4folate orthog. to dUMP
IV c.	-49 (-27)	R218, R179' H259, S219,Y261 Y146 (?) W82, F228	PHOS 3'-OH O4 PTERIN	C4-C5-C11 = 60 N5-C11-C5 = 129 C6 'R', C5 'S' PABA point towards arg 218, 179' CH_2-H4folate parallel to dUMP
IV d.	-45 (-34)	R218, R179' H259, S219,Y261 Y146 (?) W82 F64,Y146 Y233	PHOS 3'-OH O4 IMIDAZOL[d] PABA NH (PABA)	C4-C5-C11 = 84 N5-C11-C5 = 89 C6 'R', C6 'S' dUMP 'syn' PABA points away from arg 218,179' CH_2-H4folate parallel to dUMP

a. The first energy values for each complex are the AMBER interaction energies. The values within the parantheses are the energy values for the complex not including the electrostatic term, which is suspect due to uncertainties in the partial charges attributed to CH_2-H4folate. All energies are approximate due to uncertainties in the CH_2-H4folate parameterization. *b*. The first 3 rows of residues for each complex are those interacting with the dUMP, the next three are those interacting with the CH_2-H4folate. The '@' symbol specifies a particular atom associated with a given residue. *c*. The first three rows of atoms for each complex are from the dUMP, the next three are from the folate. *d*. The imidazolidine ring of CH_2-H4folate.

similar in most respects. This problem is particularly acute since AMBER is thought to overestimate the importance of electrostatic contributions to energies of interaction even when molecular charges are well parameterized. For this reason both the total energy of interaction between the CH_2-H4folate and the binary complex and the total energy less the electrostatic component are reported. The two scales lead to different energy rankings of the complexes.

The ternary structures differed on the basis of the stereochemistry of sulphur addition to the dUMP, the orientation of the CH_2-H4folate relative to the dUMP and the identity of the important residues which interacted with the two ligands. The four groups of structures are discussed below.

dUMP Stereochemistry and Conformation: The Cys 198 - dUMP Michael adduct may be formed by attack of the sulphur on either of the two faces of the pyrimidine ring resulting in two different stereoisomers of the binary adduct. Both stereoisomers were found in our modeling of the TS ternary complex (figure III). Orientations of both the syn and anti conformers of dUMP from DOCK could lead to attack on either face upon minimization, though the anti-conformers favoured the C6 'S' stereochemistry (figure III a.) while the syn conformers favoured the C6 'R' stereochemistry (figure III b.). The tendency of each conformer to be biased towards a different stereochemistry of Michael addition relates to differences in their respective shapes. On minimization from the non-covalent binary to the covalent binary dUMP-TS complex, the formerly planer pyrimidine of dUMP took on a half-chair conformation expected of a 5,6-dihydropyrimidine. Both the C6 'S' and C6 'R' isomers of the covalent binary complex placed the Cys 198 sulphur in the axial conformation. The orientations of the various CH_2-H4folate conformations from DOCK (figure V) consistently put the C11 of folate in a position that on bond formation with C5 of dUMP would lead to a trans-diaxial

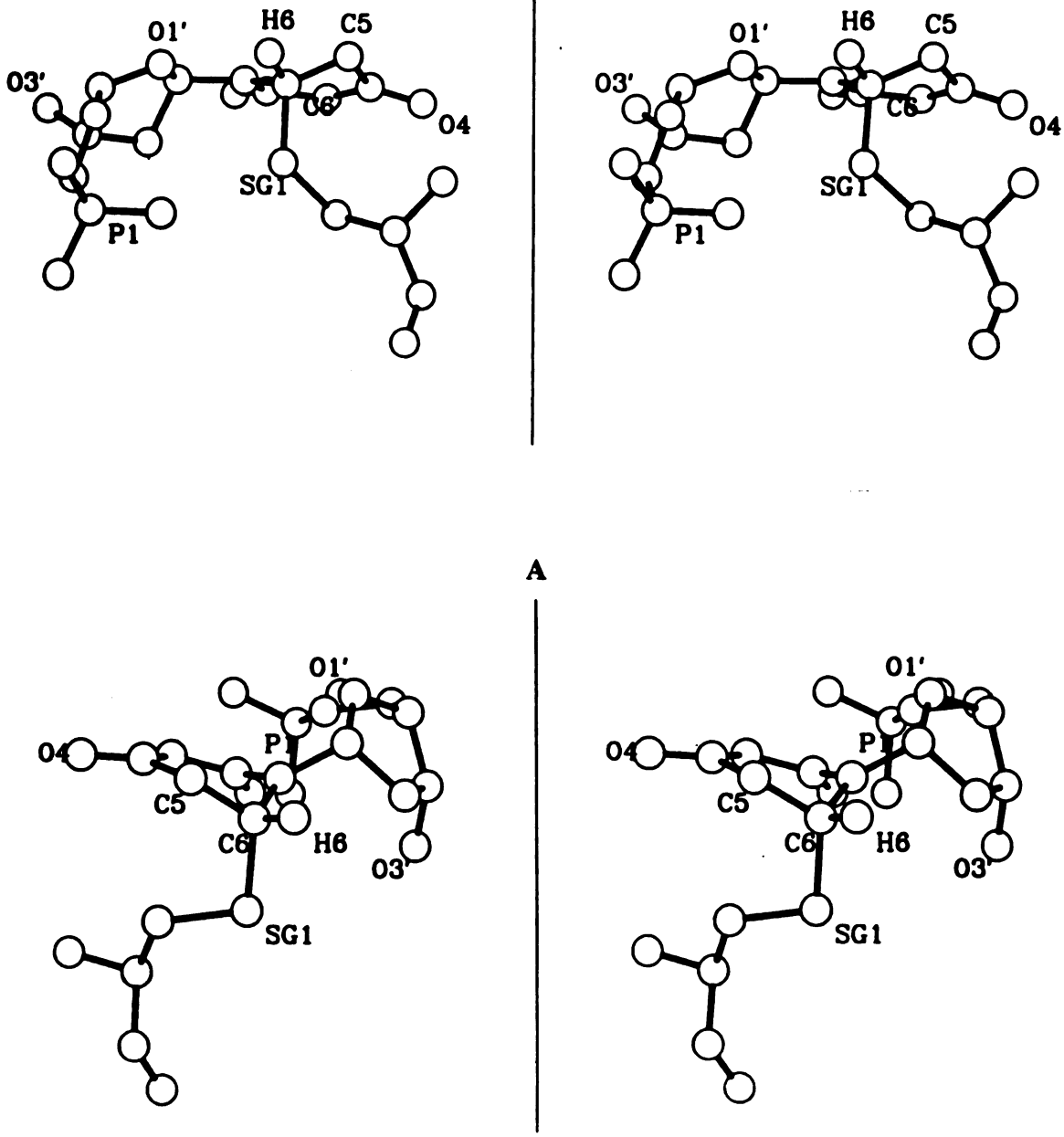


Figure III: Stereo views of the two stereoisomers of the TS-dUMP covalent binary complex: a. C6 'S' stereoisomer, b. C6 'R' stereoisomer. Each stereoisomer is formed by the cysteine sulphur of TS attacking one of two possible faces of the planar pyrimidine ring of dUMP.

relationship to the Cys sulphur. The stereochemistry of sulphur addition in the binary complex therefore effectively determines the stereochemistry of the ternary complex. Depending on the face of attack, either the C6 'S' - C5 'R' or the C6 'R' - C5 'S' stereoisomer is generated.

CH_2-H4folate Orientations: The CH_2-H4folate orientations produced by DOCK are sensitive to the stereochemistry of the binary complex. This is probably because the two different stereoisomers put the C5 of the dUMP in slightly different positions relative to the rest of the active site - the displacement of the two stereoisomers, relative to each other, is slightly more than 2 Å - causing the CH_2-H4folate conformers to explore slightly different local geometries of the active site. The C6 'S' isomer leads to orientations of the CH_2-H4folate whose major axes lie orthogonal to that of the Cys-dUMP adduct (figure V a.,b.), with the PABA moiety either pointing towards (IV a.) or away from (IV b.) Asp 221, Lys 50 and Lys 51. In either of these orientations the C11 is over the plane of the dihydro-pyrimidine ring, roughly trans-diaxial to the sulphur. The energies, as calculated by AMBER, favour orientation IV b.. The C6 'R' isomer leads to CH_2-H4folate orientations whose major axes lie parallel to that of the Cys-dUMP adduct (figure V c., d.), with the PABA moiety either pointing towards (IV c.) or away from (IV d.) Arg 218 and Arg 179'. The AMBER energies of these two possibilities are equivalent within the probable error margins of the minimization. The C11 again lies roughly trans-diaxial to the sulphur.

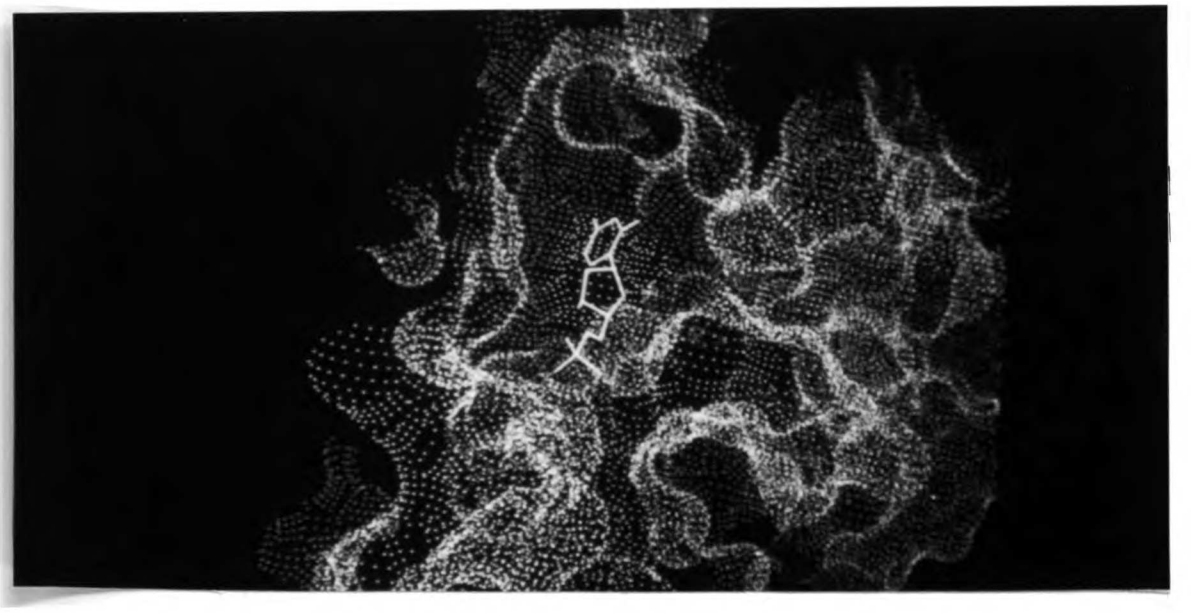


Figure IV: Molecular surface of the TS active site with a docked orientation of dUMP (white) near Cys 198 SG (green).

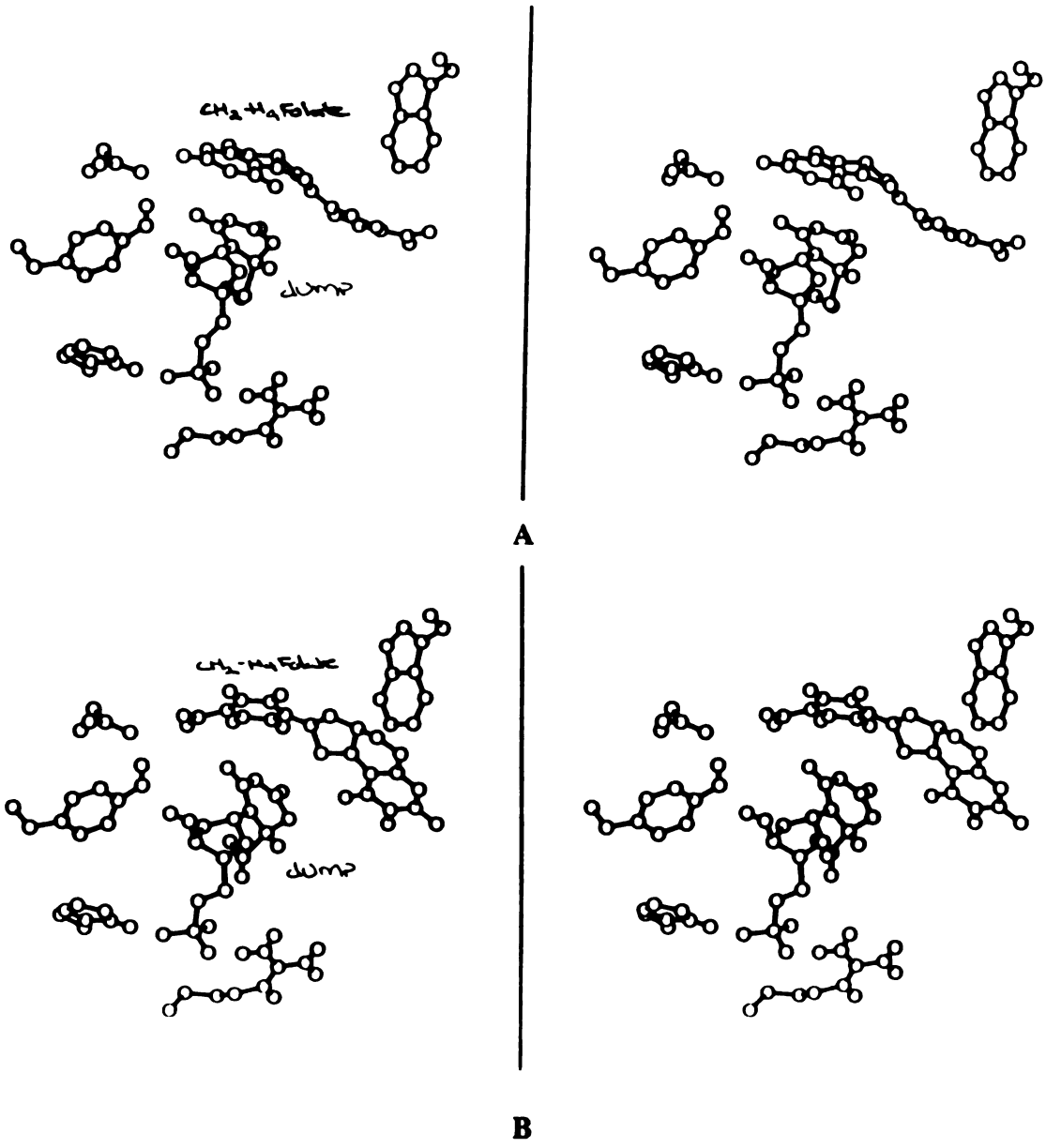


Figure V: Stereo views of the four most likely ternary complexes to come out of the docking and minimization runs showing the relative orientations of dUMP, CH_2-H4folate and some of the important amino acids: A. C6 'S', C5 'R' complex, dUMP in the 'anti' conformation, CH_2-H4folate orthogonal to the principal axis of the dUMP with its PABA moiety pointing away from Lys 50 and Asp 221. Arg 179' hydrogen-bonds to the phosphate of dUMP as does His 259, Tyr 261 hydrogen bonds the O3' hydroxyl of dUMP and the gamma carboxy of Asp 221 while Trp 82 interacts with the PABA moiety of CH_2-H4folate. B. C6 'S', C5 'R' complex, dUMP in the 'anti' conformation, CH_2-H4folate orthogonal to the principal axis of the dUMP with its PABA moiety pointing towards Lys 50 and Asp 221. The amino acid interactions shown are as in a., with the exception that Trp 82 now appears to interact with the pteridine ring system of CH_2-H4folate.

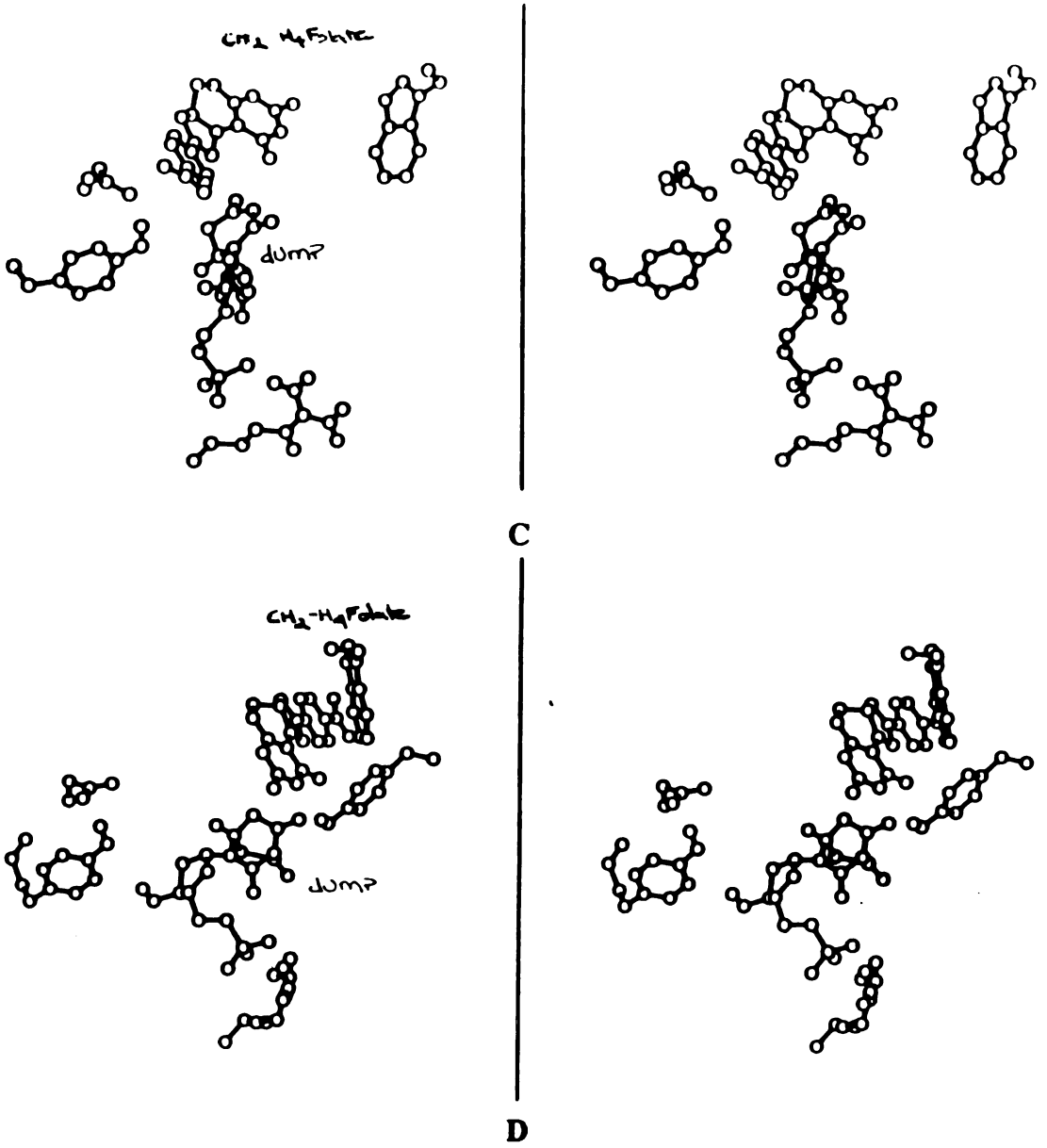


Figure V : c. C6 'R', C5 'S' complex, dUMP in the 'anti' conformation, CH_2-H4folate parallel to the principal axis of the dUMP with its PABA moiety pointing towards Arg 218 and Arg 179'. Arg 179' hydrogen-bonds with the phosphate of dUMP while Trp 82 interacts with the pteridine system. d. C6 'R', C5 'S' complex, dUMP in the 'anti' conformation, CH_2-H4folate parallel to the principal axis of the dUMP with its PABA moiety pointing away from Arg 218 and Arg 179'. Arg 179' hydrogen-bonds with the phosphate of dUMP while Trp 82 interacts with the imidazolidine moiety. Tyr 261 hydrogen bonds to the O3' hydroxyl of dUMP and Asp 221. Tyr 146 is edge on to the PABA and forms a hydrogen-bond with O4 of dUMP.

Important Residues

C6 'S' - C5 'R' Stereoisomer (IV a., b.): As predicted on the basis of crystallography (Hardy *et al.*, 1987) and incorporated in our models by the filtering criteria used in the dUMP dockings, Arg 218 and Arg 179' were found to be two of the principal anchors of the dUMP phosphate in the enzyme. AMBER also found His 259 to be important for ligating the phosphate, with an HND - OP^5 hydrogen bond distance of less than 2 Å. There is no obvious relationship between His 199, which has been proposed as a candidate for the the base activating C5 in the methyl transfer step (Hardy *et al.*, 1987) and the C5 of dUMP. The His NE to dUMP C5 distance is about 6 Å and from the relative geometries it is difficult to imagine a way for the His to take part in activating the C5 or the sulphur. Tyr 261 forms a hydrogen bond to the 3'-hydroxyl of the sugar ring of the dUMP, with an OH - 3'-OH distance of just over 2 Å. A hydrogen bond is also formed between the main chain nitrogen of Cys 198 and the O4 of the dihydro-pyrimidine, with an HN - O4 distance of just under 2 Å. This last interaction would be catalyticaly useful as it would stabilize the enolate form of the dUMP, favouring both Michael addition on the part of the Cys 198 sulphur as well as nucleophilic attack of the C5 on the C11 of CH_2-H4folate

The residues neighbouring the CH_2-H4folate depend on the orientation of the molecule. In orientation IV b. the NA2 amino group of the pteridine ring forms a hydrogen bond with Asp 221, with a NH - O distance of 2.6 Å. Gln 149 forms a hydrogen bond with the carbonyl of the PABA, with a hydrogen bond distance of 1.9 Å. In the models the main chain carbonyl oxygen of Leu 144 forms a hydrogen bond with the amido nitrogen group of the PABA, though it is unclear how predictive this is since in reality this group is part of a poly-glutamate tail and only contains one of the two amido hydrogens present in the model CH_2-H4folates. Trp 82 interacts edge on to the

face of the pteridine ring in what appears to be a quadrupole interaction of the sort commonly found in aromatic ring-aromatic ring contacts in proteins (Burley & Petsko, 1985). In orientation IV a. Asp 221 forms a hydrogen bond with an HN from the amido part of the PABA moiety, but again the relevance of this interaction is unclear. The N8 of the pteridine is hydrogen bonded to Glu 84, with a HN - O distance of less than 2 Å. Trp 82 again interacts edge on to the plane of the pteridine ring.

C6 'R' - C5 'S' Stereoisomer (IV c., d.): While the residues involved in binding the dihydro-dUMP in this stereoisomer are similar to the C6 'S' - C5 'R' isomer, there are some notable differences. Arg 218 and Arg 179' remain important residues in terms of anchoring the phosphate of dUMP, however His 259 no longer has such a role. Rather, it is involved in a hydrogen bond network with Ser 219 which terminates with the 3' hydroxyl of dUMP. Tyr 261 is also involved in hydrogen bonding to this hydroxyl. There is no obvious interaction between the O4 of the dUMP and the enzyme, although in the minimized binary covalent complex Tyr 146 hydrogen bonded to O4 with an OH - O4 distance of 1.9 Å. The distance between His 199 NE and the C5 of dUMP was approximately 6.5 A; from the graphics it was again difficult to envision a direct role for this residue in catalytically activating the C5.

Enzyme interactions with the CH_2-H4folate in this stereoisomer again depended on the orientation of the ligand. In orientation IV c. the CH_2-H4folate extensively overlapped the dUMP capping the major groove of the active site and has few visually obvious directed contacts with the enzyme. Most of its interactions seem to be of a van der Waals contact type, these include interactions between Trp 82 and Phe 228 and the pteridine ring of CH_2-H4folate as well as with the dUMP itself. Several high-scoring dockings of the CH_2-H4folate in this general orientation were found which placed its pteridine ring within 5 Å of a group of six aromatic amino acids: Phe 64, Trp

82, Tyr 146, Trp 150, Phe 228 and Tyr 233. This result is noteworthy because this group of very highly conserved amino acids (except for f3T Phage substitutions at 64 and 150 they would be completely conserved) has been postulated to play a role in catalyzing the conformational change of the CH_2-H4folate (Santi *et al.*, 1987). In orientation IV d., the important interactions included potential hydrogen bonding between Tyr 233 and an amide NH of the PABA as well as an apparent quadrupole stacking network involving the PABA, Tyr 146 and Phe 64. Trp 82 once again stacks edge on to the plane of the pteridine ring system of the CH_2-H4folate.

Briefly summarizing, the results of our modeling of the TS ternary complex were as follows. The orientations of the dUMP found by DOCK are sensitive to the conformation of the molecule with the anti, *gauche-minus* (II a.) conformer scoring the best. The anti conformers (II a., b.) in general favoured binary complex formation with the C6 'S' stereochemistry while the syn conformers (II c., d.) favoured binary complex formation with the C6 'R' stereochemistry. The three model-built conformers of CH_2-H4folate (III a., b., c.) gave similar orientations with similar scores - they all scored better than the structure produced by AM1 (III d.). We find two stereochemistries of binary covalent complex formation between dUMP and the enzyme (IV a., b.). The C6 'R' stereoisomer leads to CH_2-H4folate binding in orientations which would give C6 'R' C5 'S' ternary complexes while the C6 'S' binary stereoisomer leads to CH_2-H4folate binding to give C6 'S' C5 'R' ternary complexes. We describe four reasonable orientations for dUMP and CH_2-H4folate in ternary complexes with TS (refer to Table I). The C6 'R' C5 'S' complexes (IV c., d.) all have the CH_2-H4folate bound parallel to the major axis of the dUMP, while the C6 'S' C5 'R' complexes (IV a., b.) all have the CH_2-H4folate's bound orthogonal to the major axis of the dUMP. We have modeled interactions between the four complexes and specific residues of the enzyme.

Discussion

The fundamental thesis of the "docking" approach (Kuntz *et al.*, 1982) is that shape plays an important role in determining how two molecules position themselves so as to maximize favourable interactions. The great advantage of using shape complementarity is that it allows one to explore the whole range of rigid body orientations available to a system of fixed shapes. The disadvantage is that the docking method is ingenuous in its treatment of the energetic determinants of molecular binding. It ignores forces such as electrostatics and hydrogen bonding which are certainly important for determining how two molecules will interact, though the algorithm does crudely deal with van der Waals interactions and perhaps the hydrophobic effect as well (Kuntz *et al.*, 1982; Lesk, 1986; Naray-Szabo, 1984). While we have used molecular mechanics to evaluate the orientations produced with DOCK, the energy analysis we have employed must still be considered fairly simplistic. Molecular mechanics can only explore a very small region of conformation space - possible global conformational changes in the molecules resulting from binding will not be found using this technique. Moreover, our simulations proceeded with only a very incomplete consideration of solvent. Thus we make no claim to be able to *precisely* predict the orientations of dUMP and CH_2-H4folate in the TS active site, nor to have determined the configurations of all of the important residues. Such a precise agreement between prediction and reality, should it occur, would be fortuitous. Despite these reservations, we have been able to use the docking approach to predict a complex molecular interaction on the basis of the isolated structures of the individual component molecules. Our work leads to certain predictions regarding the nature of the interactions which will be found in the ternary complex of TS, dUMP and CH_2-H4folate which are experimentally testable. We have collected these predictions in Table II for easy perusal.

Table II
Predictions and Tests

Prediction	Test	Results
dUMP in Ternary complex is in one of two stereoisomers: 1) C6 'S' C5 'R' or 2) C6 'R' C5 'S'	Solution of ternary complex Crystal structure.	Not Available[a]
CH_2-H_4folate configurations depend on dUMP stereochemistry: in 1) the folate is oriented perpendicular to the dUMP, in 2) the folate is oriented parallel to the dUMP.	Solution of ternary complex crystal structure. NMR of ternary complex: focus on phosphate folate NOE's.	Not Available *b*
His 199 has no *direct* role in catalysis.	Site-directed mutagenesis of His to a non-basic amino acid.	His mutated to Val: enzyme retains 25% of specific activity.
Tyr 261 hydrogen-bonds to 3'-OH of dUMP.	Mutate Tyr to Phe: K_d should increase.	*b*
His 259 hydrogen-bonds to either: 1) dUMP phosphate or 2) Ser 219	Mutate His to Val or Phe: K_d goes up, probably more in 1) than 2).	*b*
Ser 219 hydrogen-bonds to either: 1) dUMP phosphate or 2) 3'-OH of dUMP	Mutate to Ala: K_d goes up in either case.	*b*
Trp 82 interacts with CH_2-H_4folate	Mutate to Phe: Study affect on UV absorbance of folate binding.	*b*
Asp 221 interacts with Tyr 261 in hydrogen bond network.	Mutate to Val or Ala: K_d of dUMP should increase: temperature/binding studies should show most affect on $\Delta\Delta S$ of binding.	*b*
Asp 221 may also interact with CH_2-H_4folate. This interaction only found in two of the four putative structures.	Mutate to Val or Ala K_d of folate should increase.	*b*

a. The experiment has been done but the results had not been published and were not available to us. *b*. The experiment had not been done at the time this chapter was writen (fall, 1988).

We briefly consider the published results on the TS ternary complex that pertain to structural issues and relate to this work.

^{19}F NMR studies suggest that that the ternary complex of TS- 5F-dUMP -CH_2-H4folate involves a trans-diaxial relationship between the C11 of CH_2-H4folate and the SG of Cys 198 (James *et al.*, 1976; Byrd *et al.*, 1977; 1978). Benkovic and co-workers have postulated that the stereochemistry of this complex is C6 R, C5 'S' (Slieker & Benkovic, 1984) While all four of our structures reproduced the trans-diaxial geometry that has been predicted, only two of the structures had the stereochemistry predicted by Benkovic. Owing to the error margins inherent in our parameterization and charge calculations for the CH_2-H4folate, it is very difficult to distinguish between the two stereochemistries based on AMBER energies. What we can predict is that the C6 'R' C5 'S' stereoisomer should have the CH_2-H4folate aligned parallel to the major axis of the dUMP, while the C6 'S' C5 'R' stereoisomer should have the CH_2-H4folate aligned orthogonal to the major axis of the dUMP (figure V).

TS is believed to undergo a conformational change during the course of the reaction which it catalyzes. Evidence for such change comes from Raman (Fitzhugh *et al.*, 1986) and UV spectral data (Lewis & Dunlap, 1981; Donato *et al.*, 1976) as well as hydrodynamic studies (Lockshin & Danenberg, 1980). Workers familiar with the three dimensional structure of TS have reported that it is very difficult to 'fit' dUMP (Hardy *et al.*, 1987) and CH_2-H4folate (Santi *et al.*, 1987) into the active site of the native enzyme in a chemically reasonable way. They have interpreted these results as being consistent with the conformational change hypothesis, suggesting that in order for its natural substrates to bind the shape of the active site of TS has to change,

presumably in a fairly major way. Our ability, however, to fit both substrates into TS in chemically reasonable geometries suggests that conformational change does not necessarily involve a serious deformation in the shape of that part of the active site responsible for substrate coordination. An example of an enzyme undergoing this sort of conformational change is triose phosphate isomerase (TIM) (Albert *et al.*, 1982). In this protein a loop closes in a hinge-like motion to form the 'roof' of the active site upon binding of substrate, leaving the shape of the original substrate recognition region relatively unchanged.

Predictions regarding important residues are constrained by the ambiguities inherent in our inability to choose between the four possible substrate configurations that we have outlined. Nevertheless, certain relationships stand out. It is difficult to envision a direct role for His 199 in catalysis from the structures, contrary to what has previously been considered (Hardy *et al.*, 1987). The modeling would seem to suggest that the side chain of this residue has only a minor role in terms of *direct* interactions with the substrates. Recent work on the T4 phage TS by Maley and co-workers (Frasca *et al.*, 1988) seems to support this hypothesis. Site-directed mutagenesis on this protein which changed the histidine to a valine resulted in an enzyme which retained 25% of the specific activity of the native form. The authors of this work have suggested that the role of His 199 may be more structural than catalytic. In almost all of the modeled structures, Tyr 261 hydrogen bonds to the 3' hydroxyl of the dUMP, as has been previously suggested (Hardy *et al.*, 1987). We would predict that mutating this residue to a Phe would raise the K_d of the enzyme for dUMP, assuming that the mutation did not disturb the conformation of the enzyme. The roles of His 259 and Ser 219 are more ambiguous. The modeling suggests that mutation of these amino acids to ones unable to donate hydrogens for hydrogen bonding would lower the binding constant of dUMP. It is unclear, however, whether this would occur due to

interactions with the phosphate or the 3' hydroxyl of dUMP. One report in the literature does implicate a histidine in phosphate binding (Rode *et al*, 1986); this hypothesis needs further investigation.

Predictions of important residues that determine folate specificity must be seen as very tentative. Trp 82, a completely conserved residue, interacts with all four CH_2-H4folate geometries and is probably a good candidate for site-directed mutagenesis in conjunction with UV spectroscopic investigations, which have already implicated protein chromaphores as taking part in the conformational change that the folate undergoes during catalysis (Donato *et al.*, 1976). Asp 221, also completely conserved, is another possible candidate for site-directed mutagenesis. Though this residue is involved in binding folate in only two of the predicted geometries (IV a. and IV b.), our modeling suggests that it participates in a hydrogen bond network involving the hydroxyl hydrogen of Tyr 261, whose hydroxyl in turn binds to the 3' hydroxyl of dUMP. Mutating Asp 221 to a residue not capable of interacting with either Tyr 261 or the CH_2-H4folate should increase the K_d of dUMP with TS and might also increase that of CH_2-H4folate. Moreover, because of the nature of Asp 221's relationship to Tyr 261, we suggest that such a mutation would have greater effects on the entropy of dUMP binding than on its enthalpy of binding.

Conclusions

This project was begun as a harsh test of the docking methodology and its ability to address the problem of the favourable co-positioning of two or more molecules in the absence of complete structural data. Thymidylate Synthase is a very large enzyme catalyzing a complex series of reactions. The structure of TS with which we have worked is of fairly low atomic resolution,[6] the enzyme is thought to undergo significant conformational change on substrate binding and the three dimensional structure of one

of its ligands, methylenetetrahydrofolate, is known only approximately. The positions of important bound waters are not available from the 3 Å structure. We have suggested a number of experimentally testable hypotheses concerning the relative geometries available to dUMP, CH_2-H4folate and TS in the ternary complex, as well as some residues on the enzyme which are important in terms of binding and activating the substrates. Our interest in formulating these hypotheses has been not only that they can be falsified, but that if they are falsified they will hopefully show not only where the method went wrong but *why* it went wrong and how methods using shape complementarity to address the molecular recognition problem might be improved in the future. It would of course be gratifying to be right in our predictions, but if we are wrong in some of them, we hope this communication can serve as a paper-trail of what we did and how it might be improved. We look forward to the imminent solution of the crystal structure of a TS ternary complex, as well as to site-directed mutagenesis experiments currently being undertaken, to resolve these issues.

Acknowledgements

We would like to thank the members of the groups of Robert Stroud, Daniel Santi and Peter Kollman for their help and patience. In particular we would like to thank Bill Montfort and Janet Finer-Moore (Stroud group), Joe Davison (Santi group), and Dave Pearlman, George Seibel and David Spellmeyer (Kollman group). We would like to thank the Computer Graphics Lab (grant RR-1081, R. Langridge director) for use of their facilities.

References

Albert, T., Gilbert, W.A., Ponzi, R.P. & Petsko, G.A. (1982) *Mobility and Function in Proteins and Nucleic Acids,* Pitman, London (Ciba Foundation symposium 93), 4-224.

Allen, F.H., Kennard, O., Motherwell, W.D.S., Town, W.G. & Watson, D.G. (1973) *J. Chem. Doc., 13*, 119.

Aoki, K. & Saenger, W. (1984) *J. Chem. Soc., Dalton 7*, 1401 - 1409.

Bash, P.A, Singh, U.C., Brown, F.K., Langridge, R. and Kollman, P.A. (1987) Science 235, 574-576.

Bolin, J.T., Filman, D.J., Matthews, D.A., Hamlin, R.C. & Kraut, J. (1982) *J.Biol.Chem. 257* (22), 13650-13662.

Burley, S.K. & Petsko, G.A. (1985) *Science 229*, 23-28.

Byrd, R.A., Dawson, W.H., Ellis, P.D. & Dunlap, R.D. (1977) *J. Am. Chem. Soc. 99*, 6139.

Byrd, R.A., Dawson, W.H., Ellis, P.D. & Dunlap, R.D. (1978) *J. Am. Chem. Soc. 100*, 7478.

Cody,V. (1986) *J. Mol. Graph. 4*(1), 69-73.

Connolly, M.L. (1986) *Biopolymers 25*, 1229 - 1247.

Desjarlais, R.L., Sheridan, R.P., Dixon, S.J., Kuntz, I.D. & Venkataraghavah, R. (1986) *J. Med. Chem. 29*, 2149.

Desjarlais, R.L., Sheridan, R.P., Seibel, G.L., Dixon, S.J., Kuntz, I.D. & Venkataraghavah, R. (1988) *J. Med. Chem. 31*(4), 722-729.

Dewar, M.J.S., Roebizh, E.G. & Healy, E.F. (1985) *J. Am. Chem. Soc., 107,* 3902-3909.

Donato, H.J., Aull, J.L., Lyon, J.A., Reinsch, J.W. & Dunlap, R.B. (1976) *J. Biol. Chem. 251*(5), 1303-1310.

Emerson, J. & Sundardingam, M. (1980) *Acta Cryst. Sect. B, 36*, 537-543.

Filman, D.J., Bolin, J.T., Matthews, D.A. & Kraut, J. (1982) *J. Biol. Chem. 257* (22), 13663-13672.

Fischer, E. (1894) *Chem. Ber. 27,* 2985-2993.

Fitzhugh, A.L., Kaufman, S., Fodor, S. & Spiro, T.G. (1986) *J. Am. Chem Soc. 108* 7422-7424.

Fontecilla-Camps, J.C., Bugg, C.E, Temple, C., Rose, J.D., Montgomery, J.A. & Kisliuk, R.L (1979) *J. Am. Chem. Soc. 101*, 6114-6115.

Frasca, V., LapPat-Polasko, L., Maley, G.F. & Maley, F. (1988) *Advances in Gene Technology: Protein Engineering and Production,* Proceedings of the Miami Winter Symposium, 149.

Gautham, N., Seshadri, T.P., Viswamitra, M.A., Salisbury, S.A. & Brown, D.M. (1983) *Acta Cryst., C (Cr. Str. Comm.), 39*, 1389-1392.

Goodford, P.J (1984) *J. Med. Chem 27* (5), 551.

Goodford, P.J. (1985) *J. Med. Chem. 28*, 849.

Gready, J.E. (1984) *J. Mol. Struct. 109*, 231.

Gready, J.E. (1985) *J. Am. Chem. Soc., 107*, 6689.

Gready, J.E. (1986) *Chemistry and Biology of Pteridines,* Walter De Gruyter and Co.

Hansch, C., Li, R., Blaney, J.M. & Langridge, R. (1982) *J. Med Chem, 25,* 777.

Hardy, L.W., Finer-Moore, J.S., Montfort, W.R., Jones, M.O., Santi, D.V. & Stroud, R.M. (1987) *Science 235*, 448-455.

Houk, K.N., Paddon-Row, M.N., Rondan, N.G., Wu, Y., Brown, F.K., Spellmeyer, D.C., Metz, J.T., Li, Y. & Loncharich, R.J. (1986) *Science 231*, 1108-1117.

Huang, C., Jarvis, L., Ferrin, T.E. & Langridge, R. (1983) UCSF MIDAS: Molecular Interactive Display & Simulation.

Jones, T.A. (1982) *Computational Crystallography*, D. Sayre, Ed., Oxford Univ. Press, Oxford, 303-317.

Kallen, R.G. & Jenks, W.P. (1966) *J. Bio. Chem. 241* (24), 5845-5850.

Kuntz I.D., Blaney, J.M., Oatley, S.J., Langridge, R. & Ferrin, T.E. (1982) *J. Mol. Bio 161*, 269-288.

Lesk, A.M. (1986) ***Acta Cryst.*** *A42*, 83-85

Lewis, C.A., Jr., & Dunlap, R.B. (1981) in *Topics in Molecular Pharmacology* (Burgen, A.S.V., & Roberts, G.C.K.,Eds.), 171-219, Elsevier, Amsterdam.

Lockshin, A. & Danenberg, P.V. (1980) *Biochemistry 19*, 4244.

Naray-Szabo, G (1984) *J. Am. Chem. Soc. 106*, 4584-4589

Paine, G.H. & Scheraga, H.A. (1985) *Biopolymers 24*, 1391 - 1436.

Paine, G.H. & Scheraga, H.A. (1986) *Biopolymers 25*, 1547 - 1563.

Pattabiraman, N., Levitt, M., Ferrin, T.E. & Langridge, R. (1985) *J. Comp. Chem. 6* (5), 432-436.

Poe, M., Jackman, L.M.. & Benkovic, S.J. (1979) *Biochemistry 18* (24), 5527.

Rode, W., Kedziersk, B., Kulikowski, T. & Shugar, D. (1986) *Chemistry and Biology of Pteridines*, Walter de Gruyter & Co.

Santi, D.V. & Danenberg, P.V. (1984) in Folates and Pterins, R.L. Blakely and S. J. Benkovic, Eds. (Wiley, New York, 1984), vol 1, 345-398.

Sheridan, R.P. & Venkataraghavan R. (1987) *J. CAMD., 1*, 243-256.

Singh, U.C. & Kollman, P.A. (1984) *J. Comp. Chem. 5*, 129.

Slieker, L.S. & Benkovic, S.J. (1984) *J. Am. Chem. Soc. 106*, 1833.

Vasquez, M. & Scheraga, H.A. (1985) *Biopolymers 24*, 1437 - 1447.

Weiner, S.J. & Kollman, P.A. (1981) *J. Comp. Chem., 106*, 765.

Weiner, S.J., Kollman, P.A., Case, D.A., Singh, U.C., Ghio, C. Alagon, G., Profeta, S. & Weiner, P. (1984) *J. Am. Chem. Soc., 106*, 765-784.

Weiner, S.J., Kollman, P.A., Nguyen D.T. & Case, D.A. (1987) *J. Comp. Chem., 7*(2), 230-252.

Wong, C.F. and McCammon (1986), *J. Am. Chem. Soc. 108*, 3830.

Appendix

The special AMBER parameters and "prep" files we used in the minimization runs are provided below.

New Parameters:

Parameters for the protein residues and substrates were taken from the parameter file "parmall.dat", distributed with AMBER. We used the united atom parameter set for all residues in the TS - substrate system, except for the dUMP and CH_2-H4folate molecules and residues Cys 198, His 199 and Tyr 146, for which we used the all-atom parameter set (Weiner *et al.*, 1987). New parameters were derived in order to allow us to minimize the Michael adduct of TS and dUMP, as well as the CH_2-H4folate.

Angular parameters:

CT-C -NA	70.	116.4	KURDMP, Camb. Crystal Database Structure (Allen *et al.*, 1973).
CT-N*-CT	70.	120.0	KURDMP, Camb. Crystal Database Structure (Emerson and Sundardingam, 1980).
S -CT-N*	50.	112.8	ZZZAOV01, Camb. Crystal Database Structure (Gautham *et al.*, 1983).
N -C -CA	70.	120.	APPROX., Consistent with AMBER parameters.
O -C -CA	70.	120.	APPROX., Consistent with AMBER parameters.
CA-CA-N2	70.	120.	APPROX., Consistent with AMBER parameters.
CA-HC-CA	35.	120.	APPROX., Consistent with AMBER parameters.
N2-CT-N2	63.	105.	APPROX., Consistent with AMBER parameters.
CT-N2-CT	63.	110.	AROMATIC Coordination.
CT-N2-H2	35.	119.	Quantum mechanics (Gready, 1984; 1985; 1986).
CA-NC-CA	85.	118.	Quantum mechanics.
NC-CA-CA	85.	123.4	Quantum mechanics.
NA-C -CA	70.	113.6	Quantum mechanics.
C -CA-N2	70.	118.2	Quantum mechanics.

Torsional parameters:

CA-CA-HC-CA	1	6.59	180.	2.
CA-HC-CA-N2	1	6.59	180.	2.
NC-NA-CA-N2	1	10.5	180.	2.

CH_2-H4folate "prep" file:

0

CH_2-H4folate, FROM MTX, METHENYL THF AND NMR (BENKOVIC)
folate.out
FOL INT 0
CORRECT NOMIT DU BEG
0.0

1	DUMM		DU	M	0	-1	-2	0.0000	0.0000	0.0000
2	DUMM		DU	M	1	0	-1	1.0000	0.0000	0.0000
3	DUMM		DU	M	2	1	0	1.0000	90.0000	0.0000
4	HN1	H	M	3	2	1		1.0100	120.0000	0.0000
5	N	N	M	4	3	2		1.0100	120.0000	180.0000
6	HN2	H	E	5	4	3		1.0100	120.0000	0.0000
7	C	C	M	5	4	3		1.2700	127.0000	180.5000
8	O	O	E	7	5	4		1.2800	90.0000	300.0000
9	C11	CA	M	7	5	4		1.2900	59.0000	0.5000
10	C12	CA	B	9	7	5		1.3000	60.0000	25.0000
11	H12	HC	E	10	9	7		1.0900	120.0000	0.0000
12	C13	CA	S	10	9	7		1.3200	62.0000	28.0000
13	H13	HC	E	12	10	9		1.0900	120.0000	180.0000
14	C16	CA	M	9	7	5		1.3400	64.0000	31.0000
15	H16	HC	E	14	9	7		1.0900	120.0000	0.0000
16	C15	CA	M	14	9	7		1.0900	66.0000	34.0000
17	H15	HC	E	16	14	9		1.0900	120.0000	180.0000
18	C14	CA	M	16	14	9		1.3800	68.0000	37.0000
19	N10	N2	M	18	16	14		1.3900	69.0000	38.5000
20	C9	CT	M	19	18	16		1.4000	70.0000	40.0000
21	H91	HC	E	20	19	18		1.0900	109.0000	294.1000
22	H92	HC	E	20	19	18		1.0900	109.0000	56.4000
23	C6	CT	M	20	19	18		1.4300	73.0000	44.5000
24	H6	HC	E	23	20	19		1.0900	112.9000	113.4000
25	C7	CT	M	23	20	19		1.4500	75.0000	47.5000
26	H71	HC	E	25	23	20		1.0900	107.4000	200.8000
27	H72	HC	E	25	23	20		1.0900	107.9000	81.1000
28	N8	N2	M	25	23	20		1.4800	78.0000	52.0000
29	H8	H2	E	28	25	23		1.0300	120.0000	0.0000
30	C8A	CA	M	28	25	23		1.5000	80.0000	55.0000
31	N1	NC	M	30	28	25		1.5100	81.0000	56.5000
32	C2	CA	M	31	30	28		1.5200	82.0000	58.0000
33	NA2	N2	B	32	31	30		1.5300	83.0000	59.5000
34	H2A	H2	E	33	32	31		0.9940	120.0000	180.0000
35	H2B	H2	E	33	32	31		0.9940	120.0000	0.0000
36	N3	NA	M	32	31	30		1.5600	86.0000	64.0000
37	H3	H	E	36	32	31		1.0000	122.0000	180.0000
38	C4	C	M	36	32	31		1.5800	88.0000	67.0000
39	OA4	O	E	38	36	32		1.5900	89.0000	68.5000

```
 40  C4A     CA     M   39 38 36   1.6000   90.0000   70.0000
 41  N5      N2     M   40 39 38   1.6100   91.0000   71.5000
 42  CM      CT     M   41 40 39   1.6200   92.0000   73.0000
 43  HM1     HC     E   42 41 39   1.0900  111.0000  267.1000
 44  HM2     HC     E   42 41 39   1.0900  111.0000  145.7000

CHARGE
  0.29670 -0.63937  0.23465  0.48227 -0.39766
  0.06667 -0.18020  0.09715 -0.08187  0.06579
 -0.07619  0.08178 -0.25822  0.11723  0.35064
 -0.40632 -0.00011  0.02446  0.07205  0.28387
  0.01840 -0.26390  0.08422  0.08517 -0.31133
  0.26269  0.49787 -0.73712  1.02496 -1.04113
  0.43685  0.34424 -0.72469  0.34074  0.63809
 -0.41517 -0.20404 -0.40387  0.09154  0.08927
  0.05384

IMPROPER
 N   C11 C   O
 N1  N3  C2  NA2
 C4A N3  C4  OA4
 C2  C4  N3  H3
 C13 C11 C12 H12
 C14 C16 C15 H15
 C15 C11 C16 H16

LOOP CLOSING EXPLICIT
 C4A C8A
 N5  C6
 CM  N10
 C13 C14

DONE
STOP
```

Gloss to Chapter II

One happy consequence of an over-bold project is that one quickly discovers the weaknesses in an approach. The modeling work in Chapter I laid bare several such with DOCK. We noticed early on that our docking searches were taking a long time. Long run times had not previously been observed with DOCK, it later turned out that they became a real problem only with large sites (such as TS) when looking at large numbers of configurations. Also troubling, DOCK results were sensitive to the record order of the spheres in the file that they were read from.* Thus, changing the ordering of the spheres led to different results. Since DOCK is an internal distance-based method, the order of spheres should make no difference. That it did pointed to what turned out to be an algorithmic bug in the program.

I wanted to design novel inhibitors of TS, and since the long run times this would require with DOCK1.0 would have been prohibitive (about 150 cpu days on the fastest machine we had in 1988), I set out to make the program faster and more reliable. To overcome the algorithmic bug, we found it necessary to largely rewrite the matching part of DOCK, based on an idea of Tack's regarding sphere pre-organization into bins. The bin code that is described in Chapter II took the lion's share of my time, and still forms the tangled viscera at the heart of all versions of DOCK beyond 1.1, and of those now under development by my colleagues. Though the matching work did have some interesting spin offs, however, it is not as important as three other innovations that eventually enabled DOCK to tackle the problem that I was most interested in - inhibitor design against thymidylate synthase.

---

* This was a discovery that Dale Bodian and I made independently in summer of 1988.

It was clear early on that the scoring scheme used in DOCK1.0 (DesJarlais et al., 1988) was never going to be fast enough. DOCK1.0 measured shape complementarity by evaluating the distances between all ligand and all receptor atoms for every ligand orientation. Since the receptor is held fixed in a docking run, however, there was no need to constantly remeasure the same set of distances. It would be much faster, I reasoned, to measure the distances from every point on a lattice in the volume of space defined by the receptor atoms once, and then store the contact values for future use with DOCK. Once this lattice had been calculated, it would be a simple matter to map DOCK generated ligand configurations to the appropriate lattice points, and then look up the resulting scores. The lattice idea was not novel - Peter Goodford had been using it since the early 80's mapping active sites (Goodford, 1985) and other workers were increasingly turning to lattice based potentials for the same reasons I was (Goodsell and Olson, 1990). The lattice scoring implementation made DOCK about 4 times faster; the notion of lattice based potentials now underlies all the work that is ongoing in the group towards making DOCK more sensitive to chemical complementarity (Meng et al., 1992).

Even with the lattice, DOCK was still too slow for thymidylate synthase. This was the fault of TS's very large active site - more that 240 spheres, leading to a combinatorial explosion in configurational possibilities. A general solution to this problem came out of a discussion with Tack about a way to tackle it for TS. Tack suggested that perhaps we would not lose that much information if we could sensibly parse the TS site into several binding regions, each of which would be composed of many fewer spheres than the full site. I tried his idea - dividing TS into three different regions and docking to each of them - and found to my delight that not only did the run times plummet, but the DOCK scores of the resulting orientations were better than if I had used the whole site. Thinking about why this might be so, I hit upon the

mathematical justification for sub-clustering spheres in large sites that is developed in equations 1-3 in Chapter II. I believe that this method is general, and allows us to DOCK molecules of almost any size. I show that DOCK can, using sub-clustering, address the "protein docking problem" very efficiently - something it could not do without clustering. Our approach to this problem is presently the definitive one in the field: DOCK is orders of magnitude faster and significantly more accurate than similar programs tackling similar problems in the current literature (Cherfils et al., 1991; Jiang and Kim, 1991; Wang, 1991).

The last important innovation in Chapter II is the focusing technique. The problem focusing addresses is that of the large number of possible configurations in a DOCK search. In such a vast space, where best to search, given limited computational resources? Focusing allows one to sparsely sample configuration space, and use the sampling results to *focus* the search on regions that are likely to result in high numbers of successes when more highly sampled. Graph 12 in Chapter II illustrates that the technique can be very powerful. Of course, the basic assumption that there is a relationship between 'hits' from lower and higher resolution searches will not always hold; where it does not, focusing will lead one astray. Nevertheless, the approach allows for searches in complex spaces that are at once global and detailed. The technique might be interesting to workers outside the field of molecular docking.

Cherfils, J., Duquerroy, S. and Janin, J., *Proteins* **11**, in press (1991)

DesJarlais, R., Sheridan, R. P., Seibel, G. L., Dixon, J. S., Kuntz, I. D. and Venkataraghavan, R., *J. Med. Chem.* **31**, 722-729 (1988)

Goodford, P. J., *J. Med. Chem.* **28**, 849-857 (1985)

Goodsell, D. S. and Olson, A. J., *Proteins* **8**, 195-202 (1990)

Jiang, F. and Kim, S. H., *J. Mol. Biol.* **201**, 79-102 (1991)

Meng, E. C., Shoichet, B. and Kuntz, I. D., *J. Comp. Chem.* accepted for publication. (1992)

Wang, H., *J. Comp. Chem.* **12**, 746-750 (1991)

Molecular Docking Using Shape Descriptors

Brian K. Shoichet, Dale L. Bodian‡ and Irwin D. Kuntz *

Department of Pharmaceutical Chemistry
University of California, San Francisco
San Francisco, California, 94143-0446

‡Department of Biochemistry and Biophysics
University of California, San Francisco
San Francisco, California, 94143-0448

* to whom correspondence should be addressed

Key Words: docking, structure prediction, tree searches, complementarity, molecular description

Running Title: Molecular Docking

Abstract

Molecular docking explores the binding modes of two interacting molecules. The technique is increasingly popular for studying protein-ligand interactions and for drug design. A fundamental problem with molecular docking is that orientation space is very large, and grows combinatorially with the number of degrees of freedom of the interacting molecules. Here we describe and evaluate algorithms that improve the efficiency and accuracy of a shape-based docking method.(Kuntz et al., 1982; DesJarlais et al., 1988) We use molecular organization and sampling techniques to remove the exponential time dependence on molecular size in docking calculations. The new techniques allow us to study systems that were prohibitively large for the original method. The new algorithms are tested in 10 different protein-ligand systems, including 7 systems where the ligand is itself a protein. In all cases, the new algorithms successfully reproduce the experimentally determined configurations of the ligand in the protein.

Introduction

Molecular docking fits molecules together in favorable configurations using their topographic features. Practically, docking has been an important technique for the modelling of protein-ligand interactions, and has been used in studies of the structural basis of biological function(DesJarlais et al., 1986; Goodsell and Olson, 1990) and drug design.(Goodford, 1984; DesJarlais et al., 1990) Theoretically, the approach is a relatively tractable instance of the general problem of combinatorial optimization, a focus of much work in recent decades.(Miller and Pekny, 1991)

One of the first practical suggestions for docking came from Crick,(Crick, 1953) who suggested that complementarity in helical coiled-coils could be modelled as knobs fitting into holes. More recently, workers have used both geometric(Kuntz et al., 1982; Zielenkiewicz and Andrzej, 1984; Connolly, 1985; Lee and Rose, 1985; DesJarlais et al., 1988) and energy-based methods(Wodak and Janin, 1978; Goodford, 1985; Wodak et al., 1987; Goodsell and Olson, 1990) to search for fruitful binding modes of ligands in receptors. The geometric methods have focused on matching descriptors of topographical features to generate favorable configurations, while the energy methods have used potential energy functions to guide their search of orientation space.

Docking is computationally difficult because there are very many ways of putting the two molecules together, and the number of possibilities that must be sampled grows exponentially with the size of the component molecules. The orientation space of two biomolecules, especially when one or more of them is a protein, is so large as to make exhaustive methods prohibitive.(Connolly, 1985) This difficulty reflects the many

interfaces and multiple minima presented by the surface of a macromolecule; for descriptor-based methods, the docking problem is nondetermined in polynomial time (NP-complete).(Kuhl et al., 1984)

We have previously developed a rigid body docking method which uses molecular descriptors (DOCK program).(Kuntz et al., 1982; DesJarlais et al., 1988) DOCK could regenerate experimentally determined configurations in several ligand-protein complexes. Like all descriptor-based methods, however, the search time of the algorithm scaled poorly as the number of features describing the molecules increased. This time dependence made docking of macromolecular complexes, for instance, unfeasible. Also, some of the heuristics used in DOCK to help reduce the number of possible matches made predicting the performance of the algorithm difficult. Lastly, the extent of a search was hard to control.

In this paper we discuss new algorithms which make our docking procedure faster and allow it to handle protein-protein systems, which had previously been prohibitively large for our method; we call the new program DOCK2. We describe modifications that are particular to our implementation of a docking program as well as changes in algorithmic approach that address general features of the docking problem. At the general level, we address the strong time dependence of the algorithms by a 'divide-and-conquer' procedure which separates macromolecules into independent geometric regions that are individually considered as possible interfaces. This modification dramatically improves the way that the docking problem scales with the size of the system being docked. We illustrate the improvement by docking 7 different pairs of proteins. Also at the general level, we outline a technique to increase automatically the number of possible configurations generated in regions of likely complementarity.

Ideally, this should allow for more efficient sampling of orientation space, moving from low density sampling in "poor" regions to higher degrees of sampling in regions that

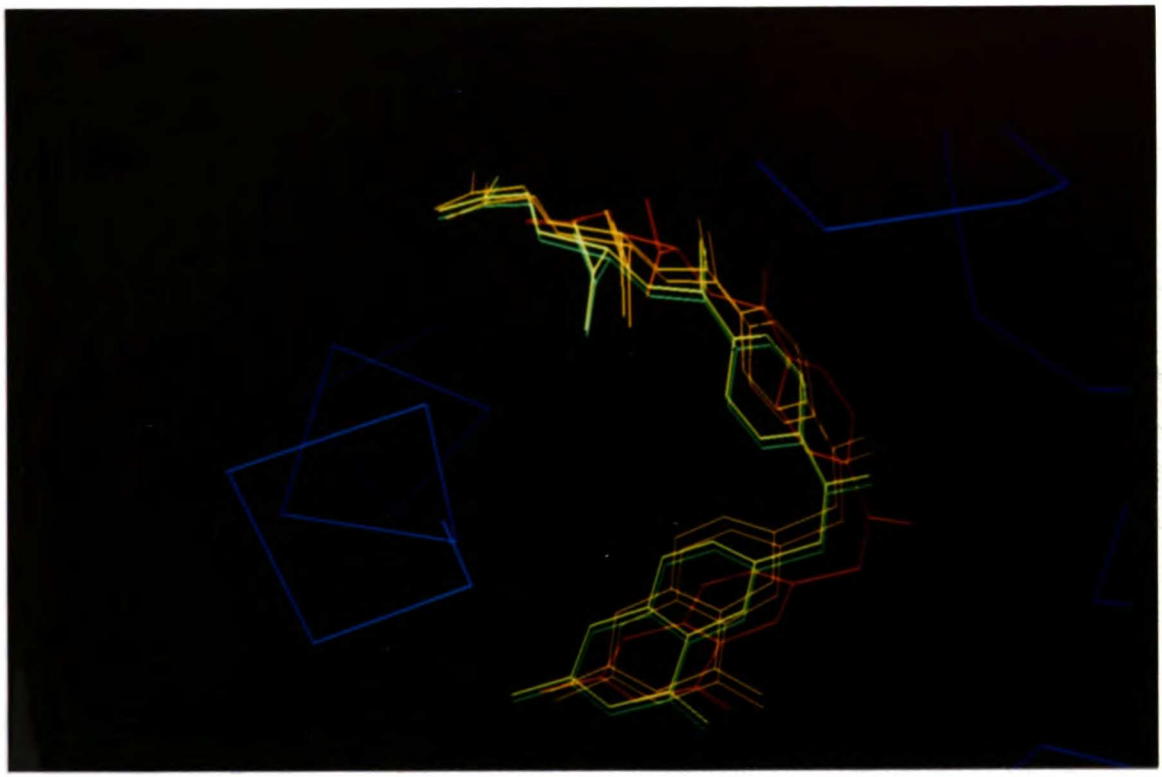


Figure 1: Several low r.m.s.d. dockings of methotrexate in dihydrofolate reductase.(Bolin et al., 1982) Crystallographic configuration in green, docked orientations in yellow, amber and red, in increasing r.m.s.d., respectively. Protein in blue. Figures 1-4 and 11 were made with the MidasPlus graphics program.(Ferrin et al., 1988)

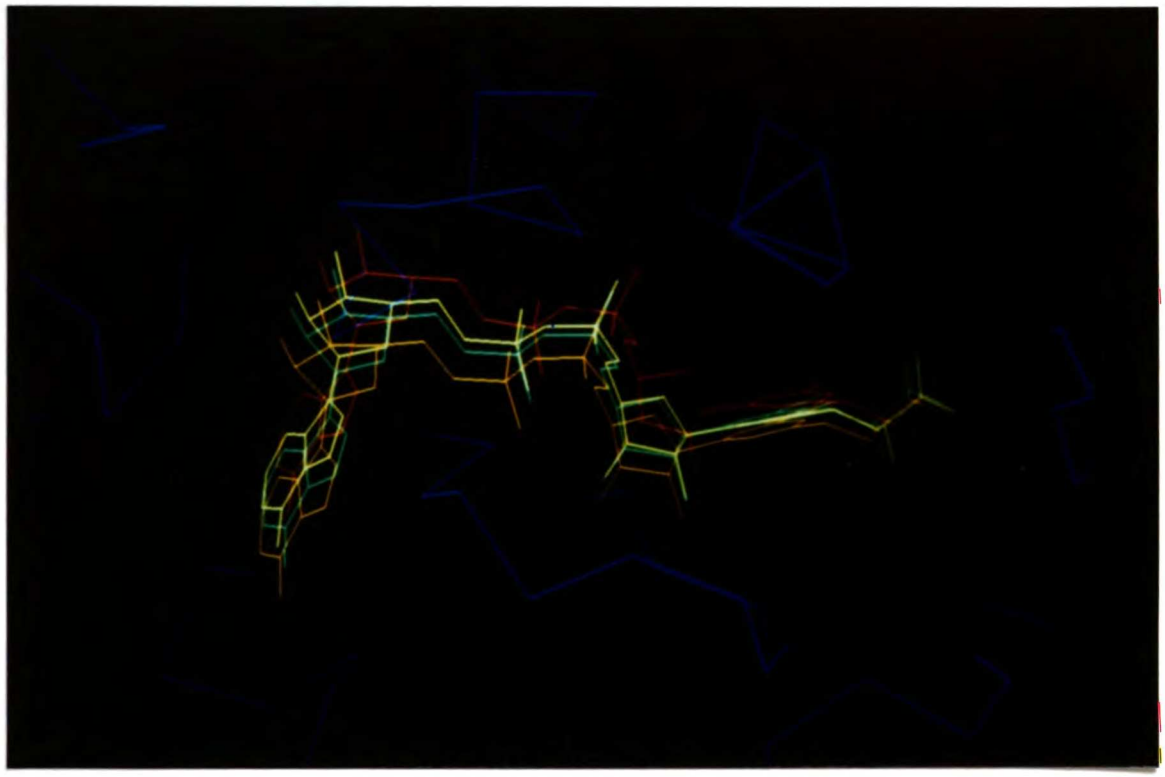


Figure 2: Several low r.m.s.d. dockings of NAD-lactate in lactate dehydrogenase.(Grau et al., 1981) Crystallographic configuration in green, docked orientations in yellow, amber and red, in increasing r.m.s.d., respectively. Protein in blue.

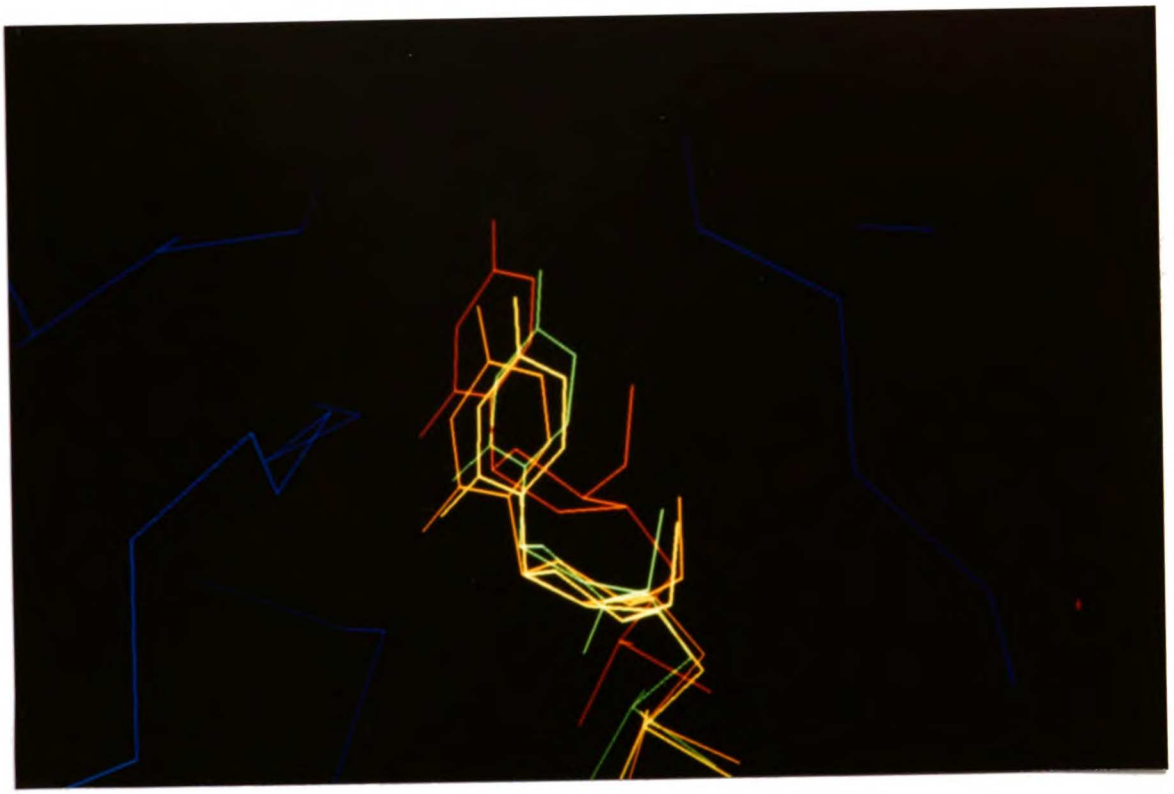


Figure 3: Several low r.m.s.d. dockings of uridine vanadate in ribonuclease.(Borah et al., 1985) Crystallographic configuration in green, docked orientations in yellow, amber and red, in increasing r.m.s.d., respectively. Protein in blue.

have produced favorable configurations. We also take up issues specific to our program. The new program is more systematic in its searches of orientation space, and also more easily controlled in depth of search by the user. We consider three different ways of selecting features for matching, and compare the success of each approach at reproducing experimental configurations. Finally, we describe a lattice-based method for evaluating the goodness of fit of the docked complexes, which significantly reduces run times. We test the new algorithms extensively in four crystallographically determined protein-ligand complexes (all structures are taken from the Protein Data Bank(Bernstein et al., 1977)): ribonuclease/uridine vanadate,(Borah et al., 1985) dihydrofolate reductase/methotrexate,(Bolin et al., 1982) lactate dehydrogenase/NAD*-lactate(Grau et al., 1981) and trypsin/PTI(Marquart et al., 1983) (Table Ia, Figures 1-3(Ferrin et al., 1988)). We show that the methods can be used to regenerate the crystallographic configurations of 6 other complexes (Table Ib, Table II), where the ligand as well as the receptor is a protein. In their bound (as they occur in the crystal complex) conformations we dock subtilisin with chymotrypsin inhibitor,(McPhalen and James, 1988) chymotrypsin with ovomucoid third domain(Fujinaga et al., 1987) and thymidylate synthase monomer with thymidylate synthase monomer(Montfort et al., 1990) to regenerate the dimer. In their unbound conformations (as they occur in isolation of their cognate ligand or receptors) we dock trypsin(Walter et al., 1982) with PTI,(Marquart et al., 1983) subtilisin(Neidhart and Petsko, 1988) with chymotrypsin inhibitor,(McPhalen and James, 1987) and chymotrypsin(Blevins and Tulinsky, 1985) with ovomucoid third domain.(Bode et al., 1985)

---

* Abbreviations used: NAD (nicotinamide adenine dinucleotide); r.m.s.d. (root mean square deviation); PTI (pancreatic trypsin inhibitor); DHFR (dihydrofolate reductase).

Receptor[a]	Ligand	Number of Atoms in Receptor[b]	Number of Atoms in Ligand
Ribonuclease (6rsa)[17]	Uridine Vanadate (6rsa)	951[c]	20
Dihydrofolate Reductase (3dfr)[18]	Methotrexate (3dfr)	1298	33
Lactate Dehydrogenase (5ldh)[19]	NAD-Lactate (5ldh)	2560	51
Trypsin (2ptc)[20]	PTI (2ptc)	1595	423

Table Ia

Receptor	Ligand	Number of Atoms in Receptor	Number of Atoms in Ligand
Trypsin[25] (2ptn)	PTI[20] (4pti)	1564	449
Subtilisin[22] (2sni)	Chymotrypsin Inhibitor[22] (2sni)	1938	513
Subtilisin[26] (1sbc)	Chymotrypsin Inhibitor[27] (2ci2)	1920	521
Chymotrypsin[23](1cho)	Ovomucoid 3rd Domain[23] (1cho)	1751	400
Chymotrypsin[28] (5cha)	Ovomucoid 3rd Domain[28] (2ovo)	1736	418
Thymidylate Synthase monomer[24]	Thymidylate Synthase monomer[24]	2143	2143

Table Ib

Table I: Test complexes and structures used for docking. a). The 4 complexes used in the most extensive testing of the algorithms, including focusing, clustering, scoring and the different graph generation methods. b). The structures used in the protein-protein docking tests DOCK2.

[a] All structures taken from the Protein Data Bank[16] have their reference code in parentheses.

[b] The number of atoms are for the structures actually used in the docking runs. These may differ slighly from the those in the pdb files in that atoms which had no density, as recorded in the pdb files, were not used in the docking runs.

[c] Though 6rsa is a neutron structure and contains hydrogen/deuterium coordinates, only heavy atoms were used in the docking runs.

The Docking Problem

The underlying notions in descriptor-based docking have their antecedents in the "lock and key" ideas of Ehrlich.(Ehrlich, 1907) The computational problem is to describe the features that define the shape of the "lock" and "key" and then to map the two sets of features together in favorable ways. There are many ways of describing molecules for this purpose.(Rose, 1978; Zielenkiewicz and Andrzej, 1984; Connolly, 1985; Leicester et al., 1988) We use spheres which are locally complementary to a molecular surface.(Kuntz et al., 1982; Connolly, 1983) However the features are described, the next task is choosing which of them to use for matching the two molecules together. This brings up a fundamental difficulty.

Matching features (descriptors) involves selecting a set of some number of points from a larger collection of possibilities. The number of possible sets of features depends combinatorially on the number of features in each set (n), and the total number describing each molecule:

$$\text{Number of sets} = {}^{n}C_{Nr} \times {}^{n}P_{Nl} = {}^{n}P_{Nr} \times {}^{n}C_{Nl} \qquad (1)$$

where N_r is the total number of receptor features and N_l is the total number of ligand features. 'C' and 'P' represent combinations and permutations. When $N_r, N_l >> n$, (1) can be rewritten:

$$\text{Number of sets} = (N_r)^n \times (N_l)^n \qquad (2)$$

Therefore, as the number of features in a set rises, the number of sets to be considered rises exponentially, as would the computation time of any method that attempts to dock molecules by looking at all possible sets.

Fortunately, for docking it is not necessary to consider sets above a certain size. Four non-planar points from each set - four features from each molecule - uniquely define a configuration involving two molecules of any shape. The computation of a docking problem thus becomes bounded, minimally but sufficiently, by $(N_r)^4 \times (N_l)^4$.

This still makes for long calculation times when N_r and N_l are large, as is the case in macromolecular docking. One can further improve matters by pruning sets as they are being constructed, a procedure that is described in detail in the following section. Even with pruning, however, the number of configurations that might plausibly be looked at for two macromolecules is still very large, and this number grows quickly as the size of the ligand and receptor increase. We return to this problem in the next section.

Having described the molecules, chosen sets of features and mapped the one onto the other, it only remains to evaluate the resulting configurations for the 'goodness of fit' between the ligand and the receptor. There are as many ways of doing this as there are docking programs; most methods use simplified potential functions or some version of shape complementarity. We describe the details of our implementation below.

Methods

We summarize the method and then take up each point in greater detail in the following paragraphs. Geometric descriptions (spheres or atoms) of local bumps and clefts on ligand and receptor surfaces guide the search of orientation space, the goal being to find orientations that map the bumps of one into the clefts of the other. We look for sets of spheres from the first molecule that have the same internal distances, within a certain tolerance, between their centers as do sets of spheres from the second molecule. We will refer to sets that pass this distance criterion as "matches". Matches are used to define rotation/translation matrixes that map the second molecule onto the first.(Ferro and Hermans, 1977) Configurations of the ligand in the receptor depend, therefore, on the locations of the sphere sets on the surfaces of the respective molecules. An orientation, once found, is subjected to a fast preliminary evaluation of complementarity based on a simple examination of atomic contacts between the receptor and the ligand (scoring). Orientations of the ligand that place it in regions of space occupied by the receptor are discarded. The configurations that pass the excluded volume filter and have enough 'good' contacts are saved for further evaluation.

Molecular Description - Spheres

We describe a receptor geometrically using spheres which are locally complementary to grooves and ridges in its molecular surface(Kuntz et al., 1982; Connolly, 1983) (Figure 4). The spheres fill the empty volume of a site, generating its negative image. The centers of these spheres may be thought of as pseudo-atoms; they are used as an irregular grid for mapping the ligand into the binding site. Spheres are generated analytically to touch the molecular surface at two points, have their centers along the surface normal to one of the points and are placed so that they do not intersect the

surface. The spheres are of different sizes and typically overlap one another within a given pocket in the protein. A collection of overlapping spheres defines a cluster. The molecular surface of a protein will typically have tens of clusters, each of which describes a potentially interesting site of interaction. The radius of a sphere reflects the concavity of a local region of the molecular surface. The larger the sphere radius, the larger and shallower the pocket that the sphere describes. Macromolecular ligands are similarly described, except that the spheres are placed *within* the molecular surface and are complementary to local ridges rather than the grooves. For smaller ligands such as methotrexate, the atom centers are used rather than spheres.(DesJarlais et al., 1986) Other points can be added to the receptor or the ligand descriptions without loss of generality. These include the center of mass, centers of rings, centers of molecular attraction, bound waters and so forth.

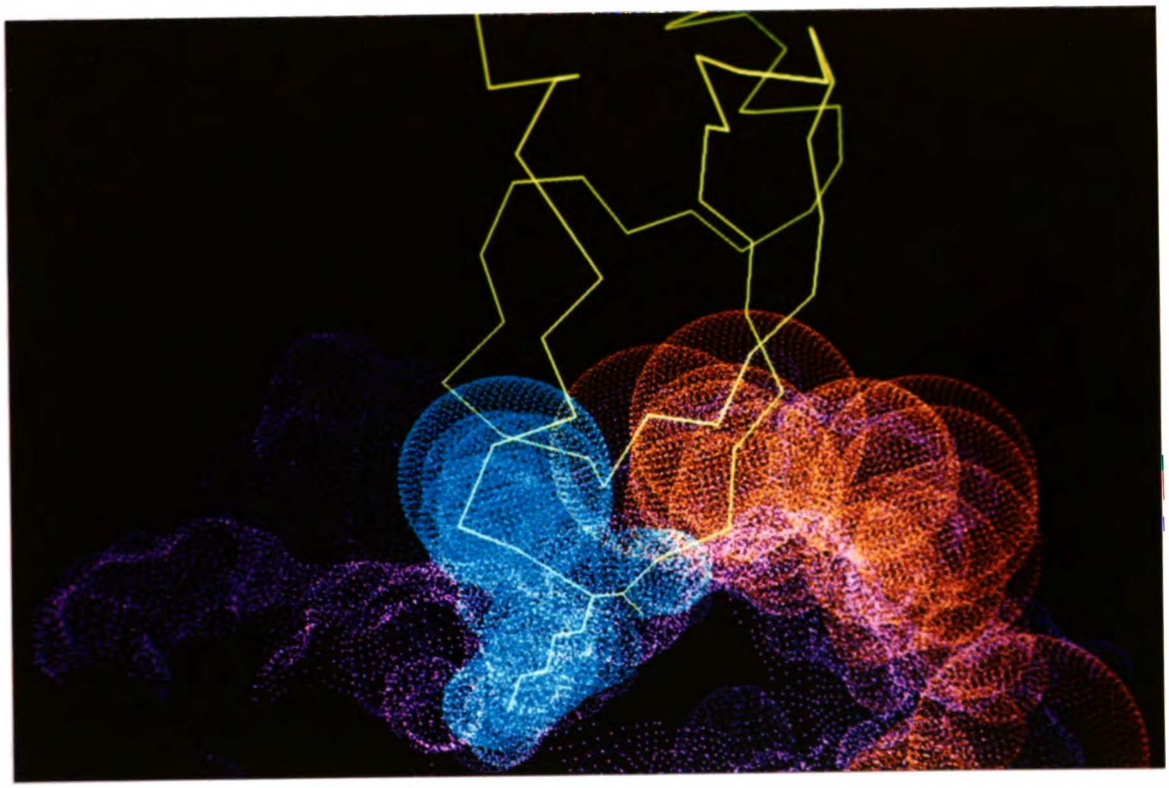


Figure 4: Trypsin spheres. Two sub-clusters are shown, in blue and red, the trypsin molecular surface is colored magenta, and a c-alpha trace of PTI is in yellow. The blue sphere set describes the trypsin specificity pocket.

Molecular Organization: Divide and Conquer

Macromolecules have many descriptors, which leads to a great number of possible dockings. PTI, for example, is described by 292 spheres in one molecule-spanning cluster, about 10 times more descriptors than in a typical drug-type inhibitor such as methotrexate. Given the third to fourth power dependence of run time on the number of descriptors,(Kuhl et al., 1984) the computation time for docking macromolecular ligands could be as much as 10^4 times longer than for small molecule ligands. In complexes of determined structure, however, the ligand is much larger than the binding site of its receptor, which suggests that most of the ligand's surface will not be involved in any given interface with the receptor. It is thus worthwhile to organize the macromolecules into geometrically distinct subsections, each of which can be matched independently. Ideally, each sub-section would describe one potential interface region of the molecule. We call this procedure "sub-clustering."

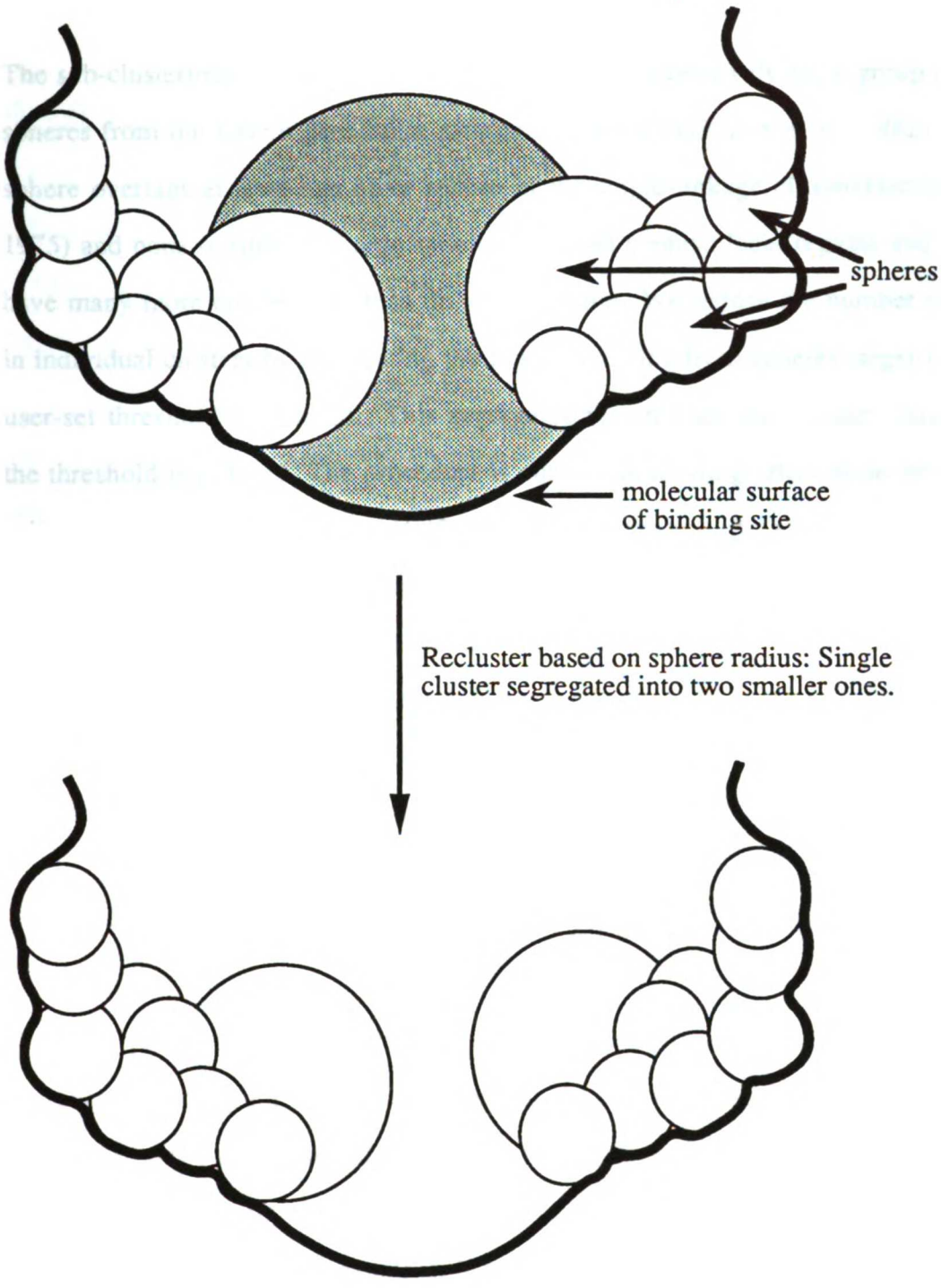


Figure 5: Sphere sub-clustering. The sub-clustering algorithm segregates groups of spheres based on their radii. In this example, a large radius sphere (shaded) is removed from the sphere set, breaking the link connecting one part of the site to the other and segregating the spheres into two sub-clusters.

The sub-clustering program (CLUSTER) begins with a relatively large group of spheres from the sphere generation program (SPHGEN).(Kuntz et al., 1982) Each sphere overlaps at least one other sphere in the single-linkage cluster(Hartigan, 1975) and none outside it. Large spheres span and connect local regions and often have many more connections than do small spheres. We reduce the number of spheres in individual clusters by eliminating the linkages arising from spheres larger than a user-set threshold (Figure 5). This segregates the spheres into smaller clusters as the threshold is reduced. The procedure is analogous to using articulation vertices to split connected graphs.(Wilson, 1985) During this process, the total number of clusters increases. Since we treat each cluster as a potential interface site, the greater number of sites increases the number of possible orientations. This effect is, however, small compared to the combinatorial advantage of restricting the total number of spheres in each site. The ratio of the number of possible matching sphere sets after and before sub-clustering is:

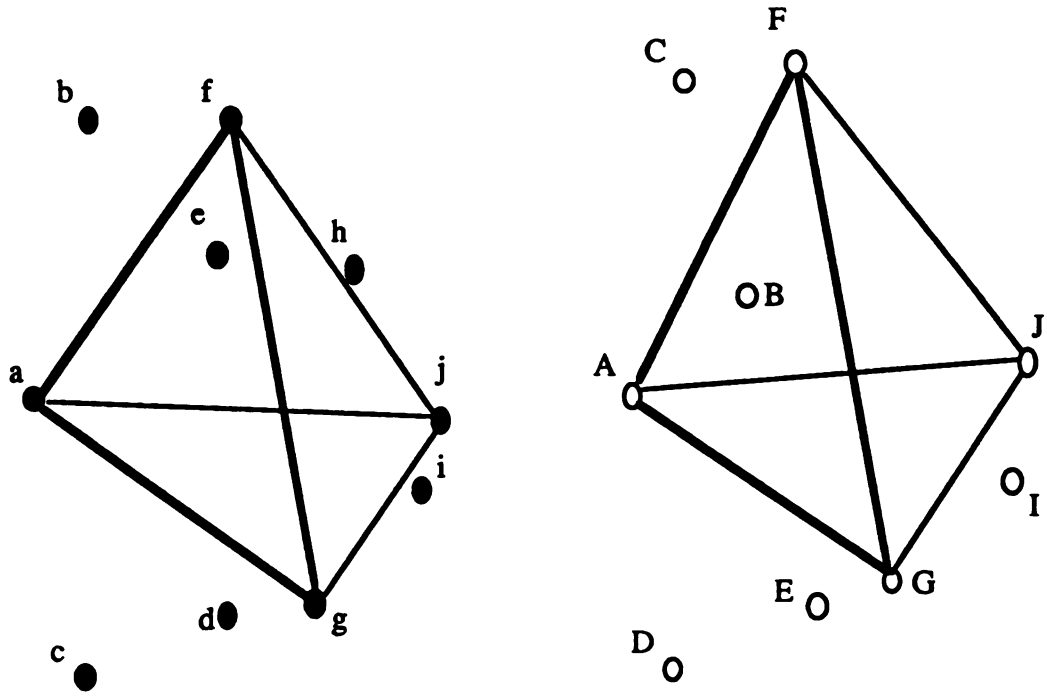
$$\text{Configs}_{\text{sub-clustered}}/\text{Configs}_{\text{unclustered}} = \Sigma_{\text{rclus}} \Sigma_{\text{lclus}} (N'_r/N_r)^n \times (N'_l/N_l)^n \quad (3)$$

N'_r and N'_l are the number of descriptors in the sub-clustered groups being matched in the receptor and ligand, respectively, for N'_r and $N'_l >> n$. rclus and lclus are the numbers of new, sub-clustered sphere sets for the receptor and the ligand. In this manner, sub-clustering significantly decreases the number of possible matches necessary to consider. For example, if sub-clustering reduces the number of spheres in a sphere to 1/4 its original size, while the number of total clusters needed to describe the molecule rises from 1 to 4, than the ratio in (3) will be $4^{-(n-1)}$, or 1/64 when n is 4. Of course, further reduction of the search may be possible if attention can be focused on one of the sub-clusters, such as the active site.

Matching - the bipartite graph

We dock ligands into receptor sites by matching subsets of ligand internal distances onto subsets of receptor sphere internal distances. Most of the possible combinations of ligand and receptor descriptors will not lead to successful dockings. It is therefore sensible to prune the matching search tree as soon as possible.

Docking may be posed as a graph theoretical problem.(Kuhl et al., 1984) If a ligand has N_l descriptors and a receptor has N_r descriptors, then the number of nodes in the docking graph *D* is N_l x N_r. An edge exists between two nodes composed of descriptors $(N_l)_i$,$(N_r)_i$ and $(N_l)_j$,$(N_r)_j$ (where i and j are ligand or receptor descriptors), from the ligand and the receptor, when the distance $N_l(i,j)$ is the same as $N_r(i,j)$, within some tolerance. A minimal match between a ligand and a receptor occurs when there is a sub-graph of *D* that is completely connected by edges and that has at least four nodes (and therefore 6 edges) in it. Four nodes must be specified to determine a rotation-translation matrix that preserves chiral information. It is physically impossible to match, simultaneously, most ligand-receptor features. Another way of saying this is that *D* is sparsely connected, which leads to our method for pruning the search of orientation space.



Receptor Ligand

Match if | Distance(i,j)- Distance(I,J) | < tolerance

1st Node: a onto A
2nd Node: j onto J, if aj-AJ < tolerance
3rd Node: f onto F, if af-AF, jf-JF < tolerance
4th Node: g onto G, if ag-AG, jg-JG, fg-FG < tolerance

Figure 6: Internal distance matching. Ligand and receptor internal distances are compared. If the internal distances do not match at a given node, the tree-search is "pruned" at this node.

As in the original program,(Kuntz et al., 1982) we use a distance matching algorithm that calculates whether the receptor and ligand descriptors share the same pairwise distances, within some tolerance level, in a build-up procedure that evaluates the growing graph as each node is added (Figure 6). A graph that fails the distance check

at some number of nodes M, i.e., which is not completely connected due to the addition of the M'th node, will also fail at all numbers greater than M; therefore we can prune the search at M. Such pruning dramatically reduces the number of ligand-receptor nodes necessary to consider. We further reduce the search space by biasing the search to long edges representing large internal distances. This is a heuristic that weights long range information more heavily than local information.

To control the search of orientation space we organize the receptor spheres based on the internal distances between pairs of sphere centers. Starting with each sphere center, all other centers in the cluster are sorted into "bins" based on their distances to the starting sphere center (Figure 7). All distances in a certain range will be placed in the same bin. The bins are of adjustable resolution - the larger the distance interval for a bin the more points it will typically contain and the fewer the number of bins overall. DOCK2 can allow overlaps between sequential bins to diminish the effect of discrete distance ranges. Bins are constructed for each receptor sphere. Ligand bins are constructed using the same procedure. The ligand and receptor bins for each pair of starting points are matched based on the distance ranges of the points within them; only those points in bins with similar ranges will be used to generate a graph. The first n-1 receptor-ligand bins (those bins representing the longest distances) that match are chosen for graph generation. Features from a given bin are tried at only one stage in the graph generation: thus the features from the second bin will always provide the third node in a graph, the original (bin-defining) pair of points defining the first pair of nodes. All centers are ultimately tried as starting points and all centers within a "longest distance" bin are tried in the generation of the matching graph, unless the graph has been pruned before *any* of the centers in the bin have been tried. Because of the "longest distance" heuristic, not all bins are tried; the method is not exhaustive. With the caveat of this heuristic, however, the method is path

independent . The number of points in each bin determines how many matches will be attempted. In general, the larger the bins the larger the number of orientations generated. The breadth of search is under user control.

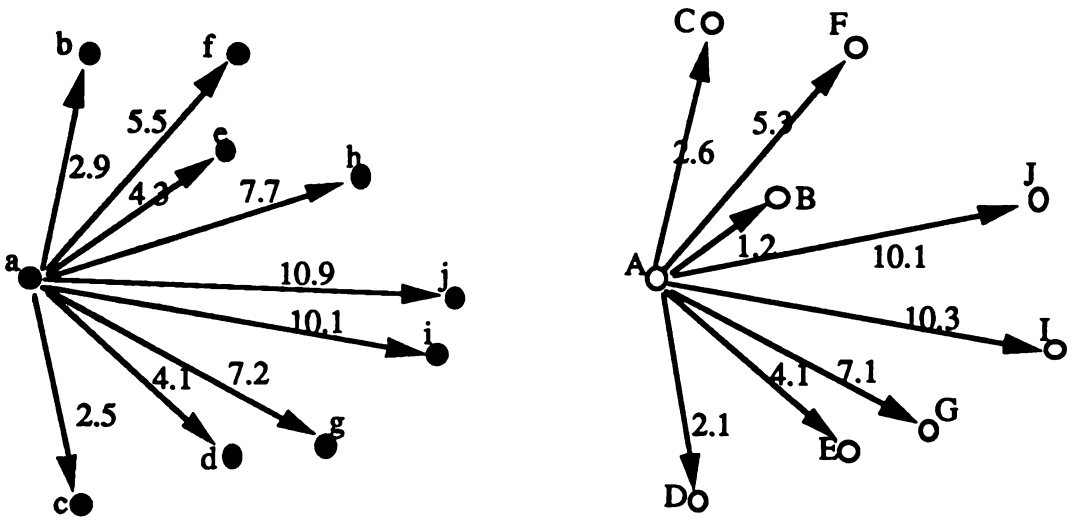


Figure 7 a.

1st Sphere:	a	1st Sphere:	A
10-11 bin:	j i	10-11 bin:	J I
7-8 bin:	h g	7-8 bin:	G
5-6 bin:	f		
4-5 bin:	e d	4-5 bin:	E
2-3 bin:	c b	2-3 bin:	D C
		1-2 bin:	B

Figure 7 b.

Node 1:	a,A
Node 2:	{j or i}, {J or I}
Node 3:	{h or g}, {G}
Node 4:	{e or d}, {E}

Figure 7 c.

Figure 7: Pre-organizing descriptors into bins. a). Descriptor distances from a seed descriptor, ligand "A" and receptor "a" ; all descriptors are used as the seed. b). Histograms of descriptor internal distances. The resolution of the histograms is user-set, in this figure they are 1 Å wide. c). Possible bipartite graphs from 7b. For this example, 16 possible graphs can be constructed (1 (possible match at node one) x 4 (possible matches at node two) x 2 (possible matches at node three) x 2 (possible matches at node 4)).

Three Graph Construction Methods.

We describe three different methods for choosing which internal distances to compare in the construction of the bipartite graphs, all of which use bin matching. All three algorithms begin by pairing a ligand descriptor with a receptor descriptor. All N_l x N_r starting pairs are tried.

Fan Algorithm: Descriptors from protein and ligand are chosen based on their distance from an initial starting descriptor in each molecule. The starting point in this method therefore implicitly defines which region of the molecule will be looked at for matching (Figure 8 a.).

An initial pair of points (a node from *D*) is picked from amongst the set of spheres or atoms describing the molecules. The bins for each molecule are independently generated based on the distances of the remaining points from the initial point. The Fan method uses the first n -1 bins, those with the longest distances from the initial point, that have distance ranges that match the second molecule's bins. These bins provide the molecular features, spheres or atoms, that are used for bipartite graph generation in the matching.

Cat's Cradle Algorithm: Descriptors at a given level of the bipartite graph generation are chosen based on their distance to the descriptor successfully used at the *previous* level of graph generation. As always, we use the longest distance heuristic. The starting point in this method is therefore less important than in the Fan procedure, and since the longest inter-point distances in the molecule will tend to be found regardless of starting point, more of the same nodes will be repeatedly used (Figure 8 b.).

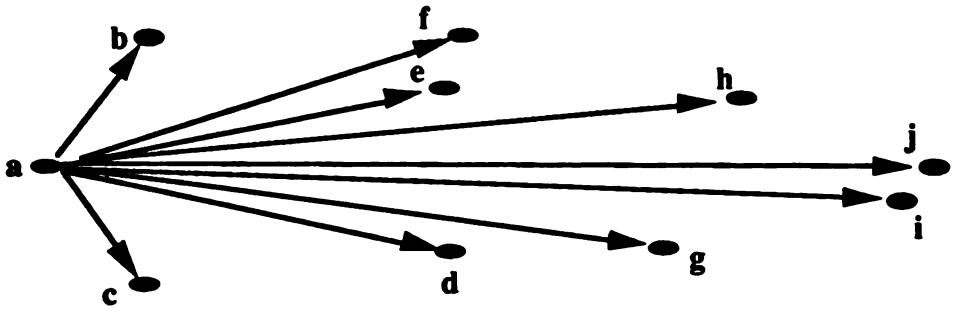


Figure 8 a .

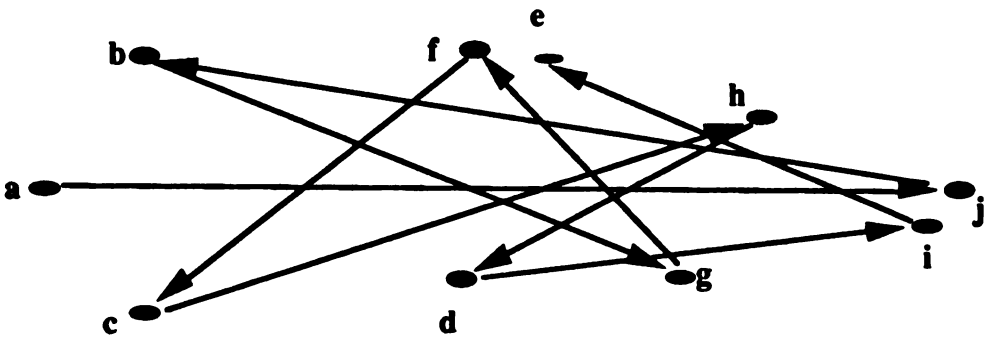


Figure 8 b.

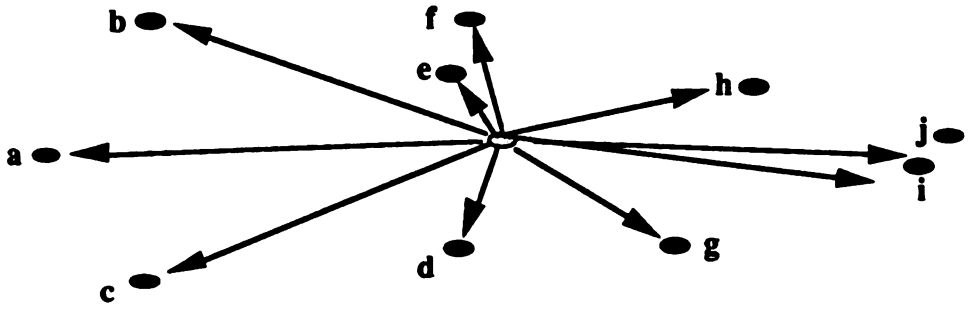


Figure 8 c.

Figure 8: Graph construction methods. a). Fan procedure. Descriptors are chosen based on their distances from an initial descriptor. b). Cat's Cradle procedure. Descriptors are chosen based on their distances from the last descriptor chosen. c). Center of Mass procedure. Descriptors are chosen based on their distances from the center of mass (light shaded circle) of the overall descriptor set.

An initial pair is picked and the bins are generated as in the Fan algorithm. Unlike Fan, only one pair of bins are selected, providing for only the second node in the graph. The next bins, providing the third node possibilities in the graph, are created based on distances from each sphere or atom selected as second nodes. Spheres/atoms in these second generation bins are biased for longest distances. Multiple third bins are similarly created based on distances from spheres or atoms selected from the second bin, and so on.

Center of Mass Algorithm: This method resembles the Fan method, except rather than choosing atoms or spheres based on their distances to the starting pair of points, centers are chosen based on their distances to the center of mass. Except for the starting centers used for the first node, therefore, the centers used in matching will always be the same (Figure 8 c).

An initial pair is chosen. Bins are generated based on distances from the center of mass of the molecule. The algorithm then proceeds as in the Fan algorithm: the bins used are the first n -1 bins from the center of mass that have distance ranges that match a set of bins from the second molecule.

Scoring on a Lattice

We score possible orientations of the ligand in the receptor based on atomic contacts between ligand and receptor structures. We calculate an atomic contact 'potential' for the receptor by constructing a cubic lattice, which fills the volume of the binding site, and evaluate every point on the lattice on the basis of its contacts with the protein atoms. This lattice is usually calculated once for any given site. A point on the lattice receives a score of one for every receptor atom that is within a user-defined

range of distances, and a highly negative score for any contact closer than the low end of this range. We allow the user to distinguish between polar and apolar contacts between the ligand and the receptor atoms by using a second "cutoff" distance parameter (Figure 9). Thus, ligand atoms are often allowed to come closer to receptor oxygen and nitrogen atoms than to other receptor atoms. The cutoff distances are set by the user - for most systems we set the polar close contact cutoff to 2.4 Å and used a range of cutoffs between 2.6 and 2.8 Å for the non-polar close contacts. In the free conformer protein-protein docking runs we set the close contact limit to 2.0 Å for both polar and non-polar contacts. For the lactate dehydrogenase/NAD-lactate runs we used 2.1 Å and 2.3 Å as the polar and non-polar cutoffs. We set the long distance cutoff for scoring a contact to 4.5 Å in all runs. Ligand orientations are scored by mapping their atoms onto the nearest lattice points and summing over all of the mapped points. Only one lattice point is used per ligand atom.

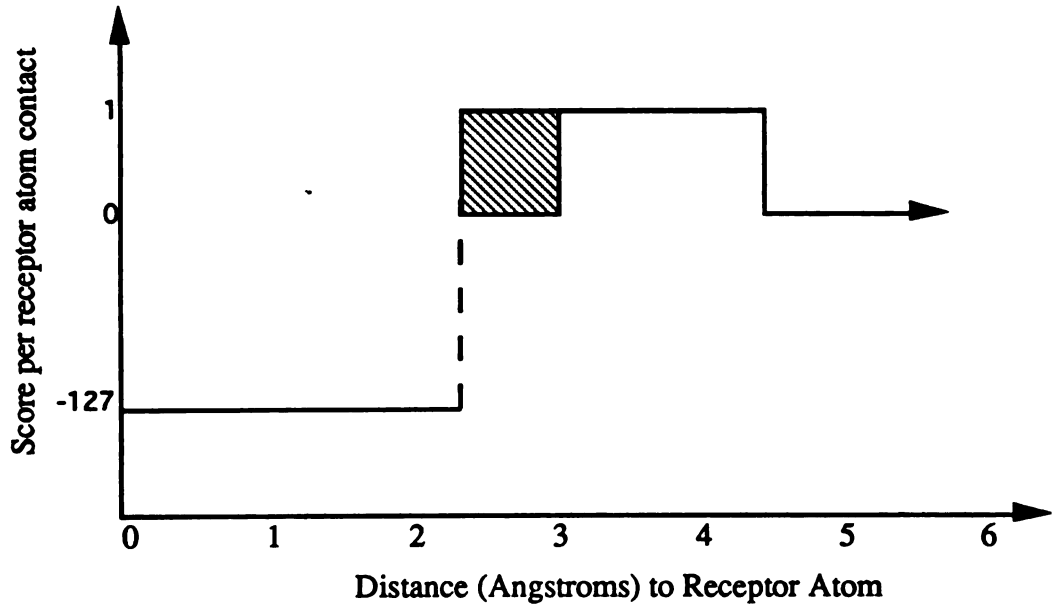


Figure 9: Lattice scoring function. A score of 1 is given to all lattice points within 2.8-4.5 Å of a receptor atom. Lattice points further the 4.5 Å from a receptor atom are give a score of 0. Lattice points closer than 2.4 Å to a receptor atom are considered 'bad contacts' and are given very negative score. Lattice points within 2.4-2.8 Å of a receptor nitrogen or oxygen atom (shaded portion of figure) are given a score of 1, points within 2.4-2.8 Å of all other atom types are considered 'bad contacts' and are given very negative scores.

The lattice-based scoring differs in three ways from the scoring function used in DOCK versions 1.1 and earlier.(DesJarlais et al., 1988) Firstly, the lattice method uses a step function for scoring: a ligand/receptor pair of atoms either contributes a score of 1 or 0 or is a "bad contact," whereas the earlier method used a partly continuous exponential function. Secondly, the lattice method is discontinuous in space since ligand atoms are mapped onto lattice points of some fixed resolution in order to be scored, while the earlier scoring used the pairwise distance for each atom pair to calculate the score of each ligand-receptor configuration. Lastly, the lattice method distinguishes between polar and non-polar contacts, while the earlier method made no distinctions based on atom type.

Sampling and Focusing

To make the sampling procedure as effective as possible for a fixed amount of computer time, we wish to emphasize regions of orientation space where two docking molecules are likely to form productive interfaces, and de-emphasize regions where this is less likely. We begin by sampling orientation space at a low bin resolution, generating a relatively small number of configurations. As the search proceeds, we monitor whether for a particular first pair of spheres/atoms any of the resulting matches produce configurations with positive scores. If any do, the bins for this set of first points are expanded by the contents of the bins immediately below them in the distance ranking and the graph generation loop is continued with these new points. This creates more possibilities for matches in the part of distance space defined by the first pair of successful spheres and atoms. Once a region of orientation space has been examined at this higher level of sampling, the search returns to its former sampling level and proceeds to the next region. More configurations are thus tried in areas that return positive scores than areas that do not. The decision to focus on a region of space from an initial level of general sampling is set dynamically by the program and does not demand human intervention. The user determines only whether to use this feature, and how many bin expansions should be performed.

Hardware

All calculations were done on SGI PI 4D/25's, 4D/70 (Silicon Graphics, Inc., Mountain View, CA) and Sun Sparc (Sun Microsystems, Inc., Mountain View, CA) workstations.

Receptor[a]	Ligand	TYPE[b]	Best Docked (To Crystal Structure, r.m.s.d. Å)[c]	Total Orientations Evaluated in Docking	Run Time (hrs:min)
Trypsin (2ptc)[20]	PTI (2ptc)	Bound	0.29	360,366	1:04
Trypsin (2ptn)[25]	PTI (2ptn)[20]	Free	0.52	9,976,471	27:11
Chymotrypsin (1cho)[23]	Ovomucoid 3rd Domain (1cho)	Bound	0.72	1,650,604	4:19
Chymotrypsin (5cha)[28]	Ovomucoid 3rd Domain (2ovo)[29]	Free	0.82	2,117,929	5:44
Subtilisin (2sni)[22]	Chymotrypsin Inhibitor (2sni)	Bound	0.14	1,511,411	5:30
Subtilisin (1sbc)[26]	Chymotrypsin Inhibitor (2ci2)[27]	Free	0.64	8,615,720	20:23
Thymidylate Synthase Monomer[24]	Thymidylate Synthase Monomer	Bound	0.33	1,886,885	14:08

Table II: Protein-protein docking results (Shoichet and Kuntz, 1991).

Results

Reproduction of Crystallographic Orientations

We were able to reproduce the experimental configuration of the docked molecules accurately and in a timely fashion in all systems (Table II, Table III). In three of the complexes we used inhibitors and receptors in their unbound conformations, those adopted by the molecules when they are crystallized independently of their cognate receptor or inhibitor. This was a stringent test of the methodology owing to the

[a] PDB[16] reference numbers in paranthesis.

[b] Two types of calculations were performed, using the bound (from the crystal complex) or the free (from the uncomplexed crystal structures) conformations of the molecules.

[c] r.m.s.d.'s measured to the crystal complex for bound docking runs and to best-fits to the crystal complex in the free conformer runs.[35]

conformational differences between the bound and the unbound forms of the molecules (Shoichet and Kuntz, 1991).

Protein/ Inhbitor	Search Algorithm	Number of matches[a]	Best rms (Å) to Crystal
Trypsin/PTI	Fan	360,336	0.29
	Cat's Cradle	354,346	0.42
	Center of Mass	224,846	4.56
	Dock1.1[b]	15,453	none found
Dihydrofolate Reductase/ Methotrexate	Fan	5,452	0.35
	Cat's Cradle	11,072	0.99
	Center of Mass	211,155	0.17
	Dock1.1	16,317	0.86
Lactate Dehydrogenase/ Methotrexate	Fan	69,717	1.55
	Cat's Cradle	2,638	1.27
	Center of Mass	386,043	none w/in 5 Å
	Dock1.1	78,526	0.89
Ribonuclease/ Uridine Vanadate	Fan	13,737	0.54
	Cat's Cradle	15,916	0.96
	Center of Mass	20,456	0.70
	Dock1.1	7,955	1.09

Table III: Comparing the search algorithms.

Having established that we can regenerate the crystallographic configurations of the in complexes, we now turn to questions of algorithm performance. We were interested establishing the relative merits of the three graph construction algorithms we tried: the Fan, Cat's Cradle and Center of Mass algorithms. We also wished to know how our molecular-organization and sampling techniques contributed to the accuracy and efficiency of the searches.

[a] The run time is proportional to the number of matches multiplied by the number of atoms in the ligand.
[b] The DOCK1.1 matching algorithm truncates the depth of search of orientation space using heuristics which make it difficult to look at the very high numbers of configurations that are possible using the bin matching algorithms.

Comparison of the Graph Construction Algorithms

The different graph construction methods, Fan, Cat's Cradle and Center of Mass, were tested in four complexes of known structure (Table III), as was an earlier version of DOCK(DesJarlais et al., 1988) (DOCK1.1). In all four cases, both the Fan and Cat's Cradle algorithms were able to reproduce accurately the crystal complex configuration. The Center of Mass algorithm was not able to reproduce the crystallographic configuration of either the lactate dehydrogenase/NAD-lactate(Grau et al., 1981) or the trypsin/PTI(Marquart et al., 1983) complex, though it was able to do so for the dihydrofolate reductase/methotrexate(Bolin et al., 1982) and the ribonuclease/uridine vanadate complexes.(Borah et al., 1985) DOCK1.1 was able to reproduce the crystal complex in the small molecule inhibitor systems, but was not able to do so in trypsin/PTI. The ability to vary the depth of search meant that the Fan and the Cat's Cradle algorithms could always produce lower r.m.s.d. configurations than could DOCK.

The Fan algorithm was usually more efficient than the Cat's Cradle algorithm. The Fan method typically produced distributions of ligand configurations that were biased towards lower r.m.s.d.'s from the crystal structure result, compared to the other two algorithms, and produced a greater number of low r.m.s.d. dockings in shorter searches (Table III). Fan also produced more acceptable orientations as a percentage of the number tried - this ratio ranged from 1/300 for dihydrofolate reductase/methotrexate (1 acceptable orientation for every 300 tried by DOCK2) to 1/1000 for trypsin/PTI, while for the Cat's Cradle procedure the ratios were worse by a factor of three.

Scoring on the Lattice

The new scoring routine improves run times by a factor of 4-5 for large sites (60 or more spheres), compared to the previous scoring method, which explicitly calculated atom-atom contacts between the ligand and the receptor. The ability to distinguish between polar and non-polar contacts significantly improves the ordering of the docked orientations as a function of score compared to the experimental result (Table IV). With polar scoring, more of the top 10 scoring dockings are within 2.5 Å of the crystallographic result than with non-polar scoring. We also notice an improved correlation between scores calculated on a polar lattice and r.m.s.d. from the crystal structure, as compared to scores calculated using a non-polar lattice. We caution, however, that there is no reason to expect even a monotonic relationship between a measure of complementarity and r.m.s.d.. Lattice generation time depends on resolution and the size of the site, but typically takes 1-2 (cpu) minutes on a SGI PI 4D/25.

Receptor	Inhibitor	Polar Lattice[a]		Neutral Lattice[a]	
		Top 10[b]	*R*[c]	Top 10[b]	*R*[c]
Ribonuclease	Uridine Vanadate	4/10	-0.39	0/10	-0.28
Dihydrofolate Reductase	Methotrexate	6/10	-0.59	0/10	-0.23
Lactate Dehydrogenase	NAD-Lactate	9/10	-0.60	2/10	-0.22
Trypsin	PTI	7/10	-0.33	4/10	-0.18

Table IV: Lattice Scoring. We compare the correlation of score with r.m.s.d. from the crystallographic result for "polar" and "non-polar" lattices.

[a] Polar Lattices distinguish between close contacts to receptor oxygen or nitrogen atoms and all other receptor atom types. Neutral lattices treat all receptor contacts equally.

[b] Number of top ten scoring orientations calculated by DOCK2 that have r.m.s.d. values to the cryatallographic result that are less than 2.5 Å.

[c] Correlation between r.m.s.d. from the crystallographic configuration as a function of score. The highest correlation is when R is -1 (high score, low r.m.s.d.)

Sub-Clustering

Sub-clustering segregates molecular features into regions which are treated independently for docking. In trypsin, for example, the initial sphere calculation produced an initial set of 102 spheres, which spanned the active site cleft. Sub-clustering divided this set into several smaller ones, the two largest having 35 and 30 spheres in them. In a similar way, the initial sphere set for the PTI was spread over the entire volume of the molecule and included 292 spheres, approximately as many spheres as there are solvent exposed atoms in the molecule. Sub-clustering produced 6 sub-clusters ranging from 40-90 spheres.

The effect of sub-clustering in the macromolecular docking calculations was dramatic; we show the results for trypsin/PTI in Table V. For a given number of matches - or run time, which is roughly proportional to the number of matches multiplied by the number of ligand atoms - the accuracy of the configurations produced was much higher in the sub-cluster runs than in the full cluster set runs. Even in runs involving extremely large numbers of matches, the full cluster dockings could not find configurations that resembled the crystal structure. Sub-clustering transforms docking from a problem that scales, minimally, as the third-fourth power of the size the molecules, to one that scales more linearly* with molecular size.

---

* It is difficult to determine accurately how the new algorithms scale with molecular size, since parametric choices (such as bin size) in the various systems can have a large effect on run time and the quality of the results. We note (table V) that it took fewer matches (and consequently less time) to arrive at a low r.m.s.d. docking of PTI in trypsin, using the sub-clustering technique, than were required for docking NAD-lactate into lactate dehydrogenase where sub-clusters were not used, even though PTI/trypsin is by far the larger system.

The results of sub-clustering in the small molecule dockings are less convincing. While the technique reduced the run time in most systems, the effect was not as great as in the macromolecular systems; good results could be achieved using the unclustered spheres. Unlike the macromolecular ligand runs, in which sub-clustering was essential to the regeneration of the experimental result, its value in the small ligand systems was case dependent. In dihydrofolate reductase/methotrexate calculations sub-clustering improved run time without sacrificing accuracy. In lactate dehydrogenase/NAD-lactate run time is improved with only a small decrease in accuracy. Also, fewer high r.m.s.d. configurations of ligand were generated. In the ribonuclease/uridine vanadate system, on the other hand, the shorter run time using in the sub-clustered spheres led to lower accuracy and poorer sampling.

Protein Inhbitor	Sub-clustered Spheres				Unclustered Spheres			
	number[a] of clusters	total[b] spheres	total[c] matches	best rmsd to crystal (Å)	number[a] of clusters	total[b] spheres	total[c] matches	best rmsd to crystal (Å)
Trypsin/ PTI	2/3	66/255	137,972	0.30	1/1	102/229	27,099,614[d]	21.7
DHFR/ Methotrexate	2/1	77/33	5,704	0.35	1/1	89/33	10,967	0.35
Ribonuclease/ Uridine Vanadate	2/1	53/20	4,389	1.79	1/1	76/20	10,022	0.54
Lactate De-hydrogenase/ NAD-Lactate	4/3	145/63	62,736	0.74	1/1	189/51	175,280	0.51

Table V: The effects of sub-clustering on run time and accuracy.

Sampling and Focusing

[a] Number of receptor/ligand clusters.

[b] Total number of receptor/ligand spheres in all clusters. Overlaps are permitted between spheres in one sub-cluster and another. Occasionally, this leads to greater totals of sub-clustered spheres than are present in the unclustered sets.

[c] Run time is proportional to number of matches multiplied by the size of the ligand.

[d] This run was not allowed to go to completion, but was stopped after those PTI spheres in the interface region with Trypsin as defined in the crystal structure (approximately 1/5 of the molecule) had been searched. The full search would have required many more matches.

DOCK2 runs that use focusing return more low r.m.s.d. ligand configurations than runs that sample orientation space at a constant level, even though the latter procedure looks at significantly more matches (Figure 10). In four test complexes, focusing increased the ratio of high scoring orientations per match number by a factor of 3-10 (results not shown).

Discussion

Molecular docking searches orientation space for favorable configurations of a ligand in a receptor. Like most search methods with many degrees of freedom, docking can only sample solutions within the space it explores. A docking search will therefore always be faced with a fundamental tradeoff between computation time and accuracy, or more correctly, adequate sampling. We are interested in reducing the time of search necessary to produce a given level of accuracy or sampling. The docking algorithm has three basic levels: molecular description, the sampling of orientation space and the evaluation of configurations. We discuss algorithm modifications at each of these levels and their effect on run time and accuracy. We measure accuracy with reference to the experimental result, the crystal structure of the protein-ligand complex, though we understand that other considerations may also be important.

Accuracy

A basic question for a docking algorithm is how long it takes to get a solution near the experimental structure. In each of our 10 test systems, the new routines reproduced the crystallographic configuration accurately in a reasonable amount of time. This is a compelling result. The test systems we chose varied in their crystallographic resolution (from 2.7 A for lactate dehydrogenase(Grau et al., 1981) to 1.9 A for

trypsin(Marquart et al., 1983)); their size (from the 20 atom uridine vanadate to the 2143 atom thymidylate synthase monomer); and their molecular determinants of binding (the ribonuclease complex relies largely on electrostatic recognition, whereas the macromolecular complexes have large hydrophobic components). Generating known complexes starting with macromolecules in their *unbound* conformations is a striking outcome that suggests that the algorithms might be used predictively.(Shoichet and Kuntz, 1991)

The accuracy of the descriptor based docking calculations reflects the ability of the spheres to identify local binding grooves and ridges. The efficiency of the calculations reflects the success of the sub-clustering and focusing techniques in concentrating the search on regions of orientation space which are likely to have high complementarity. This is the advantage of descriptor based docking over grid searches of receptor sites.(Jiang and Kim, 1991; Wang, 1991) Because grid searches are necessarily unbiased regular samplings of orientation space, they are much slower than DOCK2, which pre-identifies receptor regions of high local curvature to search in. The ability of DOCK2 to dynamically respond to search results through focusing only accentuates this difference. The advantage of the grid methods is that they will always work, given enough time, whereas descriptor-based docking relies on selecting the appropriate features of the molecules, and the avoidance of the combinatorial explosion problem. There will probably be systems which do not lend themselves to description by spheres, such as ones that have a flat protein-protein interface, or that cannot be sub-clustered into independent binding regions. In such systems, DOCK2 will not work, whereas grid based methods will. In the 10 systems we report on in this paper, however, DOCK2 can generate accurate reproduction of the crystallographic configuration in minutes or hours on a work station. Methods using

grid searches of orientation space to solve the docking problem can take days on much faster machines.(Jiang and Kim, 1991; Wang, 1991)

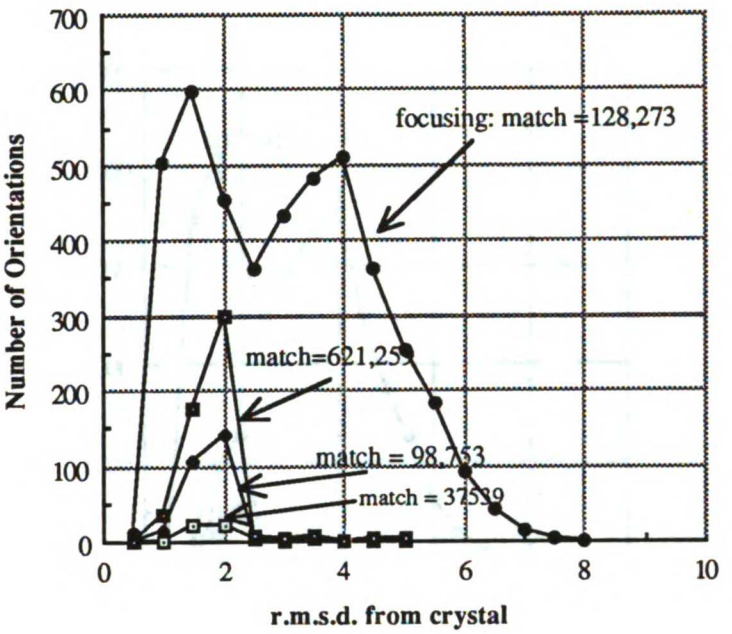
Choosing Between the Searching Algorithms

We tried three different methods for choosing which features of one molecule to map onto the those of a second. Both the Fan and Cat's Cradle algorithms reproduced the experimental results in all systems, while the Center of Mass algorithm did so in only two of four complexes it was tested against. The Center of Mass method probably fails because of the relatively few spheres it uses as matching descriptors. Since the Center of Mass matching chooses centers for bipartite graph generation based on a fixed reference point, fewer aspects of the molecule or site will be sampled in graph generation, compared with the Fan or Cat's Cradle procedures where the reference point is different for each first pair of spheres. Choosing between the Fan and Cat's Cradle algorithms is more difficult on theoretical grounds. The Cat's Cradle algorithm will more often sample the longest internal distance of a molecule or site while building the bipartite graph, and will therefore more consistently use the principal topographic features of the molecules in matching. The Fan algorithm, on the other hand, will generally sample more of the features of a molecule or site. Practically, the Fan method seems to perform more efficiently than the Cat's Cradle method, though this result might reflect our implementation and should be tested for other systems.

Scoring on the Lattice

The improvement in run time with lattice scoring more than justifies its decreased resolution compared to scoring in a continuous space. Though the scoring scheme used in the lattice implementation is simpler than in the previous versions of the program,(DesJarlais et al., 1988) scores from the two methods correlate well with

each other (results not shown). The exact numerical score for any given orientation will, of course, differ between the two metrics, as described in the Methods. Since there is no good physical reason to choose one scoring function over the other, we used the simpler function in this work. The introduction of polar differentiation in the scoring scheme improves the correlation between a configuration's score and its similarity to the crystallographic result compared to non-polar scoring. Such a correlation must, however, be interpreted cautiously. While it is gratifying that the highest scoring configurations in our test cases closely resemble the crystallographic result, we note that there are often configurations whose scores are almost as high that do not resemble it. This is most apparent in the dockings of the unbound conformations of the protease/protease-inhibitor pairs. The shape-based scoring is potentially weakest when comparing the complementarity of different putative ligands for the same receptor, which is what is done in inhibitor design applications of DOCK.(DesJarlais et al., 1990) While the method continues to prove itself useful in the design of novel inhibitors(DesJarlais et al., 1990) (Shoichet, unpublished results; Bodian, unpublished results), rankings of molecules based on their DOCK score should not be over-interpreted.


700
600
500
400
300
200
100
0
Number of Orientations
focusing: match =128,273
match=621,259
match = 98,753
match = 37539
0
2
4
6
8
10
r.m.s.d. from crystal

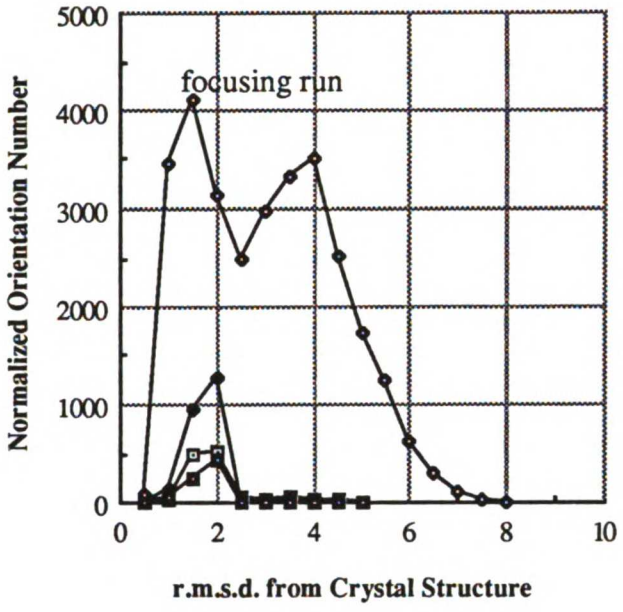


Figure 10: Focusing in trypsin/PTI. Open squares represent a small bin, low match number docking run, the closed triangles represent an intermediate match number run, the closed squares represent the maximum bin size and match number run. The closed circles represent a run with the same bin sizes as the open squares, but with focusing. a). The number of orientations generated in a docking run plotted against the r.m.s.d. of the orientations compared, to the crystallographic result. b). The same data is presented, normalized for match number.

Sub-Clustering

The fundamental change in our approach that allows us to treat macromolecular docking is our modification of the clustering algorithm. The introduction of a radial cutoff organizes and segregates molecular features into topographically distinguishable regions. By reducing the size of each cluster, we overcome the strong time dependence of docking with system size. We assume that two regions that are distinguishable are also independent. The method should be evaluated by two criteria: does it genuinely separate a molecule into physically distinct regions and does it increase the efficiency of the search without compromising its accuracy?

In the protein-protein complexes, for which the sub-clustering technique is most important and useful, the issue of how sub-clusters correspond to physical regions of the molecule can be addressed by organizing the residues of a protein along structural and functional lines. Residues of PTI, for instance, can be assigned a role either in stabilizing the tertiary fold of the molecule or in binding to trypsin(Read and James, 1986) (Figure 11 a.). Comparing PTI(Marquart et al., 1983) organized by sub-clusters (Figure 11 b.) to the structure/function organization of the molecule, one notices that the sub-clusters correspond to either structural or functional regions. The binding loop residues in Figure 11 a. are completely described by cluster 3 in Figure 11 b. The hydrophobic pocket residues in Figure 11 a. are found in clusters 1 and 2, which divide the non-binding part of PTI between them. The binding loop cluster has few overlaps with the hydrophobic regions of the molecule. Sub-clustering thus seems to do a good job of separating PTI into physically and functionally distinct regions.

A B

Figure 11: PTI residues organized by structure/function,(Read and James, 1986) and by sub-clustering. a) The structure/function organization of the molecule: the specificity loop of PTI is in green, residues important for the tertiary fold of the molecule are colored blue and the rest of the residues are in orange. b) The sub-cluster organization of the molecule. Sphere descriptors are represented by triangles. Cluster 3 is in green and clusters 1 and 2 are in orange and magenta, respectively.

Clustering improved the efficiency of the docking runs dramatically without sacrificing accuracy in all of the protein-protein complexes we tested. The results for PTI/trypsin with and without sub-clustering are compared in Table V. Without sub-clustering, macromolecular docking is not practical for our method.

The results for the small molecule test cases are less persuasive. While sub-clustering still shortened run times, the unclustered spheres always reproduced the crystallographic results in reasonable amounts of time. In lactate dehydrogenase/NAD, for example, sub-clustering meant performing 12 different docking runs, for all combinations of ligand and receptor clusters. While this did improve run times and lead to better distributions of the docked orientations around the crystallographic result (Table V), the unclustered sphere sets clearly provided an adequate description of the molecules. The different impacts of sub-clustering on the macromolecular ligand and small ligand systems might reduce to a question of size. The small molecule ligands, and their cognate receptor binding grooves, simply lack of the heterogeneous topologies that make sub-clustering necessary for the macromolecular systems. Having said this, sub-clustering *does* reduce search times in the small ligand systems, and will be a useful technique whenever one wants to target specific regions of a site for a docking search or when docking to a large receptor site.

Sampling and Focusing

In molecular docking, it is important to be able to survey the general features of configuration space, and then concentrate on those areas that offer the greatest possibilities for complementarity. The focusing algorithm guides a search by

expanding the number of molecular descriptors in regions of distance space that return favorable orientations, leading to longer searches in these regions than in regions that do not return favorable matches at the initial low density sampling. The technique is essentially a variation on the tree-search-with-pruning approach: rather than cutting off branches due to a failure of an early node, branching is increased due to an early success.

Focusing improves the efficiency of a search of orientation space. Both the number of low r.m.s.d. configurations and the number of high scoring configurations per number of matches increase dramatically with focusing, which allows for faster searches to achieve the same degree of accuracy (Figure 10). Focusing is most effective in protein-protein complexes. In a search of trypsin/PTI using small bins (low initial sampling levels) and high sensitivity to focusing signals, the number of matches necessary to achieve either a high score or a low r.m.s.d. value to the crystallographic result was reduced by a factor of 100 compared to a run where focusing was not done (Figure 12). The non-focusing run shows a monotonic increase in the best score or r.m.s.d. result achieved with the level of sampling, up to a maximum value. This is the expected result for a discrete, unbiased sampling of configuration space. The focusing run, on the other hand, achieves a result very close to the maximum almost immediately and improves only slightly as more and more orientations are looked at.

One caveat for focusing is that it often increases the number of high scoring orientations that are distant from the crystal configuration, as well as increasing the number of orientations that are close to the crystallographic result (Figure 10). This reflects the different spaces in which orientations are first generated and then evaluated. In generating a ligand-receptor configuration, we match internal distances between molecular descriptors. Focusing increases the number of descriptors to

match in a particular region of *distance* space. Two points that are the same distance to a third point will not necessarily be close to one another in Cartesian space, and focusing based on distance information can therefore lead to the inclusion of descriptors from a different region of the molecule than the one that was involved in the initial, low-level sampling match. This explains the broadening of the r.m.s.d. distributions, which are measured in Cartesian space. Notwithstanding this feature, focusing always increases the number of high-scoring configurations as a percentage of matches tried. Since the signal to focus on a particular region is score-based, this is perhaps a more consistent metric.

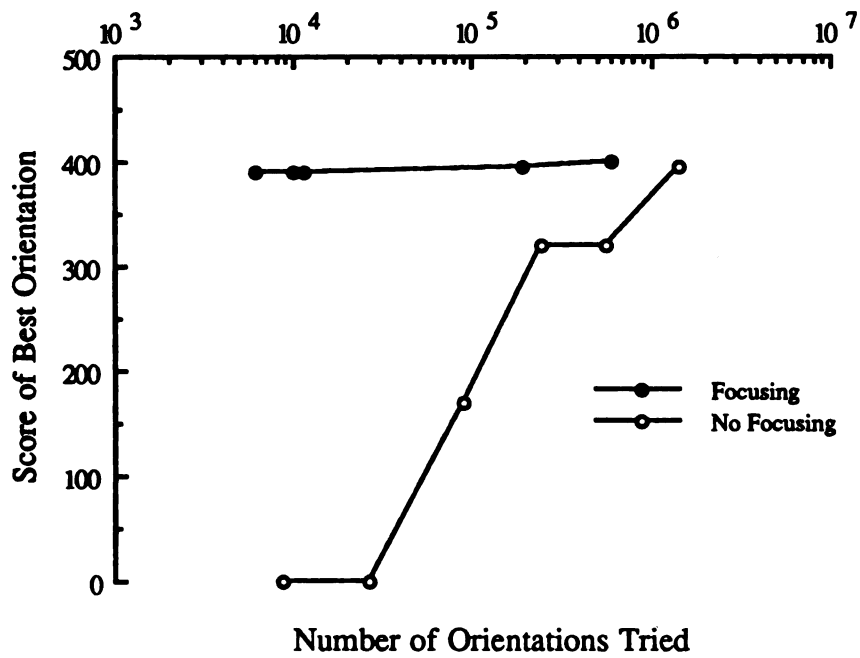


Figure 12: Sampling issues in focusing. Maximum score achieved versus number of orientations tried. Solid circles: using focusing procedure. Empty circles: no focusing.

Sampling configuration space at variable densities is a physically sound approach to a problem that can have an infinite number of solutions. We have outlined a procedure for focusing that meshes easily with our docking algorithm - other methods are certainly conceivable. The general approach is not limited to molecular docking, but should be useful in any method where low density sampling can guide high resolution searches. Such methods might include docking on a regular lattice(Wodak and Janin, 1978; Jiang and Kim, 1991) or in grid searches of conformation space, where smaller step sizes (step size here being an torsion angle, an Euler angle or a translation) would be used in low energy regions of the energy surface, and larger step sizes in high energy regions. Guided searches are implicitly implemented in Monte Carlo methods for simulating molecular dynamics,(Metropolis et al., 1953) though here it is not step size but rather time spent in a particular region of space that is modified, so that the analogy is only approximate.

Unsolved Problems

Using low resolution representations of molecules, either in the form of potential functions or topography, to guide searches of molecular interactions is a general problem in the field.(Leicester et al., 1988; van Gunsteren and Berendsen, 1990) In this paper we have shown how methods for organizing high resolution information can be used to address this issue. Both sub-clustering and focusing do not, however, use actual low resolution information as a guide. It would be conceptually appealing, and practically rewarding, to use low resolution information to prune the search tree. This would allow one to limit searches that use high resolution features of the molecules, which are the most expensive computationally, to regions of likely complementarity. To our knowledge, the general problem of using resolution to guide searches remains largely unaddressed.

The scoring function that DOCK2 uses to evaluate configurations, though improved from the one used in DOCK, is still too simplistic. Our concern previously had been that a more complex scoring function, of the sort used in molecular mechanics for instance, could only be used at the cost of reducing the amount of orientations we could look at. This should not be the case for a lattice-based scoring method, however.(Goodsell and Olson, 1990)

Finally, we have not discussed the issue of conformational flexibility. The degrees of freedom in a docking problem that allows for conformational as well as configurational sampling are potentially very large, which implies either that searching such a space would be very slow or very incomplete or both. There are ways to reduce the degrees of freedom of this problem. If one defines local regions of space as interaction zones, and only look at conformations in this region while keeping the rest of the system rigid then the issue becomes more tractable.(Ponder and Richards, 1987; Wilson et al., 1991) Alternatively, if one keeps the protein rigid and only allows the generally much smaller ligand to sample conformation space, the size of the space is similarly reduced.(Goodsell and Olson, 1990) This method suffers in situations where conformational accommodation takes place largely at the receptor, which some have argued is the general case.(Warshel and Levitt, 1976)

Applications

The modifications encoded in DOCK2 improve our ability to model biological systems (Shoichet and Kuntz, 1991) and design novel enzyme inhibitors (DesJarlais et al., 1990) (Shoichet, unpublished results; Bodian, unpublished results). The changes in matching algorithm give the user more control over the depth of search in docking

calculations, while the lattice scoring makes the program faster and leads to more sensible evaluations of orientations. The sub-clustering technique will be useful in systems where one wishes to explore particular regions of a molecule in detail, while down-playing others. Sub-clustering will also significantly improve run time efficiency in large sites, and the technique is essential for the docking of two macromolecules, an area of current pharmaceutical interest. The focusing technique will improve the efficiency of a search, judged by the number of complementary orientations per number of matches tried. Focusing will be especially useful when highly detailed searches of a particular regions are required, as will be the case when trying to reproduce or predict a biological complex, or when trying to capture particular details of a binding interaction.

Conclusions

DOCK2 can reproduce the experimental configurations of protein-ligand complexes in a wide variety of systems. We have shown how docking can be changed from a problem that scales as the fourth power of system size, to one that scales linearly with it. This is achieved by breaking down the description of the molecules into independent pieces, and by concentrating high resolution searches of orientation space on those regions that return favorable complexes at low resolution.

The success of any feature based docking scheme depends on how descriptors are chosen for matching between molecules. We have found that the Fan and Cat's Cradle internal distance algorithms work well, while the Center of Mass method does not. We have described new routines allow for variable depth searches and more efficient scoring of orientations. Compared to our previous implementations,(Kuntz et al., 1982; DesJarlais et al., 1988) the current methods allow faster and more complete

searches of orientation space. The algorithm is well suited to macromolecular docking, which the earlier methods were unable to treat.

Acknowledgements: We wish to thank Elaine Meng, Renee DesJarlais, Richard Lewis and Andrew Leach for their helpful comments throughout this work. Financial support was provided by the National Institutes of Health (GM-31497, GM-39552, and 5T32 GM08120 for DLB). We used the graphics facilities of the UCSF computer graphics laboratory, R. Langridge, director (RR-1081), for some of this work.

References

Bernstein, F. C., Koetzle, T. F., Williams, G. J. B., Meyer, E. F., Jr., Brice, M. D., Rodgers, J. R., Kennard, O., Shimanouchi, T. and Tasumi, M., *J. Mol. Biol.* **112**, 535-542 (1977)

Blevins, R. A. and Tulinsky, A., *J. Biol. Chem.* **260**, 4264 (1985)

Bode, W., Epp, O., Huber, R., Laskowski, M. and Ardelt, W. J., *Eur. J. Biochem.* **147**, 387 (1985)

Bolin, J. T., Filman, D. J., Matthews, D. A., Hamlin, R. C. and Kraut, J., *J. Biol. Chem.* **257**, 13650-62 (1982)

Borah, B., Chen, C. W., Egan, W., Miller, M. and Wlodawer, A., *Biochem.* **24**, 2058-2066 (1985)

Connolly, M. L., *Science* **221**, 709-713 (1983)

Connolly, M. L., *Biopolymers* **25**, 1229-1247 (1985)

Crick, F. H. C., *Acta. Crystallogr.* **6**, 689-697 (1953)

DesJarlais, R., Sheridan, R. P., Seibel, G. L., Dixon, J. S., Kuntz, I. D. and Venkataraghavan, R., *J. Med. Chem.* **31**, 722-729 (1988)

DesJarlais, R. L., Seibel, G. L., Kuntz, I. D., Montellano, P. O. d., Furth, P. S., Alvarez, J. C., DeCamp, D. L., Babé, L. M. and Craik, C. S., *Proc. Natl. Acad. Sci. USA* **87**, 6644-6648 (1990)

DesJarlais, R. L., Sheridan, R. P., Dixon, J. S., Kuntz, I. D. and Venkataraghavan, R., *J. Med. Chem.* **29**, 2149-2153 (1986)

Ehrlich, P., *Chem. Berichte* **42**, 17 (1907)

Ferrin, T. E., Huang, C. C., Jarvis, L. E. and Langridge, R., *J. Mol. Graph.* **6**, 13-27 (1988)

Ferro, D. R. and Hermans, J., *Acta Cryst.* **A33**, 345-347 (1977)

Fujinaga, M., Sielecki, A. R., Read, R. J., Ardent, W., Laskowski, M. J. and James, M. N. G., *J. Mol. Biol.* **195**, 397 (1987)

Goodford, P. J., *J. Med. Chem.* **27**, 557-564 (1984)

Goodford, P. J., *J. Med. Chem.* **28**, 849-857 (1985)

Goodsell, D. S. and Olson, A. J., *Proteins* **8**, 195-202 (1990)

Grau, U. M., Trommer, W. E. and Rossmann, M. G., *J. Mol. Biol.* **151**, 289-307 (1981)

Hartigan, J. A. (1975). Clustering Algorithms. New York, Wiley.

Jiang, F. and Kim, S. H., *J. Mol. Biol.* **201**, 79-102 (1991)

Kuhl, F. S., Crippen, G. M. and Friesen, D. K., *J. Comp. Chem.* **5**, 24 (1984)

Kuntz, I. D., Blaney, J. M., Oatley, S. J., Langridge, R. and Ferrin, T. E., *J. Mol. Biol.* **161**, 269-288 (1982)

Lee, R. H. and Rose, G. D., *Biopolymers* **24**, 1613-1627 (1985)

Leicester, S. E., Finney, J. L. and Bywater, R., *J. Mol. Graphics* **6**, 104-108 (1988)

Marquart, M., Walter, J., Deisenhofer, J., Bode, W. and Huber, R., *Acta Crystallogr., Sect. B* **39**, 480 (1983)

McPhalen, C. A. and James, M. N. G., *Biochem.* **26**, 261-269 (1987)

McPhalen, C. A. and James, M. N. G., *Biochem.* **27**, 6582-6598 (1988)

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H. and Teller, E., *J. Chem. Phys.* **21**, 1087-1092 (1953)

Miller, D. L. and Pekny, J. F., *Science* **251**, 754-761 (1991)

Montfort, W. R., Perry, K. M., Fauman, E. B., Finer-Moore, J. S., Maley, G. F., Hardy, L., Maley, F. and Stroud, R. M., *Biochem.* **29**, 6964-6976 (1990)

Neidhart, J. J. and Petsko, G. A., *Protein Eng.* **2**, 271-276 (1988)

Ponder, J. W. and Richards, F. M., *J. Mol. Biol.* **193**, 775-91 (1987)

Read, J. R. and James, M. N. G. (1986). Introduction to the protein inhibitors: X-ray crystallography. Proteinase Inhibitors. Amsterdam, Elsevier. 301-335.

Rose, G. D., *Nature* **272**, 586-590 (1978)

Shoichet, B. and Kuntz, I. D., *J. Mol. Biol.* **221**, 327-346 (1991)

van Gunsteren, W. F. and Berendsen, H. J. C., *Angew. Chem. Int. Ed. Engl.* **29**, 992-1023 (1990)

Walter, J., Steigemann, W., Singh, T. P., Bartunik, H., Bode, W. and Huber, R., *Acta Crystallogr., Sect. B* **38**, 1462 (1982)

Wang, H., *J. Comp. Chem.* **12**, 746-750 (1991)

Warshel, A. and Levitt, M., *J. Mol. Biol.* **103**, 227-249 (1976)

Wilson, C., Mace, J. E. and Agard, D. A., *J. Mol. Biol.* **220**, 495-506 (1991)

Wilson, R. J. (1985). Introduction to Graph Theory. Harlow, UK, Longman.

Wodak, S. J., De Crombrugghe, M. and Janin, J., *Prog. Biophys. molec. Biol.* **49**, 29-63 (1987)

Wodak, S. J. and Janin, J., *J. Mol. Biol.* **124**, 323-342 (1978)

Zielenkiewicz, P. and Andrzej, R., *J. theor. Biol.* **111**, 17-30 (1984)

Gloss to Chapter III

Chapter III describes our work on "the protein docking problem," and is a direct off-shoot of Chapter II. The work had three interesting aspects. Technically, it showed that the docking approach could treat complicated molecular interfaces. Predictively, we showed that we could dock proteins in their bound and free conformations, which suggests that many of the determinants of binding reside in those side chain and main chain atoms of the molecules that remain invariant on binding. Lastly, the inability of the 'first approximation' computational methods to reliably distinguish between native and non-native like complexes, generated by DOCK, suggests that the interactions that distinguish these configurations in nature are more subtle than is commonly supposed. To be able to simulate them computationally would certainly have required more sophisticated modeling - as for instance dynamics simulations using molecular water - though even with such techniques I doubt that we would have done much better. What we can definitively say is that, based on the work in Chapter III, we should not expect DOCK, which now and in the foreseable future will be restricted to some version of the 'first approximation' evaluations of chemical complementarity, to be able to reliably distinguish a micromolar inhibitor (a very exciting hit) from a millimolar (an uninteresting hit) inhibitor, beginning with the unbound conformations of receptor and ligand. All that we can hope for is that DOCK will allow us to derive a list of likely candidates. This at least has been my experience (chapter four).

Joel Janin, one of the first persons to tackle the protein docking problem, has provided a nice post-script to this work. In a letter to Tack dated 10/18/91, Janin wrote:

> Please find enclosed a paper soon to appear in *Proteins*. I just read your work in *J. Mol. Biol.*,* and was struck by the similar conclusions

* Where Chapter III was published, BKS.

> reached with different algorithms. In my opinion, they illustrate the fact that we still do not understand specific recognition....

It's nice to be reminded that the same issues can excite other scientists, and that they can draw the same conclusions when confronted with the same phenomena, even when it is unexpected, even disturbing.

Protein Docking and Complementarity

Brian K. Shoichet and Irwin D. Kuntz *

Department of Pharmaceutical Chemistry

University of California, San Francisco

San Francisco, California, 94143-0446

* to whom correspondence should be addressed

Key Words: docking, proteins, complementarity, structure prediction

Running Title: Protein Docking

Abstract

Predicting the structures of protein-protein complexes is a difficult problem owing to the topographical and thermodynamic complexity of these structures. Past efforts in this area have focussed on fitting the interacting proteins together using rigid-body searches, usually with the conformations of the proteins as they occur in crystal structure complexes. Here we present work that uses a rigid-body docking method to generate the structures of three known protein complexes, using both the bound and unbound conformations of the interacting molecules. In all cases we can regenerate the geometry of the crystal complexes to high accuracy. We also are able to find geometries that do not resemble the crystal structure, but nevertheless are surprisingly reasonable both mechanistically and by some simple physical criteria. In contrast to the previous work in this area, we find that simple methods for evaluating the complementarity at the protein-protein interface cannot distinguish between the configurations that resemble the crystal structure complex and those that do not. Methods that could not distinguish between such similar and dissimilar configurations include: surface area burial, solvation free energy, packing and mechanism-based filtering. Evaluations of the total interaction energy and the electrostatic interaction energy of the complexes were somewhat better. Of the techniques we tried, energy minimization distinguished most clearly between the 'true' and 'false' positives, though even here the energy differences were surprisingly small. We found the lowest total interaction energy from amongst all of the putative complexes generated by docking was always within 5 Å RMS of the crystallographic structure. There were, however, several putative complexes that were very dissimilar to the crystallographic structure but that had energies that were close to that of the low energy structure.

The magnitude of the error in energy calculations has not been established in macromolecular systems, and thus the reliability of the small differences in energy remains to be determined. The ability of this docking method to regenerate the crystallographic configurations of the interacting proteins using their unbound conformations suggests that it will be useful tool in predicting the structures of unsolved complexes.

Introduction

An interesting problem in structural biochemistry is that of protein-protein recognition. Proteins are thought to associate in specific ways involving numerous detailed interactions. While the general principles of protein association seem fairly well understood (Blundell 1981; Hendrickson *et al.*. 1987) the application of these principles to prediction of specificities is not. Thus, the thermodynamic basis for protein association is considered to be the product of several opposing energies such as solvent organization, conformational entropy, van der Waals and electrostatic interactions (Hendrickson *et al.*., 1987). Any one term, however, may be greater in magnitude than the resulting sum, while none of them can be predicted to a high degree of accuracy. Algorithms for predicting how two proteins will associate must treat two problems: that of finding reasonable configurations of association, and that of evaluating the free energies of the putative complexes. The first problem is difficult because proteins are large, complex objects that have many possible interfaces, while the second is difficult due to the inaccuracies and imprecisions of the current theoretical methods of energy evaluation.

Connolly (Connolly 1985) has posed the "protein-docking problem" as: "given the structures of any two proteins, is it possible to predict whether they associate, and if so, in what way." Most workers in the field of protein-docking (Connolly 1985; Lee & Rose 1985; Wodak *et al.*,1987; Wodak & Janin 1978; Zielenkiewicz & Andrzej 1984) have concentrated on reproducing the configurations of protein complexes as they occur in crystal structures and have thereby imposed two simplifications. The proteins are assumed to associate, which reduces the thermodynamic problem of evaluating the

equilibrium between solvated and complexed structures to one of evaluating the equilibrium between different possible configurations of the complex. Moreover, by working with the bound conformations of the proteins (as they occur in the crystal structure of the complex) one avoids the formidable problem of conformational searching. This allows for the reduction of the degrees of freedom in the protein docking problem from several thousand to six - the rotational/translational degrees of freedom of two rigid bodies. Such a simplification is not always be justifiable; nevertheless the reproduction of the crystal structure configurations is a logical first step towards being able to actually *predict* the structures of protein complexes.

Using rigid body approaches, workers (Connolly 1985; Wodak & Janin 1978; Wodak *et al.*, 1987) have reported the reconstitution of crystal complexes along with a number of configurations that do not resemble the crystal structure. Using elementary energy evaluation schemes, they have reported being able to distinguish between the geometries resembling the crystal structure and those that do not, though in their earlier work (Wodak & Janin 1978) Wodak and Janin report some ambiguities in their evaluations of docked complexes using reduced resolution models. Given the simplicity of the evaluation methods used - buried surface area (Connolly 1985; Wodak & Janin 1978; Zielenkiewicz & Andrzej 1984) along with mechanistic constraints (Wodak *et al.*, 1987) - the ability to distinguish between the configurations similar to the crystal structure and those that are distant from it is encouraging. A possible qualification of this result might be that most of the previous work has been conducted with the bound conformations of the associating molecules, which may restrict the possibilities of 'reasonable' configurations to those near that of the crystal complex. Alternatively, it is possible that too few 'incorrect' configurations were sampled to stringently test the methods. Overall, however, the impression from the

literature is that the energy evaluation part of the protein-docking problem is computationally tractable.

In this paper, we attempt to extend previous work with a rigid body docking method that is capable of generating a large number of possible configurations of two proteins. We apply the docking method to the crystal structures of the molecules in their bound and unbound conformations. This approach allows us to address some of the conformational issues unresolved in the earlier work and provides a more stringent test of both the energy evaluation schemes and the docking method that we employ. Within this context we use the same simplifications to the protein-docking problem as previous workers, namely we know that the proteins being docked associate and that to a first approximation we assume them to be conformationally rigid. To distinguish amongst the different configurations, we apply some of the commonly accepted computational methods of energy evaluation: buried surface area, free energies of solvation, mechanistic constraints, packing, electrostatic complementarity and energy minimization using molecular mechanics. This is by no means an exhaustive list. The methods we use here are meant to be representative of the common first approximation methods (see for instance: Eisenberg & McLachlan, 1988; Gregoret & Cohen, 1990; Novotny *et al.*., 1988; Connolly, 1985; Gilson & Honig, 1988 and Wodak *et al.*, 1987).

We report work on three sets of proteases and their inhibitors: Trypsin with Bovine Pancreatic Trypsin Inhibitor (BPTI)* , Chymotrypsin with Ovomucoid Third Domain and Subtilisin with Chymotrypsin Inhibitor 2 (table I). Our principal basis for choosing these

* Abbreviations: BPTI, Bovine Pancreatic Trypsin Inhibitor; RMS, root mean square; structure abbreviations as in table I.

Protein	PDB [a] Name	This Paper Name	State[b]	Cα RMS to Bound Form[c] (Å)	High RMS Residues[d] (from bound conformation)	High B-factor Atoms Not Included in DOCK Scoring
Trypsin	2PTC[e]	PTC	Bound	0.00		
BPTI	2PTC	PTI	Bound	0.00		
Trypsin	2PTN[f]	2PTN	Free	0.34	N192, K224, K239	
BPTI	4PTI[g]	4PTI	Free	0.56	K15, R17, R39,R42, G57, A58	R17: cd,ne,cz,nh1,nh1
Subtilisin	2SNI[h]	SNI	Bound	0.00		
Chymotrypsin Inhibitor	2SNI	CI	Bound	0.00		
Subtilisin	1SBC[i]	1SBC	Free	0.44	S161	
Chymotrypsin Inhibitor	2CI2[j]	2CI2	Free	0.46	L20, M59, R62, D64, K72, L73	N19; L20; E33:oe1,oe2; K37:cd,ne,nh1,nh2; Q41::oe, ne; T41:cb,cg,og1M59:o,cb,cg,sd,ce R81:cz,nh1,nh2
Chymotrypsin	1CHO[k]	CHO	Bound	0.00		
Ovomucoid 3rd Domain	1CHO	OVO	Bound	0.00		
Chymotrypsin	5CHA[l]	5CHA	Free	0.47	N7, F39, E70, K90, R145,'K170	
Ovomucoid 3rd Domain	2OVO[m]	2OVO	Free	1.16	V4, E10, M18, R21, K29, N45, K55	R21: cd,ne,cz,nh1,nh2

Table I: Models Used

systems was that the structures of both proteins in each set have been solved to good resolution in the bound and unbound forms. While all three systems are protease/protease-inhibitor complexes, we note that the structures of the inhibitors differ significantly from one another, as do two of the proteases. We will show that in

---

[a] All structures taken from the Protein Data Bank (Bernstein et al. 1977)
[b] Bound: structure from the protease-inhibitor crystal complex; Free: structure from the crystal structure of the single molecule.
[c] Residual r.m.s. from best-fit of Ca of free structure on bound structure.
[d] RMS of 1.5 Å or greater from the residue's conformation in the bound form.
[e] 2PTC (Marquart et al. 1983)
[f] 2PTN (Walter et al. 1982)
[g] 4PTI (Marquart et al. 1983)
[h] 2SNI (McPhalen and James 1988)
[i] 1SBC (Neidhart and Petsko 1988)
[j] 2CI2 (McPhalen and James 1987)
[k] 1CHO (Fujinaga et al. 1987)
[l] 5CHA 5CHA (Blevins and Tulinsky 1985)
[m] 2OVO (Bode et al. 1985)

all cases we can search orientation space and regenerate the crystallographic configurations in a timely and accurate fashion. We also generate configurations that we assume to be 'false positives' - that is, configurations that are quite different from those seen in the crystallographic experiments . Most of the methods of energy-complementarity evaluation that we used were not able to distinguish between those putative protease/protease-inhibitor complexes that are close to the configuration in the crystal complex structure and those that are distant from it. In two of the three systems we investigated, evaluations of electrostatic interaction energy using the program DELPHI (Gilson & Honig 1988) was able to distinguish between 'true' and 'false' positive dockings, but was not able to do so in the third system. Calculations of the total interaction energy of the complexes using the program AMBER (Weiner *et al.*,1984) consistently found the lowest energy docked complex to be within 5 Å of the crystallographically determined solution. None of the methods we tried could reliably distinguish amongst the docked configurations that were within 5 Å of the crystal complex. Even energy minimization yielded low energies for the ' false' positive configurations.

Approach

Rigid body docking of proteins to explore protein-protein association simplifies a problem with thousands of possible degrees of freedom to one of six degrees of freedom. Even with this simplification, however, the problem remains difficult. The number of possible orientations for two molecules as complex in their shapes as proteins is so large as to make brute force methods, looking at every distinguishable orientation, impractical (Connolly 1985). The problem of docking molecules of any complexity based on the complementarity of their features has been shown to be NP-Complete (Kuhl *et al.*,1984). It is essential, therefore, to use a directed approach that

focuses on areas of likely complementarity between the two molecules rather than sampling orientation space at random or by use of a regular grid. Our method uses geometric descriptions of local clefts and bumps on protein surfaces to guide its search of orientation space, the notion being to find orientations that map the bumps into the clefts. Even in this restricted space we can only sample amongst the possible orientations. Once an orientation is found, it is subjected to a fast preliminary evaluation of complementarity based on a simple examination of atomic contacts between the enzyme and the inhibitor. Most orientations are quickly eliminated by this filter owing to unacceptably close contacts between inhibitor atoms and enzyme atoms. Finally, we use more detailed energy evaluation methods to try to distinguish amongst the orientations that are not eliminated on the basis of contacts.

Methods

The basic method has been described elsewhere (Kuntz *et al.*. 1982; Shoichet *et al.*, 1992) as have several applications relating to protein/small-molecule docking (Desjarlais *et al.*. 1986; DesJarlais *et al.*. 1988), and will only be summarized here except where we differ from the previous work. In the first step, we replace the atoms of the ligands and the receptors with sets of spheres (see next paragraph). The spheres fill the empty volume of a receptor site, generating its negative image. In the case of the ligand, the spheres are placed inside the molecular surface forming a positive image of the molecular topography. The centers of these spheres are then used as a rudimentary grid for mapping possible configurations of the ligand (the protease inhibitor) in the active site of the receptor (the protease). We note that while we use the spheres to describe the molecules for the purposes of orientation generation, it is the full atomic resolution structures that are actually docked and

evaluated. The protease active site is mapped once for excluded and allowed volume, and this information is used in evaluating each of the possible inhibitor/protease configurations. Orientations of the ligand that place it in excluded volume regions of the receptor are thrown out. The configurations that pass the excluded volume filter and have enough 'good' contacts between inhibitor and protease are then evaluated by more elaborate including electrostatics, molecular mechanics, buried surface area, packing and free energies of solvation.

Spheres are generated analytically to touch the molecular surface at two points, have their centers along the surface normal at one of the points and be placed so that they do not intersect the surface. The spheres are of different sizes and can overlap with one another within a given pocket in the protein. A collection of overlapping spheres defines a cluster. The molecular surface of a protein will typically have tens of clusters, each of which describes a potentially interesting site of interaction. The radius of a sphere is related to the concavity of a local region of the molecular surface. The larger the sphere radius allowed, the larger and more shallow the pocket that the cluster describes and the greater the number of spheres in that cluster. The ligand is similarly described, except that the spheres are placed *within* the molecular surface and are complementary to local ridges in the molecular surface rather than the grooves.

In docking two molecules as large as proteins, we have found it necessary to modify the way we treat the organization of spheres into clusters. The number of possible orientations of a ligand in a receptor site is combinatorially dependent on the number of spheres in each cluster and is therefore subject to a combinatoric explosion, typical of sub-clique generation in graphs (Kuhl *et al.*. 1984). This is a general problem in the field. To keep a search of orientation space to a practical length, it is necessary to

restrict the number of spheres in the ligand and receptor sites to be matched. Clusters of spheres are defined as sets of spheres which overlap with at least one other sphere in the set and with no spheres outside the set. To reduce the number of spheres in clusters we reduce the maximum allowed sphere size in sphere-sphere overlaps. This has the effect of segregating the spheres into smaller clusters as the radius criterion is tightened. The total number of grooves and ridges described meanwhile increases. Since we treat each groove/ridge as a potential site, there are more orientations to be considered. This effect is, however, small compared to the combinatoric advantage of restricting the total number of spheres in each site (Shoichet *et al..*, 1992).

As a point of technical interest, we note that our new clustering method was successful at reducing run times while still producing good docked geometries. The initial sphere calculation typically produced protease active site clusters containing 100-150 spheres. Re-clustering segregated the initial cluster into several smaller ones of sizes ranging from 25-60. Whereas the original cluster covered the full site, the new clusters described more local features within it. In a similar way, the initial sphere cluster for the inhibitors was spread over the entire inside volume of the molecules and included approximately as many spheres as there were solvent exposed atoms in the molecule. This typically amounted to 300-400 spheres. Re-clustering produced several smaller clusters of size ranging from 40-90 spheres. The segregated clusters described local regions of the inhibitor. We found that as the cluster sizes grew beyond 60-80 spheres the docking calculations became prohibitively long due to a combinatorial explosion of possible sphere set choices. Test calculations were performed using the original full clusters from the unbound structures of BPTI and Trypsin. While configurations resembling the crystal structure complex results were found, the search of orientation space was very slow. The run was stopped after accumulating 45 hours of CPU time, calculations indicate that over

three years of CPU time would be required for the run with the large clusters to complete. This result highlights the importance of methods to reduce the number of possibilities without losing adequate sampling of orientation space.

To explore the possible orientations of an inhibitor in a receptor site, we use a modification of the internal distance algorithm described previously (Kuntz *et al.*. 1982). As in the original algorithm, sets of points (sphere centers) from the ligand cluster are compared with sets from the receptor cluster on the basis of the pair-wise internal distances within each set. If all internal distances match, within a preset tolerance, the ligand spheres are mapped onto the receptor spheres.

To allow for more complete searches of orientation space, we have adapted the method to organize the spheres based on the internal distances of the sphere centers (Shoichet *et al.*., 1992). Starting with any given sphere center, all the other points in the cluster are pre-sorted into "bins" based on their distance to the first point. The bins are of adjustable resolution - the wider the range of distances in each bin the more sphere points each will usually contain. All points are ultimately tried as 'first' points, while at any given position in the set generation all the points from an acceptable bin are tried. The number of points in each bin determines how many sets of points in the receptor and the ligand will ultimately be matched. In general, the larger the bins the larger the number of orientations generated. Thus the breadth of search is under user control.

Another useful modification is the ability of the current program to investigate particular regions of orientation space in greater or lesser detail depending on the characteristics of that region. More configurations are generated in regions of space that return high scoring 'fits' of the ligand in the receptor than regions that do not do

so by automatically increasing the size of the particular bins that define that region. This feature allows for detailed searches in some parts of space and cruder searches in others. The fineness of the search is set dynamically by the program and does not demand human intervention (Shoichet *et al.*., 1992).

Possible orientations of the ligand in the receptor are scored on the basis of atom-atom contacts between ligand and receptor structures. In the previous implementations of the algorithm (Desjarlais *et al.*. 1986; DesJarlais *et al.*,1988) all atom-atom contacts between receptor and ligand were summed on the basis of an exponential distance function. To increase the efficiency of the code we grid the internal volume of the receptor site on a cubic lattice and score every point on the lattice on the basis of its contacts with the protein atoms (Shoichet *et al.*, 1992). This is done once for any given site. Ligand orientations are scored by mapping their atoms onto the nearest lattice points and simply summing over all of the mapped points. For large sites, this results in a factor of 4-5 improvement in computation time at the expense of using large amounts of computer memory.

All modifications to the previous program - DOCK - are encoded in the program DOCK2, available from authors. All computations were done on IRIS 4D workstations (Silicon Graphics, Inc., Mountain View, CA). Unless explicitly stated otherwise, we used the following variable parameters in generating and scoring inhibitor orientations in protease sites. Spheres were initially generated using the program SPHGEN (Kuntz *et al.*,1982) on a molecular surface calculated by the program MS (Connolly 1983) using a 1.4 Å radius probe. Spheres were originally clustered by SPHGEN using a 5.0 Å cutoff for the maximum radius, all spheres with radii greater than 5.0 Å were discarded. Clusters were defined in two steps. First, all spheres with radii less than 2.5 Å were combined into overlapping sets as described

above. Then all spheres with radii between 2.5 and 5.0 Å were added to any cluster they overlapped but the clusters were not further coalesced (Shoichet *et al.*, 1992). For the 2SNI/CI docking runs, we used a modification of this procedure. SPHGEN, though it calculates approximately one sphere per molecular surface point, normally only lists the largest sphere that is associated with a given atom. In generating spheres for the 2SNI/CI sites, we modified SPHGEN so that it outputs all spheres calculated. This has the effect of producing many more spheres than would be produced by the one-per-atom method. We then re-clustered the full set of spheres using a maximum radius cutoff of 2.5 Å. While the latter procedure offers a more complete description of the molecular topography, both procedures are qualitatively the same. Molecular surface generation for the active site region of the proteases and the entire inhibitor typically took 4 minutes each. Sphere generation, including re-clustering, took about an hour for each molecule.

Orientations were generated using sub-cliques of 4 to 5 points each, with an internal distance tolerance of 1.5 Å for any sub-clique node. For the 2PTN/4PTI dock run an internal distance tolerance of 2.0 Å was used - tolerances of 1.5-2.0 Å were typical in our earlier studies (DesJarlais *et al.*,1988). Search times varied with the system (table II).

For scoring the orientations we used a cubic grid of 1/3 Å resolution giving approximately one million grid points/site. The close contact limit was set to 2.4-2.6 Å on the basis of the actual contacts in the crystal structure. For docking the unbound conformations of the proteases and inhibitors the close contact limit was set to 2.0 Å, which is in the range of contacts that may be found in some low atomic resolution crystal structure complexes (Grau & Rossmann 1981). The limit for a positive score

for 'good' contacts was set to 4.5 Å in all runs (DesJarlais *et al.*,1988). Grid calculation took approximately 10-25 minutes, depending on the size of the site.

To evaluate how well the method does at docking the protease and inhibitor structures, it is important to be able to measure a distance between the crystallographic configuration of the inhibitor in the protease and the putative configurations which we generate. This is a simple matter when the bound conformations of the proteases and inhibitors are used, but is slightly more difficult when one is using the unbound conformers since there are obviously no crystal structures of the unbound molecules together. In this case we report the distances of the docked orientations of the unbound inhibitor to the crystallographic orientation of the bound inhibitor and correct for the conformational difference between the bound and the unbound forms. Thus, for instance, the 0 Å RMS configurations reported in tables IV and V represent the unbound conformations of the inhibitors, mapped onto the bound conformers of the same molecules in such a way as to minimize the RMS distance between the two structures. We report two sorts of distances (tables IV and V), an RMS one and another, generated by the program CDSFIT (McLachlan 1982), which is a combination of the distance between the centers of gravity of the two orientations and an angle representing a rotation around a helical screw axis. Strictly speaking the RMS figures are misleading since they imply a statistical variance in individual atom positions though the atoms are fixed in the generation of orientations. Moreover, the RMS figures are occasionally skewed by rotations that lead to relatively large movements of the the outer surface of the inhibitor in the active site but leave the interface residues relatively unchanged. The CDSFIT distances are more informative though slightly more complex. We have included the RMS values for easy perusal of the data.

Seven different methods were used to evaluate the docked structures that had good contacts: molecular mechanics using AMBER (Weiner *et al.*,1984), buried surface area using ACCESS (Richmond & Richards 1978) as implemented by Lewis (Mitchel Lewis, personal communication), solvation free energy using the method of Eisenberg and McClachlan (Eisenberg & McLachlan 1986), solvation free energy using the method of Lewis (Mitchel Lewis, personal communication), packing using QPACK (Gregoret & Cohen 1990), mechanistic filtering based on simple enzymologic characteristics of the protease-inhibitor interaction, and electrostatic complementarity using DELPHI (Gilson & Honig 1988). AMBER minimizations were conducted under a variety of different conditions, the best results were realized using a ten angstrom pair-list cutoff with a distance dependent dielectric function of 4.5R (Pickersgill 1988), minimizing each system for 900 cycles. The united atom potentials were used in all runs. No water molecules were included in the calculations. All energies reported are from the program analysis program ANAL, distributed with AMBER. Minimization took approximately 3 hours of cpu time per configuration. Buried surface area was calculated using a 1.4 Å probe sphere and took between 3-4 minutes per configuration. Packing was analyzed for ligand-receptor interface residues only, as defined by the QPACK program itself. Packing calculations, including associated analyses, took approximately 1 minute of cpu time per configuration. Mechanistic filtering concentrated on evaluating whether or not important protease-inhibitor contacts were made in the various configurations, especially whether the active site serine in the protease was close to the appropriate scissile bond in the inhibitor and whether the respective specificity residues between the inhibitor and the protease were close together. DELPHI was run on each ligand-receptor configuration individually. A water probe size of 1.5 Å and a ionic probe size of 1.6 Å were used in all calculations. The interior dielectric constant was set to 4, the exterior to 80, and the ionic molarity to 0.1 M. Using the AMBER charge potentials distributed with the program, the

electrostatic potential of the receptor was calculated using a three step boundary focussing calculation that began with the proteins filling 20 percent of the grid, focussing to 60 percent and then focussing again to 90 percent filled. The goal of the focussing calculation was to reduce errors that occur at the boundary of the grid (Gilson *et al.*, 1988). The ligand was, for the purpose of the receptor petential calculation, considered to be uncharged. The presence of the ligand affected the potential by altering the size of the grid as well as the volume assigned a low dielectric constant. After focussing, the potential of the receptor at each ligand atom was calculated using the PHITOPDB program, distributed with DELPHI, and multiplied by the AMBER charge of that ligand atom. The resulting electrostatic energy was summed over all ligand atoms. DELPHI potential calculations took approximately 15 minutes of cpu time per configuration.

Models Used

Protein structures used in this work were take directly from the Protein Data Bank (Bernstein *et al.*,1977). Atoms with occupancy factors of 0 in were not included in contact scoring of orientations. Most docking runs involved only ligand atoms that were solvent exposed in order to decrease run times. This did not effect the orientations found. When docking the unbound protein conformations, some ligand atoms with high temperature factors in the crystal structure were not considered for contact scoring (table I). In addition, Arginine 39 in the 4PTI structure was truncated at the CD atom. The two N-terminal residues of 4PTI were not used in docking - these residues have occupancies of 0 in the bound (2PTC) structure. Methionine 18 in the Ovomucoid Third Domain structure 2OVO was truncated at its CG atom. This residue is hypervariable across species and is in a region of high mobility in the structure (Laskowski & Kato 1980); the analogous residue in the Ovomucoid molecule from the 1CHO complex structure is a Leucine. Since we used the contact

scoring as a first filter only, we were not worried that leaving these groups out would be too permissive in terms of configurations generated. Our concern was that if we left what are generally poorly determined parts of the structure in the scoring we would be too restrictive. We emphasize that these deletions were only used during the docking calculations: all atoms reported in the crystal structures were used in the energy evaluations. As a model for the unbound conformation of Subtilisin Novo (McPhalen & James 1988) we have used the Subtilisin Carlsberg structure, which is a good model for the unbound conformation of the Novo enzyme (Neidhart & Petsko 1988). For the unbound conformation of Ovomucoid 3rd Domain we use the unbound structure from Silver Pheasant (Bode *et al.*,1985), which is highly homologous to Turkey Ovomucoid molecule present in the crystal structure of the Chymotrypsin/Ovomucoid complex (Fujinaga *et al.*,1987) and crystalizes isomorphicaly to the Turkey Ovomucoid in the unbound state.

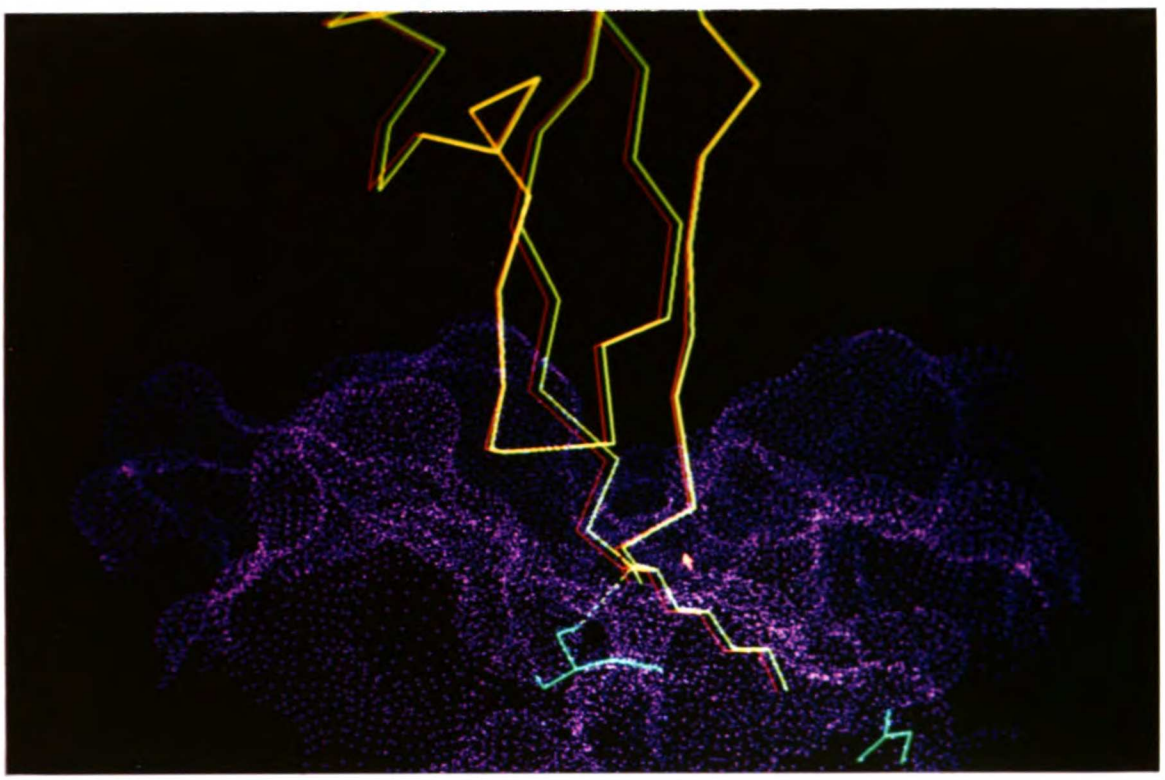


Figure I: BPTI/Trypsin docking, bound conformations. A Cα tracing of the best (by RMS) docked orientation of BPTI (in yellow) is shown, compared to the crystal structure (2PTC) configuration of the ligand (in red). The molecular surface (Connolly 1983) of the Trypsin active site is shown in magenta. The side chains of serine 195 and aspartate 189 of the Trypsin are shown in green, the specificity lysine 15 of the inhibitor is fully displayed. The distance shown is from the Oγ of the serine to the main chain C of lysine 15 of the docked ligand and is 2.96 Å.

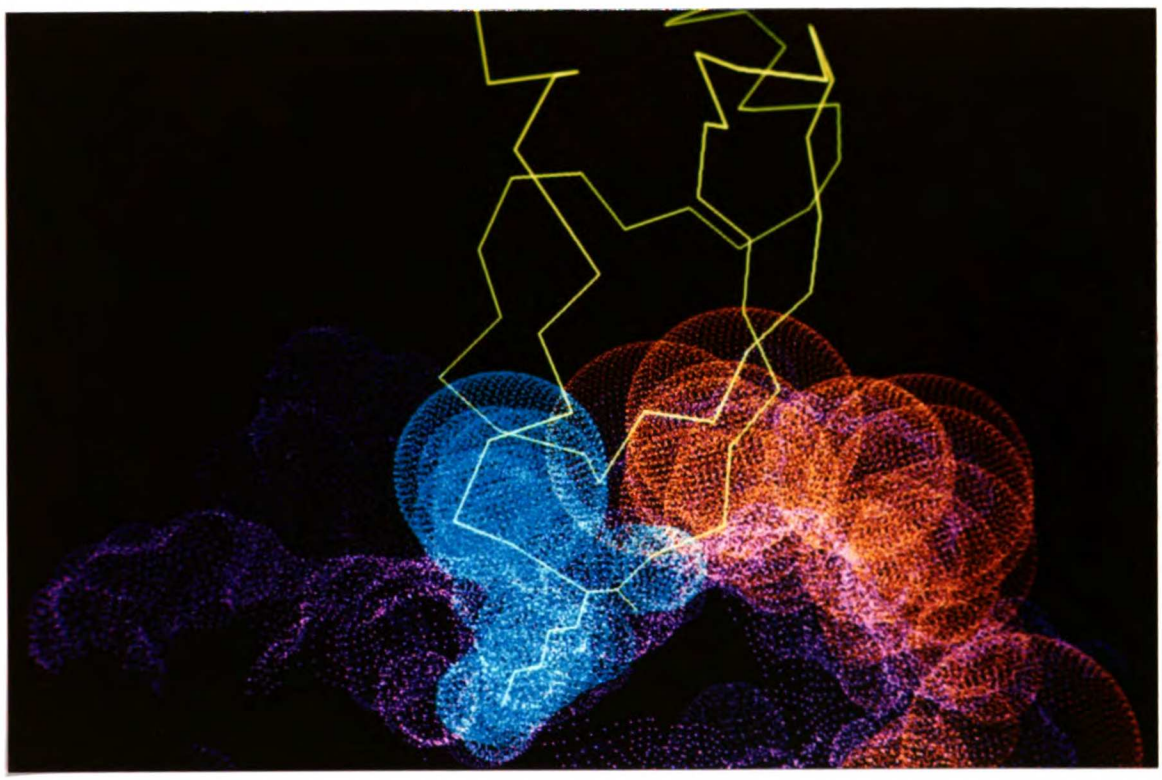


Figure II: Trypsin Spheres: The two major clusters of Trypsin (2PTC) active site spheres are shown in red and blue. The molecular surface of the Trypsin is shown in magenta, a Cα tracing of the BPTI is shown in yellow. The blue cluster describes the specificity pocket and surrounding regions of the enzyme. The blue cluster includes 35 spheres, the red 31 spheres.

Results

The algorithm is able to reproduce the crystallographic complex configurations of the three protease/inhibitor complexes in a timely fashion and to high accuracy, as summarized in table II. For all three systems we find inhibitor orientations that are within 0.75 Å RMS of the crystallographic complex structure. For Subtilisin/Chymotrypsin-Inhibitor (2SNI) system the closest inhibitor orientation is within 0.14 Å RMS of the complex structure, which is within the probable experimental error of the structure. The best docking of the bound conformations of BPTI and Trypsin is shown in figure I. To the eye the docked BPTI configuration is virtually indistinguishable from the crystal structures upon which it is overlaid. In Figure II we show the principal Trypsin sphere clusters used to generate the configuration displayed in figure I, amongst others.

Also in table II we summarize the docking results using the unbound conformations proteases and inhibitors. Once again the method does well at returning putative complexes that are close to the crystallographic complex configuration. In order to return such close complexes it is necessary to search through approximately an order of magnitude more orientations, and the run times for the unbound conformers are longer than in the bound conformation docking runs, though still reasonable. The energy minimized best docking of the unbound conformations of BPTI and Trypsin is shown in figure II. Except for conformational differences, the configuration is very close to the crystallographic result upon which it is overlaid.

Receptor	Ligand	TYPE[1]	Best Docked (To Crystal Structure, RMS Å)[2]	Total Orientations Evaluated in Docking	Run Time (hrs:min)
TRYPSIN (2PTC)	PTI (2PTC)	Bound	0.97	43,767	0:09
TRYPSIN (2PTC)	PTI (2PTC)	Bound	0.29	360,366	1:04
TRYPSIN (2PTN)	PTI (2PTN)	Free	0.52	9,976,471	27:11
CHYMO-TRYPSIN (1CHO)	Ovomucoid 3rd Domain (1CHO)	Bound	0.72	1,650,604	4:19
CHYMO-TRYPSIN (5CHA)	Ovomucoid 3rd Domain (2OVO)	Free	0.82	2,117,929	5:44
SUBTILISIN (2SNI)	Chymotryp. - Inhibitor (2SNI)	Bound	0.14	1,511,411	5:30
SUBTILISIN (1SBC)	Chymotryp. - Inhibitor (2CI2)	Free	0.64	8,615,720	20:23

Table II: Docking Runs for Protease/Inhibitor Complexes

As noted above, the docking procedure also found configurations that did not resemble the crystal structure of the complex. These configurations usually involved rotations around one or more of the principal axes of the inhibitor. While the vast majority of these (typically over 99.99%) were discriminated against on the basis of the initial contact scoring, there were some orientations sampled that passed the contact criteria but were distant from the crystal complex configuration. For brevity we shall refer to

---

[1] Two types of calculations were performed, using the bound (from the crystal complex) or the free (from the uncomplexed crystal structures) conformations of the molecules.

[2] RMS's measured to the crystal complex for bound docking runs and to best-fits to the crystal complex in the free conformer runs, as explained in the text.

these distant orientations as 'false positives', though we shall return to the possible physical relevance of such distant orientations later in the paper. Most of the 'false' positives occurred in the unbound conformation docking runs. Figure III shows some representative 'true' and 'false' positive docked configurations of the unbound conformation of BPTI in the reference frame of the Trypsin active site. The range of these configurations is extraordinary: DOCK2 finds acceptable orientations that put BPTI in upside-down, tilted on its side and twisted, amongst others, relative to the crystallographic complex. Some of these 'false' positives , such as the one shown in figure III d., which puts the carboxy-terminus of BPTI into the Trypsin specificity pocket, are easily disqualified by polarity or other simple criteria. Other 'false' positives appeared surprisingly reasonable. Figures IV and V compare a 'true' to a 'false' positive docking for the Trypsin/BPTI system. Both configurations put a lysine in the specificity pocket of Trypsin; figure VI shows that the interfaces for these two putative complexes are both extensive. Interface residues for some 'true' and 'false' positive configurations in this system are presented in table III. We attempted to distinguish amongst the 'false' and 'true' positives using standard energy evaluation techniques. The buried surface area, solvation free energy, packing and mechanism-based filtering results are summarized in table IV. The results of the energy minimization and electrostatic calculations are presented in table V and figures VII-IX. Table IV shows that there is little or no correlation between buried surface area, solvation free energy or packing at the interface and an orientation' s resemblance to the structure of the crystal complex. This is especially apparent when considering the unbound conformation dockings. Note that some of the BPTI (4PTI) orientations in Trypsin (2PTN) that are over 20 Å RMS from the crystallographic orientation have the highest amount of buried surface area and also amongst the lowest solvation free-

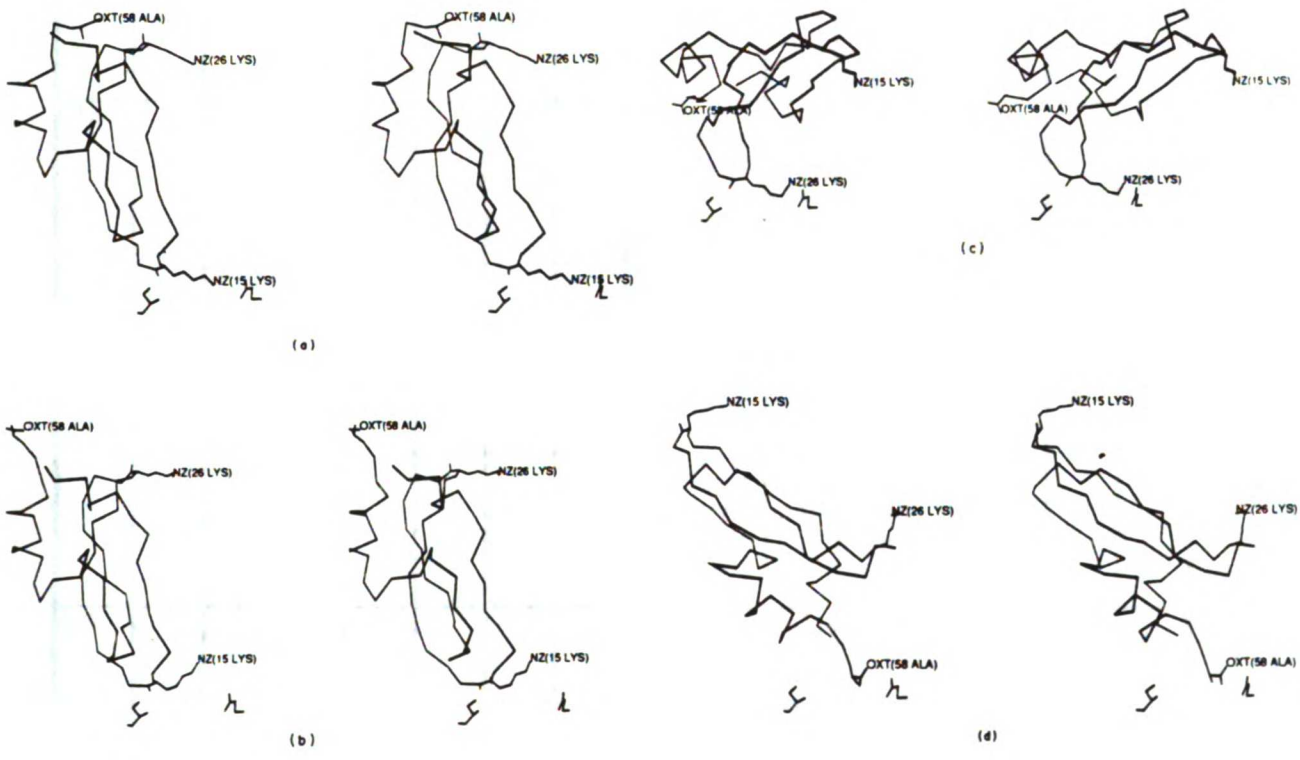


Figure III: Cα tracings of representative docked configurations of BPTI in the reference frame of the Trypsin active site. Lysines 15 and 26 of the BPTI are fully displayed, as is Ala 58, the C-terminal residue. Ser 195 (bottom left) and Asp 189 (bottom right) of the Trypsin are also displayed. The orientation of the Trypsin is kept fixed for all views. a). BPTI configuration from the crystal complex structure, 2PTC. b). Unbound conformation of BPTI (4PTI), 0.52 Å docking. c). 4PTI, 21.41 Å docking. d). 4PTI 20.40 Å docking. The conformational differences between the bound and free BPTI structures can be clearly seen by comparing a. and b. in the C-terminal and lys 15 regions.

BPTI r.m.s. from crystal (Å)	BPTI Residues														
	D3	L6	E7	P8	P9	T11	P13	C14	K15	A16	R17	I18	I19	F22	Y23
0.00							W215 G216	H57 L99 S214 W215	H57 C191 N192 G193 D194 **S195** S214 G218 C219	F41 H57 C42 C58 G193 **S195**	Y39 H40 F41 C42 W139 G140 Y149 N192 G193 D194	Y39 H57 K60	Y39		
0.52							W215 G216	H57 L99 **S195** S214 W215	H57 C191 N192 G193 D194 **S195** S214 C219	F41 H57 C42 C58 **S195**	H40 F41 C42 W139 G140 Y149 N192 G193 D194	H57 K60	Y39		
2.07						N192	L99 G216 W215	L99 H57 D102 **S195** S214 W215	S190 C191 N192 G193 D194 **S195** V213 S214 G218 C219	F41 H57 C42 C58 N192 G193 **S195**	H40 F41 W139 G140 Y149 N192 G193 D194	F41 H57 K60	Y39		
4.29								S96 N97 L99	L99 S214 W215 G216 S217	H57 L99 S214 W215	H57 C191 N192 G193 **S195** V213	H57			
21.41	S145 G146 S147	N141 K143 S144 N192 G218 C219	G218	S217 G218 K224 N221	S217										N192
20.11	Y59 K60	F41 K60 H57 C58	Y39	H40 Y149	T147 Y149 N192	T147								Y149 N192	N192 C219

Table III: Interface residues for trypsin/BPTI complexes. Interface residues are those where at least one atom of a Trypsin residue is within 4.0 A of at least one atom of a BPTI residue. BPTI residues are in the first row of every column while Trypsin contacts for the BPTI residues are in the succeeding rows. The results are from the unbound conformer docking runs.

BPTI r.m.s. from crystal (Å)	BPTI Residues														
	N24	A25	K26	A27	L29	Q31	T32	F33	V34	G36	G37	C38	R39	K41	A58
0.00									Y149	H57	H57	S96 L99	D96 N97 T98 L99		
0.52									Y149	H57	H57	S96 L99	S96 T98 L99		
2.07									Y151	H57 **S195**	H57	H57 L99 W215	N95 N97 N100 L99 S96 S98		
4.29										L99	S96	S96 N97 T98	N97		
21.41	W215 G216	N192 C219	H57 **D189** S190 **S195** S214 W215 G216 G218 C219	H57 **S195** S 214 W215	H57	L99 W215									
20.11	N192 C219	H57	C191 G193 D194 **S195** V213 S214 W215 G216	G216 S217 G218		N192 G218 C219	G146	G146 T147						Y39	N97

Table III: Interface residues for trypsin/BPTI complexes (continued from previous page)

energies. This same point holds for some of the unbound conformation dockings of Ovomucoid Third Domain (2OVO) in Chymotrypsin (5CHA) and of Chymotrypsin-Inhibitor (2CI2) in Subtilisin (1SBC) - some of the most distant orientations have the largest buried surface areas and the lowest solvation free energies. The pair-potential energies from the packing calculations are not useful. Mechanistic filtering (see Methods section) can reject some, but not all, distant configurations on simple

enzymologic grounds. Energy minimization finds the low energy structure to be from amongst the configurations to be within 5 Å of crystal structure (figure VII). There is, however, often significant energy overlap between the putative complexes that resemble the crystal solution and those that do not. Moreover, the low energy structure can only be said to be *within* 5 Å of the crystallographic complex. The crystallographic complex itself rarely has the lowest energy. We have also plotted the

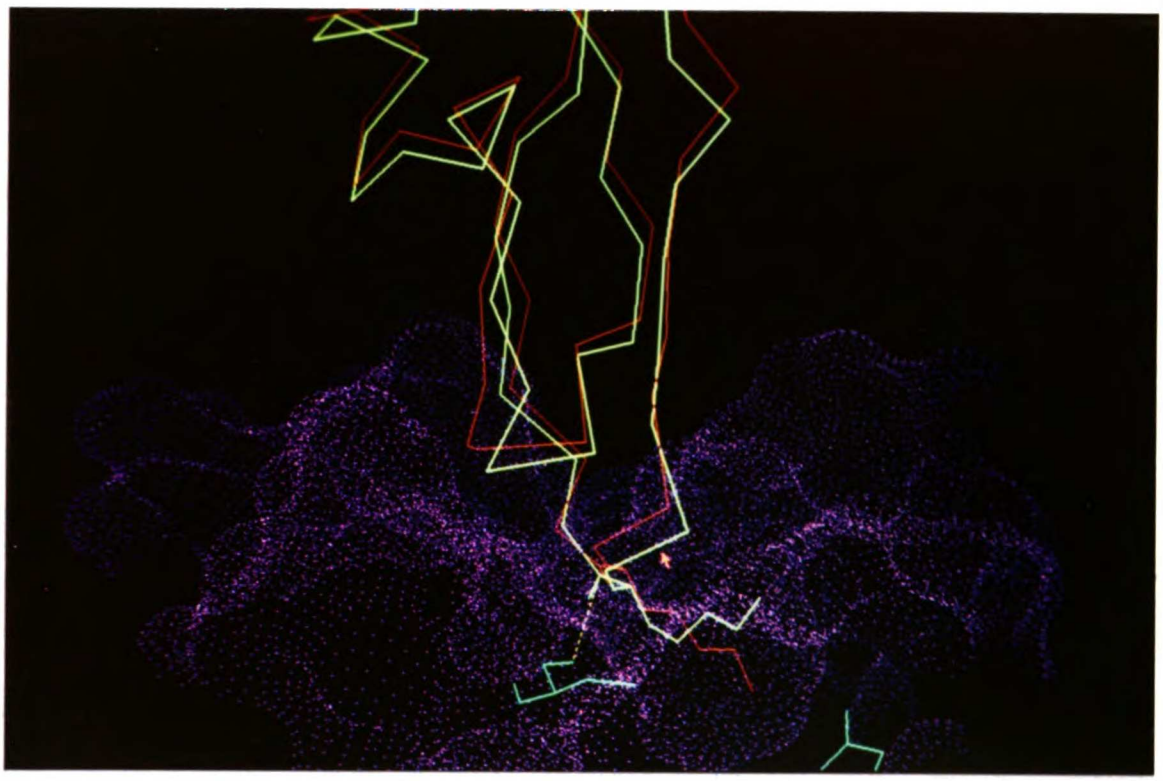


Figure IV: Energy minimized BPTI/Trypsin docking, free conformations. A Cα tracing of the best docked orientation of BPTI, by RMS, (4PTI, in yellow) is shown, compared to the crystal structure (2PTC) configuration of the ligand (in red). The molecular surface of the Trypsin active site is shown in magenta. The side chains of serine 195 and aspartate 189 of the Trypsin are shown in green, the specificity lysine 15 of the inhibitor is fully displayed. The distance shown is from the OG of the serine to the main chain C of lysine 15 of the docked ligand and is 3.38 Å. The conformational differences between the docked and bound ligands are most obvious in Lys 15.

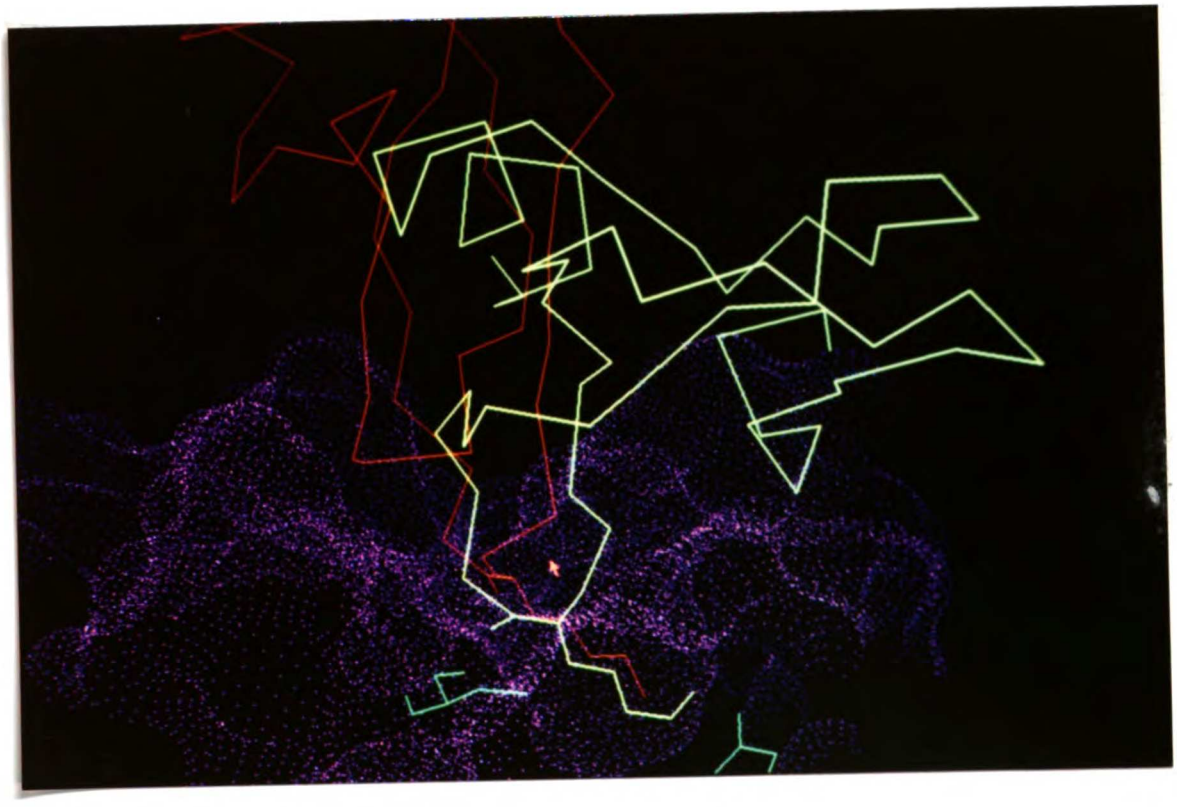
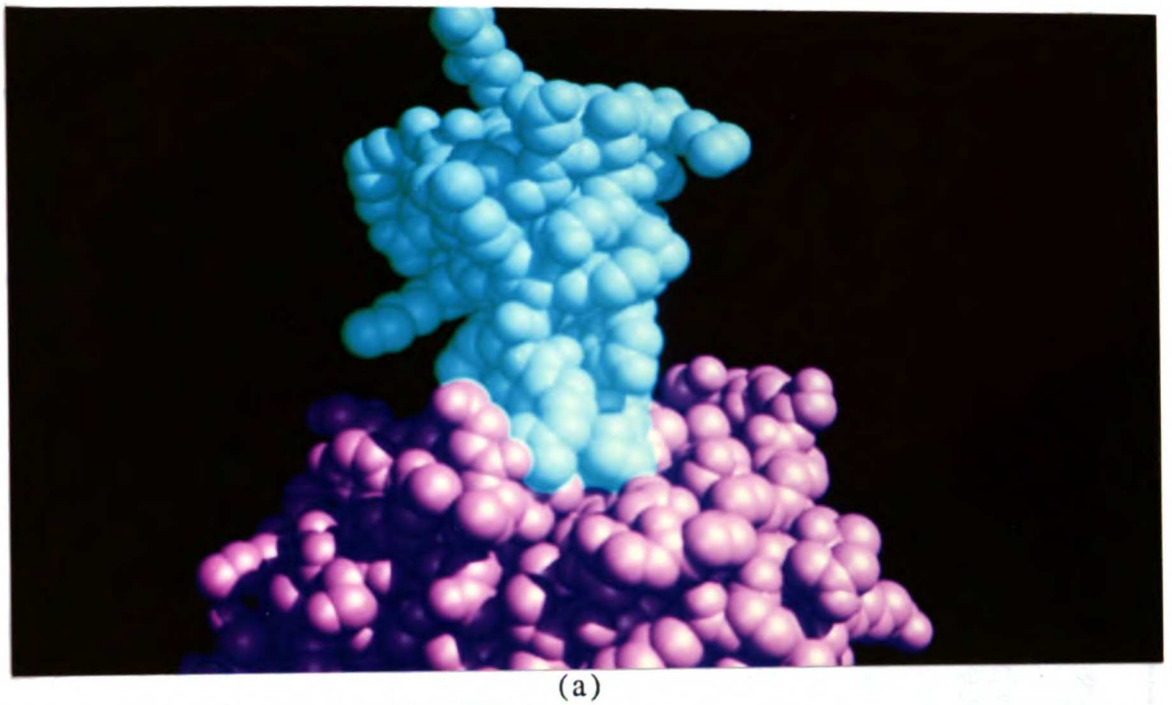
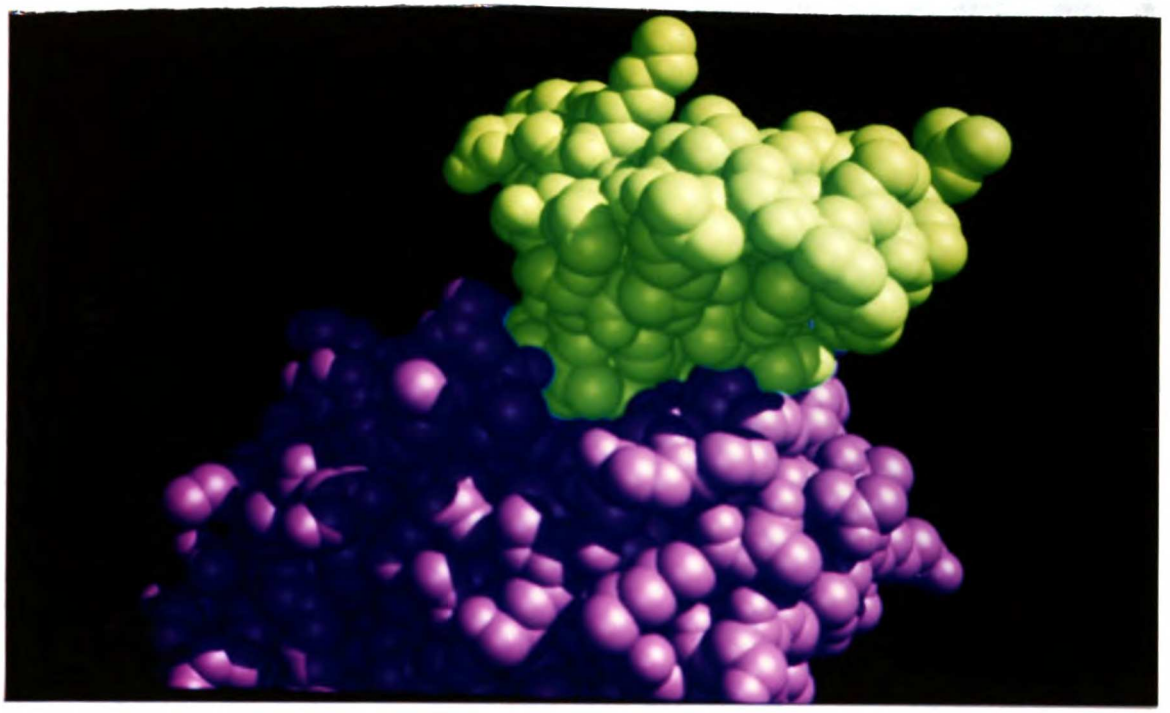


Figure V: 'False positive' BPTI/Trypsin docking, Energy minimized, free conformations. A Cα tracing of a docked 4PTI configuration 21Å away from the 2PTC configuration is shown (in yellow) in the molecular surface of the Trypsin active site (in magenta). S195 and D189 of the enzyme are shown in red, K26 of the inhibitor is shown in the Trypsin specificity pocket. The distance shown is from the OG of the serine to the main chain C of lysine 15 of the docked ligand and is 3.66 Å.

AMBER non-bonded interaction energy against the AMBER electrostatic interaction energy for these complexes (figure VIII). Electrostatic complementarity evaluations were performed on both the minimized and the unminimized complexes (figures IX a-c). In the Subtilisin and Chymotrypsin systems, electrostatic energies calculated using DELPHI seem to distinguish between configurations that resemble the crystal structure and ones that do not at least as well as AMBER does. This is not true in the Trypsin/BPTI system, however. The lowest electrostatic energy structure for both energy minimized and unminimized complexes is one in which the BPTI is over 21 Å RMS distant from the crystal structure configuration of this molecule.



(a)



(b)

Figure VI: Energy minimized BPTI/Trypsin docking, free conformations, van der Waals representations of the interfaces. a) The best docked configuration, 0.52 Å from the crystallographic configuration, BPTI in cyan, Trypsin in magenta. b) A docked configuration of BPTI over 21 Å from the crystal structure of the complex, BPTI in green, Trypsin in magenta. Both configurations have extensive, highly complementary interfaces. The orientation of Trypsin is the same in both views.

Structure	Similarity to Crystal[3]			Solvation Measurements			Packing[4]		Catalytic Competence
	rms Å	center-center dist. Å	rotation angle[5] (degree)	Δ surface area(Å)on complex formation	DELDG[6] ΔG of solvation	ASP[7] ΔG of solvation	pair potential	mean sphere size	Ser Oγ - scissile C w/in 4 Å[8]?
subtilisin/	0.00	0.00	0.00	1703	-13.41	-362.3	-0.5	92.2	NM[9]
chymotrypsin	0.14	0.09	0.47	1693	-13.30	-362.3	-0.7	93.2	NM
inhibitor	0.17	0.14	0.34	1712	-13.50	-362.4	-0.3	91.4	NM
(bound)	0.29	0.26	0.74	1701	-13.24	-362.3	-0.4	94.1	NM
SNI/CI	3.13	2.63	12.13	1767	-16.86	-363.5	-1.7	90.4	NM
	3.29	2.79	12.69	1776	-16.89	-363.6	-1.9	88.9	NM
	0.00	0.00	0.00	1355	-14.15	-355.6	-0.6	81.6	+
	0.65	0.31	3.29	1328	-12.19	-355.1	0.7	75.8	+
subtilisin/	0.72	0.49	3.14	1344	-12.39	-355.2	1.1	78.1	+
chymotrypsin	0.92	0.72	3.47	1347	-12.46	-355.1	0.6	76.7	+
inhibitor	1.33	1.06	4.26	1296	-15.60	-355.3	-0.2	87.6	+
(free)	1.51	1.41	4.11	1298	-13.53	-355.4	-0.5	85.1	+
1SBC/2CI2	1.93	1.28	8.40	1368	-15.43	-355.2	0.2	84.6	+
	10.47	6.32	51.40	1665	-17.79	-354.4	1.4	80.8	-
	10.93	8.02	46.50	1754	-17.91	-354.3	-1.1	86.1	+
chymotrypsin/	0.00	0.00	0.00	1545	-14.22	-319.7	0.0	94.7	NM
wvomucoid	0.72	0.64	2.09	1513	-12.94	-319.6	-0.6	92.8	NM
(bound)	1.17	0.92	5.15	1476	-12.79	-319.2	-0.8	91.6	NM
CHO/OVO	2.44	2.06	9.28	1608	-10.46	-321.0	-1.1	96.1	NM
	0.00	0.00	0.00	1410	17.43	-315.7	0.1	83.7	+
	0.82	0.26	4.46	1832	-17.85	-316.6	0.5	88.5	+
	0.84	0.47	4.09	1813	-19.01	-315.8	-0.1	84.3	+
chymotrypsin/	0.90	0.46	4.97	1765	-16.07	-315.6	0.5	84.4	+
Ovomucoid	0.97	0.76	4.08	1757	-15.82	-315.3	0.4	83.6	+
(Free)	1.46	1.14	5.30	1754	-15.69	-315.7	-0.1	83.5	+
5CHA/2OVO	1.98	1.37	8.32	1926	-17.06	-316.4	1.2	81.6	+
	2.47	1.81	10.62	1954	-16.55	-316.7	1.2	85.5	+
	4.26	3.16	18.50	1871	-16.69	-315.4	0.9	79.3	+
	8.61	6.70	33.82	1775	-16.24	-313.4	-1.1	83.5	-
	18.2	6.80	168.95	1856	-18.73	-317.0	0.8	77.7	-
	0.00	0.00	0.00	1465	-19.90	-292.2	-0.8	95.8	NM
trypsin	0.29	0.21	1.68	1465	-16.05	-292.0	-0.4	95.8	NM
BPTI	0.58	0.54	1.54	1467	-15.58	-292.1	+0.1	95.6	NM
(bound)	0.95	0.77	4.04	1423	-12.54	-291.8	-0.4	95.0	NM
2PTC/PTI	1.40	1.06	6.48	1421	-14.16	-291.8	+0.6	92.0	NM
	3.10	2.57	10.91	1424	-18.12	-292.1	-0.8	91.3	NM
	0.00	0.00	0.00	1411	-18.56	-289.8	1.8	83.9	+
	0.52	0.50	0.93	1447	-18.93	-290.4	1.2	83.1	+
trypsin/	0.71	0.68	1.14	1407	-18.09	-290.3	0.0	82.0	+
BPTI	2.07	1.59	8.16	1534	-18.65	-289.6	2.2	73.1	+
(free)	4.29	1.81	24.35	1211	-11.1	-292.0	0.6	79.3	-
2PTN/4PTI	20.11	4.11	174.91	1718	-18.31	-295.5	0.0	77.2	-
	20.13	2.50	173.63	1723	-18.37	-295.6	0.0	77.2	-
	21.41	14.24	130.83	1428	-17.22	-297.4	1.0	83.5	+

Table 4: Solvation, packing and enzymological evaluations of complexes

---

[3] Distances measured to the crystal complex for bound docking runs and to best-fits to the crystal complex in the free conformer runs, as explained in the text.

[4] (Gregoret and Cohen 1990)

[5] Rotation about helical screw axis (McLachlan 1982)

[6] Mitchel Lewis, personal communication

[7] (Eisenberg and McLachlan 1986)

[8] +'s indicate that the main chain carbon of the inhibitor siscile bond is within 4Å of the protease's catalytic serine oxygen, -'s indicate that the distance is greater than 4Å.

[9] NM: Not Measured

Structure	Similarity to Crystal[a]			AMBER[b] Energy (Kcals/mol)			DELPHI[c] Interaction Energy (Kcals/mol)	
	rms Å	Distance Between Centers	Rotation Angle[d] (degrees)	Inter-action	Total	Electro-static	Minimized Structures	Unminimized Structures
subtilisin/ chymotrypsin inhibitor (bound) SNI/CI	0.00	0.00	0.00	-109.5	-2856	-12.2	-10.7	-13.1
	0.14	0.09	0.47	-107.7	-2851	-12.0	-10.8	-12.6
	0.17	0.14	0.34	-110.3	-2858	-12.1	-10.4	-13.1
	0.29	0.26	0.74	-109.6	-2853	-12.2	-11.5	-13.1
	3.13	2.63	12.13	-110.9	-2857	-16.0	-8.7	-14.4
	3.29	2.79	12.69	-110.0	-2855	-15.7	-8.5	-14.3
subtilisin/ chymotrypsin inhibitor (free) 1SBC/2CI2	0.00	0.00	0.00	-92.4	-2749	-9.1	-9.5	-10.4
	0.65	0.31	3.29	-87.7	-2744	-9.0	-4.0	-11.2
	0.72	0.49	3.14	-88.4	-2745	-9.0	-4.6	-11.2
	0.92	0.72	3.47	-88.6	-2743	-8.6	-4.7	-11.4
	1.33	1.06	4.26	-87.4	-2738	-7.8	-3.0	-10.5
	1.51	1.41	4.11	-86.7	-2757	-5.7	-3.8	-7.8
	1.93	1.28	8.40	-94.4	-2763	-10.9	-8.0	-14.7
	10.47	6.32	51.40	-89.0	-2767	-4.5	-0.2	-6.7
	10.93	8.02	46.50	-86.8	-2756	-7.9	1.4	-7.3
chymotrypsin ovomucoid (bound) CHO/OVO	0.00	0.00	0.00	-122.3	-2267	-17.2	-12.4	-18.0
	0.72	0.64	2.09	-120.7	-2267	-16.8	-11.2	-18.0
	1.17	0.92	5.15	-118.3	-2274	-18.3	-11.6	-21.3
	2.44	2.06	9.28	-118.3	-2259	-14.7	-11.2	-14.0
chymotrypsin ovomucoid (free) 5CHA/2OVO	0.00	0.00	0.00	-107.5	-2221	-12.8	-4.4	-15.5
	0.82	0.26	4.46	-94.1	-2214	-9.7	-20.8	-14.1
	0.84	0.47	4.09	-103.8	-2224	-10.9	-4.3	-13.1
	0.97	0.76	4.08	-104.1	-2231	-14.7	-9.4	-16.5
	1.46	1.14	5.30	-100.5	-2226	-13.2	-9.6	-15.4
	1.98	1.37	8.32	-104.9	-2240	-9.1	-6.7	-10.8
	2.47	1.81	10.62	-106.4	-2240	-9.0	-0.4	-10.5
	4.26	3.16	18.50	-103.2	-2214	-10.8	-8.1	-14.4
	8.61	6.70	33.82	-67.4	-2192	-3.9	2.7	-4.7
	18.2	6.80	168.95	-59.5	-2182	-0.8	14.8	2.1
trypsin/ BPTI (bound) 2PTC/PTI	0.0	0.0	0.0	-123.2	-2257	-21.6	-31.3	-40.2
	0.29	0.21	1.68	-122.3	-2258	-21.5	-30.7	-40.1
	0.58	0.54	1.54	-122.5	-2262	-21.4	-30.5	-40.3
	0.95	0.77	4.04	-120.8	-2259	-21.5	-30.4	-39.8
	1.40	1.06	6.48	-119.3	-2252	-21.3	-31.8	-39.9
	3.10	2.57	10.91	-117.9	-2256	-21.8	-26.9	-39.0
trypsin/ BPTI (free) 2PTN/4PTI	0.00	0.00	0.00	-117.6	-2209	-14.6	-25.4	-23.3
	0.52	0.50	0.93	-120.6	-2217	-15.9	-32.4	-27.8
	0.71	0.68	1.14	-116.4	-2210	-15.6	-22.1	-25.8
	2.07	1.59	8.16	-131.2	-2213	-17.3	-32.4	-34.7
	4.29	1.81	24.35	-73.1	-2179	-5.6	-11.0	-13.9
	20.11	4.11	174.91	-119.1	-2211	-19.9	-25.0	-25.5
	20.13	2.50	173.63	-120.2	-2212	-19.9	-25.8	-30.5
	21.41	14.24	130.83	-109.6	-2206	-14.9	-37.2	-34.8

Table V: AMBER and DELPHI evaluations of complexes

---

[a] Distances measured to the crystal complex for bound docking runs and to best-fits to the crystal complex in the free conformer runs, as explained in the text.

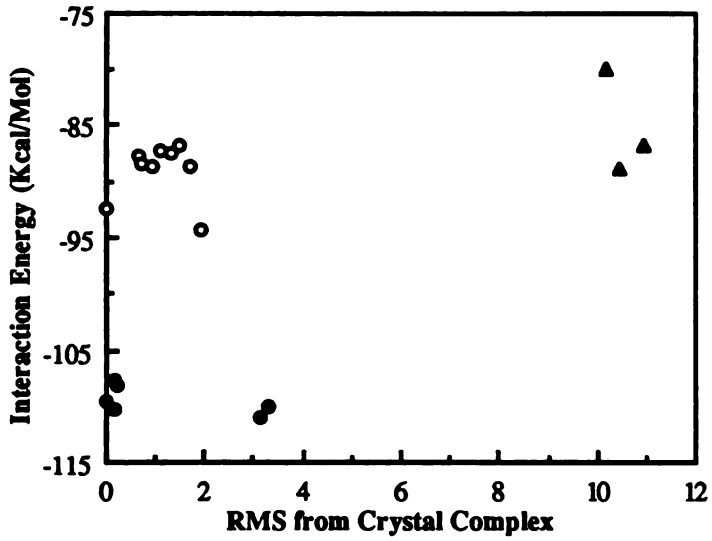
[b] (Weiner et al. 1984)

[c] (Gilson and Honig 1988)

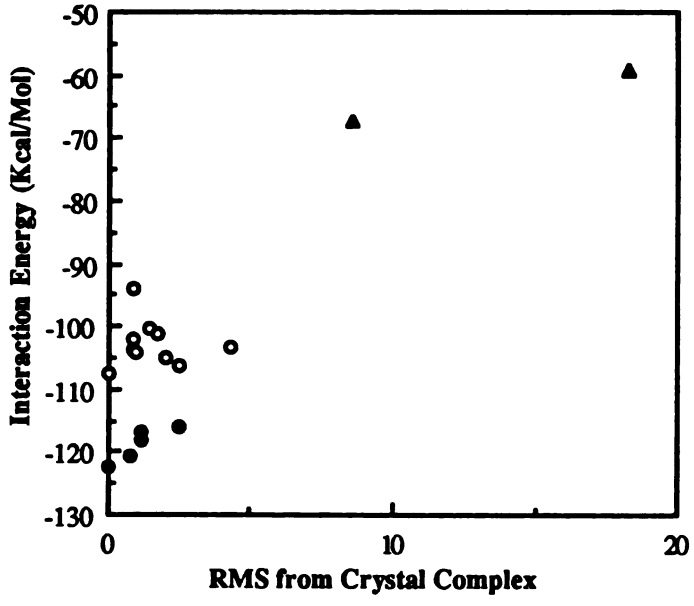
[d] Rotation about helical screw axis (McLachlan 1982)

Discussion

The protein docking problem contains two components: the exploration of configuration space and the energy evaluation of the putative complexes. The docking method addresses the first issue. It may be judged by three criteria: is it able to regenerate the crystallographic configurations, can it do so in a timely fashion and are its alternative configurations, if such are generated, reasonable? The method does well by all three criteria. For the bound conformers all docking runs proceeded smoothly and quickly. The unbound conformers demanded a deeper search of orientation space and longer run times, but the program was always able to return configurations close to the crystal complex structure. For both the bound and the unbound systems we are always able to generate orientations within 0.85 Å of the crystal configuration (table II). This is a striking result. All three inhibitors and receptors differ in conformation between their bound and free forms (Table I), and especially for the inhibitors these conformational differences seem to be most exaggerated in their specificity residues. Thus, Lysine 15 in BPTI has an RMS conformational difference of 2.1 Å between the bound and free forms of the structure, Methionines 59 in the bound and free Chymotrypsin Inhibitors differ by 3.3 Å RMS while Methionine 18 in the 2OVO free Ovomucoid structure is in the 1CHO bound structure a Leucine - a comparison of their respective positions shows them to differ by over 2 Å RMS. Given these differences, the ability of the docking method to generate the crystallographic configurations of the complexes from the unbound conformations of the molecules suggests that much of the topographical information that defines the interfaces of the complexes is present in the backbone and invariant side-chain conformations of the molecules.


-75
-85
-95
-105
-115
Interaction Energy (Kcal/Mol)
0
2
4
6
8
10
12
RMS from Crystal Complex

(a)


-50
-60
-70
-80
-90
-100
-110
-120
-130
Interaction Energy (Kcal/Mol)
0
10
20
RMS from Crystal Complex

(b)

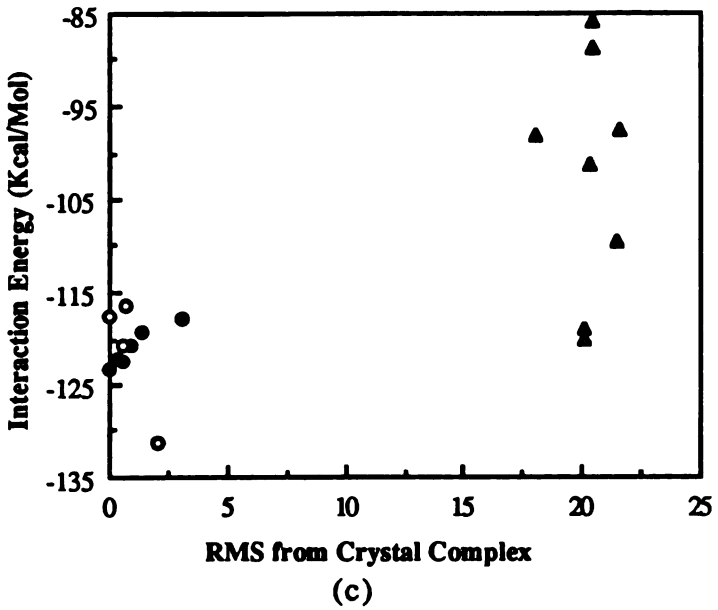


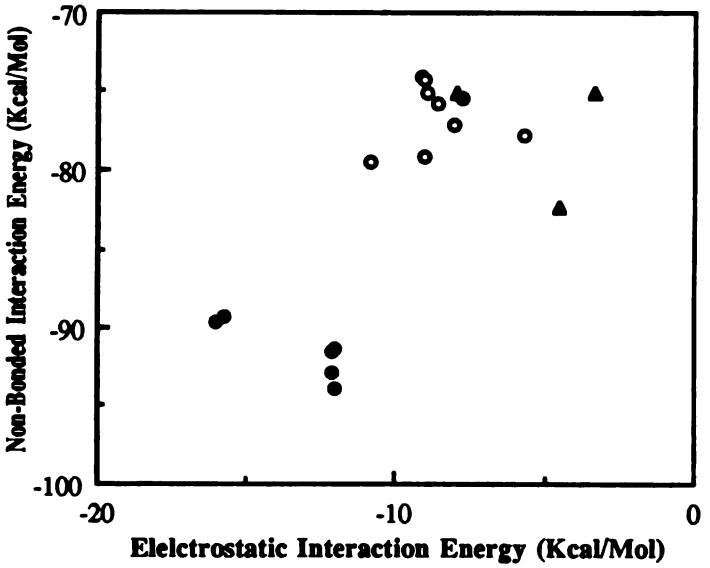
Figure VII: Interaction energies, in Kcal/Mol, of energy minimized complexes from AMBER are plotted against RMS distances, in angstroms, from the inhibitor configuration in the crystal complex. Closed circles represent bound conformer dockings, open circles unbound conformer dockings within 5 Å of the crystallographic configuration and triangles unbound conformer dockings more distant than 5 Å from the crystal structure. a. Subtilisin/Chymotrypsin-Inhibitor dockings. b. Chymotrypsin/Ovomucoid dockings. c. Trypsin/BPTI dockings.

Docking generates geometries divergent from the crystal complex structures as well as ones that closely resemble them. These distant configurations are more common for the unbound conformations of the molecules. This partly owes to the scoring criteria that we used in the unbound conformation docking runs. We were forced to reduce the close-contact limit from the 2.4-2.6 Å used in the bound conformer runs to 2.0 Å. This is a looser constraint that allows for further deviations from the crystal configuration in terms of acceptable dockings. Many of the docked configurations of the unbound structures are outrageously different from the crystallographic configuration, as is reflected by their high RMS values. Despite these differences,

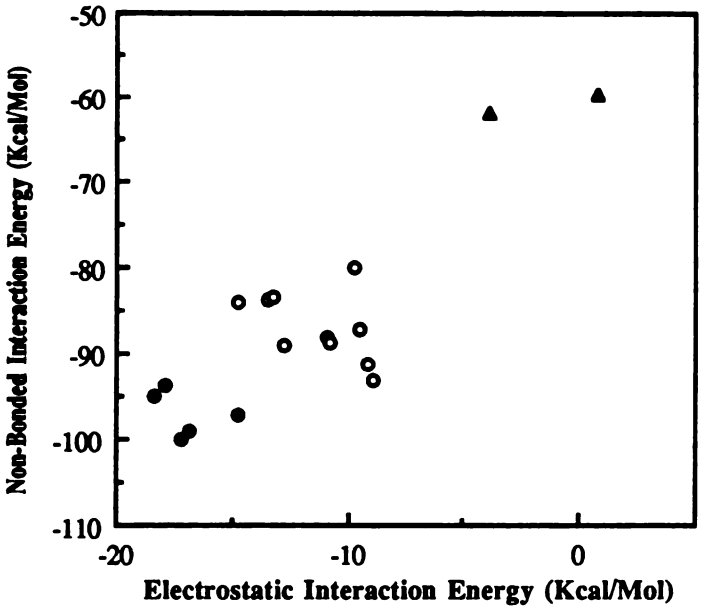
however, many of the dockings that are distant from the crystal complex structure seem surprisingly reasonable. In figure III we present an example of such a reasonable distant configurations. While the BPTI is upside down and backwards from the crystal configuration, note that it has a lysine in the Trypsin specificity pocket. Instead of lysine 15, it is lysine 26 that is interacting with aspartate 189 of Trypsin; this is a conservative substitution. On the basis of 'chemical intuition' many of the 'false' positives DOCK2 generates seem sensible. Our expectation was that standard energy evaluation techniques would be able to distinguish amongst the 'true' and 'false' positives. We were therefore surprised by the difficulties we encountered.

(a) *Buried Surface Area*

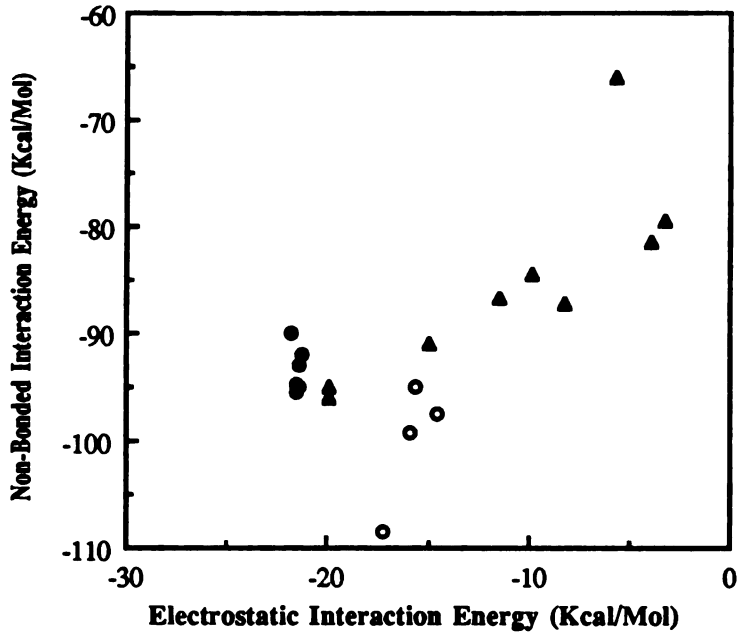
Chothia and Janin (Chothia & Janin 1975) have suggested that the burial of a significant amount of surface area is a necessary pre-condition for complex formation between two interacting proteins, and other workers have adopted the amount of surface area buried on complexation as a criterion for judging amongst putative complexes (Connolly 1985; Wodak & Janin 1978; Zielenkiewicz & Andrzej 1984). Wodak (Wodak *et al.*,1987) reports that buried surface area is highly correlated to how closely docked structures resemble the crystal configurations, and that in conjunction with simple mechanistic filters surface area burial can distinguish amongst 'true' and 'false' positives. Connolly finds that buried surface area alone is a sufficient criterion for distinguishing amongst putative complexes. While burial of surface area probably is a necessary condition for the formation of stable complexes, our results (table IV) suggest that it is not a sufficient criterion, even in conjunction with mechanistic information (see below). Most of the complexes that we have evaluated have large amounts of buried surface area, and many of the egregiously distant complexes are not only mechanistically sensible, but they bury more surface area than does the crystal complex.


-70
-80
-90
-100
-20
-10
0
Non-Bonded Interaction Energy (Kcal/Mol)
Elelctrostatic Interaction Energy (Kcal/Mol)

(a)


-50
-60
-70
-80
-90
-100
-110
-20
-10
0
Non-Bonded Interaction Energy (Kcal/Mol)
Electrostatic Interaction Energy (Kcal/Mol)

(b)



(c)

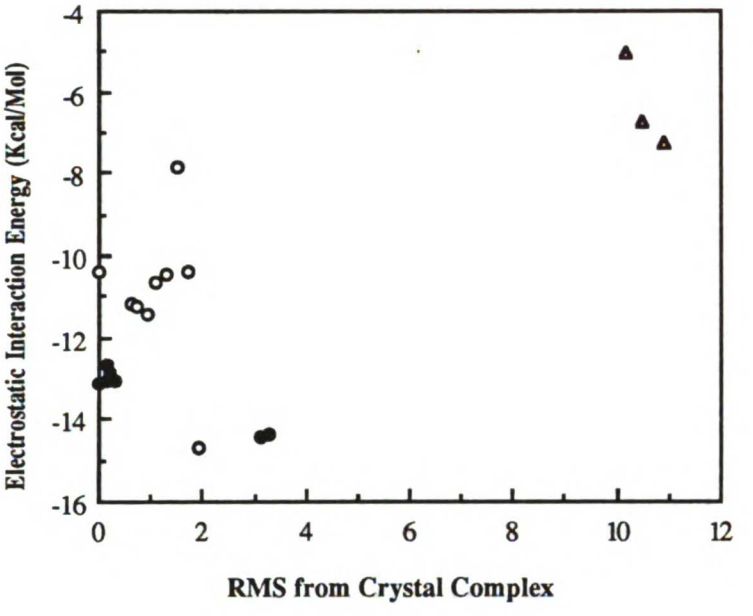
Figure VIII: Non-bonded interaction energies, in Kcal/Mol, of energy minimized complexes from AMBER are plotted against the electrostatic interaction energies from the same complexes. Closed circles represent bound conformer dockings, open circles unbound conformer dockings within 5 Å from the crystal structure. a. Subtilisin/Chymotrypsin-Inhibitor dockings. b. Chymotrypsin/Ovomucoid dockings. c. Trypsin/BPTI dockings.

(b) *Solvation Free Energy*

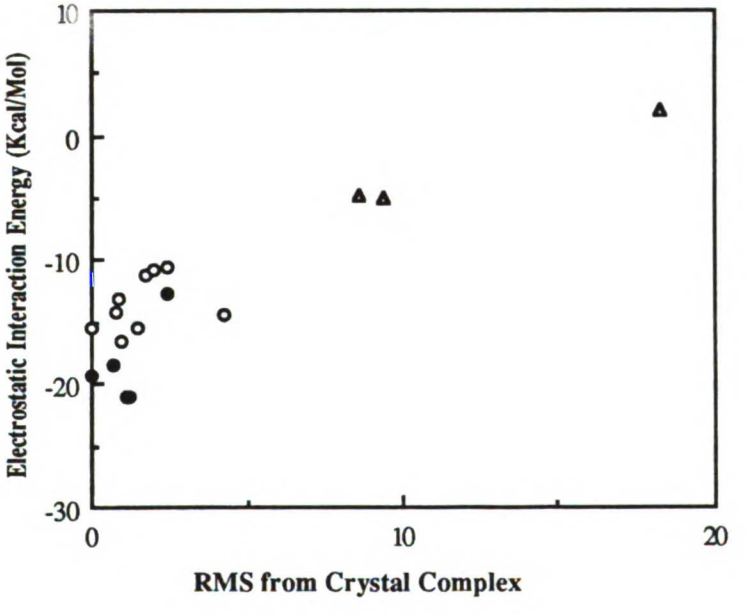
One problem with using buried surface area as a criterion for evaluating complexes is that it is not sensitive to polar complementarity at the interface. Solvation free energy techniques evaluate buried surface area based on atom type, and have been successfully used to distinguish between the native and mis-folded proteins (Eisenberg & McLachlan 1986; Novotny *et al.*,1988). We used two solvation free energy programs to evaluate the various docked complexes (Eisenberg & McLachlan 1986; Mitchel Lewis, personal communication), the latter of which was paramaterized

for protein complexes. Despite the higher sensitivity of these programs to chemical complementarity, neither program is able to reliably distinguish amongst the 'true' and the 'false' positives in the docked complexes (table IV).

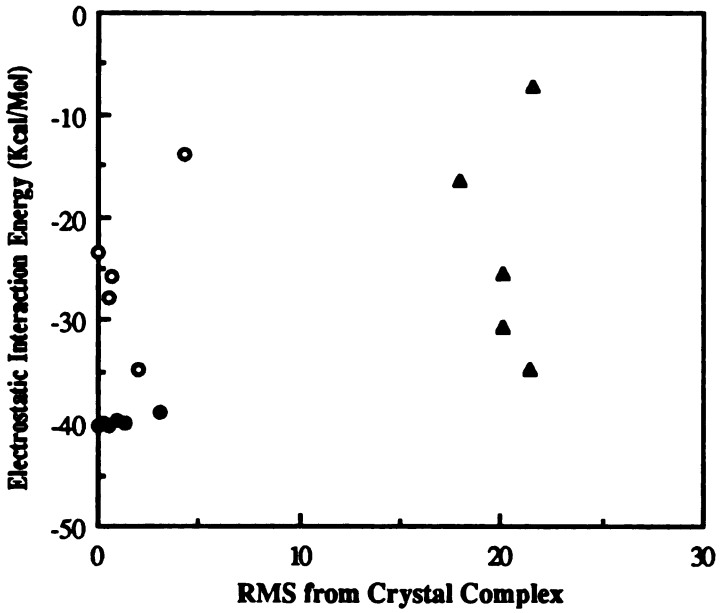
Unexpectedly, we find that there are many ways of putting a protease together with an inhibitor to bury a large amount of surface area. Surface area methods must then rely on their ability to distinguish between configurations on the basis of the chemical nature of the interface. For deciding between a correctly and an incorrectly folded protein, this distinction turns on the observation that correctly folded proteins have cores of buried hydrophobic residues. In protein-protein docking, however, it is easy to see how such a criterion might not apply. As folded proteins, the proteases and inhibitors that we have used have relatively hydrophilic surfaces. It is entirely a different problem to distinguish between a hydrophobic/hydrophilic interface and two hydrophilic ones. Thus our results, though indicating that solvation free energy methods cannot be used with confidence in protein docking, should not be taken to mean that solvation free energy cannot be used for protein folding calculations.


Electrostatic Interaction Energy (Kcal/Mol)
-4
-6
-8
-10
-12
-14
-16
0
2
4
6
8
10
12
RMS from Crystal Complex

(a)


Electrostatic Interaction Energy (Kcal/Mol)
10
0
-10
-20
-30
0
10
20
RMS from Crystal Complex

(b)



(c)

Figure IX: Electrostatic interaction energies in Kcal/Mol, as calculated for the complexes using DELPHI, are plotted against the RMS distances, in angstroms, from the inhibitor configuration in the crystal complex. Closed circles represent bound conformer dockings, open circles unbound conformer dockings within 5 Å of the crystallographic configuration and triangles unbound conformer dockings more distant than 5 Å from the crystal structure. a. Subtilisin/Chymotrypsin-Inhibitor dockings. b. Chymotrypsin/Ovomucoid dockings. c. Trypsin/BPTI dockings.

(c) *Packing*

Folded proteins tend to be well packed structures (Richards 1977), as are protein complexes (Chothia & Janin 1975). We used the packing evaluation program QPACK (Gregoret & Cohen 1990) to try to distinguish amongst the putative docked complexes (table IV). This program offers two criteria for analyzing packing: a residue-based pair-potential as well as a scheme for evaluating how evenly distributed the packing arrangements in the structure are. The pair potential results are not correlated to distance from the crystal complex and cannot be used to

distinguish amongst the docked complexes. The packing of the interfaces as judged by mean sphere size is relatively poor for all of the putative complexes of the 2PTN/4PTI, 5CHA/2OVO and 1SBC/2CI2 systems (mean residue sphere size of 79-84% as compared to 97% for typical crystal structures). This probably results from a few close contacts in the initial docked structures. On minimization the packing of the 1SBC/2CI2 complexes improves to the 89-94% (data not shown), which is in the range of the packing of the 2SNI crystal complex (92%). In neither the minimized nor the unminimized case, however, can any distinction be made between the close and the distant complexes based on packing.

(d) *Mechanism Filters*

One of the simplest methods for evaluating structures is to ask whether they correspond to what is known mechanistically about the system they represent. For the Trypsin/BPTI system, for example, it is known on the basis of enzymologic and mutagenesis studies that the scissile bond in BPTI is between residues 15 and 16, that lysine 15 in the inhibitor and aspartate 189 in the receptor are the specificity residues and that serine 195 is the chemically active residue in proteolysis. On the basis of such information it is possible to eliminate some of the more peculiar docked complexes, such as those in figures III c. and d., and mechanistic filters therefore seem useful to reduce the number of docked configurations that need to be considered. On the other hand, the mechanistic information used to derive the results presented in table IV was not able to exclude all of the distant complexes, for example that of the 10.93 Å RMS 1SBC/2CI2 configuration. Moreover, mechanistic filtering becomes much more permissive if filtering is done by residue type alone, and not residue type and residue number. In the Trypsin/BPTI case, for instance, when the mechanistic filter is that the putative complexes must have a lysine or arginine in the specificity pocket of the protease rather than lysine 15 specifically, many more of the non-native

complexes become acceptable, including the 21 Å RMS configuration shown in figure V. Mechanistic filters cannot, by themselves, be expected to reliably distinguish amongst 'true' and 'false' positives in the docking results.

(e) *Molecular Mechanics*

Energy minimization was the most computationally expensive technique that we used to evaluate the docked complexes, and was the only method that could in any sense reliably distinguish amongst the 'true' and 'false' positives (table V, figure VII). Plotting interaction energy against RMS deviation from the crystallographic complex (figures VII) makes two points. The low energy structure in all three systems is always within 5 Å of crystal structure, and there is often significant energy overlap between the putative complexes that resemble the crystal solution and those that do not. The inability to distinguish to better than 5 Å probably reflects imprecisions in the force field and problems with our choice of conditions, especially the absence of water in the simulations. The conformational differences between the bound and free structures of the proteins also reduced the resolution of the minimizations. Lys 15, for instance, is in a different conformation in the 4PTI structure than it is in the 2PTC structure. In the close configurations of unbound conformation dockings of BPTI, Lys 15 could only skirt the mouth of the P1 pocket, rather than fitting down into it as it does in the 2PTC structure. The conformation of Lys 26 of 4PTI structure is, fortuitously, such that it is able to fit down into the Trypsin specificity pocket in some of the distant configurations, which probably accounts for a large part of their favourable interaction energies. The energy overlap between the close and distant families suggests that even energy minimization finds some of the alternate binding modes to be reasonable. It is encouraging that the structure with the lowest interaction energy is always found within a close family. We caution, however, that the enegy differences between the best close and distant structures is often small and

that the magnitude of of the error in these calculations has not been established for macromolecular systems.

The use of molecular mechanics techniques for quantitative evaluation of intermolecular energies remains an active area research. There are many parametric and even strategic choices that alter the quantitative and, occasionally, the qualitative results. These include: atomic parameters, dielectric function, number of steps of minimization, non-bounded and electrostatic cutoffs, use of explicit solvent molecules, explicit counter-ions and so forth. In this project we used a linear dielectric function and did not include explicit solvent molecules or counter-ions. Under such conditions, varying the dielectric function or the number of cycles of minimizations often led to worse correlations between interaction energies of the complexes and how closely they resembled the crystallographic configuration. Thus, using a dielectric of 1.0R rather than 4.5R, minimizing for 2500 cycles rather than 900, or minimizing to a constant gradient rather than for a fixed number of cycles (results not shown), all led to poorer differentiation between the close and the distant configurations. We feel that the values for these parameters that we report here represent reasonable choices. But it is clear that other choices can have significant effects and that our results should be interpreted with caution. Comprehensive efforts at discriminating amongst alternate geometries should include explicit solvent.

(f) *Electrostatic Interaction Energy*

The highly ordered electrostatic environments of proteins (Warshel *et al.*,1989) suggests that electrostatic interactions across a protein-protein interface will be sensitive to geometric detail. We calculated electrostatic interaction energies using DELPHI (Gilson & Honig 1988) for both minimized and unminimized complexes. In the Subtilisin/Chymotrypsin-Inhibitor and the Chymotrypsin/Ovomucoid systems

(figures IX a and b) electrostatic interaction energy offers as good or better resolution between the 'true' and 'false' positives as does energy minimization. In the Trypsin/BPTI system (figure IX c), on the other hand, the correlation breaks down and the low energy structure is one that is dissimilar to the crystal structure. It is interesting that DELPHI has the most difficulty with the system where electrostatic interactions are probably the most important - the specificity residue for BPTI is a Lys, while for Chymotrypsin Inhibitor and Ovomucoid it is Met and Leu. The problems that DELPHI has in the 2PTN/4PTI system might reflect some of the same conformation difficulties that were alluded to in the discussion of the energy minimizations. In 4PTI Lys 26 is in a conformation that, in several of the low energy distant orientations, puts it in close contact with the specificity Asp 189 of the Trypsin. Lys 15, on the other hand, is in a conformation that does not allow it to interact with Asp 189 in the close orientations. We add the same cautionary note as above, that we report one particular set of of DELPHI parameters and have not made an exhaustive search of dielectric representations or ionic strength, nor have we dealt with the solvation aspects of electrostatic complementarity.

Reprise - Protein Docking

There are two features of this work that we feel might have bearing on future investigations of the protein docking problem. Firstly, despite the significant conformational differences between the bound and free forms of the receptors and ligands, especially in the specificity side-chains of the ligands, the docking method was able to regenerate the crystallographic configurations of complexes using the free conformations of the proteins. This suggests that even in their unbound forms, the invariant side-chains and backbone atoms contain enough information to specify to a good approximation, if not uniquely, the binding mode of complex. This is an encouraging result, since it implies that rigid body docking will be a useful initial

technique towards predicting a complex geometry starting from the unbound form of the components. Conformational changes will of course always be a critical and difficult factor in any such efforts. Secondly, the evaluation of putative complexes is quite difficult. Most of the simple methods are not, in our hands, able to reliably distinguish between the 'true' and 'false' positives amongst the various complexes. Even though energy minimization correctly identifies those structures within 5 Å of the crystallographic result as including the lowest energy orientation, some more distant orientations have low energies too. Further, it is unclear what is the average error associated with the potential energy function. There are more elaborate techniques that might be used to distinguish among these structures, and we have not comprehensively surveyed even the simple evaluation methods. Nevertheless, our results suggest that accurate prediction in the protein docking problem promises to be a difficult undertaking requiring careful modeling and sophisticated methods.

The difficulty that the complementarity procedures had in distinguishing between the 'true' and 'false' positive dockings probably reflects two features of the interfaces: most of them have at least some degree of genuine complementarity, and because they all involve the surfaces of folded proteins, they all look, to a certain extent, the same. DOCK2 does a good job at generating geometries for protein complexes. The sphere matching approach leads to complexes with large amounts of buried surface area while the contact scoring scheme insures shape complementarity. The complementarity differences between the close and the distant docked configurations in some cases seem genuinely subtle. The evaluation methods must be sensitive to energy differences between putative complexes that may be relatively small - a free energy difference of a few Kcals is enough to distinguish the native configuration from the non-native, as far as populations in solution or crystal structures are concerned. A slight inaccuracy or imprecision in interaction potentials summed over a large interface

will make such small differences impossible to detect. Our choice of modeling conditions may have contributed to such ambiguities. We ignored crystallographic waters in these studies, for instance, due to the ambiguities they present in the thermodynamics binding. Tightly bound waters may play an important role in complementarity at protein interfaces (McPhalen & James 1988). Predicting which few of the many waters typically present in the crystal structure of an unbound protein will play such a role in the complexed form of the protein is, however, a difficult problem.

The 'failure' of the evaluation methods may also be read in a generative light. The possibility exists that some of the distant configurations that came out of the docking represent physically sampled orientations. This suggestion was made to explain some results in earlier docking work (Wodak & Janin 1978). While most enzymologic and molecular modification studies are consistent with the crystal structure complex of Trypsin and BPTI (2PTC), for instance, there are undoubtedly other, higher energy configurations of the two molecules that are possible in solution (Elkana 1974; Rigbi 1971). One of these might be represented by the orientations of BPTI that put lysine 26 into the Trypsin specificity pocket (figures III and V), which all the evaluation methods agreed was reasonable orientation. Such a hypothesis might be tested by directed mutagenesis and enzymological techniques.

Future Applications

The ultimate application of any protein-docking method is prediction - given the structures of two uncomplexed proteins can the configuration of the complex between them be ascertained. Our results suggest that given good starting models, DOCK2 will be able to generate the correct configuration of two interacting proteins. Our

experience with simple energy evaluation methods suggests that distinguishing the 'correct' from the 'incorrect' structures, on theoretical grounds alone, will not be a simple task. The docking approach might be usefully used in conjunction with experimental approaches to suggest or refine hypotheses.

Another application for the results of this work is in the realm of testing of energy and complementarity evaluation methods for protein complexes. We have been able to generate many reasonable non-native configurations of protein-protein complexes that are indistinguishable from the native or near native complexes as judged by many commonly used energy evaluation techniques. These native, near native and non-native complexes form a data set of reasonable configurations. Workers might find this data set, and the docking approach in general, useful as discriminatory test cases for stringent trials of energy analysis methods that we have either not treated in this paper or not explored fully. The structures of the various docked complexes are available in PDB format from the authors on request.

Finally, by its ability to generate non-native complexes, the docking approach can suggest alternate binding modes in protein complexes that, while they are not present in an crystal structure and are probably higher in free energy, might still be sampled in solution and have interesting properties of their own. The docking approach lends itself to specific hypotheses regarding such non-native complexes that can be experimentally tested.

Conclusions

We have shown that a docking method can efficiently and correctly reproduce crystallographic configurations from both the bound and the unbound conformers of the

associating proteins. This suggests that much of the three dimensional information defining the interaction between the two molecules is present in each molecule individually, even in the absence of the other partner. Additionally, especially for the unbound conformations of the molecules, the method generates 'false' positives that are well packed, complementary and chemically reasonable. In contrast to earlier work on protein-protein docking (Connolly 1985; Wodak & Janin 1978; Zielenkiewicz & Andrzej 1984) we find that buried surface area is not a reliable indicator of resemblance of the docked complexes to the crystal structure results. Most of the other standard methods of complementarity analysis that we tried were similarly unable to distinguish between the 'false' and 'true' positives amongst the docked complexes. Electrostatic interaction energies were able to distinguish amongst these structures in two systems but not in a third. Total interaction energies from minimization consistently found the low energy structure, from amongst the complexes, to be one within 5 Å RMS of the crystallographic configuration. The difference between the lowest energy close and distant configurations were, however, relatively small.

Acknowledgments

We would like to thank the following people for help with various applications: Eric Fauman (CDSFIT), Celia Schiffer (solvation free energy), Mitch Lewis (solvation free energy), Lydia Gregoret (QPACK), Peter Kollman, George Seibel and Dave Pearlman (AMBER). Thanks also to Renee DesJarlais, Dale Bodian and Richard Lewis for insightful discussions.

References

Bernstein, F.C., Koetzle, T.F., Williams, G.J.B., Meyer, E.F., Brice, M.D., Rodgers, J.R., Kennard, O., Shimanouchi, T. & Tasumi, M. (1977) *J. Mol. Biol.* **112** (3): 535-542.

Blevins, R.A. & Tulinsky, A. (1985) *J. Biol. Chem.* **260**: 4264-4270.

Blundell, T.L. (1981) In *Structural Aspects of Recognition and Assembly.* (Balaban, M., Sussman, J.L. & Yonath, A., ed.), pp. 281-286, Balaban, Rehovat and Philadelphia.

Bode, W., Epp, O., Huber, R. & Laskowski, M. (1985) *Eur. J. Biochem.* **147**: 387-395.

Chothia, C. & Janin, J. (1975) *Nature.* **256**: 705-708.

Connolly, M.L. (1983) *J. Appl. Crystallogr.* **16**: 548.

Connolly, M.L. (1985) *Biopolymers.* **25**: 1229-1247.

DesJarlais, R.L., Sheridan, R.P., Dixon, S.J., Kuntz, I.D. & Venkataraghavah, R. (1986) *J. Med. Chem.* **29**: 2149-2153.

DesJarlais, R.L., Sheridan, R.P., Dixon, S.J., Kuntz, I.D. & Venkataraghavah, R. (1988) *J. Med. Chem.* **31**: 722-729.

Eisenberg, D. & McLachlan, A.D. (1986) *Nature.* **319**: 199-203.
Elkana, Y. (1974) In *Proteinase Inhibitors: Proceedings of the 2nd International Research Conference.* (Fritz, H., Tschesche, H., Greene, L.J. & Truscheit, E., ed.), pp. 445-453, Springer-Verlag, New York.

Fujinaga, M., Sielecki, A.R., Read, R.J., Ardelt, W., Laskowski, M.J. & James, M.N.G. (1987) *J. Mol. Biol.* **195**: 397-418.

Gilson, M., Sharp, K., & Honig, B. (1988) *J. Comput. Chem.* 9, 327-335. Gilson, M. & Honig, B. (1988) *Proteins.* **4**: 7-18.

Grau, U.M. & Rossmann, M.G. (1981) *J. Mol. Biol.* **151**: 289.

Gregoret, L.M. & Cohen, F.E. (1990) *J. Mol. Biol.* **211**: 959-974.

Hendrickson, W.A., Smith, S.L. & Royer, W.E. (1987) In *Biological Organization: Macromolecular Interactions at High Resolution.* (Burnett, R.M. & Vogel, H.J., ed.), pp. 235, Academic Press, New York.

Kuhl, F.S., Crippen, G. & Friesen, D.K. (1984) *J. Comp. Chem.* **5**: 24-34.
Kuntz, I.D., Blaney, J.M., Oatley, S.J., Langridge, R. & Ferrin, T.E. (1982) *J. Mol. Biol.* **161**: 269-288.

Laskowski, M.J. & Kato, I. (1980) *Ann. Rev. Biochem.* **49**: 593-626.

Lee, R.H. & Rose, G.D. (1985) *Biopolymers*. **24**: 1613-1627.

Marquart, M., Walter, J., Deisenhofer, J., Bode, W. & Huber, R. (1983) *Acta Cryst., Sect. B*. **39**: 480490.

McLachlan, A.D. (1982) *Acta Cryst.* **A38**: 871-873.

McPhalen, C.A. & James, M.N.G. (1987) *Biochemistry*. **26**: 261-269.

McPhalen, C.A. & James, M.N.G. (1988) *Biochemistry*. **27**: 6582-6598.

Neidhart, J.J. & Petsko, G.A. (1988) *Protein Eng*. **2**: 271-276.

Novotny, J., Rashin, A.A. & Bruccoleri, R.E. (1988) *Proteins*. **4**: 19-30.

Pickersgill, R.W. (1988) *Protein Engineering*. **2**: 247-248.

Richards, F.M. (1977) *Ann. Rev. Biophys. Bioeng*. **6**: 151-176.

Richmond, T.J. & Richards, F.M. (1978) *J. Mol. Biol*. **119**: 537-555.

Rigbi, N. (1971) In *Proceedings of the International Research Conference on Proteinase Inhibitors*. (Fritz, H. & Tschesche, H., ed.), pp. 74-88, Walter de Gruyter, Berlin.

Shoichet, B.K., Bodian, D.L. & Kuntz, I.D. (1992) *J. Comp. Chem*. **13**(3), in press.

Walter, J., Steigemann, T.P., Singh, H., Bartunik, H., Bode, W. & Huber, R. (1982) *Acta Cryst., Sect. B*. **38**: 1462-1472.

Warshel, A., Aqvist, J. & Creighton, S. (1989) *P.N.A.S*. **86**: 5820-5824.

Weiner, S.J., Kollman, P.A., Case, D.A., Singh, U.C., Ghio, C., Alagona, G., Profeta, S. & Weiner, P. (1984) *J. Am. Chem. Soc*. **106**: 765-784.

Wodak, S.J., De Crombrugghe, M. & Janin, J. (1987) *Prog. Biophys. Molec. Biol*. **49**: 29-63.

Wodak, S.J. & Janin, J. (1978) *J. Mol. Biol*. **124**: 323-342.

Zielenkiewicz, P. & Andrzej, R. (1984) *J. Theor. Biol*. **111**: 17-30.

Gloss to Chapter IV

This chapter introduces several novel techniques and resources that made the design effort it describes successful. The use of chemical complementarity scoring schemes through electrostatic and covalent recognition were significant advances from our earlier work, where shape scoring dominated. TS was the first system for which we used these more sophisticated schemes. Since the time of this work, these and more complicated evaluation techniques (Meng et al., 1992) have come to dominate all the modeling in our group. The solvation correction method mentioned in chapter four is also an important innovation; the form of its future implementation is a source of ongoing research in the group, the need for it is generally accepted. Consensus docking, described at the end of the chapter, should give us a better handle on structure prediction. By far the most significant development underlying the work in chapter 4, however, was our use of the Fine Chemical Directory (FCD) as a source of ligand models for inhibitor design.

Up until 1990 all the design work in the Kuntz group had depended on docking molecules from the Cambridge Structural Database (CSD). We would DOCK the CSD compounds into an enzyme site, evaluate their complementarity based on shape, and then try to design in chemical complementarity based on the CSD *templates*.

The acquisition of the FCD (Gunner et al., 1991) changed everything. The FCD consisted of commercially available compounds whose structures had been computationally generated (Rusinko et al., 1987). I was very excited to learn of this database (on the road with Tack, somewhere between Uppsala and the Arctic Circle) since it promised a way to get away from the long design and synthetic efforts that

were usually necessary with CSD database searches. Anything that docked well from the FCD could, after all, just be bought.

The strategy that finally resulted in successful inhibitors was informed at every stage by the application of theory to a specific system, TS. Most of the new algorithms that I put into DOCK to take advantage of the FCD (electrostatic scoring, solvation correction, consensus docking), sprang from a clear need imposed by the enzyme. Future progress in structure based inhibitor design will, in my opinion, depend on a serious commitment to experimental systems. The richness and complexities of biological receptors will be a constant challenge and delight for theory; we can agree with Hamlet that "there is more in heaven and earth, Horatio, than is dreamt of in our philosophy."

Another lesson of this work is that progress will be most rapid when we can restrict ourselves to simple experiments. Practically, this means trying as much as possible to buy rather than synthesize compounds. Obviously there will come a point when synthesis is unavoidable; the longer this can be delayed, however, the more critical one is likely to be in analyzing a result or interpreting a hypothesis. Synthetic efforts are often so time consuming that one is loath to design negative controls. Such controls cannot, however, be avoided without risk.

The issue of hypothesis falsification deserves special emphasis. There is little point doing an experiment, in a lab or on a computer, if you cannot expect to interpret both a negative and a positive result. In pursuing the leads in chapter 4 (figure 1), I found that I was more wrong than right in my personal predictions about whether a specific inhibitor would bind well or not. I tried compound X (figure 1) for instance, as a control, to prove to myself that the hydantoin ring was important; I fully expected that

X would not inhibit TS. Instead, I found that it was the best inhibitor that I had tested to that time. I vividly remember watching, 3 am on a Tuesday morning, 300 seconds of flat line on the spectrophotometer. I could not have breathed more than thrice in those five minutes. My negative control had completely shut down the enzyme, in a way that no inhibitor until then had. *Wrong again*, but I knew what it meant. My precious hydantoin rings could be completely discarded - this result led directly to the phenolphthalein series that has been so fruitful. By the same logic, negative results where I expected positive have been very powerful. Knowing the phenolphthalein bound to TS, I hoped that fluorescein would bind as well, since this is a very well explored family of molecules, potentially offering a great number of potential inhibitors. The lack of inhibition on fluorescein's part was, however, extremely informative; it has allowed to me to greatly restrict my models of phenolphthalein binding and will guide future work in this system.

The chapter that follows is purposefully brief, owing to page restrictions in the journal that we hope to publish it in. At the time of this writing we are awaiting, with some anticipation, the results of X-ray crystallographic experiments on the TS/phenolphthalein complex. These results will influence our conclusions. For this reason, the chapter remains less complete than I hope the paper will be.

Gunner, O. F., Hughes, D. W. and Dumont, L. M., *J. Chem. Inf. Comp. Sci.* **31**, 408-414 (1991)

Meng, E. C., Shoichet, B. and Kuntz, I. D., *J. Comp. Chem.* accepted for publication. (1992)

Rusinko, A., Skell, J. M., Balducci, R. and Pearlman, R. S. (1987). Sybyl Manual, Tripos Associates, St. Louis, Mo.

Structure Based Inhibitor Design in Thymidylate Synthase

Brian K. Shoichet*, Kathy M. Perry†, Daniel V. Santi†* and Irwin D. Kuntz*

* Department of Pharmaceutical Chemistry, University of California at San Francisco, San Francisco, California 94143-0446.

† Department of Biochemistry and Biophysics, University of California at San Francisco, San Francisco, California 94143-0448.

We used the crystal structure of thymidylate synthase (TS) and a molecular docking program to screen a structural database for chemical compounds complementary to the enzyme. Besides retrieving the natural substrate and known inhibitors, the docking program identified molecules previously unknown to bind to TS. We tested several of these putative ligands and found two different classes of novel inhibitors of TS. The crystallographic solution of complexes of several of these inhibitors bound to TS allows us to test our models to atomic resolution. These structures suggested new ligands with improved binding efficacy.

Structure-based inhibitor design has been a grail of biological chemists since the turn of the century (Fischer, 1894; Ehrlich, 1907). The hypothesis is: given a detailed structural model of a receptor, it should be possible to design a molecule that will bind to it. The elucidation of atomic resolution protein structures promised to make this goal achievable (Perutz, 1964). Despite much effort, however, a general approach to structure-based design has remained elusive (Cohen et al., 1990).

The problem has two refractory aspects. First, it is difficult to predict what the free energy of binding for a given molecule to a particular protein will be (Cohen et al., 1990). Second, the chemical diversity of the potential inhibitors is very great and largely unconstrained. Chemical/structural possibilities grow exponentially with molecular complexity; it is impossible to compute the 'best' ligand for a given site; it is very difficult to arrive at even a sensible one, without reference to known inhibitors or substrates. Recently, workers have successfully used a substrate-analogue approach with detailed structural models (Matthews et al., 1990), but only rarely has inhibitor

design been reported using the structure of a receptor alone (Goodford, 1984; DesJarlais et al., 1990).

We looked for novel inhibitors of thymidylate synthase, an important target for drug design(Santi and Danenberg, 1984). We began with the crystal structure of L. *casei* TS (Hardy et al., 1987; Finer-Moore et al., 1990) and the Fine Chemicals 3-Dimensional Database (FCD) (Gunner et al., 1991), a collection of commercially available small (generally less than 100 atoms) molecules. Partial atomic charges were calculated for 55,313 of the molecules in the database (E. Meng, unpublished results) using the Gasteiger algorithm (Gasteiger and Marsili, 1980). Using the program DELPHI(Gilson et al., 1988), we calculated the electrostatic potential of the active site. We used the program DOCK (DesJarlais et al., 1988; Shoichet and Kuntz, 1991) to explore the active site of TS with each FCD molecule, calculating an average of 10^4 orientations per molecule. DOCK scored the orientations based on three independent parameters: shape (Shoichet and Kuntz, 1991) and electrostatic complementarity to TS, as well as the ability to covalently modify the enzyme (Shoichet *et al.*, in prep). Of the 55,313 molecules searched, the best 600 by shape and electrostatic scoring , and the best 300 from the covalent list (1500 total) were saved. Ligand solvation correction (Rashin, 1990) and finally visual inspection (Huang, 1989) further thinned this list, removing many of the less likely candidates.

The natural substrate for TS, dUMP, and its inhibitory analogues, received good electrostatic scores, and were near the top of this list (Table 1). Also highly scored was pyridoxal phosphate (PLP), a known non-nucleotide inhibitor of TS (Chen et al., 1989). The program also retrieved molecules previously unknown to bind to TS (Table 1). We picked 25 of these molecules from the electrostatic and semi-covalent adduct lists for testing.

DOCK predicted that all the compounds would inhibit TS by binding in the dUMP region of the active site. Most of the 25 compounds inhibited TS with inhibitory constants in the 1-10 mM range (results not shown) and were not pursued. Three of the compounds had IC_{50}'s in the high μM range (Table 1). To gain further insight into the binding of these molecules, the co-complex structure of one of them, solisobenzone, with TS was solved crystallographically (Perry et al., 1992).

Ligand	Rank[a]		IC_{50} (μM)[b]
	Electrostatic[c]	Covalent[d]	
fdUMP	83	not defined	10
PLP	314	18	10
4-Thiouridylate	16	not defined	50
2FPA	175	9	200
4-Nitro-2FPA	157	34	350
Solisobenzone	193	not defined	900
Creatine monophosphate	1	not defined	no inhibition

Table 1: Representative Ligands from Docking Search

---

[a] Rank order amongst the 600 best scoring ligands in the electrostatic list or the 300 ligands in the covalent adduct list. Ligands were docked into the TS site, described by 67 spheres. Spheres were calculated using the SPHGEN [Kuntz, 1982 #35] and CLUSTER [Shoichet, 1992 #740] programs from the molecular surface [Connolly, 1983 #105] of the TS active site - defined as all residues within 15 A of the C198. We used a model of TS from which all bound waters and inorganic phosphate had been stripped.

[b] Inhibitory concentrations measured in a spectrophotometric assay [Wahba, 1961], using L. *casei* TS [Climie, 1990]. Values for fdUMP and PLP from Tang and Santi, unpublished results. IC_{50} values for 2FPA, 4-nitro-2FPA solisobenzone and phosphocreatine measured at 7.5 uM dUMP and 108 uM CH_2-H_4folate.

[c] Electrostic interaction energy is calculated from the sum of the products of partial atomic charges and the molecular electrostatic potential at the atomic locations for any given orientation of a ligand in the TS site in the original DOCK run. We corrected the interaction energies of the top scoring electrostatic and covalent adduct ligands based on the electrostatic component of the solvation free energy [Rashin, 1990], and ranked the orientations accordingly.

[d] Covalent adduct scores are the electrostatic interaction energy scores, corrected for solvation, of ligand molecules which have functionalities allowing them to form covalent or semi-covalent bonds with the C198 sulfhydril of TS. The following functionalities were so defined: aldehydes, ketones, esters, ß-haloketones and ß-haloaldhydes. In order to be acceptable as possible covalent adducts, the appropriate carbon had to come within 3.6 Å of the cysteine sulfhydril.

The solution of the solisobenzone complex was surprising: solisbenzone was largely not in the dUMP region of TS, inconsistent with our modeling (Perry et al., 1992). Instead, it was closer to the folate region, though partly overlapping the ribose and pyrimidine binding regions. The crystal complex had a phosphate ion in the arginine binding pocket, precluding the placement of the sulfonate moiety of the ligand in this region, as required by the DOCK model. In our docking calculations, we used a TS structure stripped of all waters and counter ions, including a phosphate ion bound in the arginine pocket, presuming that a good dUMP inhibitor could displace ionic phosphate. We re-docked solisobenzone into TS containing phosphate, using the orientations from the original docking run. The highest scoring orientations now clustered in the region of the crystal structure complex (results not shown). The electrostatic interaction energy of the crystal complex structure was much less favourable than that of the original DOCK model, going from -110 to -20 Kcals/mol, uncorrected for solvation. The shape score, on the other hand, was higher in the crystal configuration than in the original DOCK prediction.

We reasoned that molecules similar in shape to solisobenzone and chemically complementary to the solisobenzone site of TS should bind to the enzyme. MACCS-3D (Gunner et al., 1991) was used to look for FCD molecules that resembled the shape of solisobenzone. An initial survey suggested 30 candidate molecules. We docked these compounds to find which, if any, of the 30 were complementary to TS, using the coordinates of solisobenzone from the complex structure (Perry et al., 1992) as pseudo-atoms to guide the search of the TS site.

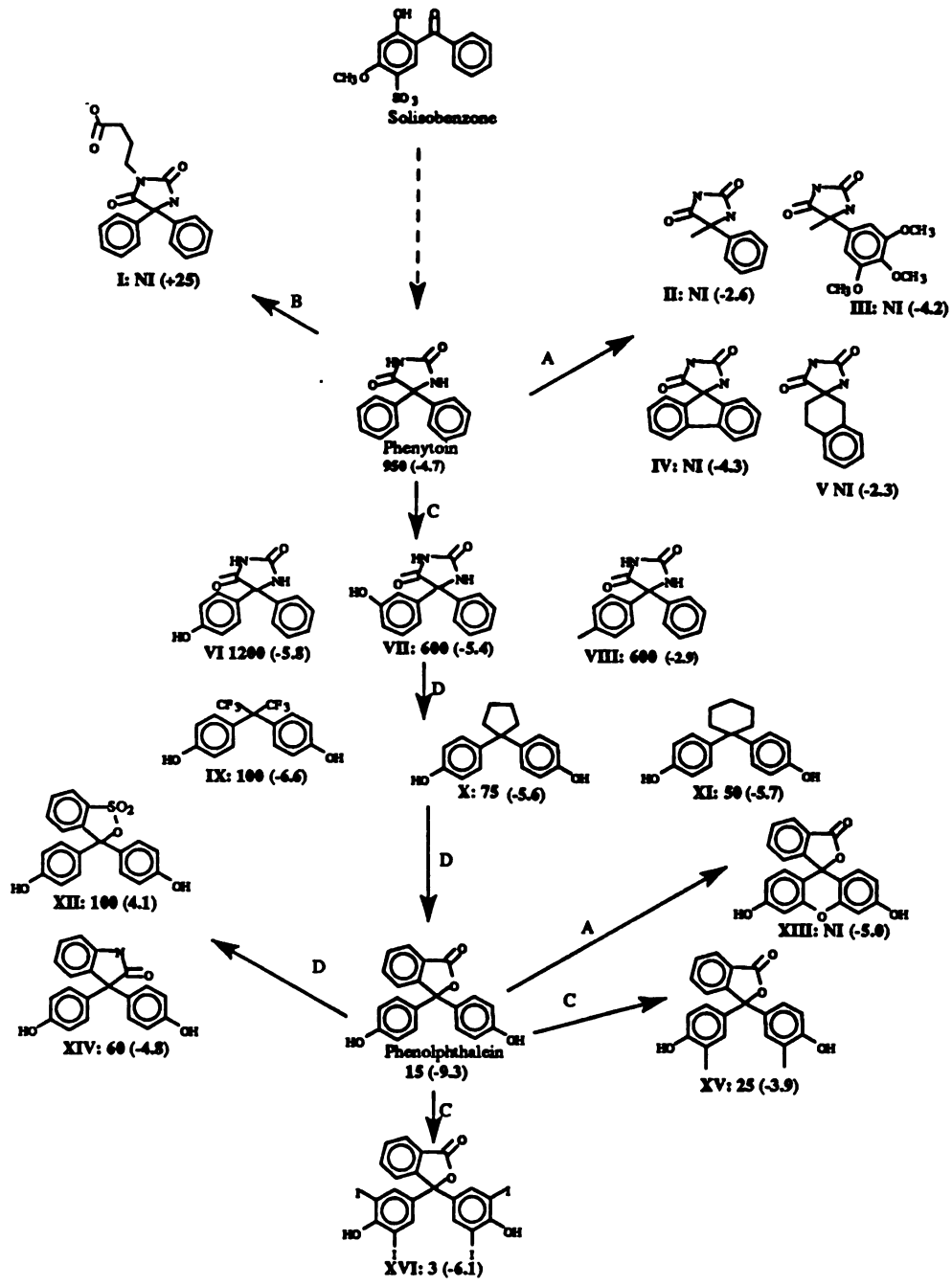


Figure 1: Design strategy for solisobenzone-site inhibitors. NI indicates no inhibition, bold face numbers are IC_{50}'s, numbers in parentheses are DOCK electrostatic interaction energies, corrected for solvation. Assays performed at 35 μM CH_2-H_4folate, 30 μM dUMP. Phenytoin was the lead secondary inhibitor in this class. A. Effect of removing or modifying the phenyl rings: compounds with only one phenyl (II and III) did not inhibit, neither did compounds which modified the 'splayed' conformation of the rings (IV, V and XIII). B. Effect of adding a negative group. C. Effect of adding substituents to the phenyl rings: polar groups at the 4 position decrease binding, at the 3 position binding is increased; a methyl group at the 4 position increases binding. D. Effect of modifying the hydantoin ring: hydrophobic groups substitute favourably for the hydantoin system. Increased size contributes to binding. Specific hydrogen-bonding or polar interactions are indicated by the decreased efficacy of XIV vs. phenolphthalein. C'. Effect of adding substituents to the phenyl rings: Large radii atom substituents are tolerated and improve binding (XVI).

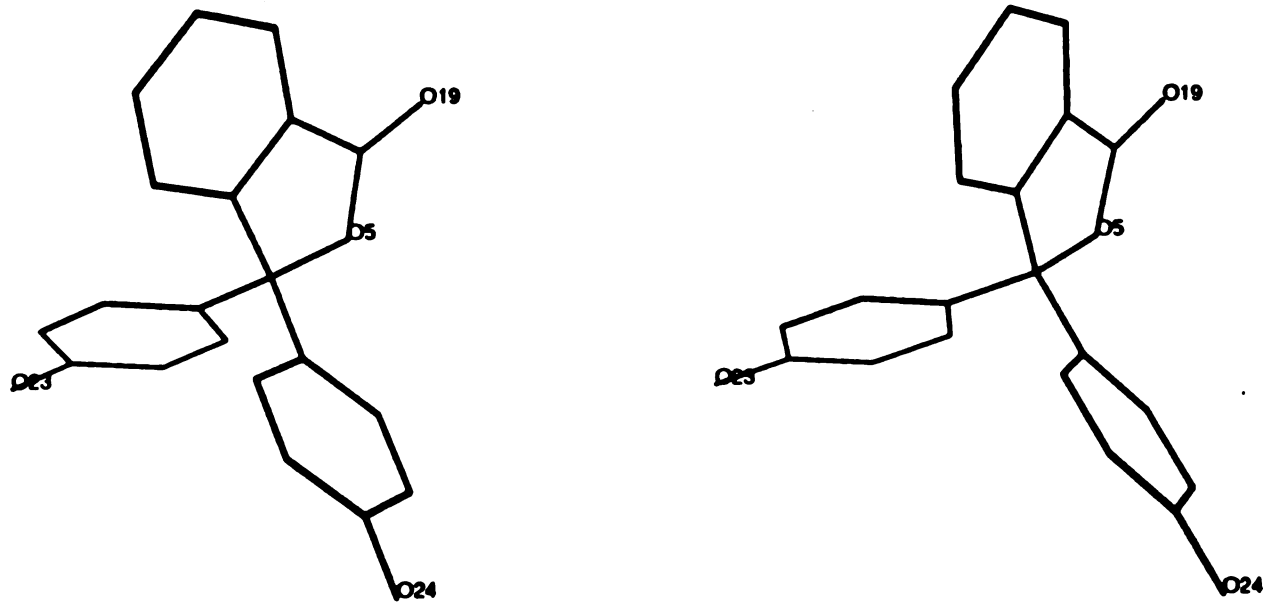
The highest scoring of these 30 'secondary' molecules were phenytoin and several of its derivatives. We tested these compounds in an enzyme assay (Wahba and Friedkin, 1961) and found that several phenytoins inhibited TS, again with IC_{50}'s in the high μM range (Figure 1). We expected the phenytoins and related molecules to be well represented in the FCD. A similarity search using MACCS found 992 molecules that were similar to phenytoin. Each of these 'tertiary' molecules was docked to assess chemical complementarity to the solisobenzone site in TS. We assayed several of these 'tertiary' molecules for inhibition. High scoring compounds usually inhibited TS, poor scoring compounds did not inhibit TS measurably (Figure 1).

The range and number of this tertiary class of molecules allowed us to investigate the aspects of the three ring system that are important to inhibition (Figure 1). Both phenyl rings are necessary for strong binding. The phenyl rings must be unbridged, allowing them to adopt a splayed conformation, since neither compounds IV, V nor fluorescein (XIII) (Figure 1) inhibit TS. The hydantoin ring is not important and can be usefully replaced with other moieties. The best of these to date is from the phenolphthalein family, where the hydantoin is replace by a phthalide ring. Hydrogen bonding substituents on the phenyl rings can favourably affect binding, more so at the meta position than the para. The decreased inhibition of the oxindole (XIV) suggests that the phthalide lactone makes specific electrostatic contacts with the protein. TS can accommodate bulky atoms at the 3 and 5 positions of the phenyl rings.

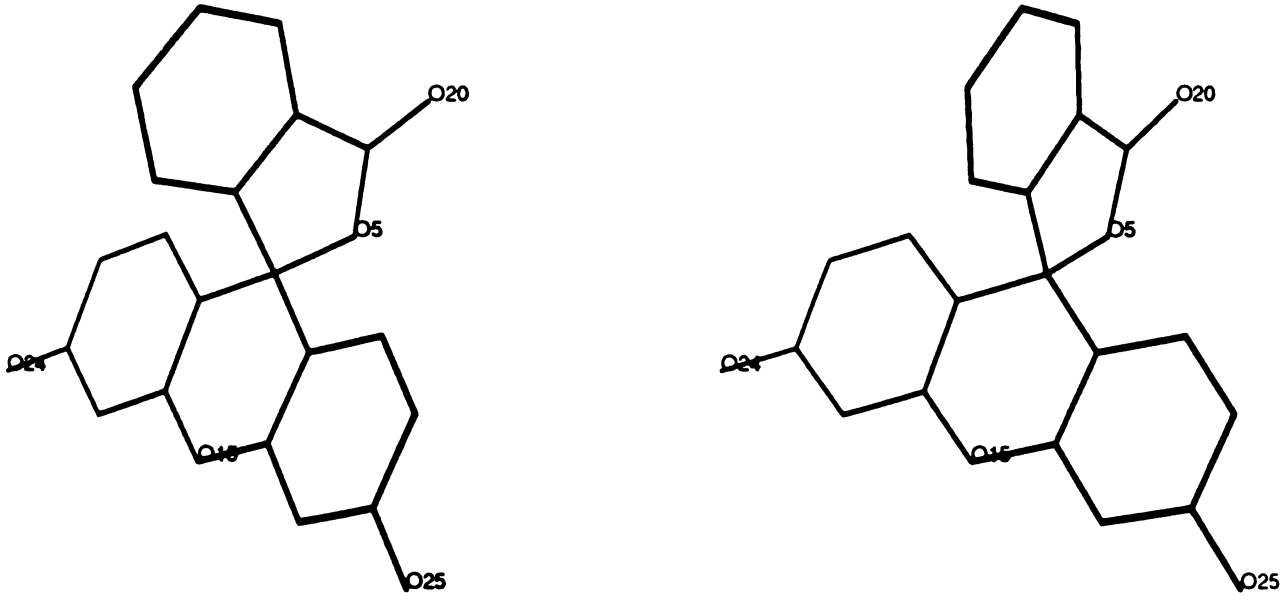
While the electrostatic scores for the 'tertiary' molecules were roughly correlated with the IC_{50}'s (Figure 1), they must be interpreted with caution. For instance, though DOCK correctly predicted that phenolphtalein would bind better than phenytoin, it also predicted that fluorescein (XIII) would about bind as well as the oxindole (XIV); instead, fluorescein does not detectably inhibit. The energy difference between a good

and a poor inhibitor can be small compared to the inaccuracies in DOCK, and we consequently cannot distinguish reliably amongst compounds with similar scores (Shoichet and Kuntz, 1991). We can more confidently exclude molecules which are unlikely to bind based on solvation (charged molecules in neutral sites, for instance) or shape.

Despite the inaccuracies inherent in the energy evaluations, docking can, under favourable circumstances, predict the binding geometry of an inhibitor in a receptor. Two conditions must be met. First, we need several ligands with different inhibitory potentials but similar structures, that can be assumed to bind similarly to the receptor. Second, we must assume that the receptor is mostly rigid. Since this assumption underlies all of our docking work, we do not multiply the constraints on our model by making it. The phenolphthalein derivatives and fluorescein satisfy the first condition well. Phenolphthalein differs from fluorescein only by one oxygen atom, which enforces planarity on the 'phenyl' rings of fluorescein (Figure 2); this small difference is presumably responsible for the lack of fluorescein's inhibitory effect. Also, tetraiodophenolphthalein is a good inhibitor of TS. We can thus insist that any favourable complex of phenolphthalein with TS also must be favourable for tetraiodophenolphthalein but unfavourable for fluorescein. This consensus analysis suggests two models for the phenolphtalein-TS complex (Figures 3 and 4, Graph 1). Model I (Figure 3a) is more consistent with the experimental results than model II (Figure 4a), but the latter could not be ruled out.



(a)



(b)

Figure 2: a. Stereo picture of phenolphthalein. b. Stereo picture of fluorescein.

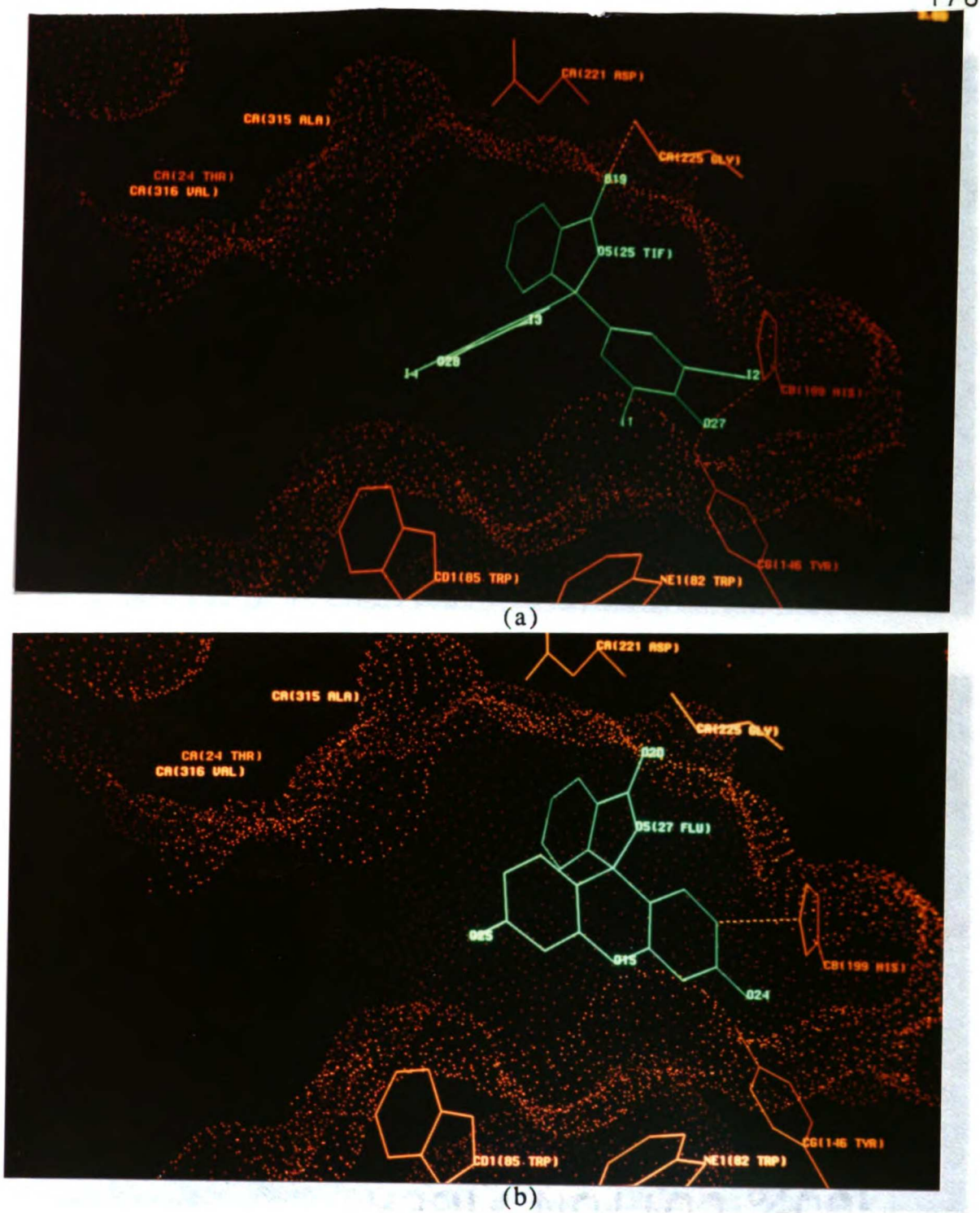
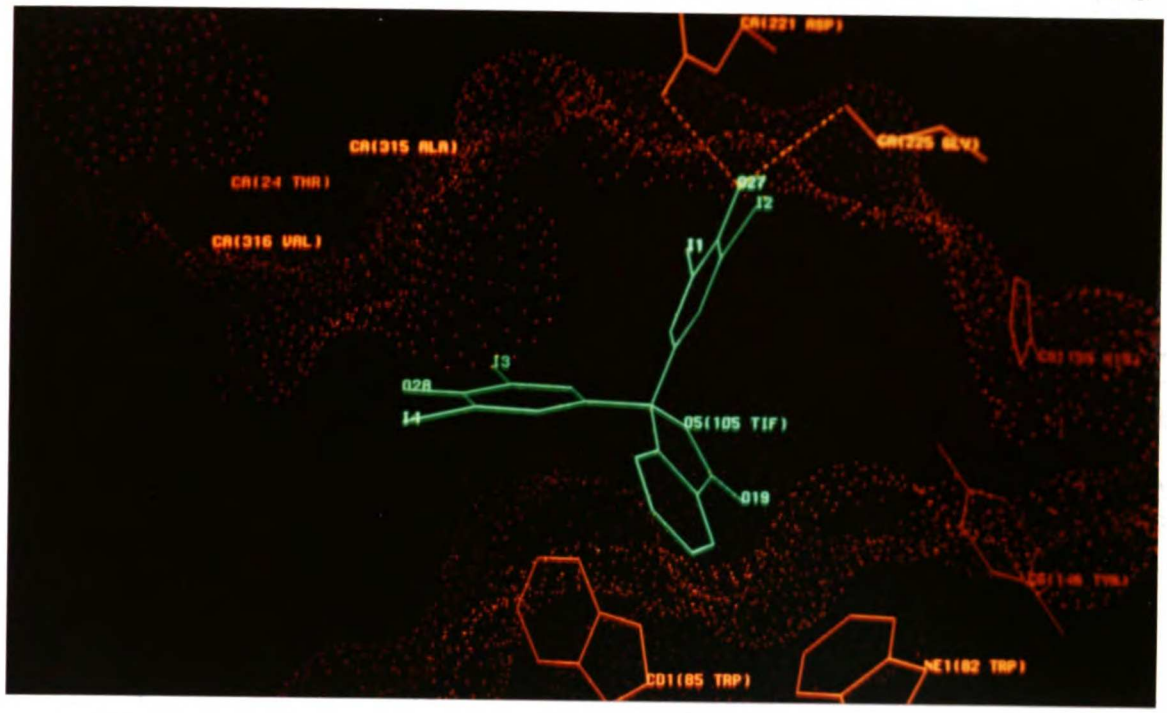
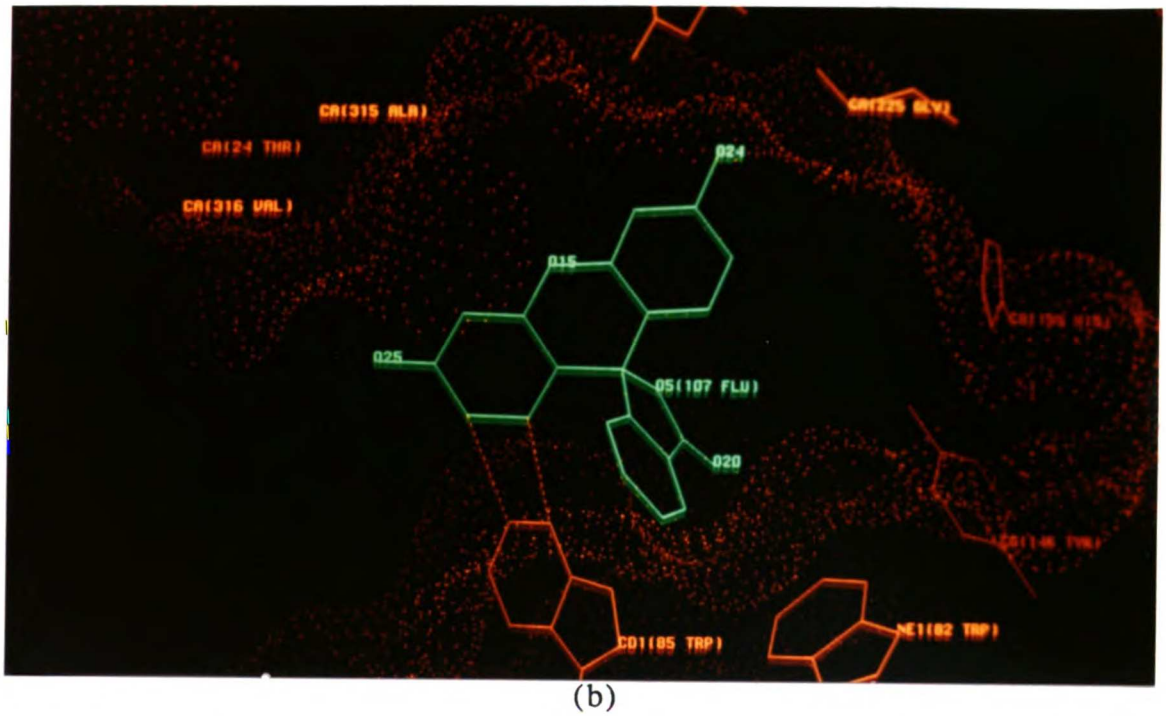


Figure 3: a. 3',5',3'',5''-Tetraiodophenolphthalein in consensus orientation 1. H199@ND1 - O27 (2.5 Å), Y146@OH - O27 (2.9 Å), G225@N - O19 (3.6 Å). b. fluorescein in consensus 1, excluded by close contacts to H199. Ligands in green, protein in red. Distances between hydrogen-bonding groups in yellow.

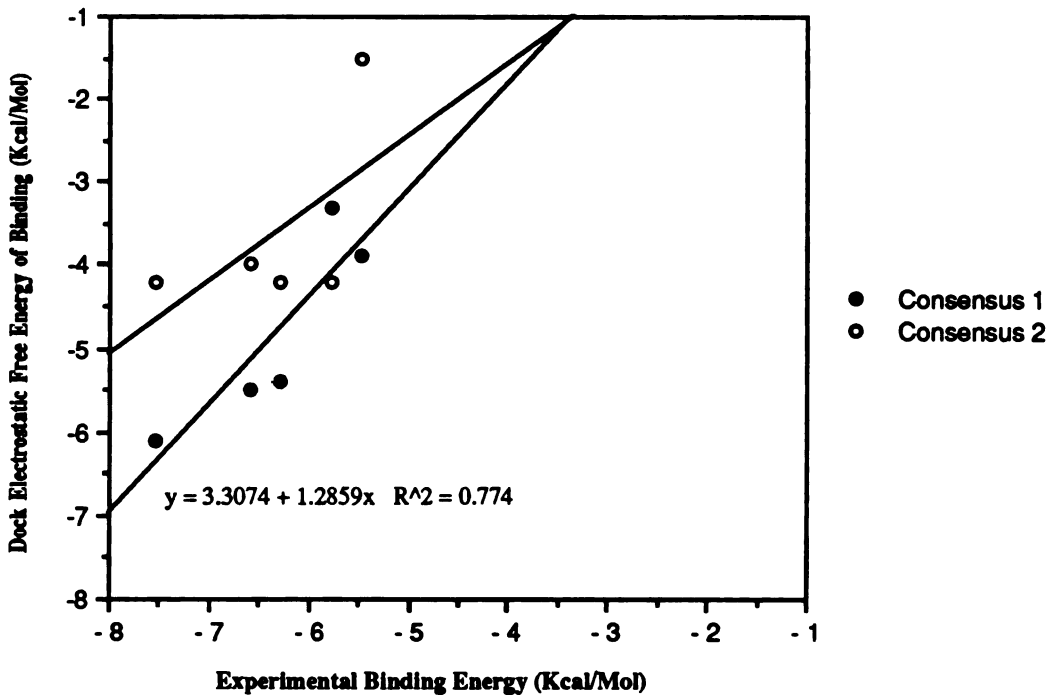


(a)



(b)

Figure 4: a. 3',5',3'',5''-Tetraiodophenolphthalein in consensus orientation 2. D221@OD2 - O27 (2.8 Å), G225@N - O27 (3.3 Å). b. fluorescein in consensus 1, excluded by close contacts to W85.Ligands in green, protein in red. Distances between hydrogen-bonding groups in yellow.



Graph I: The correlation between the predicted electrostatic component of the ΔG of binding (from DOCK), in the two consensus models, and the RTln(IC_{50}) of phenolphthalein and compounds IX, XI, XV and XVI (figure 1).

A detailed model of inhibitor-enzyme interactions, whether from modeling or crystallography, informs our search for better inhibitors. We note, for instance, that the putative alternate folate binding pocket (Montfort et al., 1990) of TS is uncompensated in our models. Y233 of this pocket is, however, within 6 Å of one of the phenol rings in consensus model 1 , suggesting that the addition of an alkyl hydroxyl group on the 3' position of the ligand will improve binding. In neither consensus model are both phenolic oxygens ligated, suggested that one can be eliminated, improving inhibition by reducing the desolvation cost of binding.

A hopeful, if unanticipated, result of this study was our ability to improve on a lead compound without having to resort to chemical syntheses. The chemical diversity in

the FCD allowed us not only to improve binding by 2-3 orders of magnitude, but also to establish which moieties in the inhibitors were important to binding. Restricting ourselves to compounds that are commercially available made the hypothesis-testing cycle rapid.

The phenolphthalein inhibitors do not resemble either dUMP or CH_2-H_4folate, and do not bind exclusively in either the pyrimidine or the folate binding sites. This result is made possible by a receptor-based approach to inhibitor design. TS's biological role is thymidine biosynthesis; its structure, however, can accomodate diverse ligands unrelated to its natural substrates.

References

Chen, S. C., Daron, H. H. and Aull, J. L., *Int. J. Biochem.* **21**, 1217-1221 (1989)

Cohen, N. C., Blaney, J. M., Humblet, C., Gund, P. and Barry, D. C., *J. Med. Chem.* **33**, 883-894 (1990)

DesJarlais, R., Sheridan, R. P., Seibel, G. L., Dixon, J. S., Kuntz, I. D. and Venkataraghavan, R., *J. Med. Chem.* **31**, 722-729 (1988)

DesJarlais, R. L., Seibel, G. L., Kuntz, I. D., Montellano, P. O. d., Furth, P. S., Alvarez, J. C., DeCamp, D. L., Babé, L. M. and Craik, C. S., *Proc. Natl. Acad. Sci. USA* **87**, 6644-6648 (1990)

Ehrlich, P., *Chem. Berichte* **42**, 17 (1907)

Finer-Moore, J. S., Montfort, W. R. and Stroud, R. M., *Biochem.* **29**, 6977-6986 (1990)

Fischer, E., *Chem. Berichte* **27**, 2985-2993 (1894)

Gasteiger, J. and Marsili, M., *Tetrahedron* **36**, 3219 (1980)

Gilson, M., Sharp, K. and Honig, B., *J. Comp. Chem.* **9**, 327-335 (1988)

Goodford, P. J., *J. Med. Chem.* **27**, 557-564 (1984)

Gunner, O. F., Hughes, D. W. and Dumont, L. M., *J. Chem. Inf. Comp. Sci.* **31**, 408-414 (1991)

Hardy, L. W., Finer-Moore, J. S., Montfort, W. R., Jones, M. O., Santi, D. V. and Stroud, R. M., *Science* **235**, 448-455 (1987)

Huang, C. (1989). University of California.

Matthews, D. A., Appelt, K., Oatley, S. J. and NG, H. X., *J. Mol. Biol.* **214**, 923-936 (1990)

Montfort, W. R., Perry, K. M., Fauman, E. B., Finer-Moore, J. S., Maley, G. F., Hardy, L., Maley, F. and Stroud, R. M., *Biochem.* **29**, 6964-6976 (1990)

Perry, K. M., Shoichet, B., Kuntz, I. D. and Stroud, R. M., manuscript in preparation (1992)

Perutz, M. F., *Scient. Am.* **211**, 64-76 (1964)

Rashin, A. A., *J. Phys. Chem* **94**, 1725-1733 (1990)

Santi, D. V. and Danenberg, P. V. (1984). Folates in Pyrimidine Nucleotide Biosynthesis. Folates and Pterins. New York, Wiley. 345-398.

Shoichet, B. and Kuntz, I. D., *J. Mol. Biol.* **221**, 327-346 (1991)

Wahba, A. J. and Friedkin, M., *J. Biol. Chem.* **236**, 11-12 (1961)

Future Directions

A "future directions" chapter is a conceit. Ideas are as Galatea to Pygmalion for scientists, and our fondest hope is that they should shake of the clay of their first molding and take on a life of their own. I cannot write of future work without admitting to a spirit of promotion.

There are several experiments that follow from the work I have presented that would, I believe, well reward the effort required to pursue them. To better insure that this is true, I will restrict myself to recommendations that are easily performed. Experimentally, I will only propose tests where negative results are as interpretable as positive ones.

Faster Code

There are quite a few docking methods in the current literature (Wodak et al., 1987; Goodsell and Olson, 1990; Cherfils et al., 1991; Jiang and Kim, 1991; Wang, 1991); ours is by far the fastest* and is the only one that has been adapted for inhibitor

---

* Our algorithm is at least one and possibly two or even three orders of magnitude faster than those of Jiang & Kim, and faster still than Chen's algorithm. Cherfils *et al.* report that their program compares to the run times of Jiang & Kim. To compare run times I did the following:

> Jiang & Kim report that run time scaled linearly with the product of number of points they used to represent each protein. They note that it took "approximately 5 days" to dock 4pti (1013 pts) into 3ptn (2469 pts) on a Vax 785. Extrapolating, I conclude that it took them 7.8 days to dock 2ptci (1203 pts) into 2ptce (3248 pts) on the same machine. With the new focusing code, it takes us 53 seconds to dock 2ptci into 2ptce, and 15.5 minutes to dock 4pti into 2ptn on a SGI PI 25 (unpublished results). A PI is about 8-12.3 times faster than a 785. Using this to convert their 785 times to PI times, their algorithm would have taken between 912-1350 minutes for 2ptci/2ptce, and between 584-860 minutes for 4pti/3ptn. Thus, we are, on the face of things, 10^3 faster for 2ptci/2ptce, and about 50 times faster for 4pti/3ptn.

There are two caveats to this argument. The first is that we did not look at as much of configuration space as Jiang & Kim, restricting ourselves to configurations of PTI in the active site of trypsin. Counterbalancing this, we achieved more 'accurate' results than they did, getting to within 1-2 degrees of the crystallographic configuration, whereas they, due to the coarseness of their grid, never did better than 10-15 degrees. Because of the nature of focusing, I do not believe that looking over the whole of trypsin would have increased our run time by more than a factor of 2 or so, though this remains to be tested.

design. As fast as it is, there is always a crying need to make it faster. Every time we improve the program's speed, we are able to search larger databases. The problem of too many molecules, an embarrassment of riches, is becoming acute. Up until 1990, we restricted our searching to a 10,000 molecule subset of the Cambridge Structural Database. With the advent of FCD runs, first adapted for for use against TS in late 1990, the DOCK group now typically finds itself docking 55,000 molecules in a go. Collaborations with industry and Chemical Abstracts may shortly challenge us to dock databases of a million or more molecules. DOCK is currently about an four times faster than it was in 1988 (not including improvements from clustering, focusing or faster computers). To address the very large databases on their own terms it will have to become at least an order of magnitude faster again. There are several obvious - and easy - ways to begin to do this.

Colouring the Matching Graph*: To sample configuration space adequately we commonly look at between 5,000 - 10,000 orientations of each FCD molecule in an active site. Reducing the number of configurations that must be sampled would speed up the program. The best way to do this is by pre-defining regions of likely complementarity in the receptor site.

The number and quality of configurations from a DOCK search depends on the similarity between receptor spheres and ligand atoms. We define similarity geometrically at match level. No matter what their chemical environment or atom type, if four ligand atoms can be superimposed on four receptor spheres DOCK considers a match to have occurred, and will produce an orientation of ligand in a

---

* This idea reflects a conversation that Tack and I had one sunny afternoon during a game of frisbee. Similar schemes have been discussed in the literature in other contexts, especially by Gordon Crippen.

receptor. This matching algorithm is easily refined to consider the chemical characteristics of the ligand atoms and of the receptor spheres. No point is served, for instance, by considering ligand orientations involving the superimposition of a quartenary nitrogen on a sphere in the center of an arginine pocket. Eliminating such matches requires three steps. First, the chemical environment of spheres and atoms must be defined. For ligand atoms this will be trivial. Spheres are slightly more difficult, but their environment should be easy to map using a program such as DELPHI (Gilson et al., 1988) or GRID (Goodford, 1985). Second, one must define matching rules. Positively charged atoms should not be matched to spheres where the molecular electrostatic potential is also positive, and so forth. Third, one must pre-colour the graph. The greatest speed ups will occur if some matches are not even attempted at sub-graph build up time, rather than recognized as they are being built and eliminated. The bin algorithm allows us to do this (chapter 2). Essentially, all that is required is that spheres be 'colour coded' into different bin types. Only bins of the appropriate 'colours' will be allowed to contribute nodes to a growing sub-graph. This is exactly analogous to the distance label filter currently implemented in the program.

New Technologies*: The recent advent of parallel computers should enable a much faster implementation of DOCK. Database docking is an inherently parallel algorithm, since every molecule is treated independently. Each molecule can be docked by a different processor in a parallel machine - this should not require very much recoding. Docking on a highly parallel computer could conceivably take no more time than that required for one processor to dock one ligand. Even if the processor

* This is mostly Tack's idea. I mention it here to emphasize its importance, and ease of implementation. A beginning graduate student or postdoc who wishes to have an immediate affect on how a docking effort precedes could do worse than to spend two weeks of their time modifying the code to permit parallel searches.

speed is 1/10 that of the workstations we typically use, we would still achieve a speed up of three to four orders of magnitude as far as the user is concerned. The drawback is, of course, that parallel machines are still expensive and experimental; this is not yet a general solution.

Algorithms

As important as efficiency improvements are, they are basically boring to the chemist. Below I outline some algorithms that might allow DOCK to make more sophisticated chemical judgements about molecular complementarity.

Solvation Corrections: The availability of the FCD has led us to include increasingly complex scoring functions in the program, beginning with the inclusion of electrostatics and covalent scoring in DOCK2.1 (chapter four), and more recently by scoring orientations according to the AMBER potential (Meng et al., 1992). After reviewing the results of the first TS runs, I realized that by going to chemical complementarity schemes we had unwittingly created the need for still another level of sophistication in our treatment, that of accounting for solvation. In nature, ligand binding is a competition between two environments: the aqueous and that of the receptor site. One cannot predict relative binding affinities without reference to both phases. I have therefore recently implemented a method for ligand solvation correction (Rashin, 1990) and applied it in the TS work (DOCK2.2, manuscript in preparation). While this correction has worked fairly well in TS, which recognizes doubly negative phosphates, and seems to have done well in trypsin, which recognizes singly positive amines, as well (C. Corwin, personal communication), it remains to demonstrate that the algorithm is general. Reasons why it might *not* be general are readily apparent. We do not, for instance, consider the desolvation cost of the active site, but only the

ligands. We also consider the ligand to be fully desolvated once bound to the receptor, which is unlikely to be true. Nevertheless, the ability of the Rashin correction to sensibly re-order ligands in systems as different as TS and trypsin suggest that it might be widely useful. To more thoroughly test the algorithm, the program should be applied to a neutral site and perhaps one even more highly charged than TS. A good neutral site might be that of chymotrypsin, for which there are numerous known inhibitors, and even several co-complex structures. A very challenging polar site might be DNA (Tack's suggestion). In chymotrypsin we would expect neutral ligands to be selected, while charged molecules would be excluded by the solvation penalty. In the DNA system, one would hope that poly-valent amines such as spermidine, despite a significant solvation penalty, would nevertheless have a negative free energy of electrostatic binding. The neutral case will almost certainly work. The DNA test case will, conversely, be a very serious challenge to the method.

Other solvation schemes are under consideration by my learned colleagues. I will not discuss them here, except to suggest that they should be tested in the same sorts of systems that I have set out for the Rashin algorithm now implemented in DOCK2.2.

Database Organization: A cherished use for DOCK, in the minds of its authors, is as an aide to creativity. By presenting the modeler with diverse chemistries in the context of a receptor site, we hope to inspire novel inhibitor design. Towards this end, we consider likely complexes involving a large number of heterologous molecules. In a typical DOCK search, we save between 200 and 600 molecules out of the over 55,000 that are considered by the program. The way DOCK chooses these 600, however, leads to overly homogeneous final lists. Most favourable modeled complexes are dominated by one or two interactions - this will be most true in highly charged systems. In TS, for instance, most nucleotides have very good electrostatic

interactions with the enzyme. This leads to repetition of what is essentially the same chemistry in the final list. Once one knows that dUMP is predicted to bind to TS, little is gained by having fdUMP, TMP, UMP, BrdUMP and so on, in the list of ligands as well. Rather, by over-representing this chemical class, one precludes different chemistries that might otherwise show up. To overcome this problem, the database should be pre-organized into families of similar molecules. Only the best molecule from any family will then be saved in a final list, though all molecules will still be docked. Each molecule in such the final list will thus represent a *family* of likely inhibitors. There are a number of algorithms that might be used to this end, including some in development by Tack and Guy Bemis* in the lab and one program from Jeff Blaney (unpublished results) that seems well behaved.

Chemical Similarity: Chemical similarity is underused in the lab. Looking for compounds that resemble a lead compound, and that can be bought or otherwise acquired without resorting to costly and time consuming syntheses, is the way for groups at UCSF to pursue follow-ups to a lead compound in the future. The ability to look for similar compounds and then *buy* them has allowed me to begin with a millimolar lead and end with what is currently a low micro-molar inhibitor (Chapter 4) in five weeks time. This approach to 'secondary' design will be fruitful, I suspect, in other projects as well.

---

* Tack and Guy's notions of the virtues of database organization go beyond mine. They suggest that a good clustering algorithm should allow one to only dock representative molecules from each family, and thus dramatically reduce run times. This turns on how far one trusts one's clustering. If one could really pick the best representative from each family than they would be correct, and the database organization would contribute enormously to our efforts, especially towards treating the large corporate databases. I suspect, however, that the 'best' representative is too slippery a concept for any clustering scheme, and, moreover, will differ for each receptor context. Comparing docking runs with and without clustering should go a long way towards resolving this question, however.

Our similarity searches are now limited to FCD look-ups using MACCS (Gunner et al., 1991). While the FCD is properly the first place to look, it is not our only resource. Chemical Abstracts Service (CAS) also offers a structure-based search facility. While CAS compounds will undoubtedly be harder to acquire than the FCD ones, it's likely that small quantities of at least some interesting compounds will be available. As an example of the possible benefits of a CAS search, I looked up phenolphthalein derivatives online. A simple search projected between 1600-3000 compounds. Looking at 50 of these, chosen at random, 25 seemed worthwhile to test.

Experiments

Non-Native Configurations: Mutating K15A in PTI is known to reduce binding affinity to trypsin by 10^7. This still leaves a micromolar inhibitor. To what does this remaining interaction energy owe? Quite possibly binding depends on the overall complementarity between PTI and trypsin as they exist in their crystallographic configuration. Another possibility, however, is that other binding modes are explored in the K15A mutant. Chapter three suggests that a configuration that places K26 in the specificity pocket of trypsin will be favourable. Mutating the K15A mutant further to K26A will test this hypothesis. If the K15A/K26A mutant binds significantly less well than the K15A mutant, this will be consistent with the model put forward by DOCK, and will provide a persuasive example of the 'deconstructive' power of modeling in the context of structure. If, on the other hand, there is no significant change in binding affinity in going from K15A to K15A/K26A, this would imply that our finding of non-native complexes is really only an artifact of still inadequate models of molecular recognition. Falsification of the non-native complex hypothesis should convince us

that more attention needs to be paid to getting energies correct, and will slip from under us the stool of 'non-native but sensible,' on which we now precariously rest.

TS Inhibitor Design: Just as I was sitting down to write this dissertation, the TS inhibitor design work shot off in a new direction (Chapter four). In five weeks I was able to drive from a millimolar inhibitor to a micromolar one. I suspect that we have not reached the limits of this series of compounds. On the next page are eleven putative inhibitors that I will test for the following reasons:

I and **II** allow us to investigate the ability of the receptor to favourably accommodate hydrogen bonding partners in the solisobenzone site, as well as further refining our consensus models. Consensus model one, for instance, suggests that **I** will be a better inhibitor than the des-amino analogue (chapter four) already tested. Consensus model two, on the other hand, suggests that **I** will bind worse than its analogue.

III-VII allow us to play with the positions flanking the phenolic hydroxyls. Given the uncertainties in our binding calculations, it is difficult to make a strong prediction for this series. The results should nevertheless give us a better understanding of the electronic and bulk effects of these substituents. **VII** will be especially informative in this regard.

VIII and **IX** allow us to test the roles of the phenolic hydroxyls. Our current binding models suggest that the hydroxyls are be important for binding - their removal should result in lower binding affinity for **IX**, even allowing for solvation corrections. Note that this prediction contradicts that which would be made by arguing analogously to

the phenytoin series results (chapter 4). Compound **VIII**, on the other hand, might still in bind well to TS. If consensus model one, where the phenolic hydroxyl is complemented by H199 and Y146 , is correct **VIII**

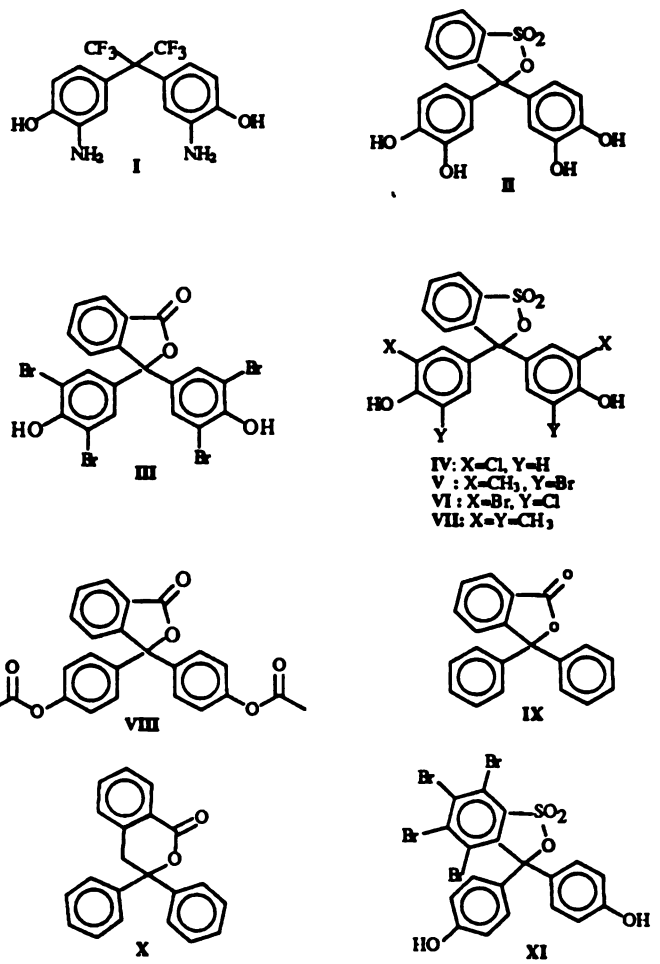


Figure I: Putative inhibitors of thymidylate synthase (untested as of this writing).

might well have improved binding compared to phenolphthalein. If consensus model two is the true binding mode, conversely, **VIII** should show reduced binding to TS.

X and **XI** allow us to test the importance of the phthalide ring to phenolphthalein binding. Neither consensus models one nor two can accommodate **XI**; if **XI** binds it falsifies the consensus configurations. I haven't tested **X** computationally, it will allow us to explore in a less dramatic way than **XI** the ability of TS to accommodate larger ring systems.

All these compounds are commercially available. Testing them will take three days of lab work.

Species Specific Design: Species specific inhibitor design is an important consideration in drug development. With the structures of two prokaryotic TS's solved to high resolution, and the human enzyme well on its way, we can consider implementing a species-specific design program. The phenolphthalein class of inhibitors might well provide a good jumping off point for such an effort. Consensus docking suggests that the phenolphtaleins bind in a pocket that differs in some important residues (E84A, W85N, V315M) between prokaryotic and eukaryotic TSs. We have preliminary evidence that the phenolphthalein structures have different specificities for human and L. *casei* TS. Along with Dale Bodian, I have developed algorithms that allows us to dock specifically to those regions of the site that differ most between two structures (Shoichet, Bodian, unpublished results).

And so it ends, "Not with a bang, not with a bang..." Or perhaps like Bottom - "I die, I die, I die, I die," - if only he would. I'll make a last plug for molecular reductionism and the receptor approach:

Peter Medawar, defending molecular reductionism generally, wrote (Medawar and Medawar, 1983):

> By representing a composite whole as a function of its component parts, we are almost automatically empowered to envision a domain of possible wholes other than that which formed the original subject of the analysis. Indeed, perhaps the simplest way to epitomize reductionism is to envision any particular frog as one realization in a universe of Possible Frogs, any one of which *might* have become real, though in fact only one did; likewise the whole world is seen reductively as one of only a whole domain of possible worlds - not necessarily the best.

For all of its hubris, such reductionism is a liberating mental discipline. By considering recognition from the viewpoint of the receptor, we free ourselves from the intellectual tyranny of inhibitor-based structure activity relationships. Correspondingly, we enfranchise the receptor as a self-contained molecular system, one that we can analyze, disassemble and interpret without reference to what Nietzche referred to as the nightmare of history. Whenever we applied our receptor-based approach to biological systems (chapters 1, 3 and 4) we were led to conclusions that the ligand binding data would not suggest. In chapter three, for instance, we found that we could sensibly dock protease inhibitors in ways that are inconsistent with the crystallographic solution, while in the chapter 4 we discovered a new class of inhibitors that do not bind as a folate or a pyrimidine, but rather seems to finds a middle site of its own. Should we call this site the phenolphthalein binding site? No doubt this would annoy the teleology of the enzymologists, but from a structural point of view it makes as much sense as refering to a pyrimidine or a folate site. By applying a Popperian dialectic to structure, we have pulled some very particular and unexpected species from the universe of "Possible Frogs."

Cherfils, J., Duquerroy, S. and Janin, J., *Proteins* **11**, in press (1991)

Gilson, M., Sharp, K. and Honig, B., *J. Comp. Chem.* **9**, 327-335 (1988)

Goodford, P. J., *J. Med. Chem.* **28**, 849-857 (1985)

Goodsell, D. S. and Olson, A. J., *Proteins* **8**, 195-202 (1990)

Gunner, O. F., Hughes, D. W. and Dumont, L. M., *J. Chem. Inf. Comp. Sci.* **31**, 408-414 (1991)

Jiang, F. and Kim, S. H., *J. Mol. Biol.* **201**, 79-102 (1991)

Medawar, P. B. and Medawar, J. S. (1983). Aristotle to Zoos. Cambridge, Harvard University Press.

Meng, E. C., Shoichet, B. and Kuntz, I. D., *J. Comp. Chem.* accepted for publication. (1992)

Rashin, A. A., *J. Phys. Chem* **94**, 1725-1733 (1990)

Wang, H., *J. Comp. Chem.* **12**, 746-750 (1991)

Wodak, S. J., De Crombrugghe, M. and Janin, J., *Prog. Biophys. molec. Biol.* **49**, 29-63 (1987)

BRODART CAT. NO. 23 012 PRINTED IN U.S.A.