Title
Modeling the energetics of protein folding

Author
Thomas, Paul Denis

Publication Date
1996

Peer reviewed|Thesis/dissertation

Modeling the Energetics of Protein Folding

by

Paul Denis Thomas, Jr.

DISSERTATION

Submitted in partial satisfaction of the requirements for the degree of

DOCTOR OF PHILOSOPHY

in

Biophysics

in the

GRADUATE DIVISION

of the

UNIVERSITY OF CALIFORNIA

San Francisco

Date **University Librarian**

Degree Conferred: . . . . . .

To Anette and Sebastian

Acknowledgments

I want to briefly mention the people and institutions that have been immediately and indispensably involved in the work that comprises this dissertation. The Preface that follows gives a more detailed and exhaustive account of my path toward and through these years.

First of all, I want to thank my advisor and mentor, Ken Dill. I mean "advisor" and "mentor" in the broadest sense. Ken has helped me learn how to do scientific research, but even more importantly, how to think about problems and communicate their solutions. Any clarity that comes through in this dissertation is because of his direct and indirect influence. Beyond that, his faith in me never seemed to waver even when my own faith was flagging. His role at times almost bordered on the psychiatric.

I also want to thank Tom James, who gave me my "first shot" at being a scientist despite my lack of any demonstrable scientific knowledge. He gave me my first view of the world of proteins and the physics used to explore that world. In Tom's lab, Brandan Borgias, Uli Schmitz and Shauna Farr-Jones were all exceptionally helpful to me, and good friends.

Tack Kuntz has offered me indispensable and very patient advice, both scientific and personal, from before I entered graduate school. I want to thank the members of my orals committee— Tack Kuntz, Fred Cohen and Dave Agard, and especially the chairman Peter Kollman— for their help in formulating the work in this dissertation; and the readers of this dissertation, Ken Dill, Fred Cohen and Tack Kuntz.

I am indebted to many members of the Dill lab both past and present, particularly to Hue Sun Chan and Klaus Fiebig for enjoyable discussions and pearls of physical wisdom; and to Rick Rodgers, Dave Yee and Danny Heap for a virtually bug-free computing environment.

The UCSF Graduate Group in Biophysics cannot go unrecognized. It is a unique program, and its breadth of research opportunities gave me a "big picture" as well as considerable freedom. The doors of the professors in the program were always wide open. I would like to thank my many friends in the program for making both the rough and smooth spots in graduate school alot smoother still, Julie Ransom for all of her support and effort on my behalf over the years, and Roger Cooke for helping to teach me how to teach.

I gratefully acknowledge the financial support of the Howard Hughes Medical Institute, which bought me the freedom to pursue the project of my choice. Much of this dissertation has already appeared in *Protein Science*, *Protein Engineering* and the *Journal of Molecular Biology*.

Finally, love and thanks go to my family— my parents, Anne and Denis; brother Christian and sister Elizabeth; grandparents Joe and Elizabeth; and especially to my wife and muse, Anette, and to our beautiful baby son Sebastian— for giving me more than words can express.

Preface

Art is science made clear.

Jean Cocteau, *Le Rappel à l'Ordre.*

From science to art and back again

I came to science the long way round. I must have been about ten or so when I dug into the darkest reaches of my dad's filing cabinet and pulled out the manila folder marked "Graph Paper, Vellum, Onion Skin, etc.". Many times I had noticed Dad reaching for that same folder while he was sitting at his desk, chewing on his pipe-end and "thinking." He would be right there— theoretically I could have touched him, but instead I would just watch him like I would watch someone who was stone asleep. He was always even farther away than that. So when one day I had made a couple of "observations" about the physical world, that folder seemed the logical place to go.

I picked out a piece of graph paper with the smallest squares— you could fit more on it that way— and sat down at his desk. The first thing I drew was a parabola describing the path of a model rocket fired at an angle. Earlier that day, I had launched my first model rocket, and I noticed in its trajectory that the downward half of the path traced a steeper angle than the upward path. A couple of weeks before that, I was rolling a ball on the beach, on the gradual slope carved by the high tide surf. I had seen the same thing there too— an asymmetric parabola. The ball came down the slope at a greater angle than I had kicked it up. Thinking back on these observations, I'm not sure what is more striking— that I noticed that these two imperfect parabolas must be pointing to the same underlying principle, or that I even expected a perfect parabola in the first place. Anyway, after I had

crafted my two penciled parabolas and surrounded them with the obligatory scrawl of descriptive words, even adding a vertical dashed line marking the apex of each curve, I was really impressed with myself. It was my first feeling of scientific discovery.

More than that, it was a step toward adulthood. I was still under the child's-eye impression that one of the bonuses of being an adult was having an understanding of *everything*. As expected, that view soon gave way to adolescent mistrust of straightforward explanations of anything. "Objective truth" was just a collection of opinions trying to curtail your freedom. Art, it seemed to me, was more honest in its uncertainty, ambiguity and subjectivity. This was no "passing phase." When I was twenty, my summer-job employer asked me what my plans were after finishing college, and I ingenuously replied, "I'm going to be a poet." That autumn I received from this same employer a newspaper clipping in the mail— it read "The diminishing writer's market." Apparently he was really worried about me. I must have sounded as serious as I believed I was.

Why, then, through all these years— through the Bachelor of Arts in English Literature, through the cafe sessions sipping cappuccino while discussing "high art" with "friends," through the attempts to write poetry, one-act plays and short stories, through all the hours spent writing papers to expose the demons that drove the really great writers— why did I continue to "nurse" this interest in the "hard" sciences? At the time, I thought I took classes linear algebra, biology, physics and chemistry just to maintain the illusion of myself as a "Renaissance Man in the Age of Specialization." But the truth was that they were just damn interesting.

When I was finishing up my honors dissertation comparing Edward Bond's play *Lear* to its Shakespearean namesake, my advisor, Professor Reginald Foakes, told me that he would like to recommend me to the faculty of his college at Cambridge, to pursue my doctorate there. That's when it hit me like an epiphany (yes, I think I had an epiphany— after reading about it in so many works of Literature it actually condescended to visit me).

I could see my life closing shut like moving backward through a tunnel— I was going to spend the rest of my life writing about what other people had written! I don't know why, but this thought literally made me sick to my stomach. Something about it felt so... *vicarious*. At the same time, I had been reading about formally undecidable propositions, about chaos theory and about molecular biology. These fields, on the other hand, were so open, so first-hand. They were still nascent enough (especially compared to the dinosaur of Literature) that they gave off the excitement of definition and discovery. I decided right there, in the corridor of Rolfe Hall, in the center of the English department, to become a scientist.

Of course, any writer will tell you that you can't have an epiphany without first laying out a drama which it can transcend. I'll just sketch it here so none of us has to put up with any weak attempts at creating any real drama. Imagine, if you can, a young adult who somewhere in his soul, despite all indications to the contrary, still believes it is possible to know it all. Why not an artist and a scientist both? Of course, you probably want to know why "artist" and "scientist" covers all the bases— well that's just intellectual snobbery, plain and simple.

He's the only one he knows who dares to cross with impunity the line dividing North (Art) from South (Science) Campus. Then he finds a man who has done just that, not just in the insulated environment of UCLA, but in the real world. Enter Dr. Robert Root-Bernstein, an accomplished biologist *and* painter. The lore was that UCLA couldn't decide whether to hire him in the Biology or the Art Department.

Now you'd think the young adult would be a bit upset to find he wasn't unique in what he perceived as his "Renaissance man" abilities. But he wasn't. In fact, he found that it wasn't really arrogance after all that was driving him. Dr. Root Bernstein's class was titled "The creative process in art and science." Underlying this apparently discordant interest in both art and science there was a single principle, just like the parabolas in the air and in the sand. The seed was planted, and even though UCLA decided not to hire Dr.

Root-Bernstein because he had neither published enough biology papers nor had enough paintings hanging in galleries, the seed would grow.

He's getting a little bored by 18th century English literature. It's his last required class for his English degree, and it's his only non-scientific class these days. *Clarissa Harlowe* was pretty good, *Tom Jones* was maybe a little better, but *Pride and Prejudice* didn't live up to its reputation. The student's papers on *Clarissa* and *Pride and Prejudice* had been lackluster at best, so when he wrote something really stunning on the scientific paradigm in *Journal of the Plague Year*, professor Georges Rousseau asked him to come to his office. After a few minutes of directed questions, Dr. Rousseau sums it up succinctly: "You are clearly not very interested in this class. But you need it to graduate. I don't need another student who doesn't want to learn. So what can I do to make you want to learn? How does this sound— I'm writing a book right now on the history of vitalism, and I could use an editor to help me with the current scientific views in biology." The student came in with a copy of *Northanger Abbey* under his arm, and left with Driesch's *The History and Theory of Vitalism.*

Three hundred students were waiting for Dr. Robert Goldberg to start the first lecture of Molecular Biology. Fewer than a hundred celebrated the end of the last day of class ten weeks later, and of these, a small group had become galvanized into a small family. One student in particular will never forget the second day of class. He hadn't done his reading (actually, he had read *Journal of the Plague Year* instead), and believed he was safely blending into an ocean of (still, at that time) about three hundred faces. That's when the law was broken. The Law of Large Lecture Halls clearly states that "There is a polarized flow of questions and answers, questions only flowing from student to professor." The corollary is "Even should this polarity be momentarily reversed, the chances that you, Paul, will be chosen are vanishingly small." But there was Paul, all of a sudden standing behind a microphone in the middle of the hall, with all six hundred eyes on him:

GOLDBERG: Design an experiment... to show that it is the bond between the beta and gamma phosphates of ATP that is hydrolyzed by a ribosome and not some other phosphate bond.

PAUL (trying the easy way out): I have no idea. Actually I didn't even do the reading.

But Goldberg smells blood and isn't about to fall for that one. He makes Paul think it out, logically, right there in front of the class. Paul forces himself to stay calm and think, even though every pore is screaming at him to bolt. Paul doesn't even know the chemical structure of ATP, what a ribosome is, or anything at all about radioactive labeling, but he still manages, not without some help, to come up with the *general idea* of labeling the phosphates. After he sits back down and the adrenaline leaves his body, he realizes that this is where the challenge is— this is creative problem solving. He isn't sure whether to rage at Goldberg or to thank him for giving such a visceral lesson in what it takes to be a scientist.

That was the first quarter of hard work that I ever did. It was glorious work— not for the facts I learned but the fact that I learned to use facts. This was the crucial dramatic backdrop for the epiphany in Rolfe Hall.

Unfortunately, it takes more than an epiphany to get you into graduate school in a subject you haven't formally studied. That's when you need someone powerful to take a risk on you. Dr. Thomas James at UCSF was willing to take that risk on me. He needed someone to do some scientific computer programming, and suggested that I learn FORTRAN before the summer of 1988 began. I took the suggestion, and spent the year after I graduated from UCLA in Tom's lab progressing from programmer to scientist. By the end of the year, I had published my first scientific paper, and had laid the theoretical groundwork that would lead to another. After that, it was just a formality to join the Graduate Group in Biophysics at UCSF.

How to choose a research project

In the simplest classification scheme, there are two kinds of research project. The first I will broadly refer to as the "safe" project: applying an existing set of techniques to a new problem. Calling this the "safe" route isn't really fair, since I have seen some of my peers embroiled in problems to which the old techniques are not easily applied. The second I call the "risky" route. It is working on a problem which is not yet solved. It is a risk because there is a chance that you will make absolutely no progress. The best way to stack the odds in your favor is to choose your research advisor wisely.

I would like to tell you that I carefully considered the pros and cons of different research projects. But the truth is that I chose from the gut. It was a conversation with Tack Kuntz that first set me thinking about an intermediate-resolution model of protein folding. He gave me Tim Havel's thesis to look over, and I have to confess it was really beyond me, filled with geometric proofs and reams of computer code. This was not how I wanted to begin. I turned to the literature, and was particularly intrigued by a 1989 paper in *Proteins* by our very own (UCSF Biophysics student) Chuck Wilson when he was with Doniach at Stanford. All I can say is that I just felt that this was the right problem to be working on.

It was about this time that I first saw Ken Dill in action. At the time, there was a group of UCSF professors spanning several departments who were linked together via a grant to fund the "Macromolecular Workbench" (MMWB). The MMWB held a (fairly formal) meeting every month in the Laurel Heights auditorium, replete with cold drinks and trays of hors d'oeuvres. Tom James was a member, so I was always informed about the meetings. At my first MMWB meeting, Ken casually proceeded to tell us about a model of proteins that was so incredibly simple. You couldn't help but be impressed by the *cajones* of a guy who could stand up there and tell you connected beads configured on a checkerboard really captured some important properties of proteins. I was transfixed.

I made an appointment to see Ken soon after that. I wanted to talk to him about developing an intermediate resolution model of protein folding. By "intermediate" I mean of lower resolution than full-atomic molecular dynamics and of higher resolution than Ken's elegant two-dimensional model. By the time I left his office, I knew that this was the guy I wanted to work with. He has a very easy-going manner, and is an excellent and thoughtful listener. But what really appealed to me was the overall philosophy behind his approach to modeling protein structure: start with the simplest model based on the right physics, and then add detail and additional parameters only if necessary. Maybe this seems like an obvious and quintessentially scientific approach, and not very original. But that's just it— it was *daringly* original. I suppose it was just the apparent complexity of the protein folding problem that made almost everyone who worked on it simply assume that its solution had to be complex too. Ken's two dimensional square proteins were not, I hear, well received at first by the scientific community. So when they paid off years later, they paid off big. Ken, and the rest of the lab through him, are still receiving the dividends to this date.

Now not all grad students get the chance to simply follow their guts. This kind of freedom is helped along if you have your own source of money. I'm not sure whether Ken had the money to support my graduate school career, but I *am* sure that having received a Howard Hughes Predoctoral Fellowship didn't hurt in his decision to take me on as a student. And in turn, I have to thank Tom James for his help on my fellowship applications— I was offered both the NSF and HHMI fellowships and actually had the luxury of choosing between them.

The story behind the science

I want to record something about the way in which science is done, as opposed to the way in which it is presented. I have been assured that "no one is going to read your dissertation anyway," so I figure my thesis is probably an appropriate place for it. The chapters of this dissertation are all either published or in the process of being published. So they are science in its presented form. Here's how it really happened.

I. Local and nonlocal interactions in proteins, and mechanisms of alcohol denaturation.

This is where I cut my teeth. I really had no idea what I was doing when I began this project. By the time I was finished, I had an understanding of statistical mechanical models and a reasonable intuition about how to think of proteins in this context. But it was a slow road.

Ken told me that Hue Sun (Chan & Dill, 1990b) had found that secondary structures (helices and sheets) arise from simply enforcing compactness in 2D lattice-bound polymers. This was a surprising and remarkable result. Secondary structures have traditionally been associated with geometrically specific hydrogen bonding between backbone amide and carbonyl groups. Hue Sun's observations suggested that regular structures are also efficient packing schemes, and that secondary structures may arise as a consequence of the nonspecific hydrophobic interactions that drive proteins to become compact in water. In helices, hydrogen bonds are *local* in sequence (between monomers four residues apart in primary sequence), while hydrophobic interactions are *nonlocal* in sequence, occurring over any arbitrary sequence separation.

This division between local and nonlocal interactions, while historically important in classical polymer research, was still somewhat foreign to the field of protein research

(though that was being rapidly changed by Ken's seminal 1990 review in *Biochemistry*). Ken suggested to me that the simplest model of local vs. nonlocal interactions in proteins would be a two-interaction model. The "HP model" introduced by Lau & Dill (1989) has a single (nonlocal) energetic interaction, a "sticking energy" between hydrophobic (H) residues. Short chains are configured on a 2D square lattice. The "Helical-HP model" adds a local interaction modeling the helical hydrogen bond. The relative strengths of the two interactions can be varied, and ensemble statistical mechanical properties of different sequences can be enumerated exactly.

Ken's goal was to show that a relatively large local helical interaction would result in a secondary structure distribution that favored longer helices over shorter ones, while in proteins the opposite was true. This would argue for the primacy of nonlocal interactions in proteins. He had thought that adding only a *small* helical energy to the HP model might bring the model secondary structure distributions closer to those observed in real proteins. He was right in the first case, i.e. that when the helical interactions are stronger than the HH contact interactions, the helix length distributions are not protein-like. But he was wrong in the second case. For these short lattice-bound polymers, even a small helical energy favors helical conformations too much. What about tertiary structures? Here too adding a helical energy just makes the distribution of different structures look less like real proteins. The straightforward message I took from all this was simple— that the Helical-HP model wasn't going to be useful for modeling native proteins. The HP model alone was both simpler and more protein-like.

It's almost embarrassing to admit how much time I spent on the early slope of the learning curve, coming up (worse yet) with what were mostly negative results. It's especially embarrassing given all the help I received from Hue Sun Chan, from computer programs to invaluable discussions. I started working with the Helical-HP model in March, 1991. I spent June through September trekking through East Africa, and came back to the model in October. Over the next full year I gradually managed to put together a

somewhat respectable draft of a scientific paper outlining our conclusions from this model. I don't want to minimize the research I did during this period, but frankly my heart wasn't really in it. And I believe that draft paper (not reproduced here, but it eventually became what is now the first part of the *Protein Science* paper here) shows it.

To some extent I can get away with blaming my personal life for my unprolific early stint in graduate school. I came back from Africa to find my parents divorced and my father in a depression that at times bordered on psychosis. Shortly after my return, I suffered a debilitating back injury (the pain persisted for no less than six months). I was in a comfortable but ambivalent relationship with a woman I could neither leave nor fully commit to. In the spring of 1992, just as I was recovering from the back injury, I suffered from serious heart arrhythmias and went through a battery of tests that finally culminated in a diagnosis of Mitral Valve Prolapse. One thing I feel I have to interject here is how supportive Ken was of me during this time. I was feeling like I was not exactly earning a good reputation in Ken's eyes— how could he possibly believe in me when I couldn't seem to get anywhere with a really straightforward project that he all but laid out for me? But he clearly understood better than I the way our personal lives can impinge on our "scientific" lives. I had subconsciously bought into the folklore that portrays scientific research as divorced somehow from the part of human nature that rolls around in the mud with the other animals.

It wasn't until August of 1992 that I had battled my way through it all. I had reforged a (still rather weak) relationship with my father. My back was healed and I was training and playing soccer regularly again. A second opinion from a cardiac specialist had assured me that my arrhythmias were quite benign. And probably most apropos to this Preface, I had finally completed a decent draft of my first research paper with Ken. Just in time— that same month I was scheduled to present my results at the "Understanding Protein Folding" mini-course in Stockholm.

The trip to Stockholm turned out to be the watershed in both my graduate career and my personal life. It was a classic case of good timing that came dangerously close to looking like fate.

I had planned to leave for Stockholm a couple of days early so I would get over the jet-lag before the conference began. But for some reason, I couldn't seem to get my poster done. I still don't know what the problem was— I thought I had prepared everything. But it took me not one but *two* extra days to finish my poster and my packing. So on the last possible day I got the last open seat on the last flight to Chicago (my mother works for American Airlines so I was flying space-available). They even had to bump me to first class (poor me!) just to fit me on the plane.

When I arrived in Chicago I checked the gate where the Stockholm leg of the flight was boarding. There was only one person waiting there, two hours early. An incredibly beautiful woman. I remember thinking to myself that I hoped she would be sitting next to me during the flight. I got that wish and then some. Just over one year later she would become my wife.

It was this woman, and not the conference itself, that sparked my research. The excitement she stirred in my heart pervaded every aspect of my life. I was filled with more energy than even *I* was used to. She was finishing up her degree in Seattle, and I would shuttle up there every weekend (usually an "extended" weekend) courtesy of American Airlines. Much of the time I was in San Francisco was devoted to my teaching assistantship— a class in cellular biophysics that required a great deal of reading to keep ahead of those sharp students. It would seem natural to infer that in this milieu my research project would suffer. But the opposite was true: it could be argued that Fall Quarter of 1992 was when I started to become a scientist.

I dove into the scientific literature with a vigor I had never known before, even recruiting my fellow lab members by starting a "journal club." It was during this time that I stumbled across an issue of *Biochemistry* that revived my flagging interest in the Helical-

HP model. It was a paper showing the conformational changes induced in lysozyme by adding trifluoroethanol to an aqueous solution. The protein lost tertiary structure while gaining helical secondary structure. I had already shown that similar transitions could occur in the Helical-HP model. It struck me that we might be able to use the model to tell us something about the physical mechanism by which this transition occurs in proteins. This was the beginning of the work that, in my opinion, redeemed the *Protein Science* paper reproduced below. Undoubtedly the reviewers of that paper would agree with me. They felt that the section on alcohol denaturation of proteins displays a clarity that can't be found anywhere else in the paper. That clarity is a direct result of my own excitement about the work it describes.

II. Statistical potentials: How accurate are they?

This is where my research really started to follow the direction I had intended three years before. It was the beginning of what I consider to be my best work. Ken approached me with an idea he had been working on with Kai Yue. It was so simple I couldn't help but become totally hooked.

For the past 17 years, several groups of researchers had been working on the problem of converting the amino acid pairings observed in known proteins into some sort of "pseudo-energetic" potentials. These potentials were at a sufficiently low level of resolution to be useful in protein folding, and had been used in several folding algorithms. But the methods used to derive these potentials were based on spurious use of statistical mechanical equations. So the question was how accurately these pseudo-energies reflect the true underlying energies. Kai's idea was to use the 2D lattice model pioneered in our lab to test these methods. It was up to me to work out how to implement such a test, and to draw some physical insights from the results.

The implementation probably should have been the more time-consuming and less interesting part of the project, but I found that really wasn't the case. The implementation of this lattice test was exceptionally interesting because it gave me the excuse to devour over one hundred papers on statistical potentials and protein folding. The other half of the implementation involved programming the computer to construct "databases" of 2D lattice model structures, and extracting "statistical energies" from these databases. I began the implementation in June of 1993, and by the time I left the country again, I had about half of the framework for applying the statistical method of Miyazawa and Jernigan (1985).

I took what was becoming a yearly trip to Europe— headed to Sweden for "several" weeks to really meet Anette's family. I intentionally use the word "several" to maintain some ambiguity— let's just say that I stayed away from the lab for longer than some might feel was "appropriate." When I returned, Ken, who has a very "hands-off" managerial style (and who had let me slip away silently to Europe the previous summer and to Seattle on all those "extended weekends" the previous fall and winter), actually had a talk with me. As always, he was very tactful and friendly, but it didn't take a genius to read between the lines. He was concerned that my absence might be a symptom of a problem that was deeper than just a desire to immerse myself in Swedish culture for a significant period of time. He told me about a former student of his who had turned out to have a drug problem. Fortunately, I was having few problems beyond working out how to illegally import Anette so I could marry her. She came in on a tourist visa on September 7. We promptly ran off to Reno and were married in the "elegant" (this is a direct quotation from the yellow pages we used to make our choice) Chapel of the Bells at 12:15 a.m. (that's right— in the middle of the night) on September 12. For this act, we would be penalized $90 by the Department of Immigration (our other options were (1) to petition to have her enter the country with the "intent to marry" or (2) to marry in Sweden and petition to have her enter the country as my spouse, either of which would take six months... the $90 was a good deal).

After the repercussions of the wedding had blown over (my mother hasn't forgiven me to this day for not having invited her), I continued my implementation of the lattice model test. By the end of the year, I had calculated Miyazawa-Jernigan potentials for databases constructed from different true potentials. Further, I had spent an inordinate amount of time just trying to understand exactly what it was that Miyazawa and Jernigan were calculating.

By spring, I had finished a draft of a paper. But Ken returned it to me promptly. In the margin were the words: "Very good analysis, but it still lacks a clear physical picture of what is going wrong with statistical potentials." He was right, of course. I was too fixated on the details of the Miyazawa-Jernigan method which I had so painstakingly gleaned, and couldn't quite catch the big picture. From the path my work took from here, it is clear that at this point I had at least the essentials of an intuitive understanding of what was going wrong with statistical potentials and how to correct them. It's just that it would take walking away from this project for about a year and diving into it with fresh eyes, before I would reach the point where I could express this understanding concisely and simply. See section V. below, *Mopping up and moving on*. In the meantime I would work on two different projects and take my oral exams.

III. An iterative method for extracting potentials from the known protein structures.

This was probably the most intellectually frustrating time in graduate school. But I always had a soft spot in my heart for come-from-behind victories. The taste of blood is sweeter when alot of it is your own. At the end of this time I got my first taste— I actually came up with something original and useful. When I was a child I had a recurring dream in which I was standing atop a huge pile of silver-dollar-sized stones (I know it sounds corny, but I'm afraid it's actually true...). The triumphant moment was when I reached

into my pocket, pulled out a stone of my own, and made the pile stand just a little bit higher. Now I've done just that, for real this time. This was the creative process.

The pressure was on. I was expected to take my oral exams sometime this academic year. I knew exactly what I *wanted* to do. I wanted to use the insights I'd gained over the past year to develop a better way of inferring protein energetics from a database of known structures. Easier said than done. It was spring of 1994 already, and I wanted to schedule my orals before fall quarter opened in September. During the next few months I would work harder than at any time in my graduate career. Sixteen hour days were the norm (no kidding).

I had several ideas. The obvious problem with the work of Miyazawa and Jernigan was the fact that they had a constant term that was added to all of their contact energies which depended on the average length of the proteins in the database. I came up with a way to normalize this contribution, to make it chain-length independent. But really that just meant I could choose any arbitrary constant rather than relying on an artificial database artifact. So that road turned out to lead nowhere.

My next idea was to calculate statistical energies that depended in some systematic way on protein sequence information. In Chapter 2 below I define a property called the "partition propensity" that can be calculated from the protein size and amino acid composition. Since the statistical energy depends systematically on this property, perhaps it would be appropriate to use different energies for proteins with different partition propensities. Unfortunately, this turned out to be of marginal help, and I hit another wall.

My next idea was that perhaps part of the problem was in assuming that interactions were pairwise independent. Since hydrophobic residues in proteins tend to cluster, while charged residues will typically only have a single oppositely charged residue in their "contact" shells, one of the obvious things to exploit is the statistical tendency of different residue types to have different numbers of certain types of neighbors. I spent quite some time figuring out how to define proper "random mixture" reference states for each observed

state. In the end, the approach was even less successful than assuming pairwise independence. Another "failure," but I gained a valuable insight. I concluded that the "random mixture" reference state, which assumes that each observed state is independent of all the others, becomes worse when the states are more closely coupled. And the number-of-neighbors states were more closely coupled than the pairing states.

Time was running out. I had to schedule my oral exam soon, if I wanted to take it before the summer quarter ended. I was pretty dejected after the failure of my latest, cleverest (I thought) idea. But it stuck with me that the real problem with statistical potentials was that pairing states were not independent of each other. How could one account for the interdependence? This was a real problem. According to the lattice model, the main culprit was excluded volume— not all residue pairs could be at their optimal interaction distances simultaneously, simply because all these residues were packed closely together and "frozen" in a single state (the native conformation). Compromise was the rule rather than the exception. But how could one know how one given interaction would influence the observed distributions of other pairs? It occurred to me that the only way to do that for a chain polymer is to actually try to make some conformations using computer conformational search algorithms. That way both excluded volume and chain connectivity could be accounted for. It was my vague notion that it might be possible to iteratively refine a potential if one could make some accounting of interdependence.

It was natural to use Klaus Fiebig's "hydrophobic zippers" (HZ) algorithm for generating conformations (Dill *et al.*, 1993; Fiebig & Dill, 1993). I had modified it in June 1993 to find conformations that were low-energy according to any arbitrary monomer contact potential. The only problem was that it could only construct conformations on a 3D cubic lattice. But Klaus was already working on an off-lattice version, so I would be able to try that at a later date. On the lattice there would still be some accounting of excluded volume.

Now what I needed was a way to change the calculated contact energies with each iteration. I had come to think of the statistical energies as a "first approximation", so it was natural to try using a "series approximation" where each iteration adds a term. Each term replaces the denominator of the standard statistical potential equation with a Boltzmann-weighted ensemble of conformations. Obviously this has the right limiting value, i.e. it approaches zero as the Boltzmann-weighted ensemble is dominated by the native conformation. But there was no way to know whether it would be a convergent series. I tested the equation with the 2D lattice model, generating non-native conformations with hydrophobic zippers. There were two remarkable results: (1) after the first iteration the energies were almost exactly equal to the statistical energies, and (2) after several iterations, the potential converged to the correct (scaled) values! I still remember my surprise— after all of the failures I had experienced over the past couple of months, I had really expected this idea to fall short.

I still don't know how I came up with the crucial insight, namely to use a Boltzmann-weighted ensemble. Obviously having worked with a statistical mechanical model set the stage, but in the end I really don't know where it came from. Evidently my intuition was much deeper than my conscious understanding of why statistical potentials did not work.

It seemed to me that the key to understanding why my method worked might be uncovered if I could explain why the first iteration energies were equal to the statistical potential. Then I would have a physical interpretation of the "random mixing" reference state. It is at times like this that I always seek out Hue Sun Chan. His extensive and thoughtful work with lattice models, as well as willingness to entertain discussion, have been indispensable to me. He really hit the mark this time. I told him that I had empirically observed that averaging HZ endstates for my model sequences yielded a "random mixing" distribution of my two monomer types. He pointed me to a paper he had written a few years before (Chan & Dill, 1990a) that had a very interesting result: an unweighted average

of compact conformations on a 3D lattice gives a random mixing distribution with respect to inter-monomer contacts. This was not intuitive, at least not to me. In a random-flight polymer, the probability of observing a given monomer-monomer contact depends on the sequence separation. But constraining the polymer to be compact changes the rules completely.

Now I started to consciously understand exactly what the iterative method was doing, and I developed what can only be described as "faith" in its ultimate success. I promptly scheduled my oral exam. It was during the ordeal of preparing for this exam that the iterative method gelled into basically the same form it has in this dissertation. The help of my orals committee, Peter Kollman (the chair), Dave Agard, Fred Cohen and Tack Kuntz, was invaluable. I also want to thank the members of the Dill group, Klaus Fiebig, Hue Sun Chan and Karen Tang in particular, for helpful suggestions and discussions. And, of course, Ken Dill himself.

Peter Kollman gave me alot of guidance in thinking about potential functions, especially about potentials of mean force in solvent systems. Dave Agard helped me think about protein folding experiments, and what kinds of potentials can be extracted from a database of protein structures. Fred Cohen helped me to put my work in the context of the current literature, and how my lattice model results might or might not apply to the problem at hand. And Tack Kuntz discussed with me the physical interpretation of a potential inferred by my method, about self-consistency and multiple solutions to an underdetermined problem.

My ideas evolved quickly during my preparation for orals. I abandoned HZ for the time being since it was clear that the off-lattice version would be slow in coming, and that configurational constraints depend crucially on what dihedral angles are allowed (the 3D lattice conformations just can't look like real proteins). I opted instead to generate conformations using the "threading" method that had been used by others (Goldstein *et al.*, 1992; Maiorov & Crippen, 1992). I realized that it was critical to include the native

conformation in the ensemble— the reason HZ had succeeded for the 2D lattice model was that the native conformation was nearly always found for short sequences. This gave the iterative process the proper form for converging to a solution that favors the native conformation over the alternatives.

Studying for the oral exam was extremely stressful but fascinating. Two days before the exam it took a Shiatsu massage to stop the spasms in my right shoulder. The exam itself, July 18 at 10 a.m., almost felt like a privilege— I had all four committee members to myself for over an hour. It would probably be best described as an extremely intense discussion in which my opinion was asked on every topic. I don't want to say it was easy, because afterwards I felt like I'd just run a marathon. It's just that it wasn't the torture session you hear about from other graduate students. Despite the difficulty of the questions, I really felt like the committee members were all on my side. Each one actually bailed me out of a hard question at some point during the exam.

But what probably helped me the most during my orals was the fact that, barely two days before the exam, I had shown that my iterative method could solve the "protein recognition problem" (a.k.a. the threading problem). Up until that point I was operating on faith alone— a faith that my committee members did not all share, despite what I thought was a pretty convincing orals proposal. What would have happened if it had not worked, I can't say. But hindsight justifies my faith even if the facts at the time did not.

The day after my orals, I finished writing my first peer review of a submitted journal article. The day after that I was on a plane bound for Sweden. And one day after that, I was waterskiing in broad daylight at 11 p.m. It was a blissful recovery period. During this time, I came up with the idea of using the Z-score as an "energy" in the Boltzmann weighted ensemble. When I returned to San Francisco I ran all sorts of threading tests of the iterative method— changing training sets and test sets, contact potentials, distance-dependent potentials. I was still living off of the original thrill of success, and I wanted to see just how far the success would go— at one point I even used

the entire database of nonhomologous structures as a training set, just to find a limit to the learning ability of the method. It didn't take long before it started getting boring. Just in time, I got just the respite I needed.

IV. A simple protein folding algorithm using a binary code and secondary structure constraints.

This was work that Shaojian Sun initiated. He and I had a discussion in early 1994, when I was first starting to use the "threading test" to evaluate some of my ideas about improving potential functions. I wanted some sort of benchmark for success levels in the threading test. With my Dill lab experience, it was almost second nature to try simply counting contacts between hydrophobic residues. This turned out to be a surprisingly effective threading potential— as effective as the statistical potentials Shaojian had used in folding algorithms. Shaojian believed that a similarly simple potential might be effective for folding proteins as well.

Shaojian wanted to use the same search algorithms (a genetic algorithm, GA, and simulated annealing, SA) and reduced representation that he had used in previous work with statistical potentials (Sun, 1993, 1995). He chose 9 small proteins of known structure and one of unknown structure (but designed to form a helical bundle). Instead of a statistical potential governing interresidue interactions, he used a single type of attraction between hydrophobic residues. He found, *for all the proteins he tested*, that if the known (or predicted) secondary structures were fixed, the lowest energy folds searched by either GA or SA were very similar to the known (or predicted) tertiary structures.

I first read a draft paper of his in early November 1994. It had been submitted and accepted with revision to *Protein Engineering*, and Shaojian had just moved on to the National Cancer Institute in Bethesda. Since I had some experience with folding

algorithms and potential functions, Ken wanted me to help address some of the reviewers' concerns. This was how I got directly involved in the project.

It didn't take me long to recognize that there might be a problem with Shaojian's results. The lowest-energy structures found by both search algorithms were actually higher in energy than a fully extended conformation (I made, checked and rechecked the calculations just to be sure). This was mostly due to some steric overlap between segments. Furthermore, for one of the proteins, there were no hydrophobic residues in one of the segments, yet in the reported structure that segment was improbably packed in roughly the same way as the known crystal structure.

I approached Ken with my findings and asked if I could redo Shaojian's study. Certainly the paper did not report the lowest-energy configurations, since extended conformations were lower. Did this mean that the results and conclusions of the paper were not valid? Or would they remain intact when the search algorithms could find truly low energy conformations?

Since the problem was likely to be with the search algorithms, I did not want to use Shaojian's FORTRAN code directly. Several months earlier, Klaus Fiebig had nearly completed a version, in the C programming language, of Shaojian's original statistical potential GA. I want to thank Klaus yet again, this time for providing me with this code—it was exceptionally clearly written, and easy to debug and modify for my purposes. Within a few weeks I had fully implemented the algorithm described in Shaojian's paper. After a few more weeks of calculations I found that the search was not very effective; it rarely converged to structures in which the secondary structure elements were closely packed. Nevertheless, these structures were significantly lower in energy than those found by Shaojian's searches.

Although the lowest-energy structures I found were not generally similar to the known structure, I did not feel it was fair to condemn the work. Instead I modified the GA search to include a second phase, in which dihedral angles were slightly perturbed. This

turned out to be crucial. The searched structures were much more compact. What was really exciting, however, was that for almost all of the proteins, the lowest-energy structure was very similar to the known structure. So it was clear that the primary conclusion of the original paper could stand. However, unlike in the original paper, there were some failures of the algorithm. I found, however, that these failures were explicable, and pointed the way to improvements. I tried several ways of improving the GA search itself, with little success. I then moved on to using different search strategies, and to improving the treatment of beta-strands (allowing some conformational freedom and better hydrogen-bonding potentials).

My "repetition" of Shaojian's work, calculations on several proteins that he had not considered, and full analysis of all the work, plus modification of the search strategy, required me to rewrite the entire body and conclusions of the paper. We added me as co-author. Nevertheless, since the primary conclusions remained, *Protein Engineering* did not send the paper back to the reviewers, despite the fact that we didn't return a final manuscript until June 1995.

V. Mopping up and moving on

My work on the iterative method had solidified my physical understanding of the problems with statistical potentials. It was time to return to this work and forge it into a simple, clear and insightful paper. I had originally decided to concentrate on a single method of calculating statistical potentials (Miyazawa and Jernigan, 1985) in order to avoid making the paper too complex. But I reconsidered, and I believe that including distance-dependent effects makes the physical picture a great deal clearer. I reconstructed the paper, emphasizing related research and trying to tie together disparate papers in the literature. I want to thank Ken for all of his help in writing this paper— if there's any clear physical

understanding that comes through, it's due to him. I submitted the paper to *Journal of Molecular Biology* in September, and it was accepted in late November.

It was into this context that on September 15, 1995, Paul Joachim Sebastian Thomas was born. We call him Sebastian. I won't run through the droning series of clichés that every parent launches into when asked about their first child. I'll just say that my way of seeing the world, and my place in it, has opened wider than I could ever have conceived, literally. I'll wear the title of Father infinitely more proudly than that of Doctor of Philosophy.

During the course of about a year, I had put the iterative potential work on the back burner. But that didn't mean it wasn't cooking at all. What I needed to decide on was a good presentation format. Originally I wanted to use a set of proteins that encompassed all the known, nonhomologous protein structures. But one of my primary objections to the field of "protein threading" is that there is no standard set of test proteins, so there is no way to compare different methods. I felt it would be most useful if my method were directly compared to that of Maiorov and Crippen (1992). These methods are similar in spirit— they both seek to calculate a potential by using information from both native and nonnative folds of proteins. I also felt it was important to address some of the arbitrary decisions made in any database-derived potential, such as (1) which proteins to use in a training set, (2) how to classify amino acids and (3) whether to use a contact- or distance-based potential. These studies involved a great deal of computer time, and were mostly done on a Silicon Graphics Indigo.

I had originally wanted to include some work on the "threading-through-motif problem" (Bryant & Lawrence, 1993), but a collaboration with Steve Bryant evolved too slowly for my dissertation schedule. This would have been more "cutting edge" in terms of the question it addresses. But I believe it is a step best taken after the method has been properly introduced. So, in the end the paper took on a more "exploratory" tone, which I

believe is particularly appropriate given the iterative method's novel advantages of accuracy, speed and flexibility.

The last six months of graduate school wasn't just a time to finish mopping up, it was a time to start moving on. In alot of ways, this was the "reward" part of graduate school. I gave talks, and was wined and dined and generally flattered in Philadelphia by Scott Dixon at SmithKline Beecham Pharmaceuticals, in Baltimore by George Rose at Johns Hopkins Medical School, and right here in Concord by John Hearst, a U.C. Berkeley chemistry professor whose startup company Steritech offered an intriguing alternative. By electronic mail I had corresponded with Michael Nilges at the European Molecular Biology Laboratory in Heidelberg, and received an application for a postdoctoral fellowship in Germany. I want to thank all these people for their interest in my work, and for giving me an extremely difficult choice as to where to begin the next phase of my career, and of my life. My life at this point was so much broader than it had ever been— it was harrowing to make a decision based not only on what I wanted for myself, but on what both Anette and I wanted for us, and for Sebastian. And that's the only way I would have it. If there's a point hidden somewhere in this rambling Preface, it's that the artifice of scientific endeavor is inseparable from the commonly uncommon life that created it. The science I learned was the art of making science clear.

San Francisco, November 1995

References

Austen, Jane. (1833). *Northanger abbey*. By Miss Austen, ... with a biographical notice of the author. In two volumes. Philadelphia: Carey & Lea [Griggs & Dickinson, printers].

Austen, Jane. (1894). *Pride and prejudice* by Jane Austen ; with a preface by George Saintsbury and illustrations by Hugh Thomson. London: G. Allen.

Bond, Edward. (1972). *Lear.* London: Eyre Methuen.

Bryant, S.H. & Lawrence, C.E. (1993). An empirical energy function for threading protein sequence through the folding motif. *Proteins* **16**, 92-112.

Chan, H.S. & Dill, K.A. (1990a). The effects of internal constraints on the configurations of chain molecules. *J. Chem. Phys.* **92**, 3118-3135.

Chan, H.S. & Dill K.A. (1990b). Origins of Structure in Globular Proteins. *Proc. Natl. Acad. Sci. USA* **87**, 6388-6392.

Defoe, Daniel. (1754). *The history of the great plague in London, in the year 1665. Containing, observations and memorials of the most remarkable occurrences, both public and private, that happened during that dreadful period. By a citizen,...* London: F. and J. Noble.

Dill, K.A. (1990). Dominant Forces in Protein Folding. *Biochemistry* **29**, 7133-7155.

Dill, K.A., Fiebig, K.M. & Chan, H.S. (1993). Cooperativity in protein-folding kinetics. *Proc. Natl. Acad. Sci. USA* **90**, 1942-1946.

Driesch, Hans. (1914). *The history and theory of vitalism.* Transl. by C. K. Ogden. Rev. and in part rewritten ... by the author. London: Macmillan.

Fiebig, K.M. & Dill, K.A. (1993). Protein core assembly processes. *J. Chem. Phys* **98**, 3475-3487.

Fielding, Henry. (1749). *The history of Tom Jones, a foundling*. London: Printed for A. Millar.

Goldstein, R.A., Luthey-Schulten, Z.A. & Wolynes, P.G. (1992). Protein tertiary structure recognition using optimized Hamiltonians with local interactions. *Proc. Natl. Acad. Sci. USA* **89**, 9029-9033.

Lau, K.F. & Dill, K.A. (1989). A Lattice Statistical Mechanics Model of the Conformational and Sequence Spaces of Proteins. *Macromolecules* **22**, 3986-3997.

Maiorov, V.N. & Crippen, G.M. (1992). Contact potential that recognizes the correct folding of globular proteins. *J. Mol. Biol.* **227**, 876-888.

Miyazawa, S. & Jernigan, R.L. (1985). Estimation of effective interresidue contact energies from protein crystal structures: quasi-chemical approximation. *Macromolecules* **18**, 534-552.

Richardson, Samuel. (1868). *Clarissa Harlowe* by Samuel Richardson. A new and abridged ed. by Mrs. Ward. London, New York: G. Routledge.

Sun, S. (1993). Reduced representation model of protein structure prediction: statistical potential and genetic algorithms. *Protein Science* **2**, 762-785.

Sun, S. (1995). Reduced representation model of protein structure prediction: statistical potential and simulated annealing. *J. Theoretical Biol.* **172**, 13-32.

Wilson, C. & Doniach, S. (1989). A computer model to dynamically simulate protein folding: studies with crambin. *Proteins* **6**, 193-209.

Authorship

Paul Thomas is listed as a co-author on the paper entitled "A simple protein folding algorithm using a binary code and secondary structure constraints," by S. Sun, P.D. Thomas, and K.A. Dill. So as required by University guidelines, this page is just to affirm that Paul played a major role in the calculations and the writing of that paper. He contributed original research to that paper fully at the level expected of Ph.D. thesis work.

Ken Dill

Modeling the Energetics of Protein Folding

Paul Denis Thomas, Jr.

Abstract

Reduced-representation models of proteins have been developed in an effort to simulate protein folding on a computer. The energy functions required to drive these simulations are complicated averages over atomic interactions, including interactions between elements of the protein chain as well as interactions with water. How can we obtain such energy functions, and how can we know whether they accurately reflect nature's underlying energies? To address these questions, I use simple 2-dimensional lattice models to study the physical principles in question, and then apply these principles to more realistic models of proteins. The first lattice model, the "Helical HP" model, is used to assess the relative importance of local and nonlocal interactions in proteins; nonlocal interactions are found to be dominant. The second lattice model, the "AB" model, is used to test the premises by which nonlocal interresidue interactions in proteins are commonly estimated. The simplicity of the AB model allows a physical understanding of why the current methods are not accurate. Based on this understanding, I propose an improved method of estimating protein energetics from the known protein structures. This method succeeds in finding rigorously accurate energies in the lattice model, and finds a relatively simple solution to the more realistic "protein recognition problem." Finally, I present a model for protein folding that uses a simple potential function based on rudimentary physics rather than complicated database-derived energy parameters. The simple model is surprisingly successful in predicting the structures of several small proteins, and points the way toward systematic improvements.

Table of Contents

List of Tables

List of Figures

Chapter 1

Chapter 2

Chapter 3

Chapter 4

Introduction

Background

Despite attempts by numerous researchers for over three decades, the protein folding problem is still unsolved. It has been described as one of the most important problems remaining in modern theoretical biology. The experiments of Christian Anfinsen in the 1960s on the protein Ribonuclease A showed that all of the information necessary to produce the complicated, specific three-dimensional structure of a protein is contained in just the amino acid sequence alone (Anfinsen, 1973). It should be possible, then, to predict the structure of a protein given only its amino acid sequence. This is the protein folding problem. Recently, several proteins have been found to require other proteins to assist in folding, but most single-domain, globular proteins still appear to obey Anfinsen's "thermodynamic hypothesis." In this dissertation, the word "protein" refers to such single-domain, globular proteins.

Predicting the folded structure of a protein using a computer requires three elements: a model of the protein, a method for changing the protein conformation, and a model of the energetics that drive protein folding. According to the thermodynamic hypothesis, the conformation that is lowest in free energy is the conformation that the real protein will be predicted to assume in nature. However, computer simulations of protein dynamics using atomic-resolution force-fields are computationally intensive and yield information about processes which occur in picoseconds to nanoseconds, while proteins fold in milliseconds to hours (Kim & Baldwin, 1982). Methods capable of simulating

processes occurring on the time scale of protein folding are necessarily much lower in resolution. There is therefore need of lower-resolution (e.g., on the level of amino-acids) energy functions to drive these algorithms (Wilson & Doniach, 1989; Skolnick & Kolinski, 1990; Sun, 1993; Kolinski & Skolnick, 1994; Monge *et al.*, 1995).

The work in this dissertation represents an attempt develop accurate, low resolution energy functions for proteins. Below the lofty goal of *de novo* protein structure prediction, the motivations behind attempting to develop low resolution, additive energy functions for proteins are numerous. Interresidue energy functions have been developed to help aid in the evaluation of model protein structures (Chiche *et al.*, 1990; Bowie *et al.*, 1991; Lüthy *et al.*, 1992; Wilmanns & Eisenberg, 1993), to identify the native fold or putative folding motif of an amino acid sequence among many alternatives (Hendlich *et al.*, 1990; Jones *et al.*, 1992, Bryant & Lawrence, 1993), to identify possible folds for a sequence of unknown structure (Bowie *et al.*, 1991; Sippl & Weitckus, 1992), to predict docking of protein structures (Pellegrini & Doniach, 1993), and to find amino acid sequences compatible with a desired structure (Godzik *et al.*, 1992).

Throughout this dissertation, the theme is to make protein energetics as simple as possible, and add more terms and further details only when the simpler models fail. There are two advantages to this approach over most current approaches, which have hundreds to thousands of energy parameters: (1) improvements to an energy function can be made systematically, and (2) simpler models allow a clearer physical understanding.

Overview

This dissertation has four main parts. Chapter 1 describes a simple lattice model of proteins, which is used to study the balance of local and nonlocal interactions in proteins. The "Helical HP" model was found to be particularly useful in modeling the effects of non-

aqueous solvents such as alcohols, urea and guanidine, on the folds adopted by proteins. We conclude from this study that nonlocal interactions in proteins are dominant. This conclusion forms the basis for the assumption, in subsequent chapters, of the primacy of nonlocal interactions in protein folding.

Chapter 2 describes a study of the most common current approach for obtaining low-resolution protein folding potentials. The approach, called the "statistical potential," is based on observing pairing frequencies of amino acids in the known protein structures (Bernstein *et al.*, 1977), and applying the Boltzmann relation to obtain energy-like quantities (Tanaka & Scheraga, 1976; Miyazawa & Jernigan, 1985; Sippl, 1990). I use an exact, 2-dimensional lattice model to perform a consistency check of the method. This simple model allows me to give clear physical descriptions of why the method fails to extract quantitatively correct interresidue interaction energies from a database of native structures.

Armed with the knowledge of why statistical potentials fail, I developed a new method for extracting more accurate pairwise amino acid interaction energies from the known protein structures. This work is described in Chapter 3. An exact lattice model test shows that the method works, in principle. The method is then applied to finding a solution to the protein recognition problem. The method finds an energy function that is at least as good at solving this problem as any previously published potential, but which has significantly fewer parameters. I then use the method to systematically study the assumptions commonly made when extracting potentials from protein structures.

Chapter 4 finally applies a lesson learned in the previous chapters to the problem at hand: protein folding. The lesson is clear though certainly not original: the hydrophobic effect is the dominant driving force (Kauzmann, 1959; Dill, 1990). In Chapter 4 we propose an exceedingly simple potential function for protein folding based on (1) a single type of attraction between hydrophobic amino acids, (2) hydrogen bonding between β-strands, and (3) the correct disulfide bonds. We fix the secondary structure to make the

conformational search tractable. We find that this energy function predicts structures of 10 small proteins that are generally very similar to the known structures. The results are comparable to folding algorithms having many orders of magnitude more energy parameters. Because of the simplicity of the model, we learn as much from its failures as its successes. We discuss possible systematic improvements.

References

Anfinsen, C.B. (1973). Principles that govern the folding of protein chains. *Science* **181**, 4096.

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). The protein data bank: a computer-based archival file for macromolecular structures. *J. Mol. Biol.* **112**, 535-542.

Bowie, J.U., Lüthy, R. & Eisenberg, D. (1991). A method to identify protein sequences that fold into a known three-dimensional structure. *Science* **253**, 164-170.

Bryant, S.H. & Lawrence, C.E. (1993). An empirical energy function for threading protein sequence through the folding motif. *Proteins* **16**, 92-112.

Chiche, L., Gregoret, L.M., Cohen, F.E. & Kollman, P.A. (1990). Protein model structure evaluation using the solvation free energy of folding. *Proc. Natl. Acad. Sci. USA* **87**, 3240-3243.

Dill, K.A. (1990). Dominant forces in protein folding. *Biochemistry* **29**, 7133-7155.

Godzik, A. & Skolnick, J. (1992). Sequence-structure matching in globular proteins: application to supersecondary and tertiary structure determination. *Proc. Natl. Acad. Sci. USA* **89**, 12098-12102.

Hendlich, M., Lackner, P., Weitckus, S., Floeckner, H., Froschauer, R., Gottsbacher, K., Casari, G. & Sippl, M.J. (1990). Identification of native protein folds amongst a large number of incorrect models. The calculation of low energy conformations from potentials of mean force. *J. Mol. Biol.* **216**, 167-180.

Jones, D.T., Taylor, W.R. & Thornton, J.M. (1992) A new approach to protein fold recognition. *Nature* **358**, 86-89.

Kauzmann, W. (1959). Some factors in the interpretation of protein denaturation. *Adv. Protein Chem.* **14**, 1-63.

Kim, P.S. & Baldwin, R.L. (1982). Specific intermediates in the folding reactions of small proteins and the mechanism of protein folding. *Ann. Rev. Biochem.* **51**, 459-89.

Kolinski, A. & Skolnick, J. (1994). Monte Carlo simulations of protein folding. I. Lattice model and interaction scheme. *Proteins* **18**, 338-352.

Lüthy, R., Bowie, J.U. & Eisenberg, D. (1992). Assessment of protein models with three-dimensional profiles. *Nature* **356**, 83-85.

Miyazawa, S. & Jernigan, R.L. (1985). Estimation of effective interresidue contact energies from protein crystal structures: quasi-chemical approximation. *Macromolecules* **18**, 534-552.

Monge, A, Lathrop, E.J.P., Gunn, J.R., Shenkin, P.S. & Friesner, R.A. (1995). Computer modeling of protein folding: Conformational and energetic analysis of reduced and detailed protein models. *J. Mol. Biol.* **247**, 995-1012.

Pellegrini, M., & Doniach, S. (1993). Computer simulation of antibody binding specificity. *Proteins* **15**, 436-444.

Sippl M.J. & Weitckus, S. (1992). Detection of native-like models for amino acid sequences of unknown three-dimensional structure in a data base of known protein conformations. *Proteins* **13**, 258-271.

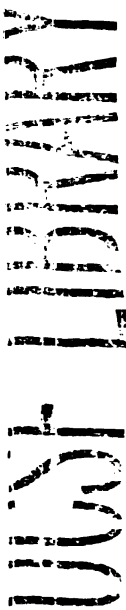
Sippl, M.J. (1990). Calculation of conformational ensembles from potentials of mean force. An approach to the knowledge-based prediction of local structures in globular proteins. *J. Mol. Biol.* **213**, 859-883.

Skolnick, J. & Kolinski, A. (1990). Simulations of the folding of a globular protein. *Science* **250**, 1121-1125.

Tanaka, S. & Scheraga, H.A. (1976). Medium- and long-range interaction parameters between amino acids for predicting three-dimensional structures of proteins. *Macromolecules* **9**, 945-950.

Wilmanns, M. & Eisenberg, D. (1993). Three-dimensional profiles from residue-pair preferences: identification of sequences with beta/alpha-barrel fold. *Proc. Natl. Acad. Sci. USA* **90**, 1379-1383.

Wilson, C. & Doniach, S. (1989). A computer model to dynamically simulate protein folding: studies with crambin. *Proteins* **6**, 193-209.



Chapter 1

Local and Nonlocal Interactions in Globular Proteins, and Mechanisms of Alcohol Denaturation

Paul D. Thomas and Ken A. Dill

This material originally appeared in *Protein Science*, vol. **2**, pp. 2050-2065 (1993).

The coauthor directed and supervised this work.

Abstract

How important are helical propensities in determining the conformations of globular proteins? Using the 2-dimensional lattice model and two monomer types, H (hydrophobic) and P (polar), we explore both nonlocal interactions, through an HH contact energy, ε, as developed in earlier work, and local interactions, through a helix energy, σ. By computer enumeration, the partition functions for short chains are obtained without approximation, for the full range of both types of energy. When nonlocal interactions dominate, some sequences undergo coil-globule collapse to a unique native structure. When local interactions dominate, all sequences undergo helix-coil transitions. For two different conformational properties, the closest correspondence between the lattice model and proteins in the Protein Data Bank is obtained if the model local interactions are made small compared to the nonlocal HH contact interaction, suggesting that helical propensities may be only weak determinants of globular protein structures in water. For some HP sequences, varying σ/ε leads to additional sharp transitions (sometimes several) and to "conformational switching" between unique conformations. This behavior resembles the transitions of globular proteins in water to helical states in alcohols. In particular, comparison with experiments shows that whereas urea as a denaturant is best modeled as weakening both local and nonlocal interactions, trifluoroethanol (TFE) is best modeled as mainly weakening nonlocal HH interactions and slightly enhancing local helical interactions.

Keywords

HP lattice model, compact conformations, hydrophobic interaction, helical propensities, alcohol denaturation

Introduction

What is the relative importance of local interactions (helical propensities) compared to nonlocal interactions (mainly solvent-mediated and hydrophobic interactions) in determining the native structures of globular proteins? One view has held that the local interactions are the main determinants of structure, while the nonlocal interactions just provide nonspecific stability to the compact state (Anfinsen & Scheraga, 1975). That is, interactions among adjacent and near-neighbor peptide units tend to drive proteins to configure into helices at certain points in the sequence, which ultimately end up as helices in the native structure. In this view, helices should be identifiable by factors that are local within the sequence, and this should strongly specify how they can pack to form the native structure. Recently an alternative view has developed that nonlocal interactions may be a major determinant not only of the stabilities, but also of the structures, of globular proteins (Dill, 1990; Chan & Dill, 1991). In this view, a major factor in determining where helices form in native structures are the sequence positions of hydrophobic monomers. That is, predicting helices (and also sheets) in native structures is more a matter of finding the correct global hydrophobicity patterns than of finding good local helical propensity patterns, even though the latter is known to be non-negligible. The present work aims to explore this question in more detail by using a simplified model that has an exact partition function, and to study the consequences of different balances between local and nonlocal interactions. We refer to this as the helical-HP model.

We then apply the model to the denaturation of proteins by alcohols and other agents. The ability of alcohols to induce α-helical conformations in proteins was first noted in optical rotatory dispersion (ORD) experiments by Tanford *et al.* (1960) on β-lactoglobulin. Tamburro *et al.* (1968) studied the effects of trifluoroethanol (TFE) on the conformations of the ribonuclease S-peptide; TFE was found to stabilize the small peptide

in the same (α-helical) conformation that it adopts in the native protein. Since then, alcohols have been used widely to examine the conformational (particularly helical) propensities of peptides (Nelson & Kallenbach, 1986; Lehrman *et al.*, 1990; Segawa *et al.*, 1991; Sönnichsen *et al.*, 1992), and to induce conformational changes in intact proteins (Stone *et al.*, 1985; Wilkinson & Mayer, 1986; Dufour & Haertlé, 1990; Jackson & Mantsch, 1992; Buck *et al.*, 1993; Fan *et al.*, 1993). The primary physical mechanisms by which alcohols effect these changes are still unresolved. The helical-HP model reproduces some basic characteristics of experimental denaturation of proteins by alcohols and other denaturants, and allows an interpretation of their mechanisms of action.

The Helical-HP Model

The helical-HP model is of short chains, which are specific sequences of H (hydrophobic) and P (polar) monomers, configured as self-avoiding walks on 2 dimensional lattices. A monomer represents a single amino acid residue. Figure 1 shows two possible conformations of a 16-monomer chain on the lattice. Each monomer occupies one site on the lattice, and monomers that are consecutive in the sequence must be next to each other on the lattice. No two monomers may occupy the same lattice site. We use a search-tree algorithm to enumerate all of the possible conformations of a chain of given length (Chan & Dill, 1989). For the 16-monomer chains modeled in this paper, there are 802,075 possible conformations on the 2D square lattice. The relative populations, $p(A)$ / $p(B)$, of any two conformations A and B are given by the Boltzmann distribution:

$$\frac{p(A)}{p(B)} = \exp\left[-\left(\Delta G_A - \Delta G_B\right)/kT\right] \qquad (1)$$

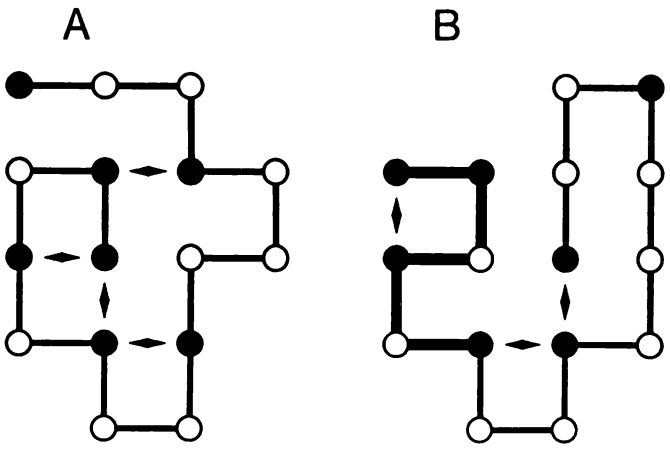


Figure 1. 2D Helical-HP model conformations, for a sample sequence (HHPHPHPPHPPPHPPH). H (hydrophobic) residues are gray, P (polar) residues are white; chain connectivity (covalent bond) is indicated by connecting lines. Conformation (A) contains 4 HH contacts (black diamonds) and no helical bonds, $\Delta G = 4\varepsilon$. Conformation (B) contains 3 HH contacts and 2 helical bonds (bold lines), $\Delta G = 3\varepsilon + 2\sigma$.

where ΔG_A and ΔG_B represent the Gibbs free energies of conformations A and B, respectively, relative to a common reference, k is Boltzmann's constant and T is the absolute temperature.

In a previous treatment, referred to as the HP model (Lau & Dill, 1989; Chan & Dill, 1991a; Shortle *et al.*, 1992), a single type of energy was considered, accounting for the favorability of each HH contact in a given chain configuration. This interaction is nonlocal in the sense that HH contacts involve two H monomers that need not be near neighbors in the sequence; they are mediated through space, rather than being mediated along neighboring covalent bonds. Conformational properties of the HP model have been studied in detail. The HP model shows certain features of globular proteins (Lau & Dill, 1989; Lau & Dill, 1990; Chan & Dill, 1991a; Lipman & Wilbur, 1991; Shortle *et al.*, 1992; Miller *et al.*, 1992; Unger & Moult, 1993; O'Toole & Panagiotopoulos, 1993; Camacho &

Thirumalai, 1993a; Camacho & Thirumalai, 1994). For example, increasing the strength of the HH interaction in the HP model leads to a relatively sharp transition from a denatured ensemble of open conformations to a small number of compact conformations (often only one or a few) that are composed of secondary structures and have cores mainly composed of H monomers. The conformations of lowest energy in the HP model are different for different sequences. Native states are relatively insensitive to mutations, particularly on the surface; there is high "convergence," i.e., many different sequences fold to a given native state; and there are favored folding pathways.

We now consider a more general model, the helical-HP model, in which the free energy is a sum of two types of energy. First, as before, we consider a *contact interaction*, $\varepsilon \leq 0$, for each HH contact. The net exposed surface of H monomers in the 2D model protein (i.e. contacting either P monomers or solvent) is decreased by two perimeter units for every contact made between two H monomers. Physically, the HH contact primarily represents the burial, or desolvation, of hydrophobic amino-acid side-chains in real proteins. For hydrophobic contacts in proteins, desolvation can also involve hydrogen bonding and steric effects, so these effects are also represented by the model HH contact interaction. If m is the number of HH contacts in a model conformation, then the total nonlocal contact energy of that conformation is $m\varepsilon$. (Because we are not concerned in this paper with the temperature dependence of the model, we interchangeably refer to contact energies as contact free energies.) It should also be noted that although the HH contacts need not be local in sequence, they may be. In this way, a 2D helix can be stabilized by both intra-helix (i, i+3) HH contacts and HH contacts between monomers in helices and monomers in other parts of the chain (e.g., in Figure 1B).

Second, we introduce an additional energy term to account for local interactions. Local interactions are those among near-neighboring residues in the sequence. Local interactions drive turns (Dyson & Wright, 1991) and α-helices in peptides (Zimm &

Bragg, 1959), and are the predominant sites of hydrogen bonding (Stickle *et al.*, 1992). The helical-HP model represents local interactions as an energy favoring helix formation. Physically, this model interaction corresponds to the intramolecular (i, i+4) hydrogen-bonds found in peptide α-helices, as well as other local conformational and steric effects. Two turns of a 2D lattice helix are shown in Figure 1B. The reason this particular lattice configuration is considered to be helical is because it is the only configuration on the 2D square lattice that has the same topology (i.e. contact map) as a 3 dimensional α-helix (Chan & Dill, 1989). When six consecutive monomers of a 2D chain are in such a structure, a helical energy of 2σ is assigned to the conformation, one contribution of σ for each of the two (i, i+3) contacts. The minimum length of a helix is defined to be six residues, i.e. two helical bonds, since a single (i, i+3) contact is more properly defined as a turn. We refer to each helical (i, i+3) contact in the model as a "helical bond." A helical bond is assumed to be favorable, i.e. $\sigma \leq 0$, independently of the HP sequence. The sequence independence of the model helical interaction is an approximation based on the experimental observation that helical propensities in peptides vary by approximately a factor of 8 ranging over the 20 amino acids, excluding proline (O'Neil & DeGrado, 1990; Scholtz & Baldwin, 1992; Lyu *et al.*, 1990), whereas hydrophobic interactions vary by a factor of 100 or more among the amino acids (Nozaki & Tanford, 1971; Chothia, 1976; Fauchère & Pliska, 1983; Rose *et al.*, 1985). It would be straightforward to include a sequence dependence of helicity in this model, but it only adds additional parameters that would obscure its simplicity for the purpose of the present study. Every added helical unit (2 added residues) adds σ to the energy sum; the local energy is then $n\sigma$ for a conformation having n helical bonds.

The total energy of a conformation, then, is:

$$\Delta G = m\varepsilon + n\sigma \qquad (2)$$

relative to an open reference conformation that has no HH contacts and no helical structure. This energy accounts for the intra-chain interactions; the conformational entropy is treated through the enumeration process. For a given sequence of H and P monomers, we evaluate the free energy of each of the 802,075 conformations according to equation (2). The "native state" of a sequence is defined as the conformation, or conformations, that have the lowest free energy for particular values of σ and ε. To illustrate the energy contributions, consider the conformations shown in Figure 1. Conformation A contains four HH contacts but no helical bonds; it has energy $\Delta G_A = 4\varepsilon$. Conformation B contains three HH contacts and two helical bonds; it has energy $\Delta G_B = 3\varepsilon + 2\sigma$. The population ratio of conformer A relative to conformer B will be:

$$\exp\{-[4\varepsilon - (3\varepsilon + 2\sigma)]/kT\} \qquad (3)$$

The strengths of the two types of interaction are varied by changing the values of σ and ε. Conformations A and B will be present in equal populations, in this example, when $\Delta G_A = \Delta G_B$, i.e. when $\sigma/\varepsilon = 1/2$. Relative to this value, if σ/ε is decreased, conformation A will be favored; if σ/ε is increased, B will be more populated. Thus, in general, the native state too will depend on the ratio σ/ε; i.e. the native state of a given sequence generally depends on the balance between local and nonlocal interactions. Different native states for the same sequence will likewise be present in equal populations at integral fractional values of σ/ε.

In physical terms, the quantities ε and σ represent the degrees to which external factors such as solvents and temperature control: (1) the affinity between two hydrophobic amino-acids, and (2) the propensity to form helices, respectively. For example, as noted in

more detail below, alcohols such as ethanol and TFE in water weaken the hydrophobic interaction and strengthen the helical propensity. Increasing concentrations of guanidine hydrochloride (GuHCl) and urea in water weaken both types of interaction (Robinson & Jencks, 1965; Nandi & Robinson, 1984). GuHCl can denature helical peptides (Shoemaker *et al.*, 1988), implying that increasing its concentration in water would correspond, in our model, to a decrease in both ε and σ.

Predictions of the Model

First, consider two limiting cases. When the contact interactions are weak (ε small), the model describes helix-coil processes: we find that all sequences become helical as σ becomes large. For $\varepsilon = 0$ when $\sigma < 0$, the helix is the conformation of minimum energy and therefore defines the native state (i.e., is the conformation of minimum free energy) for all sequences. On the other hand, when there are no local interactions ($\sigma = 0$), the helical-HP model reduces to the HP model explored in previous studies (Lau & Dill, 1989; Lau & Dill, 1990; Chan & Dill, 1991a; Lipman & Wilbur, 1991; Shortle *et al.*, 1992; Miller *et al.*, 1992; Unger & Moult, 1993; O'Toole & Panagiotopoulos, 1993; Camacho & Thirumalai, 1993a; Camacho & Thirumalai, 1994).

The two types of interactions, local and nonlocal, can cooperate to stabilize a given conformation. Figure 2 shows, for two different HP sequences, the effect of strengthening the HH attraction, (i.e. making ε more negative) while maintaining a constant small helical energy (σ). The HH contact interaction helps to increase the stability (i.e. the fractional population) of the 16-residue helix conformation of the sequence labeled A, but

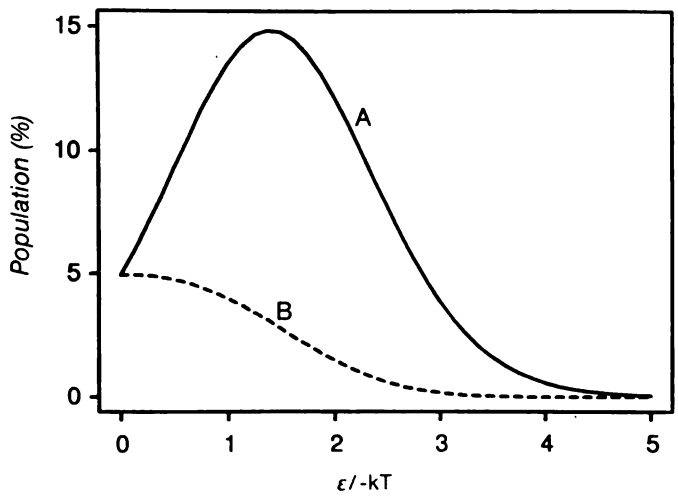


Figure 2. Population of helical conformer vs. hydrophobic interaction, ε, for sequences (A) (HPPHHPHHPHPPHHHH), and (B) (HHPHPPPHPHHPHPHH). For A the helix is stabilized by decreasing ε to about -1.5kT because several HH contacts are made in this conformation. Thus HH contact energy leads to stabilization of the same conformation favored by helical energy. Decreasing ε further leads to destabilization of the helix because the few conformations having more HH contacts become more stable. B, on the other hand, is destabilized by any decrease in ε because many other conformations make more HH contacts. Population is calculated according to equation (4); for both curves, σ = -2.25 kT.

decreases it for sequence B. The fractional population of any single conformation can be calculated for given values of σ and ε as the Boltzmann weight for that conformation, divided by the partition function, Z (which is the sum of the Boltzmann factors of all possible conformations):

$$p_{frac}(conf) = \frac{\exp\left[\left(m_{conf} - m_{native}\right)\varepsilon + \left(n_{conf} - n_{native}\right)\sigma\right]}{\sum_{m=0}^{m}\sum_{n=0}^{N} g(m,n)\exp\left[\left(m - m_{native}\right)\varepsilon + \left(n - n_{native}\right)\sigma\right]} \qquad (4)$$

where the subscripts *native* and *conf* refer to the properties of the native conformation (for that ratio of σ/ε) and the conformation of interest, respectively; m is the number of HH

contacts, n is the number of helical bonds; M is the maximum possible number of HH contacts for the given sequence and N is the maximum possible number of helical turns for a given chain length. The degeneracy function $g(m,n)$ specifies the number of conformations in the ensemble having particular values of m and n. The denominator is the partition function, Z, for the helical-HP model. The properties of the native state are used as the reference state by convention such that the minimum-energy state has a Boltzmann weight of 1. For sequence A, the helical conformation contains several HH contacts, and increasing the energy associated with these contacts stabilizes that conformation. Sequence B in that same conformation makes only one HH contact, so increasing ε only destabilizes the helix in favor of the many other conformations that have more HH contacts. Experiments on helical peptides show a similar result: helices are stabilized if they result in the burial of nonpolar surface (Tanford, 1968; Chou *et al.*, 1972; Richards & Richmond, 1978). Curve A also shows that when a helix is stabilized by HH interactions, further strengthening of the HH interactions can ultimately lead to transitions to even more stable conformations having more HH contacts and fewer helical bonds.

For a given HP sequence, the ratio σ/ε determines which conformation(s) are native. Figure 3 shows examples of "native state phase diagrams," i.e., maps of the native (minimum-energy) states, and the transitions among them, for different values of σ and ε. The conformations shown on the phase diagrams are the most populated (native) conformations, but other conformations are also present in lesser concentrations, depending on their Boltzmann weights. In the lower left-hand corner of each diagram, which corresponds to small interaction energies, all conformations have nearly the same populations; this region corresponds to the denatured state of the chain. As the interaction energies are increased (by increasing the distance from the origin), the native conformation(s) are "frozen out" of the ensemble because their Boltzmann weights dominate the partition function. By increasing the interaction energies from the origin

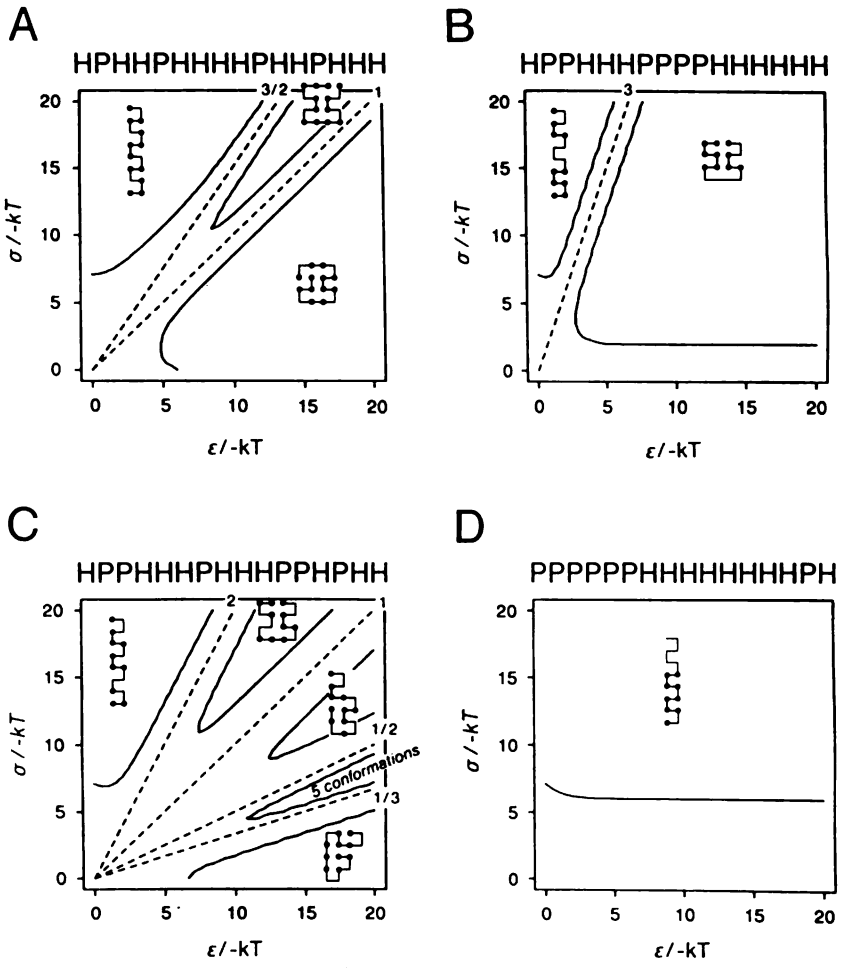


Figure 3. Native state phase diagrams vs. σ and ε. Each region is shown with a line-representation of the native 2D conformation that is most stable (i.e. native) in that particular range of σ and ε. Dashed lines indicate transitions, and are labeled with the value of σ/ε; solid lines enclose the regions where the conformation(s) shown comprise over 95% of the total population (by equation (3)). For small σ and ε, there is a large ensemble of denatured conformations. For large σ and ε, sequence (A) undergoes two "conformational switching" transitions as σ/ε is increased (i.e. sweeping out an angle from the lower right-hand corner to upper left). Sequence (B) shows only a single transition (at high σ and ε) between helix and helix bundle; this sequence has no unique native conformation at $\sigma/\varepsilon = 0$. Sequence (C) has many transitions, one to a non-unique native state favored for $1/3 < \sigma/\varepsilon < 1/2$. Sequence (D) has only a helix-coil transition, because it cannot form a core of contacting H residues.

along a line making an angle θ with the x-axis of these phase diagrams, the native state is stabilized for a balance of energies $\sigma/\varepsilon = \tan\theta$. Transition points (dotted lines) are points of equal populations of different native states, and, as mentioned above, therefore occur at integral fractions of σ/ε.

These phase diagrams show that the model predicts "conformational switching." That is, in some particular "solvent" (i.e. particular values of σ and ε), the sequence has one native state but then changing the "solvent" causes a transition to a different native state. Thus the model can flip-flop between two different states with changes in external conditions. At any given value of σ/ε there may be more than one native conformation with the same number, m, of HH contacts and the same number, n, of helical bonds (e.g., in Figure 3C). This degeneracy of native conformations is the most common situation when the value of σ/ε is small. On the other hand, when σ/ε is large, the helical conformation dominates regardless of the sequence.

How many HP sequences fold to a *unique* native conformation? It depends on the relative interaction strengths, σ/ε. By unique, we mean that only one configuration, out of 802,075 possible, has the minimum free energy. We find the native conformations of each possible sequence of HP copolymers of length 16, over all possible values of σ/ε. The number of distinct sequences (i.e. the size of "sequence-space") is 32,896 if we take into account mirror-image equivalence (HP chains have no polarity, e.g. the tetramer sequence PHPH is equivalent to HPHP). Figure 4A shows the number of sequences that fold to a unique native structure. When local helical forces are dominant (σ/ε is large), all sequences fold to a unique conformation, namely the 16-residue helix. When HH contact interactions

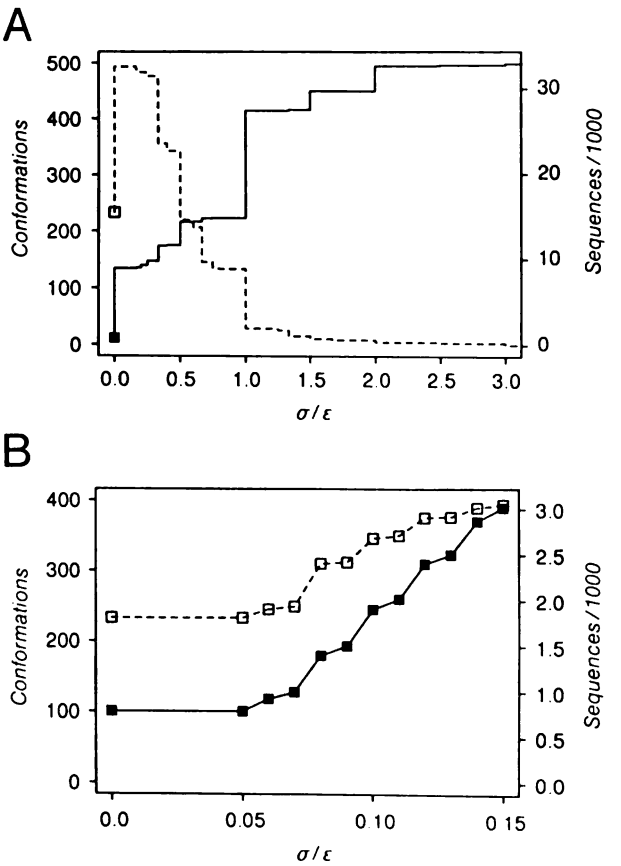


Figure 4. Uniqueness of native structure, vs. σ/ε. (A) Solid line (right scale) is the number of sequences in all of sequence-space that fold to unique native states (size of single-sequence set). It increases with σ/ε because all sequences fold to a helix for large σ. Dashed line (left scale) is the number of different conformations to which sequences fold (size of single-conformation set), which decreases to 1 for large σ. These curves are discontinuous because transitions between native states occur at a finite number of integral fractions of σ/ε. (B) With "minimum stability" criteria to make the low-σ/ε end of each curve more continuous. Note that these criteria are applied only for $\sigma/\varepsilon \leq 0.15$. New sequences are added to the list of "protein-like" sequences when the unique native conformation becomes stable (see section "Comparison of the Helical-HP Model with Properties of Proteins" for details), so for this range both curves are increasing.

are dominant, then fewer sequences map to a greater number of different unique native states.

At intermediate values of σ/ε, the dominant unique native conformations are "helical bundles": two helices packed side by side antiparallel to each other with a common "core" of contacting H residues. (The 2-helix bundle is the 2-dimensional equivalent of a bundle of multiple helices in 3 dimensions). The membrane-spanning regions of many integral membrane proteins take on an α-helical conformation. The nonpolar environment of the membrane interior would lead to a reduced hydrophobic driving force and increased intramolecular hydrogen-bonding driving force (since there will be no solvent hydrogen-bonds available in the membrane interior) relative to water. The membrane-spanning helices almost certainly associate into helical bundles, with the more hydrophobic groups exposed to the membrane (Rees *et al.*, 1989), suggesting that the hydrophobic interactions between side-chains are less favorable than interactions with the acyl chains of membrane molecules. In our model, this corresponds to a contact interaction between P monomers, which is equivalent to the HH interaction (by simply exchanging H's for P's in the sequence and *vice versa*). For such a balance of local and nonlocal interactions, i.e. where local helical interactions dominate but HH (or, equivalently, PP) contacts are still favorable, we find that most HP sequences form either helices or helical bundle conformations.

How often does conformational switching involve a transition between unique native conformations; i.e., from one unique conformation to another unique conformation? The sequence shown in Figure 3A has a unique native conformation at all values of σ/ε (except exactly at transition points), when both types of interaction are strong. This type of behavior turns out to be quite general: more than half of the sequences which have a unique native conformation at one value of σ/ε will also have a unique but different native conformation if σ/ε is increased or decreased through a transition point. The chain does not

pass through a disordered ensemble of conformations between the two states, but rather undergoes a simple two-state transition, toggling between states in response to appropriate changes in external conditions. These sequences are "conformational switches." For other sequences, the behavior can be more complex. Experimentally, peptides have been designed which can act as conformational switches, changing conformation in response to changes in external conditions (Mutter & Hersperger, 1990). The polypeptide Gramicidin A has been shown to undergo a conformational change from double-helical dimer to helical monomer upon insertion into a cell membrane (Killian, 1992). For many HP sequences, the helical-HP model predicts a globule-helix transition for such a change in external conditions, i.e. a large increase in σ and a decrease in ε.

Comparison of the Helical-HP Model with Properties of Proteins

Our aim in this section is to compare properties of the helical-HP lattice model with properties of native protein structures from the Brookhaven Protein Data Bank (PDB), in order to determine which relative strengths of local (helical) and nonlocal (HH contact) interactions cause the model to resemble real proteins most closely. To do so, we must first choose which model sequences are most "protein-like." In this regard, we consider only unique native states predicted by the model, since native states of real proteins are generally unique. Put simply, if over a certain range of σ/ε a given helical-HP sequence has only a single conformation having the minimum energy (according to equation 2), it will be considered to be "protein-like" over that range; if it has multiple (degenerate) native conformations, it will not. We construct a set of all "protein-like" sequences for each range of σ/ε (see Figure 4). Because the native states of each helical-HP sequence vary with σ/ε, different balances of local and nonlocal interactions yield different sets of unique native

conformations. Physically, the quantity σ/ε can be thought of as a general measure of the importance of local interactions relative to the importance of nonlocal interactions, in determining the conformations of proteins. A large σ/ε corresponds to the dominance of local helical propensities, while a small σ/ε corresponds primarily to the dominance of nonlocal hydrophobic interactions. Since in this section we compare our lattice model results to real proteins in water, we are essentially finding the balance of σ/ε for our model proteins which mimics aqueous solvent conditions for real proteins.

For comparison with real proteins, we choose a representative sample of 113 proteins from the PDB, derived from Appendix 3 of *Protein Architecture* (Lesk, 1991). Since we do not know how these known proteins "sample" all the possible protein sequences and structures, we consider the two possible unbiased ways to represent sets of model conformations. The "single-conformation" set contains each unique native conformation once. That is, no matter how many sequences fold to a given conformation, it is represented once. By contrast, in the "single-sequence" set, each sequence that folds to a unique native conformation is represented once; so a given conformation is represented in proportion to the number of sequences that have that same conformation. Hence the single-sequence set is larger than the single-conformation set, and grows with increasing σ/ε (Figure 4). Probably neither of these sequence ensembles truly represents the proteins in the PDB sample, but they provide two unbiased limiting cases. The PDB is not a single-conformation set because nearly identical native structures often appear more than once, such as the globins, due to different sequences that have essentially the same structures. The PDB is probably closer to a single-sequence set because every polypeptide sequence with known structure is contained in it, but it too may be a biased sample of all proteins because the PDB contains relatively small, crystallizable molecules.

The native state of a helical-HP sequence is defined as the conformation(s) of lowest energy. On either side of a transition point, there will be a different native state. But only in the limit of infinite interaction energies will there be an exact transition point between stable states. As is evident in Figure 3, for finite σ and ε there is a range of σ/ε surrounding each transition point in which the native state is not highly stable; the size of this range depends on the absolute values of σ and ε relative to kT. In addition to the requirement that a particular helical-HP sequence have a unique native conformation to be considered "protein-like," we add a stability requirement. We choose ε = -9.5 kT and a fractional population of 95% (by equation (3)), which we refer to as our "minimum stability" criteria. Thus, for example, the helical-HP sequence shown in Figure 3B would not meet the minimum stability requirement for approximately $\sigma/\varepsilon < 0.2$. Figure 4B plots the number of sequences which are "protein-like" on this basis, for small σ/ε.

To compare the model to real protein structures, we examine two properties: (1) the ratio of helix to sheet secondary structure in each set, and (2) the distribution of tertiary structural similarities, according to a property defined below. These distributions do not involve cross comparisons of lattice model and real protein structures. Rather, we obtain a distribution function for the lattice model and one for real proteins, and then compare the two distribution functions.

(1) Secondary Structures in Proteins

Lattice model studies have shown that helices and sheets can be driven by compactness, for example as a consequence of HH interactions (Chan & Dill, 1990). For many sequences, the HP model leads to native states with much secondary structure. But while the 2D lattice model predicts amounts of helix and sheet similar to those found in the

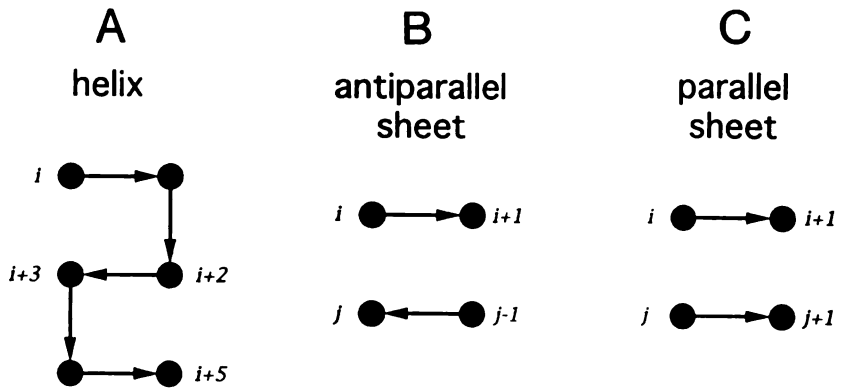


Figure 5. Secondary structures on the 2D square lattice. (A) helix, at least two sequential non-covalent contacts between residues [(i, i+3), (i+2, i+5), ..., (i+2n, i+2n+3)]; (B) antiparallel sheet [(i, j), (i+1, j-1), ..., (i+n, j-n)]; (C) parallel sheet [(i, j), (i+1, j+1), ..., (i+n, j+n)].

Protein Data Bank (Chan & Dill, 1990), recent off-lattice studies show that the secondary structures driven by compactness alone are structurally diverse; without hydrogen bonding, only a small fraction of the conformations that have the helix topology ((*i*, *i*+3) contact repeats) are specifically α-helices (Gregoret & Cohen, 1991; Hao *et al.*, 1992; Yee *et al.*, 1994). Thus compactness gives stability, but not structural specificity, to secondary structures in globular proteins. But it is not yet known exactly how much of the free energy stabilizing α-helices derives from compactness vs. hydrogen bonding, because this depends strongly on what criteria are used to distinguish an α-helix from a non-helical conformation (Yee *et al.*, 1994). The problem is substantial: the amount of α-helix predicted by different published helical criteria can vary from 3 - 50% for a given chain conformation. That is, if loose criteria are used to define helices, then compactness can account for most of their driving force, but if helices are defined by stringent criteria, then hydrogen bonding must also be invoked to account for their observed populations.

In the present study we compare secondary structures from the 2D lattice model with those from proteins in the Protein Data Bank. Because of the difficulties noted above, we focus here on the shapes and relative areas of distribution functions rather than on absolute amounts of secondary structures, since the latter are so sensitive to subjective criteria. While there are ambiguities in defining real protein secondary structures, there is no ambiguity in defining them on 2D square lattices. The definitions of the two-dimensional secondary structure types follow from properties of the contact map and are given in Figure 5 (Chan & Dill, 1990).

What balance between local and nonlocal interactions is needed to bring the model into closest correspondence with real proteins? This is addressed in Figures 6 and 7. For comparison, Figure 6A shows the length distributions of protein helices and sheets derived from the PDB by Kabsch and Sander (1983). The corresponding model distributions are shown in Figure 6B, for different values of σ/ε. It has been noted before (Dill, 1990) that if helical propensities were the dominant forces responsible for the structures in globular proteins, proteins would have more long helices and few short ones: helix stability should increase with length. This prediction is seen in the model distributions shown in Figure 6B: when local helical interactions are significant ($\sigma/\varepsilon > 1/2$), more helices are longer and fewer are shorter in the native states of the model. But when there is less helical driving force ($\sigma/\varepsilon < 1/6$), the model shows a monotonically decreasing helicity with length, as is observed in proteins in the PDB. Hence the distribution functions in Figure 6 have shapes most similar to those for real proteins if the free energy to form one helical bond is from approximately 0-15% of the free energy involved in forming one HH contact.

Figure 7 shows how the weighted areas under these distribution functions, which are simply equal to the total number of residues participating in each type of secondary structure, depend on σ/ε. The use of the area under the curve helps reduce arbitrariness of

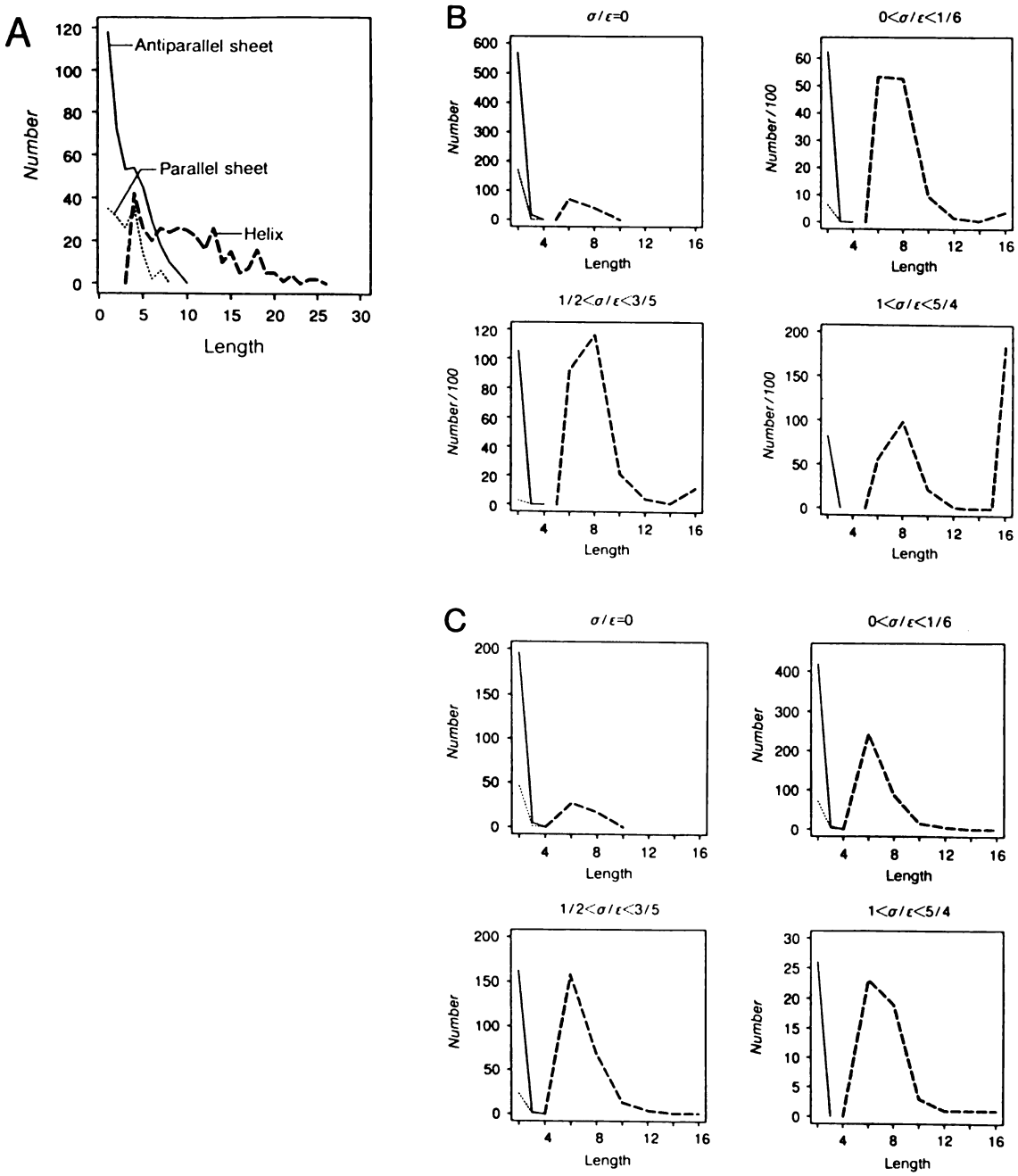


Figure 6. Distributions of secondary structures vs. length. (A) Kabsch & Sander (1983) distributions obtained from the PDB. (B) model using single-sequence sets over four sample intervals of σ/ε, and (C) model using single-conformation sets.

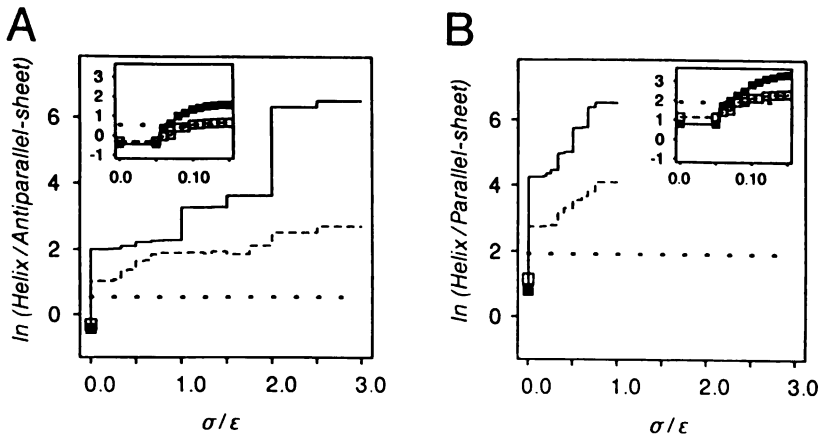


Figure 7. Ratios of secondary structure types vs. σ/ε. (A) shows the ratio of (total number of residues in helices) / (total number in antiparallel sheets) for the model and the PDB (dotted flat line). (B) shows helices / parallel sheets. These ratios compare the relative amounts of different secondary structure types. The solid line corresponds to the single-sequence sets, and the dashed line represents the single-conformation sets. The insets show small σ/ε using the additional "minimum stability" requirement for "protein-like" conformational states to smooth the low-σ/ε end of the curves. In all cases the model values intersect the PDB values for $\sigma/\varepsilon < 0.1$. Note the logarithmic scale; helical secondary structure dominates at higher σ/ε.

decisions about whether helices observed in real proteins are continuous or broken, and it helps compensate for the limited data due to the shortness of the chains in the lattice model. Figures 7A and 7B show the helix / antiparallel sheet ratio, and the helix / parallel sheet ratio, respectively, as functions of σ/ε in the model. The value for the Kabsch & Sander PDB study is shown as a horizontal line. Both the single-sequence and the single-conformation sets are shown in each figure; they show significant agreement, and mainly differ only in magnitude.

From Figure 7 it is evident that $0.5 < \sigma/\varepsilon < 0.1$ is the range for which the model most accurately reproduces the helix / parallel sheet and the helix / antiparallel sheet ratios of proteins in the PDB. One notable feature of the model curves is the large step increase in the helix / sheet ratio for even the smallest increment in σ/ε above zero. This arises because when even a small nonzero local term is added to the free energy, it breaks the degeneracy between conformations with a given number of HH contacts that have different numbers of helical bonds. As a result, all conformations in the $\sigma/\varepsilon > 0$ sets which are not present in the $\sigma/\varepsilon = 0$ set *must* contain at least one helical unit (six residues). Using the minimum stability requirement, these helix-containing conformations are added to the set gradually with increasing σ/ε and we obtain smooth curves for $\sigma/\varepsilon < 0.15$ (Figure 7, insets).

The helix / parallel sheet ratio (Figure 7B) is even more sensitive than the helix / antiparallel sheet ratio to the helical propensities. But because of the shortness of the model chains, there are relatively few ways to form parallel sheets when six of the sixteen residues are already involved in a helix. That there is so little freedom to form such sheets accounts in part for the large increase in the helix / parallel sheet ratio for $\sigma/\varepsilon > 0$. This effect, however, is offset by the inevitable double-counting of secondary structure on a low-resolution 2D lattice. Because of the simple 2D definitions of secondary structure (Figure 5), a helical residue can also be counted as participating in a sheet when another part of the chain folds back to contact the helix. For example, in Figure 1B, residues 1-6 are in a helix and residues 2-3 also form a parallel sheet with residues 15-16.

We conclude that the model native secondary structure distributions are most similar, in both shapes and relative areas, to those in the PDB when the model local interaction is assumed to be much smaller than the nonlocal interaction ($\sigma/\varepsilon < 0.1$).

(2) Tertiary Structures and the QQ Plot

Having compared secondary structure distributions between the model and real proteins, we now examine a property involving tertiary structure comparisons. We first briefly describe a method for obtaining a distribution of pairwise dissimilarities between tertiary structures, and then a general method for comparing different distributions. Finally, we use these methods to compare our model conformation distributions to the PDB. For this tertiary structure comparison as well, the model bears closest resemblance to real proteins when the helical interactions in the model are set to be near zero.

The pairwise structural dissimilarity distribution, described in more detail elsewhere (Yee & Dill, 1993), is based on a measure, *d(R, S)*, of pairwise dissimilarities of two polymer or protein conformations, R and S. *d(R, S)* is a number that ranges from 0 to 1, 0 indicating the structures are identical, and 1 indicating they have the greatest possible structural dissimilarity. The dissimilarity of two chain conformations is computed from their weighted distance maps. For a chain of length L, the weighted distance map is an L^2 matrix in which each element w(i, j) equals the distance between the positions of residues i and j (Cα coordinates are used for protein comparisons; lattice monomer sites are used for the model) raised to the inverse power, p ($p > 0$):

$$w(i,j) = d(i,j)^{-p} \qquad (5)$$

The weighted distance map has the property that residues that are close together in space are weighted more heavily than residues that are distant in space. Here we use $p = 2$, so the weights include contributions from residues other than nearest neighbors. The comparison of two distance maps, to get the score *d(R, S)*, is made by sliding one map across the other, and finding the alignment of highest similarity.

Using this measure, *d(R, S)*, of dissimilarity between conformations R and S, we find the pairwise dissimilarities among all the conformations in a given set. The *d(R, S)*

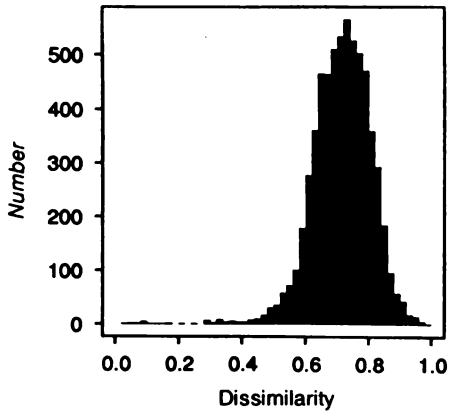


Figure 8. Distribution of pairwise tertiary structure dissimilarities for 113 proteins from the PDB (see text for details).

distribution is shown in Figure 8 for $n(n\text{-}1)/2 = 6328$ pairwise comparisons of 113 representative proteins in the Protein Data Bank. It represents the relatedness among the tertiary structures of known proteins. Correspondingly we obtain a $d(R, S)$ distribution for the native conformations predicted by the 2D helical-HP model for different intervals of σ/ε (Figure 9 insets). For each interval of σ/ε, distributions were generated for both the single-conformation and single-sequence sets.

How similar are the two distribution functions, of the pairwise dissimilarities among proteins, and of the pairwise dissimilarities among native states in the helical-HP model? We make this comparison using a quantile-quantile (QQ) plot (Chambers *et al.*, 1983). This type of plot compares the shapes of two distribution functions. The range of values on the x and y axes of the QQ plot is 0 to 1. Point (x, y) on the QQ plot is the pair ($d(R, S)$ value of one distribution, $d(R', S')$ value of the other distribution) at which the areas under the two distribution functions are equal. That is, a point (x, y) on a QQ plot is represented by:

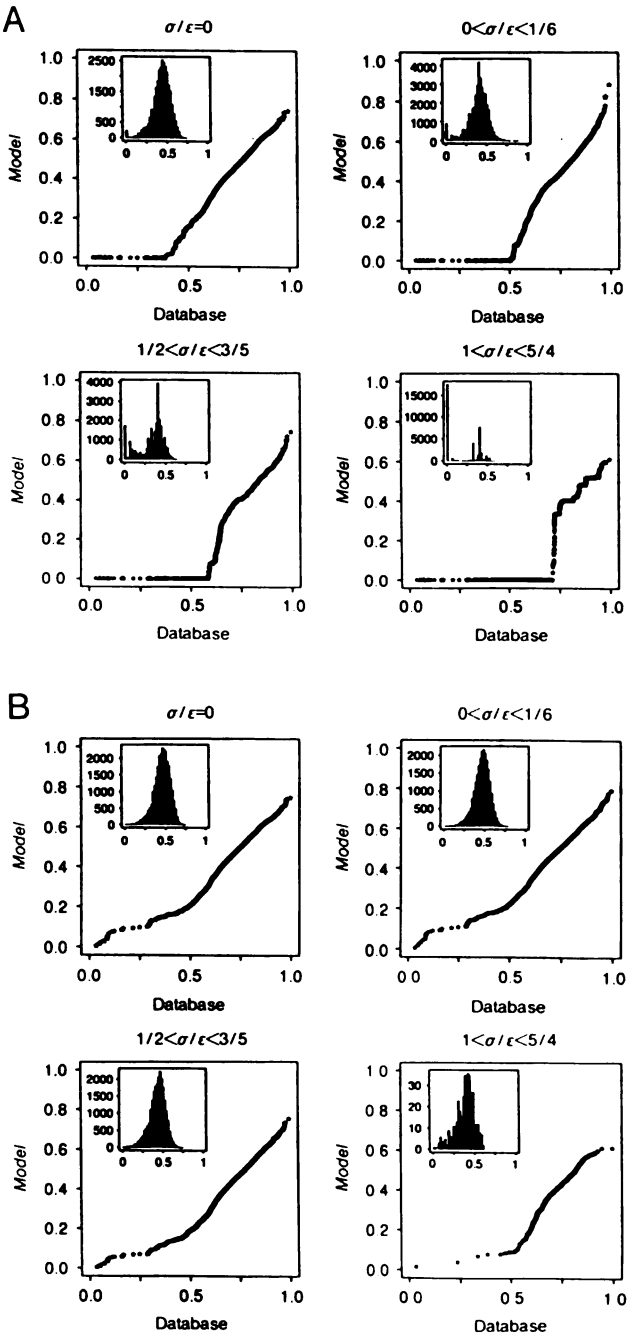


Figure 9. Model pairwise dissimilarity distributions (insets). Their agreement with the PDB distribution is shown as QQ plots. The more linear the plot, the more similar the model and PDB distributions. (A) Single-sequence and (B) single-conformation sets. The greatest similarity is for $\sigma/\varepsilon = 0$ in (A), and $\sigma/\varepsilon < 3/5$ in (B). These distributions are constructed by comparing each conformation in the set over a given range of σ/ε pairwise with every other conformation in that set. Note the peak at score 0 in (A), which reflects different sequences folding to identical conformations.

$$\int_0^x A \cdot ds = \int_0^y B \cdot ds \qquad (6)$$

where A is the distribution function represented on the x-axis of the plot, B on the y-axis, and ds is a small score interval.

The QQ plot has several properties that make it useful for comparing distribution functions. First, if two distributions have identical shape, the QQ plot is linear. For example, the comparison of two gaussian distributions will yield a linear QQ plot irrespective of their relative widths and displacements. If the QQ plot is linear, the slopes and the intercepts have a simple interpretation. The slope corresponds to the relative widths of the distributions; e.g. a slope of 3 indicates that distribution B is wider than distribution A by a factor of 3. The y-intercept corresponds to the shift of the mean of one distribution relative to the other: it is the point on distribution A that corresponds to 0 on distribution B. Figure 9 also shows QQ plots for the 2D helical-HP lattice model vs. the PDB for the given intervals of σ/ε.

We focus on three properties of the QQ plot: (1) linearity (resemblance of the shapes), (2) slope (resemblance of the widths), (3) slope-intercept (resemblance of the mean values). These three properties of the QQ plot most quantitatively express the resemblance of the pairwise structural dissimilarity distribution of the lattice model native conformations with the PDB distribution, and are plotted vs. σ/ε in Figure 10. The RMSD of the QQ plots from linearity (Figure 10A) is probably the best measure of deviation between distributions, since if linearity does not hold, the other measures lose their simple interpretations.

Figure 10A shows that the model most nearly resembles the PDB for $\sigma/\varepsilon = 0$ for the single-sequence set, and for $\sigma/\varepsilon < 1$ for the single-conformation set. This result holds true

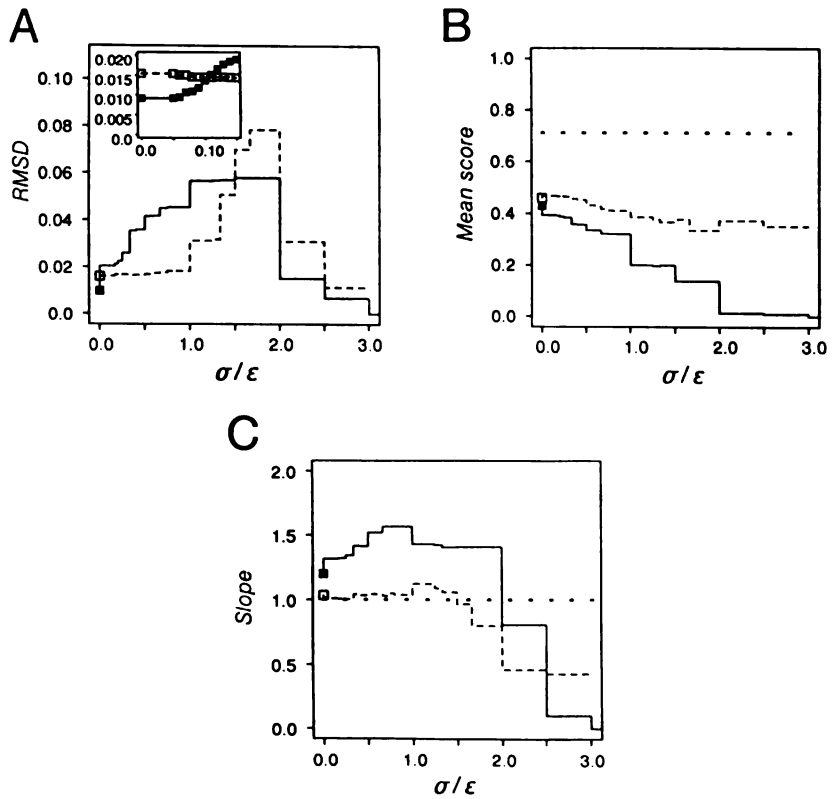


Figure 10. Comparing the QQ plots. (A) RMSD of QQ plots from linearity. The inset uses the "minimum stability" requirement for defining protein-like conformational states, which yields a "continuous" curve for $\sigma/\varepsilon \leq 0.15$. An RMSD of 0 indicates that two distribution shapes are identical; larger RMSDs indicate less similarity between distribution shapes. (B) Mean dissimilarity scores of the distributions. The PDB distribution has a mean of 0.71. (C) Slopes of linear fits to the QQ plots. A slope of 1 indicates that the two distributions have identical widths; smaller slopes indicate narrower distributions. Solid and dashed lines represent single-sequence and single-conformation sets respectively.

even when the more stringent requirement of minimum stability is used to define the conformation sets (Figure 10A inset). As noted above, the single-sequence set may be more representative of the sequence distribution in the PDB. The single-conformation set is much less sensitive to small changes in σ/ε, since conformations are lost gradually with

increasing σ/ε (Figure 4A). The fall of the RMSD to zero for large σ/ε does not indicate an increase in similarity of the distributions; it simply reflects the degenerate case. The reason for this fall in RMSD is apparent from Figure 4A: most sequences converge to a small number of different conformations. The single-sequence dissimilarity distribution collapses to a peak at 0 and the QQ plot is fit well by a line of slope 0, while the single-conformation distribution becomes undefined since there is only one conformation.

The mean scores graphed in Figure 10B simply confirm the expected result that sets of 16-residue conformations on a 2D lattice do not have as much structural variation as 3D proteins of lengths from roughly 50 to 200 residues. The mean score for the model decreases with increasing σ/ε, as native conformations become increasingly helical and tend to resemble each other more closely. What may be more surprising, however, is that the model dissimilarity distributions predicted by the model for $\sigma/\varepsilon < 1$ have approximately the same width as the PDB (Figure 10C). This result indicates that the model displays a *range* of structural variation similar to the PDB.

Thus the dissimilarity distribution for the single-sequence set at $\sigma/\varepsilon = 0$ differs from the PDB distribution primarily in having a lower mean score. If we were to shift the PDB distribution to match this lower mean score, the low-score tail of the distribution would show negative dissimilarity scores. Significantly, the relative area of this low-score region almost exactly matches the relative area of the 0-score peak of the single-sequence distribution for $\sigma/\varepsilon = 0$. This tail indicates some clustering of the proteins in the PDB into families (Yee & Dill, 1993), which the low-resolution model reflects as different sequences having identical conformations. In this way, the single-sequence set for $\sigma/\varepsilon = 0$ captures a prominent feature of the PDB which is absent from the single-conformation sets.

We conclude that the closest resemblance of the model to real proteins is obtained when the nonlocal interactions dominate and the local interactions in the model are small.

Thus adding a helical interaction to the HP model does not improve its ability to resemble properties of proteins in the PDB. Even if the 2D model intrinsically overestimates the amounts of helix, then this artifact alone would account for why no helical propensities need be added to the HP model to reach the amounts of helix observed in real proteins. But this explanation would fail to account for the agreement of the distributions of secondary structural types, the decreasing frequency of helices with length, and the pairwise tertiary structural similarities. Apart from the disagreement of the chain-length dependent mean score, the distribution of model native structures for $\sigma/\varepsilon = 0$ resembles closely that of proteins, including some clustering into families seen in the single-sequence set. Because the model has the low resolution of a square lattice, is restricted to short chains, and is 2-dimensional, it is obviously not a microscopically accurate representation of real proteins. But while this precludes any quantitative results, two different measures, of secondary and tertiary structures, agree that a striking similarity with real proteins is displayed by our model when local interactions are very small compared to nonlocal interactions. These results imply that the relative distributions of helix and sheet, and the general distribution of tertiary structural topologies, of globular proteins in aqueous solution are dictated more strongly by the nonlocal hydrophobic interactions than by helical propensities.

The Mechanism of Alcohol Denaturation

Globular proteins can be denatured or stabilized by agents in solution which mediate nonlocal or local interactions or both. For example, whereas urea and guanidine hydrochloride disrupt both hydrophobic clustering and secondary structure, denaturation by alcohols is more complex. Also, peptides adopt different secondary structures depending on the nature of the solvent (Zhong & Johnson, 1992). In this section, we use

the helical-HP model to explore the different classes of structures that could arise when different types of agents, particularly alcohols, act on globular proteins.

How might alcohols affect peptide and protein conformations? Several physical mechanisms have been proposed to explain these conformational effects, but the relative importance of each has not yet been resolved. Focusing on helix induction by TFE, Nelson and Kallenbach (1986) have compared charge and hydrogen-bonding mechanisms. The dielectric constant of TFE is approximately one-third that of water, so charge interactions should be more important in TFE. By varying the dielectric constant, Nelson and Kallenbach found a negligible increase in peptide helix stabilization in TFE, and concluded that hydrogen-bonding may be the more important mechanism of the two. Nuclear magnetic resonance (NMR) studies by Llinás and Klein (1975) have shown that TFE is a slightly stronger proton donor than water, but is a much weaker proton acceptor. Polypeptide backbone groups have both donors and acceptors. Since the effect of TFE on proton acceptance dominates, adding TFE to aqueous solution will primarily decrease the solvent's ability to compete with peptide carbonyl acceptors. In TFE, the peptide amide donors should therefore favor making hydrogen bonds with peptide carbonyls. Therefore intramolecular hydrogen bonds should be strengthened by the addition of TFE to an aqueous solution. Because the α-helical conformation forms backbone-backbone hydrogen bonds, it will become increasingly favored by addition of TFE to an aqueous solution.

The effects of alcohols on polypeptides have been found to be strongly sequence-dependent (Lehrman *et al.*, 1990; Segawa *et al.*, 1991; Sönnichsen *et al.*, 1992). The tendency to induce helical structure has been found in nearly all studies, but the amount of alcohol needed varies widely depending on the sequence and the length of the polypeptide. Three peptides corresponding to β-sheet-containing regions of plastocyanin show no appreciable increase in helical content in up to 90% TFE (Dyson *et al.*, 1992). For some

peptides, the amount of TFE required for the coil-to-helix transition correlates with the helical propensity according to secondary-structure prediction methods such as those of Chou and Fasman (Chou & Fasman, 1974). For proteins, the effects of alcohol appear to be more complex. Myelin Basic Protein (MBP) in 92% TFE shows approximately 47% helical structure as judged from far-UV circular dichroism (CD) data (Stone *et al.*, 1985), while ubiquitin shows almost 100% helical structure under similar conditions (Wilkinson & Mayer, 1986). More recent NMR studies of ubiquitin in 60% methanol (dielectric constant = 51), however, indicate that much of the native secondary structure is preserved despite the CD results suggesting a substantial increase in helical content (Harding *et al.*, 1991). Additional studies on the fragments of MBP (Stone *et al.*, 1985), and other proteins (Lehrman *et al.*, 1990; Segawa *et al.*, 1991), show that the sum of CD spectra for proteolytic fragments of a protein is not the same as the spectrum for the intact protein. Hence nonlocal interactions appear to play some role in determining what conformations are induced by alcohols.

The stability of ribonuclease is decreased by addition of alcohols and tetraalkonium salts; the magnitude of this decrease depends only on the number of effective methylene groups of the compound, i.e. on the hydrophobicity of the solvent (von Hippel & Wong, 1965). Further, the effect of ethanol on the thermodynamics of thermal ribonuclease denaturation displays the same temperature dependence as partition experiments on small hydrophobic solutes (Brandts & Hunt, 1967). NMR and near-UV CD studies of alcohol-induced states of proteins indicate a disruption of well-defined tertiary structure (Dufour & Haertlé, 1990; Harding *et al.*, 1991; Buck *et al.*, 1993; Fan *et al.*, 1993; Alexandrescu *et al.*, 1994). The alcohol denaturation of ubiquitin (Wilkinson & Mayer, 1986), myoglobin and chymotrypsinogen appear to be well described by a "preferential solvation model" in which alcohols are proposed to interact most strongly with hydrophobic groups on the protein (Arakawa & Goddette, 1985). Therefore in addition to its effects on hydrogen

bonding, TFE and other alcohols may also affect the strength of nonlocal interactions in proteins.

Thus it is known how the helicity of some proteins changes with concentration of agents such as TFE and urea. Now by studying how the properties of lattice model proteins changes as the strengths of helical and HH contact interactions are changed in fixed proportions, we can explore the mechanism of action of these agents. The helical-HP model contains both local and nonlocal mechanisms in their simplest forms. The strengthening of intramolecular helical bonds of proteins in TFE is modeled as an increase in the helical bond energy σ. The weakening of hydrophobic interactions is modeled as a decrease in the HH contact interaction energy ε. In the previous sections of this paper, we established that globular proteins in aqueous solutions are best represented in our model by taking the ratio of local to nonlocal interaction strength, σ/ε, to be approximately zero. Now let us suppose the effect of increasing the concentration of some agent, such as alcohol, is to strengthen helical bonds (σ more negative) and weaken HH interactions (ε less negative) in the ratio μ. For example, for an agent that has $\mu = 1$, every increment of its concentration leads to 1 unit of free energy stabilization of helical interactions and 1 unit of free energy destabilization of the HH interaction. An agent that has $\mu = 0$ is one that acts entirely to weaken HH interactions. An agent with $\mu >> 1$ is a pure helical bond stabilizer. An agent with $\mu << -1$ is a pure helical bond destabilizer. Figure 11A shows a native state phase diagram for a helical-HP sequence, and the lines with different slopes μ correspond to the ways that different agents will lead to the traversal of different stable conformations.

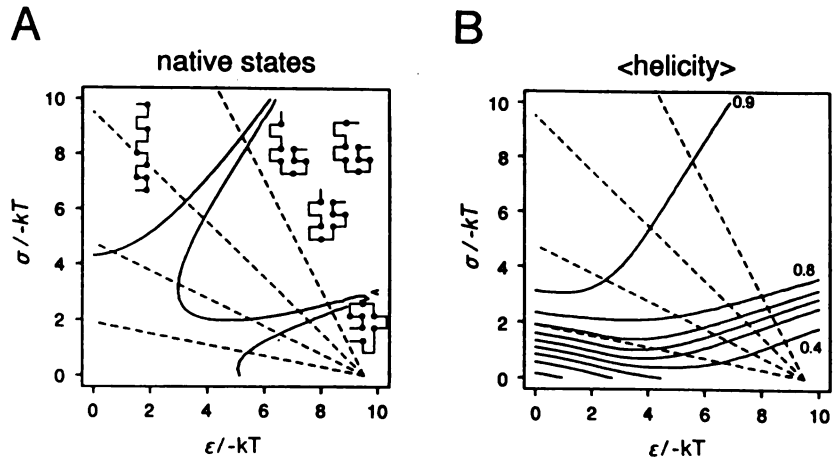


Figure 11. Model solvent effects on σ and ε. (A) The dashed lines show different possible models for how an incremental change in solvent concentration affects σ and ε, in the ratios (from bottom to top) μ = 1/5, 1/2, 1, 2 for the native state phase diagram of the sequence (PHPPPHPPHPHPHHHP). The alcohol-induced conformations are less stable than the aqueous native state for μ < 1. (B) The corresponding <helicity> phase diagram, calculated from equation (7).

The average helicity, $\left\langle \frac{h}{L} \right\rangle$, for any helical-HP sequence can be calculated as a function of the energies σ and ε, using:

$$\langle helicity \rangle = \left\langle \frac{h}{L} \right\rangle = Z^{-1} \sum_{m=0}^{M} \sum_{n=0}^{N} \sum_{h=0}^{L} \frac{h}{L} g(m,n,h) \exp\left[(m - m_{native})\varepsilon + (n - n_{native})\sigma \right] \qquad (7)$$

where h is the number of residues participating in helices, L is the chain length, and all other symbols are as defined for equation (3). Figure 11B shows a sample <helicity> phase diagram for a helical-HP sequence.

Figure 12A shows the model denaturation curves for the sequence in Figure 11. The various curves represent different balances (μ) in the increase in helical bond strength relative to the decrease in HH interaction strength. These curves are "slices" through the <helicity> phase diagram (Figure 11B). For comparison, experimental helicities measured

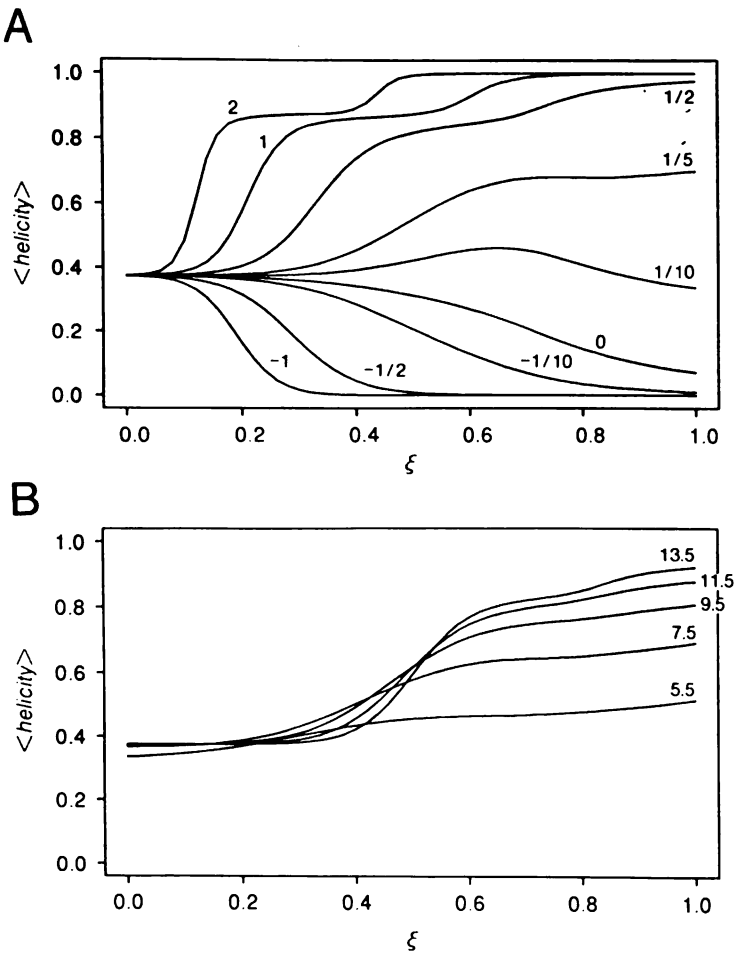


Figure 12. Model denaturation curves for different solvents. Same sequence as in Figure 11. (A) Model solvent μ's are shown for ε = -9.5kT, and range from -1 (weakens helix and HH in equal proportions) to +2 (helix is strengthened twice as much as HH is weakened). The reaction coordinate ξ corresponds to ε = -9.5kT at ξ = 0, and ε = 0 at ξ = 1, which traces the dotted lines in Figure 11 from lower right-hand corner to upper left. (B) shows the effect of different values of ε, in units of -kT, in "aqueous solution" (σ/ε = 0), i.e. of changing the location of the starting point, but not the angle μ, for the traces. Here again ε = 0 at ξ = 1, and each curve is labeled with the value of ε at ξ = 0.

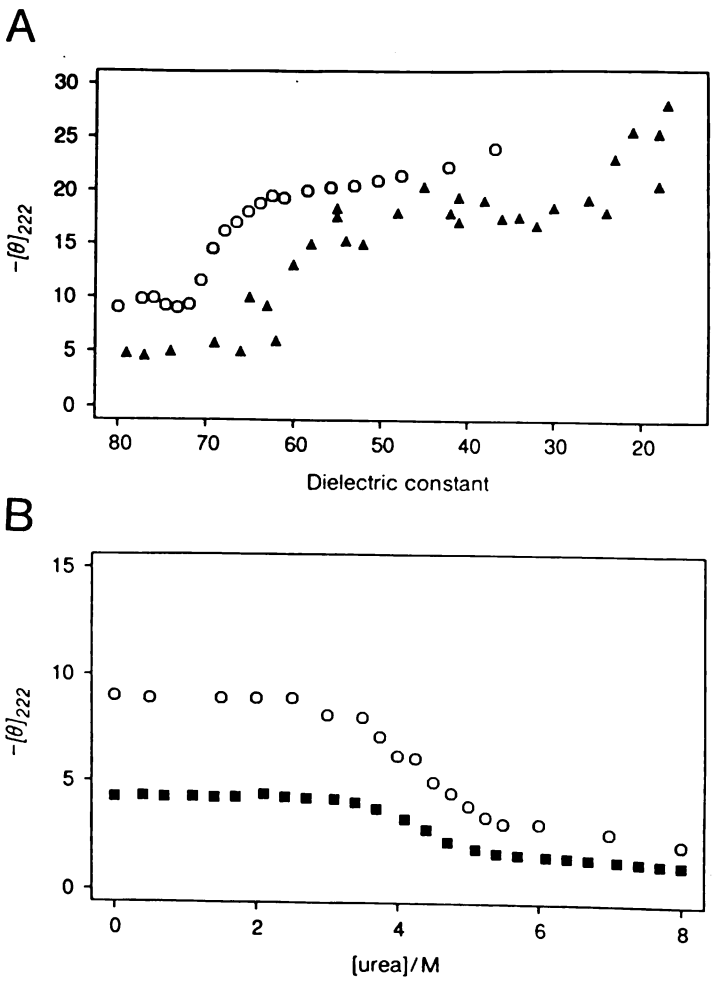


Figure 13. Experimental protein denaturation monitored by CD. (A) The molar ellipticity at 222 nm as a function of the dielectric constant of the solvent, which allows display of the effects of various different alcohols (Wilkinson & Mayer, 1986). The TFE denaturation of hen-egg-white lysozyme is shown with circles, and the denaturation of ubiquitin using methanol, ethanol, isopropanol and butanol is shown with triangles. (B) The molar ellipticity at 222 nm as a function of urea concentration for hen-egg-white lysozyme (circles) and FK binding protein (squares). Data for ubiquitin are adapted from Wilkinson & Mayer (1986); those for lysozyme from Buck *et al.* (1993); those for FKBP from Egan *et al.* (1993). Note that the units for $[\theta]_{222}$ are different for the different proteins.

by far-UV CD during alcohol and urea denaturation of different proteins are shown in Figure 13. The alcohol-denaturation data are best fit by approximately $1/2 < \mu < 1/5$, suggesting that the primary effect alcohol is on the nonlocal interactions rather than on the local helical interactions. That is, in the action of alcohols on the conformations of globular proteins, our model predicts that the hydrophobic contribution to the free energy is 2-5 fold more important than the helical propensities. We refer to agents with these values of μ as model alcohols.

Let us compare denaturation by urea with denaturation by alcohols. Fig 13B shows the experimental CD data for lysozyme and FK-Binding Protein (FKBP). The model best fits these curves for a range of values $\mu \leq 0$. Thus, whereas TFE and other alcohols strengthen helical propensities, urea weakens them. According to this model comparison, urea acts on proteins mainly by weakening the intrachain hydrophobic interactions, and probably also by weakening helical propensities, but to a lesser degree, perhaps through effects on hydrogen bonding.

The shapes of the transition curves depend not only on μ but also on the value of ε chosen to represent the pure "aqueous" solution. This parameter controls the sharpness of the curves (see Figure 12B). We also find that different HP sequences yield different denaturation curves. However, for nearly all sequences having a unique native conformation at $\sigma/\varepsilon = 0$, the qualitative behavior is similar to Figure 12, depending mainly on the amount of helix in the native conformation.

We find that different native structures respond differently to alcohol. That is, effects are sequence-dependent. Adding model alcohol to a "sheet" lattice protein causes it to become helical (Figure 14A). Even though the helical assistance from the model alcohol is weak, it is sufficient to stabilize helices when the alcohol disrupts the HH interactions that hold the sheet together. The sheet-to-helix transition has been observed experimentally

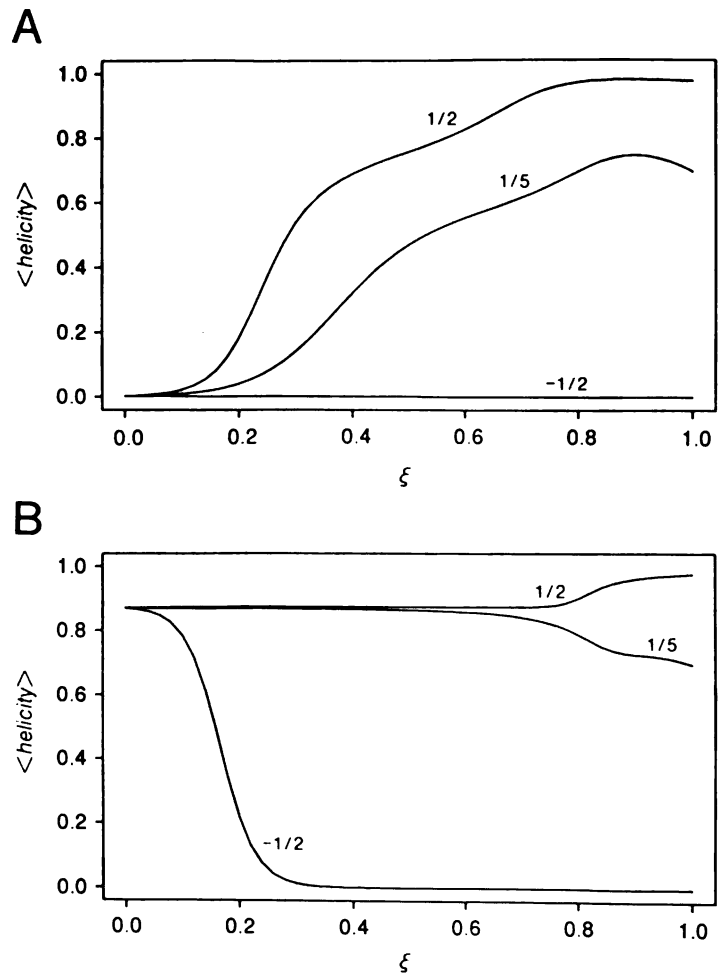


Figure 14. **(A) A model sheet protein** becomes helical in model alcohol ($1/5 < \mu < 1/2$), but denatures in model urea ($\mu = -1/2$). Sequence (HPHHPHHHHHHPHHHH) has a sheet conformation in aqueous conditions ($\sigma/\varepsilon = 0$). **(B) A model helical bundle protein**, sequence (HHPHHHHPHHPHHHHH), does not change helical content in model alcohol, but denatures in model urea. The reaction coordinate ξ corresponds to $\varepsilon = -9.5kT$ at $\xi = 0$, and $\varepsilon = 0$ at $\xi = 1$, as in Figure 11.

for several β-sheet proteins, such as concanavalin A (Jackson & Mantsch, 1992) and β-lactoglobulin (Tanford, 1960; Dufour & Haertlé, 1990). If TFE caused helix formation primarily by its action on hydrogen bonding alone, then it would be difficult to explain why

β-sheet proteins should become helical, since the total number of hydrogen bonds (intramolecular and peptide-solvent) should not change substantially in the sheet-to-helix transition. For a helical protein, on the other hand, the model predicts that alcohol will have essentially no effect on helical content, up to the very highest alcohol concentrations (see Figure 14B), with perhaps some slight decrease at the very highest concentrations. Consistent with this prediction, myoglobin, which is mostly α-helix, is found to have little change in helical content in 0% to 100% chloroethanol (Jackson & Mantsch, 1992).

Figure 12A also shows that for some HP sequences, model alcohol denaturation can induce a two-state conformational transition to a relatively stable intermediate state. This intermediate state still contains a significant population of conformations other than the native for approximately $\mu \leq 1/2$; for larger values of μ, the intermediate helical states are of comparable stability to the model "aqueous" native state ($\sigma/\varepsilon = 0$). A putative folding intermediate for hen-egg-white lysozyme in 50% TFE has recently been characterized by CD, 2D NMR and NMR deuterium exchange experiments (Buck *et al.*, 1993). The transition to this intermediate appears to be two-state, and NMR indicates significant conformational averaging in the intermediate state. A potentially intermediate form of β-lactoglobulin has also been observed at about 20% ethanol which shows a different binding stoichiometry and only a small change in ellipticity at 222 nm relative to the aqueous protein (Dufour & Haertlé, 1990). Monellin adopts a relatively stable non-native conformation in both 50% ethanol and 50% TFE in which one of the β-strands converts to an α-helix while the native α-helix remains intact (Fan *et al.*, 1993). Two helical regions have been identified from the 2D NMR spectrum of α-lactalbumin in 50% TFE at low pH (Alexandrescu *et al.*, 1994); one of these regions forms an α-helix in the native

conformation as well, while the other does so only in the presence of TFE. We find similar behavior for the model helical-HP sequence shown in Figure 11A. A conformational transition occurs at $\sigma/\varepsilon = 1/3$, in which the C-terminal residues become helical in model alcohol. This state contains three native conformations, all of which preserve the N-terminal helix found in the native state as well as the C-terminal model alcohol-induced helix. The three conformations differ in the relative orientation of the two helices; i.e. the secondary structure is well-defined, but the tertiary structure is not.

Despite its simplicity, the helical-HP model displays features of real proteins in solutions with alcohol cosolvents, including CD curves and conformational transitions for some sequences. These data are qualitatively fit for a range of values $1/5 < \mu < 1/2$, which suggests that the dominant effect of alcohols on aqueous protein solutions is on the hydrophobic interaction. Because both H and P residue types have the same propensity to form helices, the model predicts that all sequences will ultimately become helical in alcohol solutions. The observation that some peptides do not adopt a helical conformation in TFE (Dyson *et al.*, 1992) would require a generalization of the model in which different residues could have different helical propensities.

Conclusions

We describe the helical-HP lattice model of protein conformations. It contains two interaction parameters: ε, representing the nonlocal contact interactions that favor HH contacts, and σ, representing a local propensity for helix formation. For all the possible sequences for chains of length $L = 16$, we vary these two energy parameters over their full ranges. When the HH contact energy is the dominant interaction, different sequences

collapse to different unique native conformations that are compact and globular. When the helix interaction is dominant, the lowest energy state of all sequences is the helical conformation. What relative strength of helical and HH contact energies causes the lattice model to have structural properties resembling those of real proteins? We consider two structural properties. First, we consider the relative amounts of different secondary structure types. HH-driven compactness alone, in the absence of significant local interactions, leads to model distributions of secondary structures that most closely resemble those of proteins in the PDB. Even a small increase in the helical interactions in the model leads to much more helix than is observed in real proteins. The second property we consider is the distribution of structural similarities among all proteins. The 2D lattice model gives distributions that closely resemble those of real proteins. For this property, too, the model most closely resembles real proteins if the helical interaction is very small. Of course, there are important caveats due to the simplicity of the model— it is 2-dimensional, restricted to lattice conformations, low-resolution, and involves only short chains. Nevertheless, the close similarities with real proteins suggest that intrinsic helical propensities may not be a strong driving force for the folding of globular proteins.

When both HH and helical interactions are present, chains are found to undergo *conformational switching transitions* often from one unique conformation to another in response to changes in external "solvent" conditions. Conformational switching involves transitions between different native states, and results from a tradeoff between helix formation and maximizing HH contacts. This model describes the effects of different types of denaturing agents on protein conformations, including the transitions of sheet-like and other proteins to helical conformations in alcohols. Comparisons with experiments suggest that while urea weakens nonlocal hydrophobic interactions and probably helical propensities, alcohols such as TFE act mainly by weakening hydrophobic interactions but with some small strengthening of helical propensities.

Acknowledgments

We thank David Yee and Dr. Hue Sun Chan for helpful discussions and for providing some computer programs. We also thank Dr. Christopher M. Dobson for helpful discussions and suggestions. Support for this work was provided by the NIH. P.D. Thomas is a Howard Hughes Medical Institute Predoctoral Fellow.

References

Alexandrescu, A.T., Ng, Y.-L. & Dobson, C.M. (1994). Characterization of a Trifluoroethanol-Induced Partially Folded State of α-Lactalbumin. *J. Mol. Biol.* **235**, 587-599.

Anfinsen, C.B. & Scheraga, H.A. (1975). Experimental and Theoretical Aspects of Protein Folding. *Adv. Protein Chem.* **29**, 205-300.

Arakawa, T. & Goddette, D. (1985). The Mechanism of Helical Transition of Proteins by Organic Solvents. *Arch. Biochem. Biophys.* **240**, 21-32.

Brandts, J.F. & Hunt, L. (1967). The Thermodynamics of Protein Denaturation. III. The Denaturation of Ribonuclease in Water and in Aqueous Urea and Aqueous Ethanol Mixtures. *J. Am. Chem. Soc.* **89**, 4826-4838.

Buck, M. Radford, S.E. & Dobson, C.M. (1993). A Partially Folded State of Hen Egg White Lysozyme in Trifluoroethanol: Structural Characterization and Implications for Protein Folding. *Biochemistry* **32**, 669-678.

Camacho, C.J. & Thirumalai, D. (1993a). Kinetics and Thermodynamics of Folding in Model Proteins. *Proc. Natl. Acad. Sci. USA* **90**, 6369-6372.

Camacho, C.J. & Thirumalai, D. (1994). Minimum Energy Compact Structures of Random Sequences of Heteropolymers. *Phys. Rev. Lett.* **71**, 2505-2508.

Chambers, J.M, Cleveland, W.S., Kleiner, B., & Tukey, P.A. (1983). *Graphical Methods for Data Analysis*. Duxbury Press, Boston.

Chan, H.S. & Dill, K.A. (1989). Compact Polymers. *Macromolecules* **22**, 4559-4573.

Chan, H.S. & Dill K.A. (1990). Origins of Structure in Globular Proteins. *Proc. Natl. Acad. Sci. USA* **87**, 6388-6392.

Chan, H.S. & Dill, K.A. (1991a). Sequence-Space Soup of Proteins and Copolymers. *J. Chem. Phys.* **95**, 3775-3787.

Chan, H.S. & Dill, K.A. (1991b). Polymer Principles in Protein Structure and Stability. *Ann. Rev. Biophys. and Biophys. Chem.* **20**, 447-449.

Chothia, C. (1976). The Nature of Accessible and Buried Surfaces in Proteins. *J. Mol. Biol.* **105**, 1-14.

Chou, P.Y., Wells, M. & Fasman, G.D. (1972). Conformational Studies on Copolymers of Hydroxylpropyl-L-glutamine and L-leucine. Circular Dichroism Studies. *Biochemistry* **11**, 3028-3043.

Chou, P.Y. & Fasman, G.D. (1974). Prediction of Protein Conformation. *Biochemistry* **13**, 222-245.

Dill, K.A. (1990). Dominant Forces in Protein Folding. *Biochemistry* **29**, 7133-7155.

Dufour, E. & Haertlé, T. (1990). Alcohol-Induced Changes of β-lactoglobulin-- Retinol Binding Stoichiometry. *Prot. Eng.* **4**, 185-190.

Dyson, H.J. & Wright, P.E. (1991). Defining Solution Conformations of Small Linear Peptides. *Ann. Rev. Biophys. and Biophys. Chem.* **20**, 519-538.

Dyson, H.J., Sayre, J.R., Merutka, G., Shin, H.-C., Lerner, R.A. & Wright, P.E. (1992). Folding of Peptide Fragments Comprising the Complete Sequence of Proteins: Models for Initiation of Protein Folding II. Plastocyanin. *J. Mol. Biol.* **226**, 818-835.

Egan, D.A., Logan, T.M., Liang, H., Matayoshi, E., Fesik, S.W. & Holzman, T.F. (1993). Equilibrium Denaturation of Recombinant Human FK Binding Protein in Urea. *Biochemistry* **32**, 1920-1927.

Fan, P., Bracken, C. & Baum, J. (1993). Structural Characterization of Monellin in the Alcohol-Denatured State by NMR: Evidence for β-Sheet to α-Helix Conversion. *Biochemistry* **32**, 1573-1582.

Fauchère, J.-L. & Pliska, V. (1983). Hydrophobic Parameters Π of Amino Acid Side Chains from the Partitioning of N-acetyl-amino Acid Amides. *Eur. J. Med. Chem.- Ther. Chem.* **18**, 369-375.

Gregoret, L.M. & Cohen, F.E. (1991). Protein Folding-- Effect of Packing Density on Chain Conformation. *J. Mol. Biol.* **219**, 109-122.

Harding, M.M., Williams, D.H. & Woolfson, D.N. (1991). Characterization of a Partially Denatured State of a Protein by Two-Dimensional NMR: Reduction of the Hydrophobic Interactions in Ubiquitin. *Biochemistry* **30**, 3120-3128.

Hao, M.H., Rackovsky, S., Liwo, A., Pincus, M.R. & Scheraga, H.A. (1992). Effects of Compact Volume and Chain Stiffness on the Conformations of Native Proteins. *Proc. Natl. Acad. Sci. USA* **89**, 6614-6618.

Jackson, M. & Mantsch, H.H. (1992). Halogenated Alcohols as Solvents for Proteins: FTIR Spectroscopic Results. *Biochim. Biophys. Acta* **1118**, 139-143.

Kabsch, W. & Sander, C. (1983). Dictionary of Protein Secondary Structure: Pattern Recognition of Hydrogen-Bonded and Geometrical Features. *Biopolymers* **22**, 2577-2637.

Killian, J.A. (1992). Gramicidin and Gramicidin-Lipid Interactions. *Biochim. Biophys. Acta* **1113**, 391-425.

Lau, K.F. & Dill, K.A. (1989). A Lattice Statistical Mechanics Model of the Conformational and Sequence Spaces of Proteins. *Macromolecules* **22**, 3986-3997.

Lau, K.F. & Dill, K.A. (1990). Theory for Protein Mutability and Biogenesis. *Proc. Natl. Acad. Sci. USA* **87**, 638-642.

Lehrman, S.R., Tuls, J.L., & Lund, M. (1990). Peptide α-Helicity in Aqueous TFE: Correlations with Predicted α-Helicity and the Secondary Structure of the Corresponding Regions of Bovine Growth Hormone. *Biochemistry* **29,** 5590-5596.

Lesk, A.M. (1991). *Protein Architecture* [*Practical Approach Series*]. IRL Press, New York.

Lipman, D.J. & Wilbur, W.J. (1991). Modelling Neutral and Selective Evolution of Protein Folding. *Proc. Roy. Soc. London, Series B* **245**, 7-11.

Llinás, M. & Klein, M.P. (1975). Charge Relay at the Peptide Bond: A Protein Magnetic Resonance Study of Solvation Effects on the Amide Electron Density Distribution. *J. Am. Chem. Soc.* **97**, 4731-4737.

Lyu, P.C., Liff, M.I., Marky. L.A., Kallenbach. N.R. (1990). Side Chain Contributions to the Stability of Alpha-Helical Structure in Peptides. *Science* **250**, 669-673.

Miller, R., Danko, C.A., Fasolka, M.J., Balazs, A.C., Chan, H.S. & Dill, K.A. (1992). Folding Kinetics of Proteins and Copolymers. *J. Chem. Phys.* **96**, 768-790.

Mutter, M. & Hersperger, R. (1990). Peptides as Conformational Switch: Medium-Induced Conformational Transitions of Designed Peptides. *Angewandte Chemie* **29**, 185-187.

Nandi, P.K. & Robinson, D.R. (1984). Effects of Urea and Guanidine Hydrochloride on Peptide and Nonpolar Groups. *Biochemistry* **23**, 6661-6668.

Nelson, J.W. & Kallenbach, N.R. (1986). Stabilization of the Ribonuclease S-Peptide α-Helix by Trifluoroethanol. *Proteins* **1**, 211-217.

Nozaki, Y. & Tanford , C. (1971). The Solubility of Amino Acids and Two Glycine Polypeptides in Aqueous Ethanol and Dioxane Solutions. *J. Biol. Chem.* **246**, 2211-2217.

O'Neil, K.T. & DeGrado, W.F. (1990). A Thermodynamic Scale for the Helix-Forming Tendencies of the Commonly Occurring Amino Acids. *Science* **250**, 646-651.

O'Toole, E.M. & Panagiotopoulos, A.Z. (1993). Effect of Sequence and Intermolecular Interactions on the Number and Nature of Low-Energy States for Simple Model Proteins. *J. Chem. Phys.* **90**, 3185-3190.

Rees, D.C., DeAntonio, L. & Eisenberg, D. (1989). Hydrophobic Organization of Membrane Proteins. *Science* **245**, 510-513.

Richards, F.M. & Richmond, T. (1978). Solvents, Interfaces and Protein Structure. In *Molecular Interactions and Activity in Proteins* (G.E. Wolstenholme, Ed.) pp. 23-45, Ciba Foundation Symposium 60, Excerpta Medica, Amsterdam.

Robinson, D.R. & Jencks, W.P. (1965). The Effect of Compounds of the Urea-Guanidinium Class on the Activity Coefficient of the Acetyltetraglycine Ethyl Ester and Related Compounds. *J. Am. Chem. Soc.* **87**, 2462-2470.

Rose, G.D., Geselowitz, A.R., Lesser, G.J., Lee, R.H. & Zehfus, M.H. (1985). Hydrophobicity of Amino Acid Residues in Globular Proteins. *Science* **229**, 834-838.

Scholtz, J.M. & Baldwin, R.L. (1992). The Mechanism of Alpha-Helix Formation By Peptides. *Ann. Rev. Biophys. and Biomol. Struct.* **21**, 95-118.

Segawa, S.-I., Fukono, T., Fujiwara, K. & Noda, Y. (1991). Local Structures in Unfolded Lysozyme and Correlation with Secondary Structures in the Native Conformation: Helix-Forming and -Breaking Propensity of Peptide Segments. *Biopolymers* **31**, 497-509.

Shoemaker, K.R., Fairman, R., York, E.J., Stewart, J.M. & Baldwin, R.L. (1988). Circular Dichroism Measurement of Peptide Helix Unfolding. In *Peptides: Proc. of the Tenth Amer. Peptide Symposium* (Marshall, G.R., Ed.) pp. 15-20, ESCOM, Leiden.

Shortle, D., Chan, H.S. & Dill, K.A. (1992). Modeling the Effects of Mutations on the Denatured States of Proteins. *Protein Science* **1**, 201-215.

Sönnichsen, F.D., van Eyk, J.E., Hodges, R.S. & Sykes, B.D. (1992). Effect of TFE on Protein Secondary Structure: An NMR and CD Study Using a Synthetic Actin Peptide. *Biochemistry* **31**, 8790-8798.

Stickle, D.F., Presta, L.G., Dill, K.A. & Rose, G.D. (1992). Hydrogen Bonding in Globular Proteins. *J. Mol. Biol.* **226**, 1143-1159.

Stone, A.L., Park, J.Y. & Martenson, R.E. (1985). Low-Ultraviolet CD Spectra of Oligopeptides 1-95 and 96-168 Derived from Myelin Basic Protein of Rabbit. *Biochemistry* **24**, 6666-6673.

Tamburro, A.M., Scatturin, A., Rocchi, R., Marchiori, F., Borin, G. & Scoffone, E. (1968). Conformational Transitions of Bovine Pancreatic Ribonuclease S-Peptide. *FEBS Lett.* **1**, 298-300.

Tanford, C., De, P.K., & Taggart, V.G. (1960). The Role of the α-Helix in the Structure of Proteins: Optical Rotatory Dispersion of β-Lactoglobulin. *J. Am. Chem. Soc.* **82**, 6028-6034.

Tanford, C. (1968). Protein Denaturation. *Adv. Protein Chem.* **23**, 121-282.

Unger, R. & Moult, J. (1993). Genetic Algorithms for Protein Folding Simulations. *J. Mol. Biol.* **231**, 75-81.

von Hippel, P.H. & Wong, K.-Y. (1965). On the Conformational Stability of Globular Proteins: The Effects of Various Electrolytes and Non-Electrolytes on the Thermal Ribonuclease Transition. *J. Biol. Chem.* **240**, 3909-3923.

Wilkinson, K.D. & Mayer, A.N. (1986). Alcohol-Induced Conformational Changes of Ubiquitin. *Arch. Biochem. Biophys.* **250**, 390-399.

Yee, D. & Dill, K.A. (1993). Families and the Structural Relatedness among Globular Proteins. *Protein Science* **2**, 884-889.

Yee, D.P., Chan, H.S., Havel, T.F. & Dill, K.A. (1994). Does compactness induce secondary structure in proteins? A study of poly-alanine chains computed by distance geometry. *J. Mol. Biol.* **241**, 557-573.

Zhong, L. & Johnson, W.C., Jr. (1992). Environment affects amino acid preference for secondary structure. *Proc. Natl. Acad. Sci.* **89**, 4462-4465.

Zimm, B.H. & Bragg, J.K. (1959). Theory of the Phase Transition Between Helix and Random Coil in Polypeptide Chains. *J. Chem. Phys.* **31**, 526-535.

Chapter 2

Statistical Potentials Extracted from Protein Structures: How Accurate Are They?

Paul D. Thomas and Ken A. Dill

The coauthor directed and supervised this work.

Abstract

"Statistical potentials" are energies widely used in computer algorithms to fold, dock, or recognize protein structures. They are derived from: (1) observed pairing frequencies of the 20 amino acids in databases of known protein structures, and (2) approximations and assumptions about the physical process that these quantities measure. Using exact lattice models, we construct a rigorous test of those assumptions and approximations. We find that statistical potentials often correctly rank-order the relative strengths of interresidue interactions, but they do not reflect the true underlying energies because of systematic errors arising from the neglect of excluded volume in proteins. We find that complex residue-residue distance dependencies observed in statistical potentials, even those among charged groups, can be largely explained as an indirect consequence of the burial of nonpolar groups. Our results suggest that current statistical potentials may have limited value in protein folding algorithms and wherever they are used to provide energy-like quantities.

Keywords

protein folding; knowledge-based potential; Boltzmann ensemble; residue partitioning; protein structure recognition.

Introduction

Our purpose here is to evaluate "statistical potentials" and the premises that underlie them. Statistical potentials are widely used as empirical energy functions to judge the quality of proposed protein structure models (Lüthy *et al.*, 1992; Wilmanns & Eisenberg, 1993), to identify the native fold or correct folding motif of an amino acid sequence among many incorrect alternatives (Hendlich *et al.*, 1990; Jones *et al.*, 1992; Bryant & Lawrence, 1993), to identify possible folds for a sequence of unknown structure (Bowie *et al.*, 1991; Sippl & Weitckus, 1992), to predict docking of protein structures (Pellegrini & Doniach, 1993), to find amino acid sequences compatible with a desired structure (Godzik *et al.*, 1992), and to simulate protein folding (Wilson & Doniach, 1989; Skolnick & Kolinski, 1990; Sun, 1993; Kolinski & Skolnick, 1994).

Statistical potentials are putative energies that are derived from amino acid pairing frequencies observed in known protein structures. The idea was first proposed by Tanaka and Scheraga (1976). Miyazawa and Jernigan (1985) took a major step forward in including terms to explicitly consider solvent effects. Sippl (1990) and others (Hendlich *et al.*, 1990; Jones *et al.*, 1993) extended these methods to include dependence on pairwise separation of residues in space and along the sequence. Bryant and Lawrence (1993) developed a log-linear statistical model to analyze protein structures separately, rather than using simple sums over distributions of residues in all proteins. More recently, statistical potentials have been refined by adding other statistical terms involving residue triplets (Godzik & Skolnick, 1992), dihedral angles (Nishikawa & Matsuo, 1993; Kocher *et al.*, 1994), solvent accessibility and hydrogen-bonding (Nishikawa & Matsuo, 1993).

The basic idea behind statistical potentials is simple. We illustrate the idea using an idealized example. Suppose large numbers of the 20 amino acids were somehow to distribute themselves in a gas phase at temperature T. If the interactions are purely pairwise, the distributions can be described by the equilibrium pairwise density $\rho_{ij}(r)$

between any two amino acid types i, j = 1, 2, ..., 20 at distance r. In this case, the interaction free energy, $w_{ij}(r)$, can be calculated from the observed densities by the Boltzmann relation:

$$w_{ij}(r) = -kT \ln\left(\frac{\rho_{ij}(r)}{\rho^*}\right) \qquad (1)$$

where k = Boltzmann's constant and ρ^* is the reference state pair density at infinite separation where the particle interaction is zero. This example shows how to infer pairwise energies from the average spatial distributions of amino acids in this idealized gas phase example.

But protein crystal structures (and NMR structures) are not gas phases of amino acids in dynamic equilibrium. Certain assumptions and approximations are usually made to obtain energy-like quantities from protein structures. First, amino acid pair density functions $\rho_{ij}(r)$ are constructed by summing the static densities observed in different proteins from the Brookhaven Protein Data Bank (PDB, Bernstein *et al.*, 1977) rather than averaging different states of the same protein. Second, it is necessary to choose a reference state (the pair density corresponding to zero-energy). Miyazawa and Jernigan (1985) introduced the use of the "random-mixing approximation," which assumes that in the absence of interactions, the amino acids and solvent molecules would be uniformly distributed throughout the available volume. In a random mixture the number of contacts between different monomers depends only on the relative concentrations of those monomer types. For example, because alanines are more common in proteins than methionines, a random mixture will have more Ala-Ala contacts than Met-Met. Finally, using the Boltzmann equation supposes an equilibrium between the observed pairing state and the reference state. Each amino acid pair is assumed to be independent of all the other pairs in the molecule. For weakly interacting particles in the gas phase, this is a reasonable approximation. However, one of the most remarkable features about proteins is the

extremely close packing of residues (Richards, 1977). In addition, amino acids are covalently linked in specific sequences. These are the premises we test below. We do not test the assumption that interactions are pairwise additive, nor do we treat local interactions (e.g. dihedral angle potentials).

There are two main approaches to calculating amino acid pair potentials. In one approach, the interactions between amino acids are assumed to be short-ranged, and are approximated using a "contact potential" (Tanaka and Scheraga, 1976; Miyazawa and Jernigan, 1985). In the Miyazawa and Jernigan (1985) formulation, a contact potential w_{ij} is an average of amino acid pairings over distances shorter than some cutoff distance r_c:

$$w_{ij} = -kT \ln \left(\frac{\int_0^{r_c} \rho_{ij}(r) dr}{\int_0^{r_c} \rho_{ij}^*(r) dr} \right) \qquad (2)$$

Miyazawa and Jernigan (1985) recognized that protein folding involves desolvating two monomers i and j before forming a contact between them. To account for this approximately, Miyazawa and Jernigan invented a hypothetical two-step process of contact formation (Figure 1). In the first step, monomers i and j, which are regarded as being solvated in the denatured state, go through a self-pairing (i with i, j with j, solvent with solvent). In the second step, the i-i and j-j pair "bonds" are broken and i-j bonds are made. These steps involve applying equation (2) four times— for breaking two contacts and forming two contacts.

In the Miyazawa and Jernigan approach, each of the two steps, desolvation and mixing, is based on a different random mixture reference state. For desolvation, the reference state is a uniform mixture of solvent molecules and amino acid residues. For mixing, which involves moving amino acids within a compact globule, the reference state weights residue positions in terms of their degree of burial.

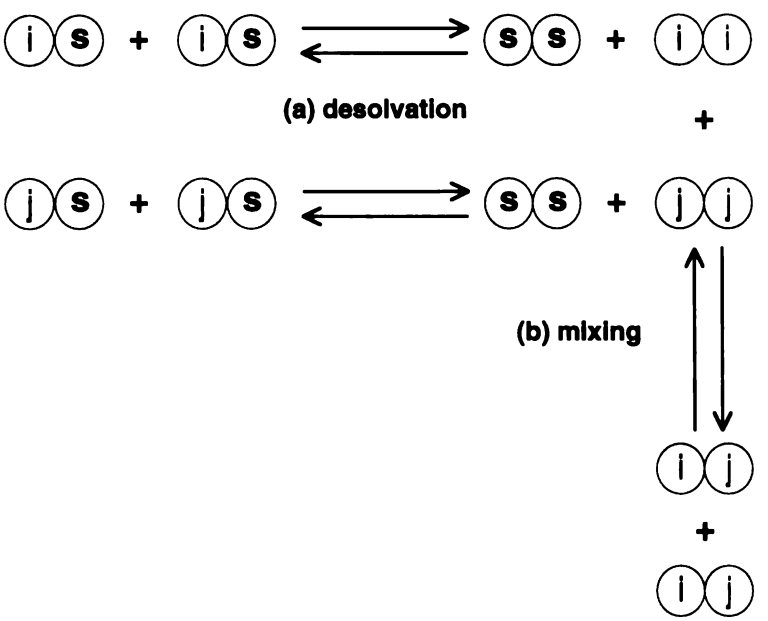


Figure 1. Hypothetical process for extracting contact energies between residues of type *i* and *j* from contact distributions in proteins. (a) "Desolvation" of two *i*-solvent contacts to form *i-i* and solvent-solvent contacts: extracted contact energy = $w_{ii} + w_{00} - 2w_{i0}$, where 0 denotes solvent and w_{xy} are defined by equation (2) using a random mixture of solvent and residues as the reference state. (b) "Mixing" of *i-i* and *j-j* contacts to form two *i-j* contacts: extracted contact energy = $2w_{ij} - w_{ii} - w_{ij}$, where w_{xy} are defined by equation (2) using a random mixture of residues, weighted according to average degree of burial in protein structures, as the reference state.

The second class of statistical pair potentials allows for distance-dependence of the interactions. For such potentials, equation (1) has been applied to individual small distance intervals. In this case, another normalization is also needed. Sippl (1990) solved the problem of how to calculate the expected "uniform density" reference state for distance-dependent potentials. The reference density of pair distances at each distance interval depends not only on the frequencies of the residue pairs in question, but also on the total number of pair distances observed at that distance. For example, more pairs of amino acids are separated by 10Å than by 80Å, because of the small sizes of proteins. Because

dividing up frequencies both by pair type and distance interval results in many parameters with few proteins to define them, Sippl (1990) developed a "sparse data correction" which corrects for the energies calculated using equation (1) by an uncertainty factor.

Testing the premises of statistical potentials

Statistical potentials are arguably intended to mimic the natural energies that drive amino acids to form contacts. How well do statistical potentials, "extracted" from native structure databases, reflect the "true" underlying energies? It is not known, because there is no independent knowledge of nature's true underlying energies. Here we devise a test that circumvents that problem. We generate different "model PDBs" using an exact lattice model for which the underlying energies can be specified exactly. We then observe the monomer pairing frequencies in each model PDB, and compare the extracted energies to the true energies. We define the term "true energy" to mean the actual contact free energy that causes the protein to fold into its given native state, and the term "extracted energy" to mean the energy-like quantity that is obtained from observing the monomer pairing frequencies in the database of native structures and using the assumptions described above. It is not important that the lattice model is not a perfect mimic of real proteins. We are simply performing a consistency check of the methods that generate statistical potentials. We aim to learn how much error is introduced by their neglect of chain connectivity, amino acid sequence and excluded volume.

The "AB-model" consists of chains of two monomer types, A and B, having lengths L=11 to 18 on 2 dimensional square lattices. We specify a true contact potential, involving 3 interaction energies (for AA, AB, and BB contacts) which defines the total energy for any conformation of any AB sequence on the 2D lattice. Monomers are in contact if they are non-bonded nearest neighbors on the lattice. Since there are only 3

energies, we are able to explore all possible sets of unique contact potentials, and all the unique native structures for each given energy function.

Our consistency check runs as follows. (1) We select a set of true contact free energies, which we denote with capital letters E_{AA}, E_{AB} and E_{BB}. (2) For each chain length, we perform an exhaustive search of conformational space, and find the native states (lowest true energy) of all 2^L sequences. (3) We make a "database" of the unique native structures. (4) We use two representative statistical potential extraction methods, one contact-based and one distance-dependent, to extract statistical energies from this lattice model database. We denote the extracted energies with lowercase e_{AA}, e_{AB} and e_{BB}. The contact potential is extracted by the method of Miyazawa and Jernigan (1985) and the distance-dependent potential by a simplified version of the method of Sippl (1990) that considers all residues in the short chains to be of the same "topological level" (i.e. there is no dependence on sequence separation). We selected these two methods as representative of the methodology of statistical potentials because they are the most widely referenced in the current literature.

Our approach involves no sampling problems since we exhaustively enumerate the complete database of all sequences having a unique fold. The data are not sparse, as they are for real proteins. We have approximately 1200 to 60000 contacts to define 3 parameters (400 to 20000 observations per parameter). For comparison, Miyazawa and Jernigan used about 27000 contacts to define 210 parameters, or about 130 observations per parameter. For the distance-dependent potentials, we include the sparse-data correction of Sippl (1990), though for most of the lattice model databases we consider there are enough pair distances to approach the uncorrected values. We do not address questions of database size or sampling problems. Rather, our purposes here are (1) to investigate the principles of statistical potentials, and (2) to assess how accurately current statistical potentials might reflect the real amino acid contact energies in proteins.

Results

First we explore the simplest version of the AB model, where the true potential involves only a single energy, namely the HP (Hydrophobic-Polar) model: E_{HH}:E_{HP}:E_{PP} = -1:0:0. In this model, contacts between H monomers are favorable relative to solvent contacts, while HP and PP contacts are energetically equivalent to solvent contacts. The folding properties of this model are known in some detail (Lau & Dill, 1989; Lau & Dill, 1990; Chan & Dill, 1991a; Lipman & Wilbur, 1991; Shortle *et al.*, 1992; Miller *et al.*, 1992; Unger & Moult, 1993; O'Toole & Panagiotopoulos, 1993; Camacho & Thirumalai, 1993a; Camacho & Thirumalai, 1993b; Chan & Dill, 1994; Chan *et al.*, 1995; reviewed in Dill *et al.*, 1995). Figure 2 shows that the Miyazawa and Jernigan procedure correctly determines that the HH contact interaction is dominant and attractive. However, the extracted energies are not equal to the true energies. Two of the main errors introduced by the extraction process are: (1) the extracted energies e_{HP} and e_{PP} are found to be non-zero, whereas the true energies E_{HP} and E_{PP} are zero, and (2) all the extracted interactions depend on chain length, whereas the true energies do not. These errors arise from the approximations made in the extraction procedures for statistical potentials, as described below.

The problem: interactions are not independent

For the HP model, the extraction process infers that the HP interaction is more favorable than the PP interaction even though the true energies are zero for both HP and PP. The problem is not that HP contacts are more common than PP in the database; the problem is the assumption that the pairwise interactions are independent. The HH

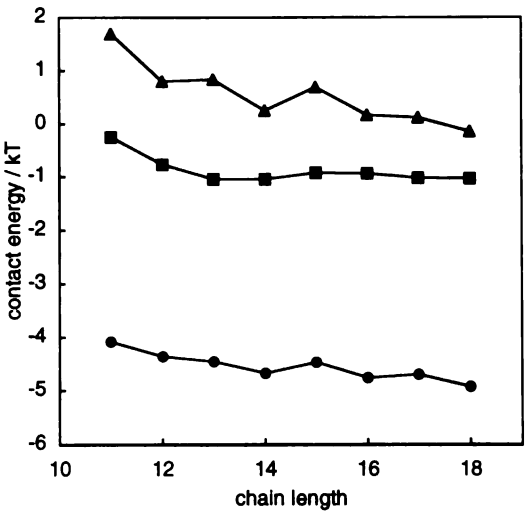


Figure 2. Extracted statistical potentials for the HP model, vs. chain length: e_{HH} (circles, true potential was -ε, where ε>0), e_{HP} (squares, compare with the "true" potential of 0), e_{PP} (triangles, compare with the "true" potential of 0).

interaction, which dominates in this model, indirectly affects HP and PP pairing frequencies.

In the HP lattice model, favorable HH contacts are made preferentially, which has two effects on the observed frequencies of other contacts: (1) because each structure makes only a limited number of interresidue contacts, HH contacts deplete the total number of contacts available for any other interresidue contact (HP and PP contacts), and (2) HH contacts deplete the supply of H contact surfaces that are available for any other contacts with H monomers (HP and H-solvent contacts). But the random mixture reference state supposes that all contact types are uniformly available. Thus in the model database, PP, HP and H-solvent pairs are underrepresented relative to the reference state, so *forming* these contacts appears as an unfavorable contribution to the extracted contact energies (*breaking* them will appear favorable). The extracted HP contact energy therefore includes two large terms of opposite sign, an unfavorable HP contact forming term and a favorable

H-solvent contact breaking term (Figure 1). The extracted PP contact energy, on the other hand, is dominated only by the unfavorable term for forming a PP contact; P residues are about equally solvated in both the native and reference states. The net result is that, as a result of the true HH interaction, the extracted HP contact energy appears more favorable than the extracted PP interaction. The HP and PP interactions are "coupled" to the HH interaction.

The coupling of interactions among different types of residue pairs is also illustrated by considering distance-dependent potentials (Figure 3). While the true potential in the HP model is just a first-neighbor HH contact interaction (i.e. a favorable "spike" at a distance of one lattice unit), the extracted potentials erroneously give a distance dependence. The extracted interactions are favorable over some distance ranges and unfavorable over others. The reason for this incorrect and complex distance dependence is because of the assumed uniform distributions in the extraction procedure reference states, as described below.

The incorrect extracted potentials come from two types of coupling. First, in both the distance dependent and contact potentials, the extracted energy of a monomer pair at a given distance is influenced by *other pairs* at the *same distance*. For the HP lattice model, the high density of HH pairs at short distances causes a correspondingly low density (relative to uniform) of HP and PP pairs at those distances. As a result, the extracted HP and PP potentials are erroneously unfavorable at short range. Second, for distance dependent extracted potentials, the energy of a monomer pair at a given distance of spatial separation is influenced by the *same pair* at *different distances*. For instance, there is a high density of HH pairs at short distances, due to the true HH attraction. The *total* density of distances between HH pairs is the same in both the database and the uniform distribution reference state, so the higher (than uniform) concentration of HH pairs at short distances causes a compensating depletion (relative to uniform) at longer distances. The concentrations at different distances are treated as independent by equation (1), but independence is a poor approximation.

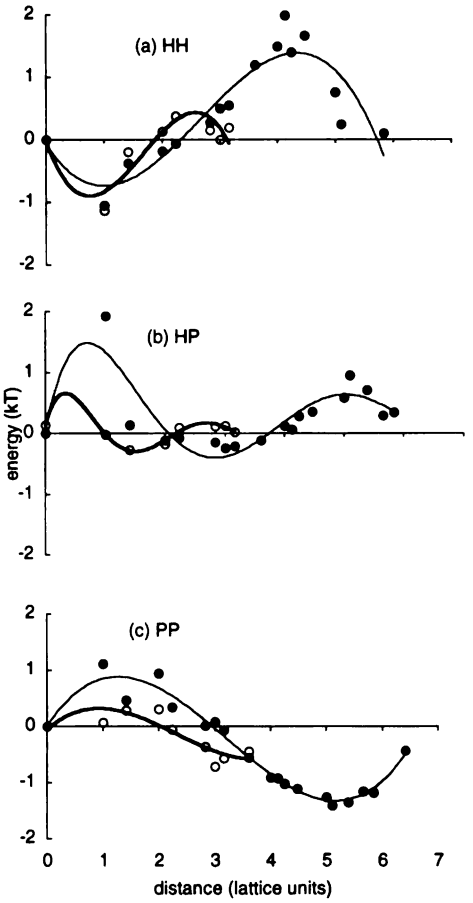


Figure 3. Distance-dependent statistical potentials extracted from 2D HP structure databases using the method of Sippl (1990), including sparse data correction (σ = 1/50). Lines are polynomial fits to the data to guide the eye. The extracted potentials for chain lengths of 11 (open circles) and 18 (filled circles) are shown. (a) extracted HH interaction (compare with the "true" potential— a favorable "spike" at a distance of 1 lattice unit, and 0 for other distances). (b) extracted HP interaction (the "true" potential is 0 for all distances). (c) extracted PP interaction (the "true" potential is 0 for all distances).

In summary, the complex extracted distance dependence of HP and PP interactions follows from the HH interactions. H monomers cluster, giving many short HH distances and few long HH distances. The HH attraction drives P monomers to the protein surface, resulting in many long PP distances. The HH attraction similarly causes many intermediate

HP distances, between interior H's and exterior P's. Thus the HP and PP pairing frequencies are not independent of the HH interaction.

This coupling of interactions is not an artifact of our model. Comparing Figures 4 and 5 shows the same coupling in potentials extracted from the Protein Data Bank. We classified the residues in proteins into just two types, interior or exterior, and then used the method of Sippl (1990) to extract a potential between these two geometrically defined "residue types." This test shows that the interior-interior pairing "energy" (Figure 4a) is nearly identical to that calculated by Hendlich *et al.* (1990) for the pair Ile-Val (Figure 5a), two hydrophobic residues. Apparently the 30 different extracted energy parameters for the different distance intervals are mainly just reflecting that isoleucines and valines tend to be in protein interiors. Moreover, Figure 3 shows that a simple nearest-neighbor HH attraction in the lattice model is sufficient to give the same functional form as that for Ile-Val pairs in proteins. Hence the apparently complex distance dependence among amino acids in proteins may reflect little more than that hydrophobic residues attract each other.

More strikingly, Figure 3 shows that the HH contact interaction is also sufficient to give an extracted PP potential (Figure 3c) that has roughly the same functional form as for both Lys-Asp (unlike charges) and Asp-Asp (like charges) (Figures 5b and c) as extracted from the PDB by Hendlich *et al.* (1990). That is, the extracted charged residue interactions in proteins are not mainly due to electrostatics; it appears to be mainly because charged residues are driven to the protein surface by the nonpolar attractions of other amino acids. This explains the observation that more Coulomb-like charge-charge interactions are extracted from the PDB when statistics are compiled only on surface residues (Kocher *et al.*, 1994). Considering only surface residues is equivalent to using a different reference state for the potentials that takes into account the average effect of the hydrophobic residues on the observed charge-charge distributions.

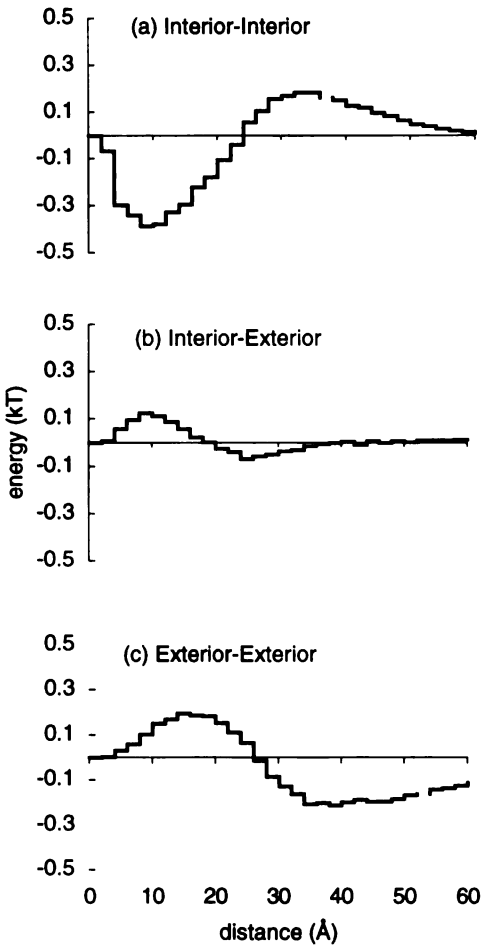


Figure 4. Distance-dependent potentials extracted from real protein structures using only 2 monomer types, interior or exterior: (a) Interior-Interior (compare with Figure 5a), (b) Interior-Exterior, (c) Exterior-Exterior (compare with Figures 5b and c). Potentials are extracted for the same sequence separations in the same set of proteins, and using the same equations (including sparse data correction, scaling for 3 parameters rather than 210) as in Figure 5. Interior/exterior positions are determined using the program ACCESS (Lee & Richards, 1971). The full backbone and sidechain center of mass is used for the calculation, with a probe radius of 2.0Å to compensate for the single-atom sidechain representation. If the accessible surface area of a carbon atom positioned at the sidechain center of mass (C^{α} for glycine) exceeds 30Å^2, the residue is considered to be exposed.

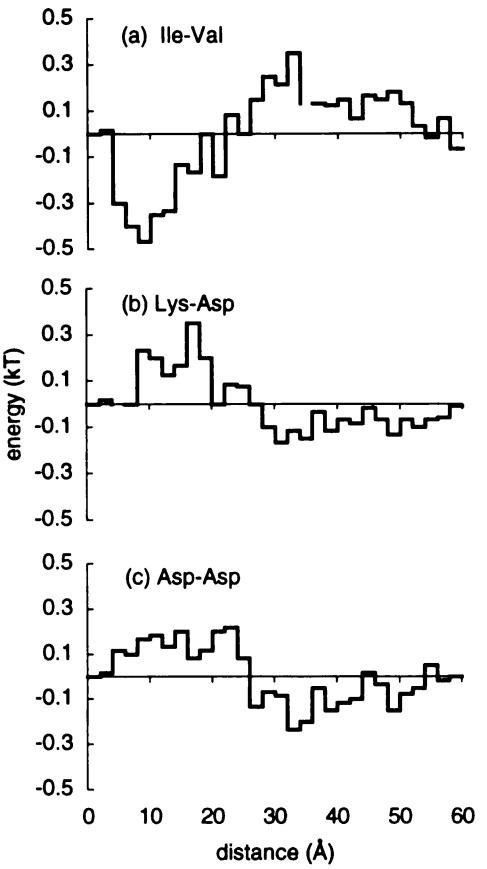


Figure 5. Distance-dependent potentials extracted by Hendlich *et al.* (1990) for residue pairs separated in primary sequence by 61 to 100: (a) valine-isoleucine, (b) lysine-aspartate, (c) aspartate-aspartate. Note the similarity between (b) and (c).

Some statistical potentials have tens of thousands of parameters for pairing frequencies as a function of distance and sequence separation. But most of those parameters may be redundant, all reflecting mainly that nonpolar amino acids form a core surrounded by polar residues. To test this possible redundancy, we performed the same threading test as in Hendlich *et al.* (1990), but using only a single "energy" parameter that accounts for contacts between hydrophobic residues. Table 1 compares our single-parameter results with the threading potential of Hendlich *et al.* (1990). Table 1 shows that

this single hydrophobic contact "energy" parameter correctly identifies the native conformation as having the lowest energy in the nearly as many cases (37/65) as the much more complex potential of Hendlich *et al.* (41/65). Most of the failures in the simple model are also failures in the complex model. We conclude that: (1) most of the information about protein energetics contained in complex statistical potentials is simply hydrophobic clustering propensity, as has been noted before (Casari & Sippl, 1992; Bryant & Lawrence, 1993), and (2) the threading test (with no insertions and deletions) is not a particularly challenging test of energy functions. Our results for the threading test are consistent with those of Bryant and Amzel (1987), who found that counting hydrophobic contacts can distinguish between correctly and grossly misfolded proteins.

Table 1. A count of hydrophobic contacts succeeds in the "threading test" nearly as often as a more complex potential for identifying native structures. Position of the correct native conformation in a list of threaded conformations sorted by the total distance-dependent statistical energy of Hendlich *et al.* (1990) (column 2), or just by counting nonpolar contacts (Ala, Cys, Ile, Leu, Met, Phe, Trp, Tyr, Val) (column 3). A contact is defined as a C^{β} - C^{β} distance of less than 8Å. N_{HH} is the number of nonpolar contacts in the native conformation, and ΔN_{HH} is the difference between the number of nonpolar contacts in the native conformation and in the "lowest energy" non-native threaded conformation.

Protein PDB code	Extracted Potential	Hydrophobic Contact Count		
	Position	Position	N_{HH}	ΔN_{HH}
4SBV A	1	1	227	15
3ADK	1	1	167	24
2STV	5	1	154	2
1HMG B	71	803	88	-34
1GCR	1	1	188	43
2ALP	1	1	202	30
3WGA A	1	1	147	29
2SGA	6	1	181	45
2LZM	1	1	183	38
4DFR A	1	1	178	14
1LH4	1	1	189	10
1MBD	1	1	161	20

Protein PDB code	Extracted Potential	Hydrophobic Contact Count		
	Position	Position	N_{HH}	ΔN_{HH}
2SOD O	1	1	136	38
2LHB	1	1	183	17
2HHB B	1	1	172	26
2PKA B	5	3	116	-9
2HHB A	1	1	155	4
3FXN	1	1	173	46
1LZ1	1	1	174	42
2AZA A	1	1	120	8
2CCY A	2	2	140	-1
1RN3	1	1	99	8
1BP2	1	1	115	2
1PP2 R	1	1	118	0
155C	1	1	99	13
1PAZ	1	1	152	29
2PAB A	1	1	119	0
1HMQ A	1	115	81	-16
2C2C	1	16	92	-11
1CPV	1	1	121	2
1ACX	2	1	116	7
1REI A	1	1	95	9
2CDV	1	1548	41	-24
2SSI	1	7	119	6
4CYT R	1	53	59	-18
1WRP R	1	160	77	-16
1PCY	1	1	101	10
1HVP A	5	1	89	3
3FXC	2	1	102	13
2GN5	26	108	65	-13
1HIP	11	24	83	-9
2B5C	1	1	62	7
1CC5	235	7	86	-8
351C	1	11	73	-5
2PKA A	1	42	62	-10
3ICB	1	1	66	19
2ABX A	65	12	56	-9
1HOE	125	4	69	-3
1CTF	71	1	101	17
1SN3	1	1	70	8
1CSE I	1	3	60	-4
2EBX	1	2	32	-1
2MT2	1	1	71	10
4PTI	1	9	52	-5
1OVO A	1	35	39	-11
1FDX	64	1	71	1
5RXN	2571	1	44	0
1CRN	24	8	45	-6
1BDS	8	4	42	-2
2RHV 4	74	9690	9	-28
1PPT	10	360	17	-10
1INS B	301	186	25	-13
1GCN	4563	666	11	-10
1MLT A	107	340	20	-11
1INS A	483	686	14	-14

Interior-exterior partitioning: Effects of protein size and composition

Whereas the true potentials between amino acids cannot depend on the chain length, the extracted potentials do (Figures 2 and 3). For the distance-dependent potentials extracted from our 2D HP lattice model (Figure 3), the functional form is similar for chain lengths of 11 or 18, but the *scale* of the distance dependence is different. The scale is a geometric description of the average location of monomers relative to the core or surface. For example, H residues are overrepresented in the cores of our model structures, which are larger for the 18-mers than for the 11-mers; accordingly, the extracted HH energy changes from attractive to repulsive between a distance of $\sqrt{5}$ and $\sqrt{8}$ lattice units for the 18-mers, but only between $\sqrt{2}$ and 2 lattice units for the 11-mers. Similarly, the PP interaction reaches a minimum at 3 lattice units for the 11-mers, but about 5 lattice units for the 18-mers; this corresponds to the difference in average conformational diameters. Analogous geometric properties may account for the results reported by Hendlich *et al.* (1990) that a potential extracted from a database of smaller proteins performs slightly better at recognizing the native conformations of other small proteins than a potential extracted from a database of proteins of all sizes.

In the case of the extracted contact potentials, the chain length dependence (Figure 2) is due primarily to the desolvation terms (Figure 1), which take a form similar to that of a transfer from a solvated to a buried state. As Janin (1979) first noted, apparent interior-exterior "partition energies" (extracted from frequencies of amino acids in buried and exposed positions in proteins) will depend on the surface-to-volume ratio of the protein because the greater volume inside larger proteins more readily accommodates nonpolar monomers. This explains the systematic errors noted by Miyazawa and Jernigan (1985) in trying to predict surface-to-volume ratios using their extracted contact energies. Miyazawa and Jernigan assumed their extracted energies would not depend on the surface-to-volume ratios since the true energies do not depend on this quantity.

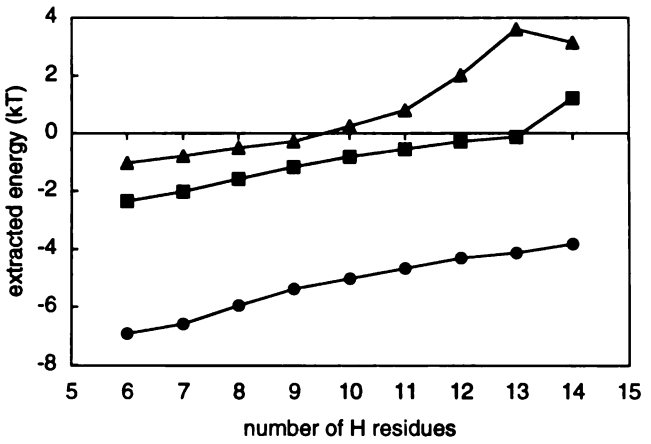


Figure 6. Extracted potentials depend on composition. The extracted HH (circles), HP (squares) and PP (triangles) energies for 2D lattice chains, L=18, vs. the number of H residues (n_H) in the sequences.

Extracted partition energies also depend on amino acid composition even though the true potentials cannot. For the lattice model, extracted energies become more positive with increasing content of H monomers (Figure 6). This is because there are more H monomers than needed to fill the core (2D lattice 18-mers can have at most 5 monomers completely sequestered from solvent), so additional H-monomers must be at least partially exposed to solvent. Because of coupling, the extracted HP and PP energies also increase with increasing nonpolar content; with more H residues in the sequence, the chances increase that a given conformation will make only HH contacts. For example, out of all 164 structures in the database of 18-mers with 14 H monomers, 162 make only HH contacts and no HP or PP contacts.

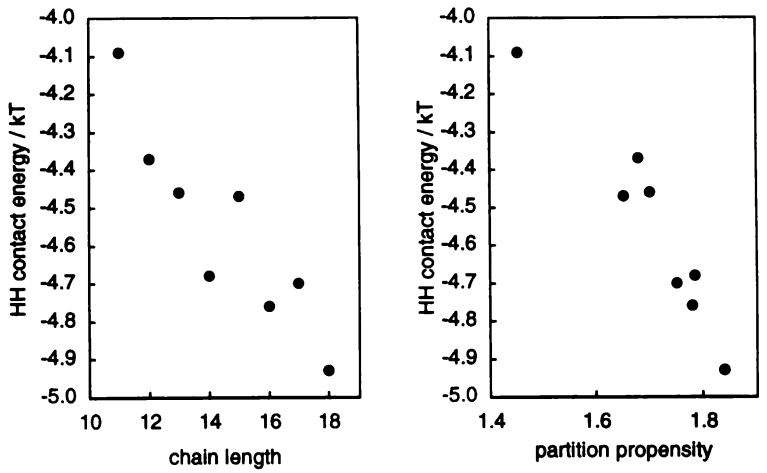


Figure 7. Extracted energies for the 2D HP model depend on (a) chain length, (b) average "partition propensity" of the structures in the database.

To account approximately for the effects of surface-to-volume ratio and composition on extracted energies, we define the "partition propensity" (π) of a given protein or set of proteins as:

$$\pi = \frac{2n_c}{q_H n_H} \tag{3}$$

where n_c is the total number of contacts in a given protein, q_H is the average coordination number of a hydrophobic (H) residue (taken from Miyazawa and Jernigan, 1985) and n_H is the number of H residues in the protein. Physically, $2n_c$ is the number of coordination sites involved in all residue-residue contacts (roughly proportional to the total buried surface of the molecule), and $q_H n_H$ is the total number of H coordination sites (roughly proportional to the total hydrophobic surface).

The partition propensity is therefore a crude measure of how effectively hydrophobic surface can be buried in a given structure. If a protein has a low partition propensity, it has more hydrophobic residues than are needed to fill a core, so some will not be able to partition effectively into a core. A high partition propensity means there is a

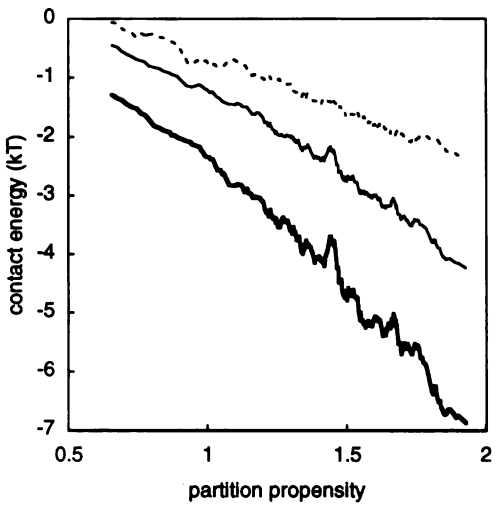


Figure 8. Extracted energies depend on partition propensity for proteins in the PDB. Hydrophobic residues (H) are Ala, Cys, Ile, Leu, Met, Phe, Tyr, Trp and Val; others are classed as P. The HH energy is shown by the bold line, HP by the solid line and PP by the dashed line.

large buried core, and few hydrophobic residues in the sequence to fill it. Figure 7b shows how the extracted HH contact energy depends on the partition propensity in the lattice model.

The same dependence of extracted energies on partition propensity is found for real proteins (Figure 8). We sorted a group of 346 non-homologous protein chains (Hobohm & Sander, 1994) from the PDB by the partition propensity of each protein, labeling each residue type as H (hydrophobic) or P (others). We made 326 sets of 20 proteins by sliding a window along the sorted list of proteins. That is, the *i*th set contains the *i*th through *i*+19th proteins in the sorted list. Figure 8 plots the extracted contact energies of each set against the average partition propensity of the set. The extracted energy between hydrophobic groups becomes significantly more favorable with increasing partition propensity, from about -1.3kT to -6.9kT, while the extracted PP energy increases from about 0kT to -2.4kT. The statistical potentials depend systematically on the size and amino acid composition of the proteins in the given set.

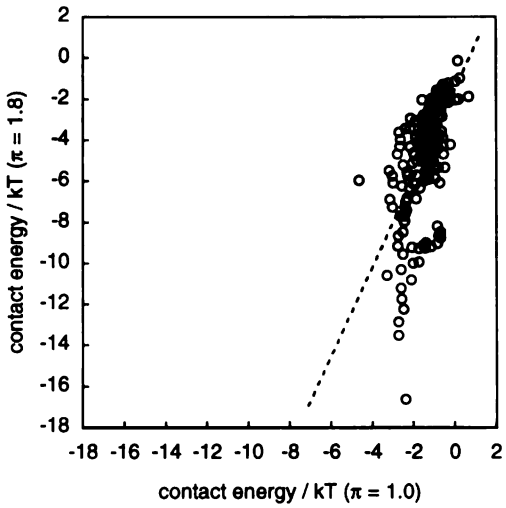


Figure 9. Contact energies extracted from two different sets of 69 proteins in the PDB. The energies calculated for the set of highest partition propensity (π=1.8) are plotted against the energies for the same pairs calculated for the set of lowest partition propensity (π=1.0). The slope of the linear best fit is 2.2, with an intercept of -1.6*kT*.

Consider two different 69-protein databases: one containing proteins with an average partition propensity $\pi = 1.8$ and the other having $\pi = 1.0$. Figure 9 plots the 210 contact energies extracted from one database against the same ones from the other database. The correlation between the two energy sets is only modest ($R = 0.62$). Even the rank ordering of contact pairings can be affected by protein size and composition. For example, for $\pi = 1.8$ the extracted Phe-Lys contact energy is *more* favorable than Ile-Val, whereas for $\pi = 1.0$ Phe-Lys is *less* favorable than Ile-Val. Coupling effects are different in these two protein sets, because of the different degree of burial of the hydrophobic residues. Phe is usually completely buried in the proteins having higher partition propensity, so a Phe-solvent contact is very unfavorable by equation (2), and any contact with Phe appears very favorable because it breaks a Phe-solvent contact. However for proteins having more

hydrophobic residues than their cores can accommodate, Phe is often sacrificed to the surface, and Phe-solvent contacts are less unfavorable. If the relative strengths of the extracted contact energies depend on average properties of the proteins in the database, such as protein size and composition, then how can we know which, if any, set of proteins gives extracted energies that are similar to nature's underlying energies? Clearly simply increasing the number of proteins in a database does not "average out" the coupling of the extracted energies. However, there may be ways to construct contact pair density functions or reference states that can take into account properties such as the partition propensity.

The Boltzmann distribution: Does it apply?

Among the deepest questions underlying statistical potentials is whether the Boltzmann distribution law, equations (1) or (2), is appropriate for converting pair frequencies in proteins to energies. What is the meaning of temperature in these expressions? The Boltzmann distribution law applies to a single closed system held at fixed temperature that can populate different energy levels. In the gas-phase amino acid pairing example above, increasing the temperature would cause Ala-Trp contacts, for example, to resemble a random distribution. But the PDB is *many* proteins, not a *single* system. The database is fixed— each protein has no degrees of freedom that are affected by temperature changes. The amino acid pairings in lysozyme, for example, are the same at $T = 300$K as at $T = 0$K. Consequently, the extracted potentials contain no information about protein stability— extracted energies will be the same whether the native conformations of the proteins are stable by 10^{-5} or 10^{5} kcal/mol. Should we use $T = 0$ K, $T = 300$ K, or some other temperature?

Nevertheless, some evidence supports the use of a Boltzmann distribution. In particular, for some protein "substructures" (e.g. proline residues), frequencies of different "states" (e.g. *cis*- and *trans*- peptide conformations) observed in the PDB correlate with the

frequencies expected from thermodynamic behavior. Observed frequencies in proteins are converted to "energy" differences using the Boltzmann relation, and then plotted against the energy differences (expected from either theory or experiment) between different states according to thermodynamic behavior at T=300K. If the correlation is linear, the slope can be adjusted by choosing the best-fit value for T in the Boltzmann relation (i.e. so the slope of the plot is 1). This yields an effective "temperature" for that substructure in proteins relative to the temperature of a true Boltzmann ensemble. This has been done for the following protein properties. Extracted partitioning energies correlate (Miyazawa & Jernigan, 1985; Rose *et al.*, 1985; Lawrence *et al.*, 1987; Miller *et al.*, 1987) with oil/water transfer experiments (e.g. Nozaki & Tanford, 1971; Fauchère & Pliska, 1983). Also well predicted are distributions of (ϕ, ψ) dihedral angles (Pohl, 1971; Kolaskar & Prashanth, 1979), charged residues (Bryant & Lawrence, 1991), *cis-* and *trans-* prolines (MacArthur & Thornton, 1991), the sizes of empty cavities (Rashin *et al.*, 1986) and residue stabilization of secondary structure elements (Serrano *et al.*, 1992) . Remarkably, for all of these different protein substructures, the apparent "temperature" is of the same order of magnitude, between about 150K and 600K.

Finkelstein *et al.* (1993) have proposed a theory to explain why substructures have about the same frequencies in proteins as they would have in thermodynamic equilibrium. They argue, based on a random heteropolymer model of proteins, that the number of random sequences having a native structure that contains any given substructure depends exponentially on the energy of that substructure. This model predicts that the "temperature" in the Boltzmann relation should be the same for all types of substructures, and roughly equal to room temperature.

We test here the Boltzmann distribution assumption for the interior-exterior partitioning of residues. We find that the apparent temperature for the extracted partition energies depends systematically on the average partition propensity (equation (3)) for the set of proteins. We divided the 346 PDB protein structures, sorted by partition propensity,

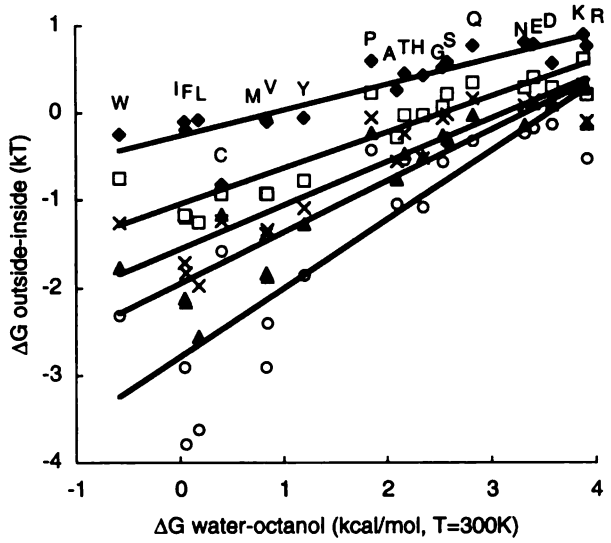


Figure 10. Extracted exterior-interior partition energies of the 20 amino acids, for protein sets having different propensities, vs. experimental water-octanol transfer energies (Fauchère & Pliska, 1983). The lines are regression fits to (from top to bottom) π=1.0 (filled diamonds, R=0.90, m=0.28), π=1.3 (open squares, R=0.92, m=0.42), π=1.5 (crosses, R=0.92, m=0.50), π=1.6 (filled triangles, R=0.92, m=0.58), and π=1.8 (open circles, R=0.88, m=0.79).

into 5 sets. For each set, we define the extracted exterior-interior partition energy for each amino acid i:

$$\Delta G_i = G_{inside} - G_{outside} = -kT \ln\left(\frac{n_{ir}}{n_{i0}}\right) \qquad (4)$$

where n_{ir} is the number of contacts between residue type i and other residues (these are "interior" sites), and n_{i0} is the number of contacts with solvent (residue "type" 0). Both n_{ir} and n_{i0} are estimated as in Miyazawa and Jernigan (1985). Physically, n_{ir} corresponds roughly to the average fractional surface area of residues of type i that is buried in protein interiors, while n_{i0} corresponds roughly to the average fractional exposed surface area. Figure 10 plots the partition energies extracted from each protein set against the experimental energies for transferring each amino acid from water to octanol.

We find that (1) extracted energies correlate with experimental partition energies, consistent with the use of the Boltzmann expression, but (2) there is no single temperature that is relevant. The first point is supported by the high correlation (R = 0.88 to 0.92) for all 5 protein sets. The relevant temperature is determined by the slopes of these plots. In Figure 10 the temperature relevant to the set π = 1.8 is T = 300K / [(0.59)(0.79)] = 640K, and to the set π = 1.0 is T = 300K / [(0.59)(0.28)] = 1800K, based on slope factors of 0.79 and 0.28 relative to oil/water partitioning at 300K. This result suggests that with respect to interior-exterior residue partitioning the proteins in the PDB may not be well-modeled by the random heteropolymer assumption of Finkelstein *et al.* (1993). Figure 10 shows that the effective temperature of interior-exterior partitioning depends on the length, composition and compactness of the proteins in the database, while the random heteropolymer model results are independent of protein length (due to cancellation of length-dependent terms) and composition (due to sequence averaging).

What is the physical meaning of the apparent "temperature" of a single protein structure or a database of structures? Here is an analogy, based on the buried/exposed partitioning of amino acids. Nonpolar sidechain surface is buried in its "ground state" and exposed in its "excited state." In a Boltzmann distribution, the amount of surface in the excited (exposed) state will increase with increasing temperature. But for a large protein having few enough hydrophobic monomers (high partition propensity) that it buries all its hydrophobic residues in the core (the ground state), the Boltzmann analogy gives an apparent temperature of 0 Kelvin with respect to hydrophobic residue partitioning. On the other hand, small proteins with many nonpolar residues will be "hotter" because those proteins are "forced," by surface-to-volume and composition constraints, to expose hydrophobic monomers. Hence proteins and databases can differ in their "temperatures" of interior-exterior partitioning.

Can statistical potentials correctly recognize native structures?

The results above indicate that statistical potentials may not quantitatively reflect the true energies that cause amino acid pairings in proteins. Here we ask if they succeed in a more modest test: do they correctly identify a native structure among a set of decoys? In this case, the value of the temperature is unimportant. The temperature just scales all interactions to the same degree, so while it affects the absolute stability, it does not affect the rank orderings of different structures. Therefore, for a statistical potential to succeed in correctly rank-ordering the energies of different structures, it only needs to be approximately correct to within an arbitrary scaling constant $C > 0$. In the AB lattice model we require only that: $e_{AA} \cong CE_{AA}$, $e_{AB} \cong CE_{AB}$ and $e_{AA} \cong CE_{AA}$.

To test the accuracy of the extracted energies in structure prediction, we now turn to the 3-energy AB model. For chains of length L=14 of two monomer types A and B, we create databases of minimum energy structures for different "true" contact potential functions, and extract statistical contact energies from each database. We then compare the extracted energies e_{AA}, e_{AB} and e_{BB} to the true energies E_{AA}, E_{AB} and E_{BB}. The most important test is whether, for all sequences in a given database, the true native conformation is identified as having the lowest value of the extracted energy over all possible conformations of that chain sequence. Unlike structure recognition tests that have been performed for real proteins, here we can exhaustively explore the conformational space for each AB sequence.

Table 2 shows the contact energies extracted from databases constructed using different true potentials. To facilitate comparison, both the true and extracted potentials are scaled relative to kT such that the strongest attractive contact interaction has an energy of -5 units. Table 2 shows that in all cases where the three true contact energies are different, the extracted contact energies have the same rank ordering as the true energies. For example, for the true potential E_{AA}:E_{AB}:E_{BB} = -5:-4:-1, the extracted potential e_{AA}:e_{AB}:e_{BB} = -5:-3:+0.8

correctly infers that the AA interaction is the most favorable and the BB interaction is least favorable (although it incorrectly infers a *repulsive* BB contact energy).

Table 2. AB model test. Left column is the scaled set of true energies used to generate the lattice model database for $L = 14$. Second column is the scaled set of energies found by the statistical potential extraction procedures. Column 3 is the number of unique folding sequences in the database for each true potential. Column 4 shows how often the extracted potential correctly identifies the native structure within the database (i.e. the structure having the lowest extracted energy is also the structure with the lowest true energy).

true	extracted	Number of	Prediction
E_{HH} : E_{HP} : E_{PP}	e_{HH} : e_{HP} : e_{PP}	sequences	success (%)
-5 : -5 : -1	-5 : -3.7 : +1.4	173	84
-5 : -4 : -1	-5 : -3.0 : +0.8	1388	74
-5 : -3 : -1	-5 : -2.4 : 0.0	1374	64
-5 : -2 : -1	-5 : -2.1 : -0.5	1726	99
-5 : -1 : -1	-5 : -1.5 : -1.0	913	97
-4 : -5 : -1	-1.1 : -5 : +1.8	1059	93
-3 : -5 : -1	-0.8 : -5 : +1.8	1046	95
-2 : -5 : -1	-0.5 : -5 : +1.4	1060	95
-1 : -5 : -1	+1.3 : -5 : +1.3	918	90
-5 : -1 : -5	-5 : +1.2 : -5	915	90
-5 : -1 : -4	-5 : -0.2 : -3.6	2264	99
-5 : -1 : -3	-5 : -0.6 : -2.5	1922	92
-5 : -1 : -2	-5 : -1.1 : -2.1	2040	100
-5 : -3 : +1	-5 : -2.6 : +2.5	1514	96
-5 : -3 : +2	-5 : -2.6 : +4.7	1467	99
-5 : -3 : +3	-5 : -2.6 : +6.8	1465	99

For the 3-interaction model, the extracted potentials qualitatively reflect the true potential. But this is not always sufficient for structure prediction since the extracted contact energies are usually different from the true energies, and sometimes even opposite in sign. As a result, the total extracted contact energy does not always correctly predict the native conformations of AB sequences. For only one of the model databases are the native structures correctly predicted for 100% of the sequences. Sometimes statistical potentials are calculated using a compact reference state rather than the unfolded state of Miyazawa and Jernigan. But this reference state is still a random mixture, and statistical energies calculated using the compact reference are essentially just shifted by a constant positive amount from the values in Table 2. For our lattice model, we find that such a shift generally decreases the success of the potential in identifying correct native conformations.

The extraction procedures fail to reproduce the correct rank ordering of interactions for degenerate potentials, in which two interactions are equal (Figure 2, Table 2). In these cases, the interactions that are equal in the true potential are not equal in the extracted potential (except when the true AA and BB interactions are equal, since the three-interaction potential is symmetric and sequence space is symmetric). This incorrect "splitting" of degenerate energies results from the coupling between different interactions, as we noted for the HP model. For the "true" potential E_{AA}: E_{AB}: E_{BB} = -5:-5:-1, the extracted energies are e_{AA}:e_{AB}:e_{BB} = -5:-2:+1. The AA interaction appears stronger than the AB interaction because few A residues are exposed to solvent (so any A contact is very favorable). Forming an AA contact breaks two A-solvent contacts while forming an AB contact breaks only one, so the extracted AA energy is more favorable. In addition, the sign of the extracted BB interaction is wrong because BB contacts are less common than in a random mixture since Bs prefer to form AB contacts.

For the 14-monomer AB model, the correct native conformations are predicted for 64% to 100% of the sequences in a given database. But Figure 11 shows that the success in structure prediction generally decreases with chain length, even over very short chain

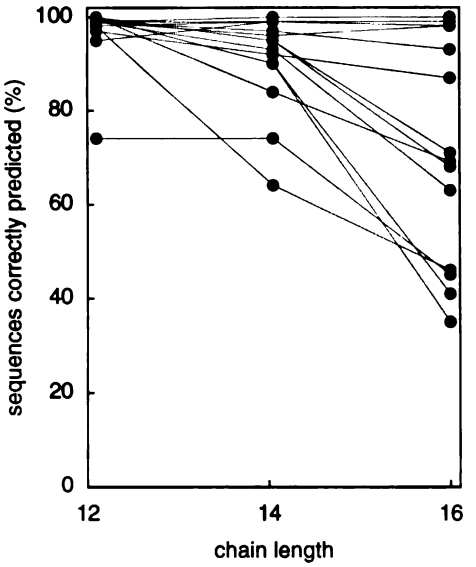


Figure 11. Percent correct structure prediction by the HP model statistical contact potential vs. chain length for 2D model. The structure of a sequence is predicted correctly if the true native conformation is also lower in extracted energy than any other conformation. Each set of connected points represents a different potential in Table 2. For the 12-mer and 14-mer chains, sequence space searching is exhaustive; for the 16-mer chains, random AB sequences are sampled until 1000 sequences having unique native structures are found for each true potential.

lengths from 12 to 16 monomers. Since real proteins are much longer than our lattice chains (having hundreds to thousands of interresidue contacts), Figure 11 suggests that extracted statistical energies may have limited success in identifying the native states of protein sequences among a set of reasonable alternative structures.

Conclusions

We test some of the principles that underlie statistical potentials. Statistical potentials are energy-like quantities that are extracted from protein structure databases,

based on certain assumptions. They have been used to model the true energies that cause proteins to fold, dock with ligands, and recognize other proteins. We test the premises behind statistical potentials using exact lattice models, and we verify our conclusions, where possible, with tests on the PDB.

We conclude that the principal weakness in all the current statistical potentials is their assumption that the frequencies of each type of amino acid pair, such as Ala-Leu, are independent of other types of pairs. In a relatively small, compact object such as a protein, the space taken by other amino acids is a strong constraint on the possible positions of each given pair. The clustering of hydrophobic amino acids is probably a stronger determinant of the statistical potentials among charged groups than electrostatics are. We find that, whereas true potentials cannot depend on chain length or composition, extracted potentials do. This too appears to be mainly a consequence of the burial of hydrophobic surface in small compact objects of differing sizes and compositions.

There are a few caveats in interpreting our lattice model results. First, excluded volume is a more stringent constraint in two dimensions than in three, simply because of the reduced number of possible spatial neighbors of each residue. Furthermore, sequence effects may be more pronounced in our short-chain model. As a control, we have constructed a database of low-energy configurations of long (60-monomer) HP chains on a 3D lattice, and found results that are similar to those from the 2D HP model. Second, there are only two monomer types, so that the dominant interaction (e.g. the HH interaction) will tend to be the primary determinant of all the observed pair distributions. The real energetics of protein folding are probably more complex.

For real protein structures, we demonstrate that the use of the Boltzmann distribution law to convert interior-exterior residue partitioning frequencies to energies, which defines a database "temperature" relative to octanol-water partition energies, is not firmly grounded. The choice of a relevant temperature is strongly dependent on the choice of proteins in the database. We define a quantity we call the "partition propensity" of a

given protein, which determines the relevant "temperature" in a Boltzmann equation. It may be possible to use this quantity to weight the burial frequencies observed in each protein to obtain database-independent partition energies.

We find that statistical potentials provide only a qualitative first approximation to the true energies that drive protein folding. The present study suggests ways to derive more quantitative potentials from the known protein structures (Thomas & Dill, in preparation).

Acknowledgments

PDT is a Howard Hughes Predoctoral Fellow. We thank Kai Yue for suggesting this approach.

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). The protein data bank: a computer-based archival file for macromolecular structures. *J. Mol. Biol.* **112**, 535-542.

Bowie, J.U., Lüthy, R. & Eisenberg, D. (1991). A method to identify protein sequences that fold into a known three-dimensional structure. *Science* **253**, 164-170.

Bryant, S.H. & Amzel, L.M. (1987). Correctly folded proteins make twice as many hydrophobic contacts. *Int. J. Peptide Protein Res.* **29**, 46-52.

Bryant, S.H. & Lawrence, C.E. (1993). An empirical energy function for threading protein sequence through the folding motif. *Proteins* **16**, 92-112.

Camacho, C.J. & Thirumalai, D. (1993a). Kinetics and thermodynamics of folding in model proteins. *Proc. Natl. Acad. Sci. USA* **90**, 6369-6372.

Camacho, C.J. & Thirumalai, D. (1993b). Minimum energy compact structures of random sequences of heteropolymers. *Phys. Rev. Lett.* **71**, 2505-2508.

Casari, G. & Sippl, M.J. (1992). Structure-derived hydrophobic potential. Hydrophobic potential derived from X-ray structures of globular proteins is able to identify native folds. *J. Mol. Biol.* **224**, 725-732.

Chan, H.S. & Dill, K.A. (1991a). Sequence-space soup of proteins and copolymers. *J. Chem. Phys.* **95**, 3775-3787.

Chan, H.S. & Dill, K.A. (1991b). Polymer principles in protein structure and stability. *Ann. Rev. Biophys. and Biophys. Chem.* **20**, 447-449.

Chan, H.S. & Dill, K.A. (1994). Transition states and folding dynamics of proteins and heteropolymers. *J. Chem. Phys.* **100**, 9238-9257.

Chan, H.S., Bromberg, S. & Dill, K.A. (1995). Models of cooperativity in protein folding. *Phil. Trans. R. Soc. Lond. B* **348**, 61-70.

Chothia, C. (1976). The nature of accessible and buried surfaces in proteins. *J. Mol. Biol.* **105**, 1-14.

Dill, K.A. (1990). Dominant forces in protein folding. *Biochemistry* **29**, 7133-7155.

Dill, K.A., Bromberg, S., Yue, K., Fiebig, K.M., Yee, D.P., Thomas, P.D. & Chan, H.S. (1995). Principles of protein folding-- A perspective from simple exact models. *Protein Science* **4**: 561-602.

Fauchère, J.-L. & Pliska, V. (1983). Hydrophobic parameters P of amino-acid side chains from the partitioning of N-acetyl-amino acid amides. *Eur. J. Med. Chem.-- Chim. Ther.* **18**, 369-375.

Finkelstein, A.V., Gutin, A.M. & Badretdinov, A.Ya. (1995). Boltzmann-like statistics of protein architectures. Origins and consequences. *Sub-Cellular Biochemistry* **24**, 1-26.

Finkelstein, A.V., Gutun, A.M. & Badretdinov, A.Ya. (1993). Why are the same protein folds used to perform different functions? *FEBS Lett.* **325**, 23-28.

Godzik, A. & Skolnick, J. (1992). Sequence-structure matching in globular proteins: application to supersecondary and tertiary structure determination. *Proc. Natl. Acad. Sci. USA* **89**, 12098-12102.

Hendlich, M., Lackner, P., Weitckus, S., Floeckner, H., Froschauer, R., Gottsbacher, K., Casari, G. & Sippl, M.J. (1990). Identification of native protein folds amongst a large number of incorrect models. The calculation of low energy conformations from potentials of mean force. *J. Mol. Biol.* **216**, 167-180.

Hobohm, U. & Sander, C. (1994). Enlarged representative set of protein structures. *Protein Science* **3**, 522-524.

Janin, J. (1979). Surface and inside volumes in globular proteins. *Nature* **277**, 491-492.

Jones, D.T., Taylor, W.R. & Thornton, J.M. (1992) A new approach to protein fold recognition. *Nature* **358**, 86-89.

Kauzmann, W. (1959). Some factors in the interpretation of protein denaturation. *Adv. Protein Chem.* **14**, 1-63.

Kocher, J.P., Rooman, M.J. & Wodak, S.J. (1994). Factors influencing the ability of knowledge-based potentials to identify native sequence-structure matches. *J. Mol. Biol.* **235**, 1598-1613.

Kolaskar, A.S. & Prashanth, D. (1979). Empirical torsional potential functions from protein structure data. *Int. J. Peptide Protein Res.* **14**, 88-98.

Kolinski, A. & Skolnick, J. (1994). Monte Carlo simulations of protein folding. I. Lattice model and interaction scheme. *Proteins* **18**, 338-352.

Lau, K.F. & Dill, K.A. (1989). A lattice statistical mechanics model of the conformational and sequence spaces of proteins. *Macromolecules* **22**, 3986-3997.

Lau, K.F. & Dill, K.A. (1990). Theory for protein mutability and biogenesis. *Proc. Natl. Acad. Sci. USA* **87**, 638-642.

Lawrence, C.E. & Bryant, S.H. (1991). Hydrophobic potentials from statistical analysis of protein structures. *Methods Enzymol.* **202**, 20-31.

Lawrence, C., Auger, I. & Mannella, C. (1987). Distribution of accessible surfaces of amino acids in globular proteins. *Proteins* **2**, 153-161.

Lee, B.K. & Richards, F.M. (1971). The interpretation of protein structures: estimation of static accessibility. *J. Mol. Biol.* **55**, 379-400.

Lipman, D.J. & Wilbur, W.J. (1991). Modelling neutral and selective evolution of protein folding. *Proc. Roy. Soc. London, Series B* **245**, 7-11.

Lüthy, R., Bowie, J.U. & Eisenberg, D. (1992). Assessment of protein models with three-dimensional profiles. *Nature* **356**, 83-85.

MacArthur, M.W. & Thornton, J.M. (1991). Influence of proline residues on protein conformation. *J. Mol. Biol.* **218**, 397-412.

Miller, R., Danko, C.A., Fasolka, M.J., Balazs, A.C., Chan, H.S. & Dill, K.A. (1992). Folding kinetics of proteins and copolymers. *J. Chem. Phys.* **96**, 768-790.

Miller, S., Janin, J., Lesk, A.M. & Chothia, C. (1987). Interior and surface of monomeric proteins. *J. Mol. Biol.* **196**, 641-656.

Miyazawa, S. & Jernigan, R.L. (1985). Estimation of effective interresidue contact energies from protein crystal structures: quasi-chemical approximation. *Macromolecules* **18**, 534-552.

Nishikawa, K. & Matsuo, Y. (1993). Development of pseudoenergy potentials for assessing protein 3-D-1-D compatibility and detecting weak homologies. *Protein Eng.* **6**, 811-820.

Nozaki, Y. & Tanford , C. (1971). The Solubility of Amino Acids and Two Glycine Polypeptides in Aqueous Ethanol and Dioxane Solutions. *J. Biol. Chem.* **246**, 2211-2217.

O'Toole, E.M. & Panagiotopoulos, A.Z. (1993). Effect of Sequence and Intermolecular Interactions on the Number and Nature of Low-Energy States for Simple Model Proteins. *J. Chem. Phys.* **90**, 3185-3190.

Pellegrini, M., & Doniach, S. (1993). Computer simulation of antibody binding specificity. *Proteins* **15**, 436-444.

Pohl, F.M. (1971). Empirical protein energy maps. *Nature New Biology* **234**, 277-279.

Rashin, A.A., Ionif, M. & Honig, B. (1986). Internal cavities and buried waters in globular proteins. *Biochemistry* **25**, 3619-3625.

Richards, F.M. (1977). Areas, volumes, packing and protein structure. *Annu. Rev. Biophys. Bioeng.* **6**, 151-176.

Rose, G.D., Geselowitz, A.R., Lesser, G.J., Lee, R.H. & Zehfus, M.H. (1985). Hydrophobicity of amino acid residues in globular proteins. *Science* **229**, 834-838.

Serrano, L., Sancho, J., Hirshberg, M. & Fersht, A.R. (1992). Alpha-helix stability in proteins. I. Empirical correlations concerning substitution of side-chains at the N- and C-caps and the replacement of alanine by glycine or serine at solvent-exposed interfaces. *J. Mol. Biol.* **227**, 544-559.

Shortle, D., Chan, H.S. & Dill, K.A. (1992). Modeling the effects of mutations on the denatured states of proteins. *Protein Science* **1**, 201-215.

Sippl M.J. & Weitckus, S. (1992). Detection of native-like models for amino acid sequences of unknown three-dimensional structure in a data base of known protein conformations. *Proteins* **13**, 258-271.

Sippl, M.J. (1990). Calculation of conformational ensembles from potentials of mean force. An approach to the knowledge-based prediction of local structures in globular proteins. *J. Mol. Biol.* **213**, 859-883.

Skolnick, J. & Kolinski, A. (1990). Simulations of the folding of a globular protein. *Science* **250**, 1121-1125.

Tanaka, S. & Scheraga, H.A. (1976). Medium- and long-range interaction parameters between amino acids for predicting three-dimensional structures of proteins. *Macromolecules* **9**, 945-950.

Unger, R. & Moult, J. (1993). Genetic algorithms for protein folding simulations. *J. Mol. Biol.* **231**, 75-81.

Wilmanns, M. & Eisenberg, D. (1993). Three-dimensional profiles from residue-pair preferences: identification of sequences with beta/alpha-barrel fold. *Proc. Natl. Acad. Sci. USA* **90**, 1379-1383.

Wilson, C. & Doniach, S. (1989). A computer model to dynamically simulate protein folding: studies with crambin. *Proteins* **6**, 193-209.

Chapter 3

OPERA: An Iterative Method for Extracting Accurate Potentials from Known Protein Structures

Paul D. Thomas and Ken A. Dill

The coauthor supervised this work.

Abstract

We present a new method for extracting, from a database of known protein structures, information about the energetics that formed those structures. We call the method Observed-Predicted Ensemble Ratio Adjustment (OPERA). Putative interaction "energies" are adjusted iteratively until the predicted structure of each protein in the database is the same as that observed experimentally. The OPERA method finds correct interaction energies when applied to an exact lattice model. We then apply the method to finding a pairwise interresidue potential that solves the "protein recognition problem." We explore the set of parameters required to find a solution to the problem, and test the power of different parameter sets in predicting the native conformations of new proteins.

Introduction

We describe a new method for deriving effective interresidue interaction "potentials" from the protein structures in the Protein Data Bank (PDB) (Bernstein *et al.*, 1977). Database-derived potentials are energy-like quantities widely used in computer algorithms for protein folding (Wilson & Doniach, 1989; Skolnick & Kolinski, 1990; Sun, 1993; Kolinski & Skolnick, 1994, Gunn *et al.*, 1994; Monge *et al.*, 1995), threading (Bowie *et al.*, 1991; Sippl & Weitckus, 1992; Maiorov & Crippen, 1992, 1994; Hendlich *et al.*, 1990; Jones *et al.*, 1992; Bryant & Lawrence, 1993; Kocher *et al.*, 1994; Flöckner *et al.*, 1995), and docking (Pellegrini & Doniach, 1993). Numerous methods for deriving residue pair potentials from known protein structures have been proposed (Tanaka & Scheraga, 1976; Miyazawa & Jernigan, 1985, 1994; Wilson & Doniach, 1989; Sippl, 1990; Hendlich *et al.*, 1990; Casari & Sippl, 1991; Bryant & Lawrence, 1993; Godzik & Skolnick, 1992; Maiorov & Crippen, 1992, 1994). The most common, which we call the "statistical potential," applies the Boltzmann relation to the pairing frequencies of amino acids observed in known protein structures. Maiorov and Crippen (1992) have developed a different approach that solves a set of linear equations to find parameters that solve the "threading problem," i.e. that best identify when a particular amino acid sequence is appropriately superimposed on its correct structure when tested against a zoo of structures that contains the correct structure plus other decoy conformations.

We have identified some problems with the way the energy-like statistical potentials are obtained from pairing frequencies (Thomas & Dill, in press). Our purpose here is to describe an improved method for obtaining such energy-like quantities. To describe the problems, we define two terms: the "true" energies and the "extracted" energies. The extracted energies are those which are based on observing the pairing frequencies ρ_{ij}, of

amino acid type i with amino acid type j, and taking the logarithm to obtain dimensionless energy-like quantities, e_{ij}:

$$e_{ij} = energy/kT = -\ln\left(\frac{\rho_{ij}}{\rho_{ij}^{*}}\right) \qquad (1)$$

where ρ_{ij}* represents the pairing frequencies in the reference state. The true energies are defined as the actual free energies, E_{ij}, that nature uses to drive proteins to fold up as they do. If the statistical potential procedures and assumptions were perfect, then the extracted potentials would equal the true energies, $e_{ij} = E_{ij}$. For real proteins, of course, the true energies are not known accurately. Using an exact lattice model, with known true energies, to test the internal consistency of the extraction procedure, we have found that energies extracted by the standard methods used for statistical potentials are not equal to the true energies.

The problems with current statistical potentials derive from an implicit neglect of excluded volume, chain connectivity and sequence information (Thomas & Dill, in press). For example, the standard extraction methods assume that the frequency of pairing alanine with tryptophan is independent of the frequency of pairing glycine with lysine, i.e., in general that the frequency of pairing of residue types (*ij*) is independent of the frequency of pairing of different types (*kl*). But our tests showed that pairing frequencies are not independent. In a compact chain molecule like a globular protein, the drive for certain amino acids to be in contact can have significant effects on the positions of other amino acids in the chain. As a result, extracted statistical energies can have the incorrect magnitudes and even incorrect signs (Thomas and Dill, in press). To fix these problems requires accounting for the actual excluded volume, sequences and chain connectivities in the reference state, rather than assuming the reference state is a sea of disconnected volumeless "ideal-gaslike" amino acid particles. Here we propose an iterative method to repair these problems.

An iterative method to extract potentials from structures

There are two main differences between our method and other approaches for extracting statistical potentials. First, rather than using a sea of amino acids as a reference state, we use the fully connected chain with the correct amino acid sequence in different conformations as a reference state. Second, because this type of reference state precludes an analytical expression, our method involves iteration to converge to a matrix of pairwise amino acid scores. We call the procedure "observed-predicted ensemble ratio adjustment" (OPERA). The potential scores e_{ij} are adjusted at each iteration by using the ratio between observed and predicted structural ensembles. When the potential predicts that the observed (native) structure is sufficiently favored "energetically" over alternative structures, the iterative process is completed. A flow chart for the procedure is shown in Figure 1.

We start with a first guess value for the statistical potential score matrix, $e_{ij}^{(0)}$, where the superscript (0) indicates that this is the "zeroth" approximation value, and e_{ij} is the "energy" score for the pairing of amino acid types i and j. We compute the Boltzmann average (over the different conformations) of the pairing frequencies using this weighting function. For example, consider the case where there are two conformations. Suppose there are two types of contact interactions, type A and type B. Type A has a zeroth-approximation "score" of -1 and type B has a score of -2. Conformation 1 has two contacts, one of type A and one of type B, for a total score of -3. Conformation 2 also has two contacts, both of type A, for a total score of -2. In the Boltzmann average, the weight of conformation 1 is e^3 and the weight of conformation 2 is e^2. The weighted pairing frequency for type A is: $(1e^3 + 2e^2) / (e^3 + e^2) = 1.27$.

In general, on this first pass, the pairing frequencies obtained in this way for the predicted ensemble of conformations will not equal the pairing frequencies observed in the observed native conformation. In the example above, if conformation 1 were the native

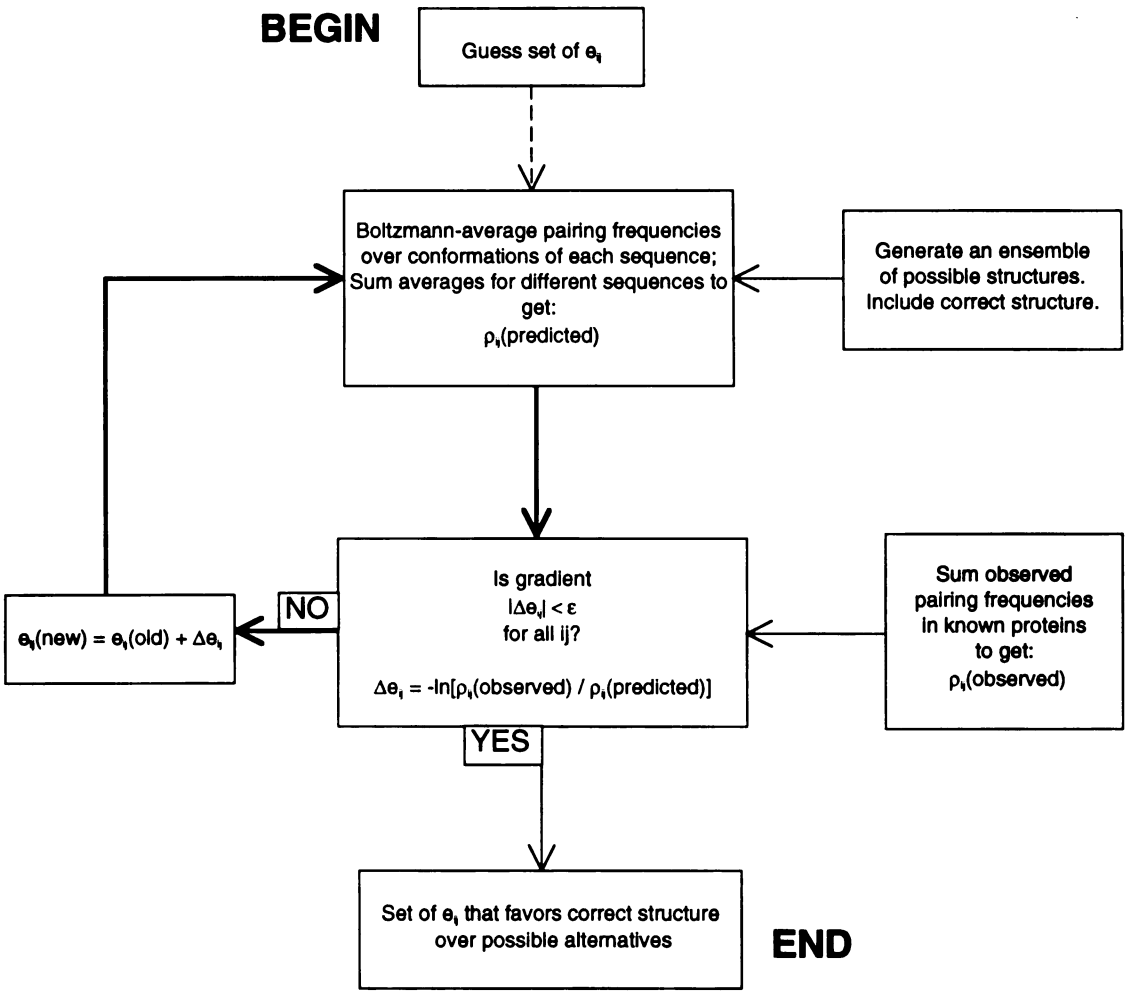


Figure 1. Flow chart for determining pairwise amino acid interactions in proteins.

conformation there would be exactly one observed A-type contact. We construct a measure, Δe_{ij}, of the error between the true energies and the guessed energies:

$$\Delta e_{ij} = -\ln\left(\frac{\rho_{ij}(observed)}{\rho_{ij}(predicted)}\right) \qquad (2)$$

Note that Δe_{ij} is equivalent to Equation (1) used to derive statistical potentials, where the Boltzmann-averaged pairing frequencies are used as the reference state. Our "first approximation" scores are equal to the zeroth approximation plus the correction:

$$e_{ij}^{(1)} = e_{ij}^{(0)} + \Delta e_{ij}$$

and in general:

$$e_{ij}^{(n)} = e_{ij}^{(n-1)} + \Delta e_{ij} \qquad (3)$$

Again we compute Boltzmann averages using the new e_{ij}'s and repeat this cycle until it converges; i.e., until $|\Delta e_{ij}|$ is smaller than some preset tolerance. Provided this process converges and is adequately parameterized, it will arrive at a final set of $e_{ij}^{(n)}$ that can fully discriminate, for a given sequence, its correct fold among the set of possible conformations. Specifically, if the native conformation is significantly favored by the scoring function over all alternative conformations, it will dominate the ensemble. If this is simultaneously true for all different conformations in the database, Δe_{ij} will approach 0 and the process is converged. We also wish to point out that if the zeroth approximation assumes all interactions $e_{ij}^{(0)}$ to be equal, the reference state is equivalent to a random mixing distribution and the correction terms Δe_{ij} are the quantities currently used in statistical potentials (plus some constant term).

An exact lattice model test

In this section we use lattice model examples, for which we can know exactly the true energies, to show that the OPERA method converges to produce extracted energies equal to the true underlying energies. We use a 2-dimensional square lattice model of chains of length $L = 14$ having two monomer types A and B, for which we can specify any

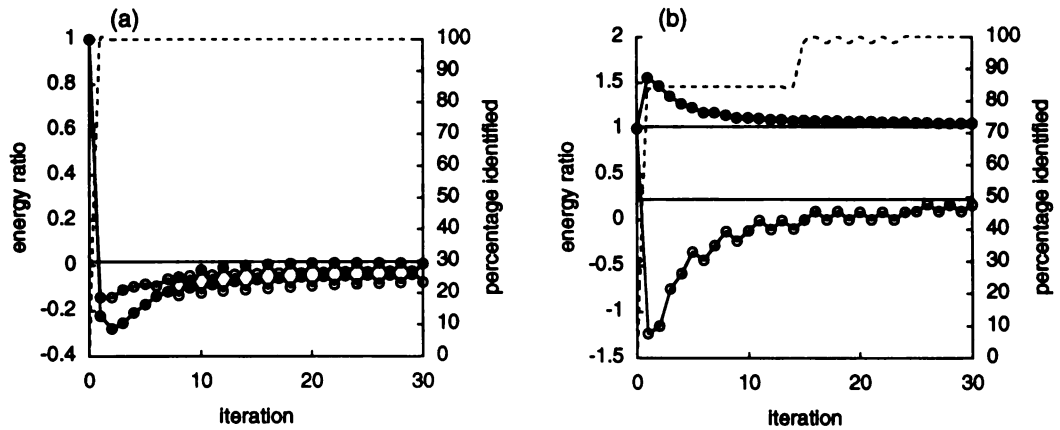


Figure 2. An exact lattice model test of the OPERA method. Database-derived contact energy ratios, AB/AA (filled circles) and BB/AA (open circles) are plotted on the left axes. Solid horizontal lines represent the "true" values to which the derived values should converge. The dotted lines (right axes) give the percentage of sequences whose native structures are correctly identified by the derived contact energies. (a) "True" potential AB/AA = 0 and BB/AA = 0. (b) "True" potential AB/AA = 1 and BB/AA = 0.2.

"true" energy we choose, and for which our decoys are the full conformation space of non-native conformations obtained by exact enumeration. For each such true potential we generate a model "PDB." We performed this test on several different lattice structure databases generated by different sets of true potentials. We then use the OPERA method to obtain extracted potentials.

The two examples in Figure 2 show that the extracted potentials converge to true values in a few iterations. For all the lattice model true energies, the method successfully converges to 100% correct discrimination of the right structure from the decoys. The number of iterations required to succeed depends on the true potential.

There are two different standards to which extracted potentials can be held. The least stringent standard requires only that extracted potentials correctly *identify* native states, even if they do not give completely correct relative energies. The more demanding

standard requires that extracted potentials give energy ratios correctly. (Extracted potentials cannot do better than this; they cannot give the absolute energies correctly, because there is always an undeterminable arbitrary constant multiplier for each energy. This is intrinsic to the concept of statistical potentials (Thomas & Dill, in press)). Figure 2 shows that the OPERA method passes this more demanding test; it correctly gives ratios of the different e_{ij}'s in addition to discriminating the correct from the decoy structures.

The gapless threading problem

Now we test the OPERA method on a more realistic problem: gapless threading of real proteins. The problem is to correctly identify the native structure in a zoo of native plus decoy conformations, where no insertions or deletions are used. This problem has been attacked by other methods with considerable success (Sippl & Weitckus, 1992; Jones *et al.*, 1992; Maiorov & Crippen, 1992, 1994; Bryant & Lawrence, 1993; Kocher *et al.*, 1994; Flöckner *et al.*, 1995), but relative performance is difficult to assess since each of these methods was tested using different definitions of test proteins and of decoy structures. Here our purpose is twofold. First, we show that the OPERA method can be 100% successful on the gapless threading test-- we use the protein sets defined Maiorov and Crippen (1992) since our method also requires both a training set and a test set of proteins. Second, because the OPERA method is relatively fast, we study several general properties of database-derived potentials, including how their success rate depends on the number and nature of the parameters.

Following Hendlich *et al.* (1991) and Maiorov and Crippen (1992), we take the interaction centers to be the amino acid C^{β} (a virtual C^{β} is used for Gly). Whereas we have arbitrarily chosen to use Boltzmann-averaged weighting, there is little justification for

preferring it over other weighting schemes. A weight need not depend exponentially on the "interaction energy score." Since the energies can be determined only to within an arbitrary scaling constant, "Z-scores" (Bowie *et al.*, 1991) are equally useful for scoring. The Z-score is the number of standard deviations of a given conformation above or below the mean threaded energy, and therefore is invariant to such scaling factors. We use $\exp[-\gamma(\text{Z-score})]$, where γ is an arbitrary constant analogous to temperature, as a weighting factor in $\rho_{ij}(predicted)$ in Equation (2). Essentially, γ determines the size of the relative score margin between the native and decoy structures that is required for convergence; a larger γ will result in a larger margin.

A minimal parameter set

Maiorov and Crippen (1992, 1994) divide the protein database into two sets, a "training set" and a "test set." The "energy" parameters were obtained by solving a set of linear equations corresponding to the gapless threading problem for their 37 training set proteins. These same scores were then used to predict a larger set of 49 compact test proteins that were not in the training set. They found that all of the test proteins were also correctly predicted. But, as in all database-derived potentials, Maiorov and Crippen had to define the "energetic" parameters in their potentials. They chose to use a specific distance-dependent functional form of interaction, and to divide the 20 amino acids into different classes depending on sequential separation. Their method might have succeeded with a different set of parameters, but computer time prohibited them from more extensive studies.

OPERA is sufficiently fast that we can explore many different parameter sets. We consider various ways of classifying amino acids into n groups, and thus $n(n+1)/2$ parameters. We chose not to distinguish between different sequential separations, nor do

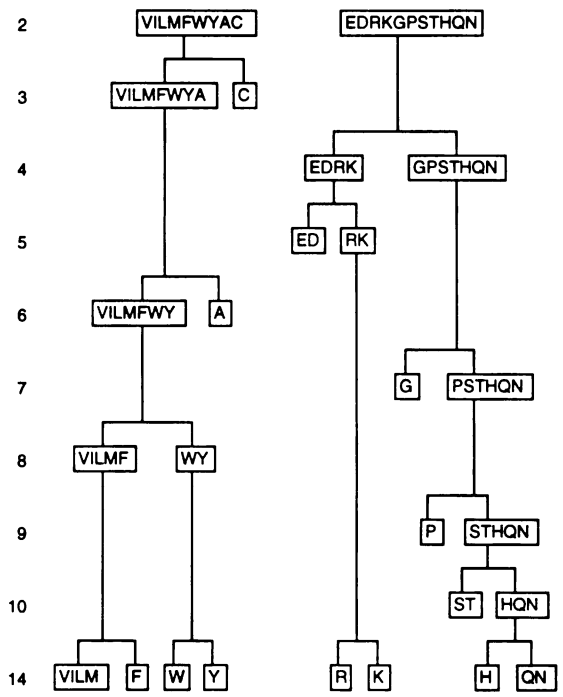


Figure 3. Reduced classifications of the amino acids. The total number of classes is given on the left. The "tree" diagram shows the division of larger classes into smaller classes to increase the total number of amino acid classes.

we include backbone-sidechain interactions. We increase the number of amino acid classes (Figure 3) until all the training set proteins can be correctly predicted. Figure 4 shows that dividing amino acids into two classes (three parameters)-- hydrophobic and other-- correctly identifies 19/37 = 51% of the training set proteins. Five classes (15 parameters: hydrophobic, cysteine, positively charged, negatively charged, and other) correctly identifies 31/37 = 84% of the training set. All 37 training set proteins are correctly identified using 10 classes (55 parameters), about half the number used by Maiorov and Crippen (1992). When applied to the full set of test proteins, our 10-class potential, like

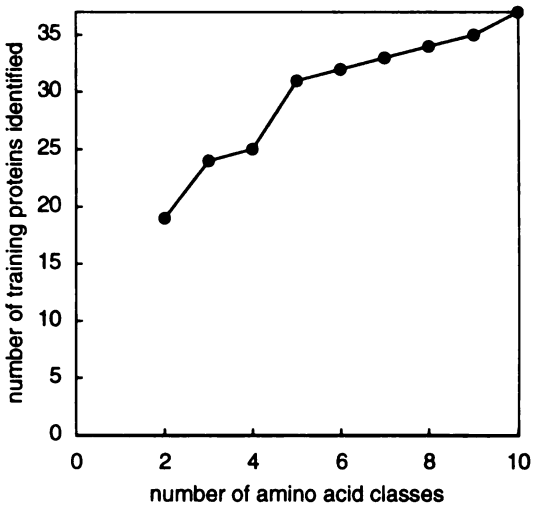


Figure 4. Number of amino acid classes required to "learn" the training set proteins. When there are fewer than ten classes, some of the training set proteins are not correctly identified by the OPERA contact potential. Classifications are given in Figure 3.

that of Maiorov and Crippen (1992), succeeds for all proteins defined as compact. In addition, it succeeds for two "non-compact" proteins [as defined by Maiorov and Crippen (1992)], labx.A and lcyc.3, for which the Maiorov-Crippen potential fails.

The number of parameters required to solve the threading problem, as well as the parameter values themselves, depend on not only the set of training proteins, but also the set of alternative structures. While our parameter set solves the simple gapless threading problem for the sets of proteins and alternative structures of Maiorov and Crippen (1992), we do not expect it to solve, for all proteins, more complicated versions of the threading problem. For example, including insertions and deletions (Bryant & Lawrence, 1993) greatly increases the number of alternatives, making a solution much more challenging.

The jackknife test

Maiorov and Crippen point out that their training set proteins are the most challenging proteins in their threading test. Are the native structures of these proteins correctly predicted if they are excluded from the training set? Because the OPERA method is relatively fast, we are able to perform a full jackknife test of our potential function. We recalculate our 55-parameter potential 37 times, each time excluding one of the training set proteins. We then try to predict the native conformation of the excluded protein. Significantly, more than one-quarter (10/37) of the 37 proteins are not predicted correctly when excluded from the training set. Thus the exact choice of training set proteins can be very important. Most of the proteins that fail the jackknife test are either very hydrophobic (1crn), or stabilized by forces that are not considered in our simple potential -- protein-heme interactions (351c, 1cc5, 5cyt.R, 2cdv) or significant interactions with other protein chains (2ovo, 1cse.I, 2ssi, 2hmq.A and 2pab.A).

Choosing a parameter set

Using the same set of training proteins, how well do different parameter sets perform in predicting non-training-set proteins? We systematically change the definitions of the energetic parameters in the potentials, and judge the relative success of different potentials by their ability to predict proteins not in the training set. Specifically, we study the effects of grouping amino acids into classes, and the effects of different distance-dependent interactions. Using the Maiorov and Crippen training set of 37 proteins, we attempted to thread a more challenging test set. This set of 121 proteins comprises the representative set of Hobohm and Sander (1992) having less than 25% sequence identity, which also have a relatively small radius of gyration [as defined by Maiorov and Crippen (1992)]. The set of decoy structures is still the same as in Maiorov and Crippen (1992).

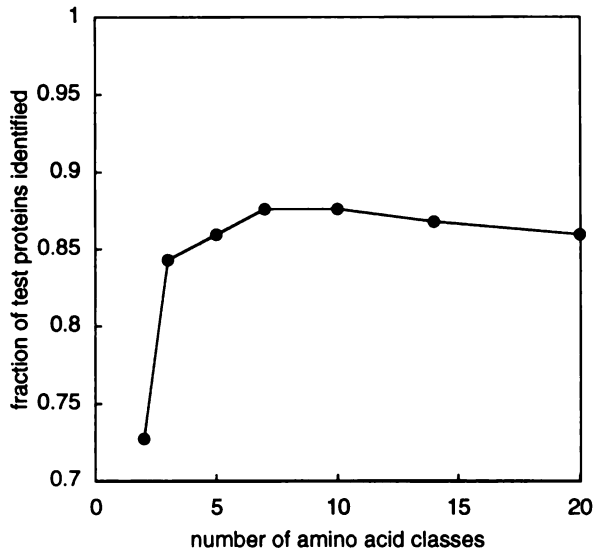


Figure 5. Fraction of test set proteins correctly identified by OPERA contact potentials having different numbers of amino acid classes.

Figure 5 shows the success rates in recognizing test proteins, using the various collections of amino acid classes listed in Figure 3. Hydrophobic and polar classes alone successfully thread about 72% of the proteins. Stepping up to three classes-- hydrophobic, polar, and cysteine-- succeeds at the 84% level. The use of seven to ten classes succeeds at about the 88% level. It is interesting that adding more parameters beyond ten classes *diminishes* the success rate slightly. Our results show that for the gapless threading problem, assuming that amino acids fall into 20 classes is no more successful than assuming that they fall into five classes, if no distance dependence is included. Beyond ten classes, added parameters largely serve to "learn" special case protein structures; they specialize rather than generalize.

How is this conclusion affected if a statistical potential includes an adjustable distance dependence? We found that for ten or fewer amino acid classes, allowing different interaction strengths for difference distance intervals generally does not change the

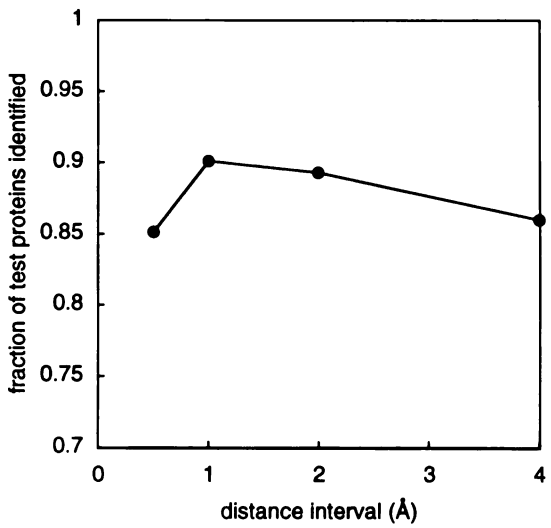


Figure 6. Fraction of test set proteins correctly identified by distance-dependent OPERA potentials having different distance bin sizes. The bins span from 5 to 9Å.

predictive power. However with 20 classes, distance information helps, up to a point. There are characteristic interaction ranges for specific amino acid pairs in proteins that are lost when averaged into amino acid classes.

What are the optimal parameters in a distance-dependent potential? There are two commonly used ways of changing the number of parameters in database-derived potentials: (1) changing the width of distance bins over which statistics are collected, and (2) changing the upper limit of the interaction distance between two amino acids. Figure 6 shows a study in which distance-dependent interactions between 5 and 9 Å were collected into bin sizes of 0.5, 1.0, 2.0, or 4.0 Å widths. The optimal distance bins width was 1.0 Å ; smaller bins apparently exceed the resolution of statistical potentials in our method, which is roughly what one would expect based on the resolution of the experimentally determined structures in the PDB. What is the longest meaningful distance for distance-dependent potentials? We now fix each distance bin to be 2 Å . Figure 7 shows the effect of different

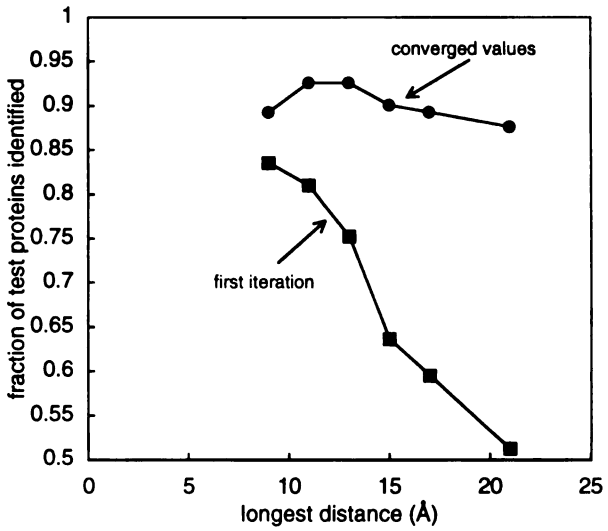


Figure 7. Fraction of test set proteins correctly identified by distance-dependent OPERA potentials having different upper limits on the interaction distance. The success level of the first iteration values (squares, approximately equal to the "statistical potential") decreases with increasing distance, while the success of the converged values (circles) peaks at 11-13Å. The lower limit is 7Å in all cases, and the bin size is 2Å.

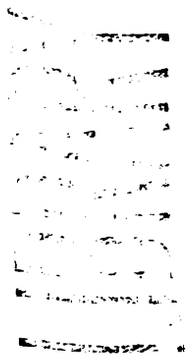
upper limits for the distance dependence. Using the first iteration value of the OPERA potentials, corresponding approximately to current methods, the predictive power consistently diminishes for longer ranges. For the converged values, the best upper limit, with around 93% prediction success, is 10-14 Å.

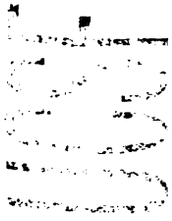
Conclusions

We have developed an iterative method, which we call OPERA (Observed-Predicted Ensemble Ratio Adjustment), for extracting potentials from protein structural

databases. It differs from other methods in using decoy conformations of the native sequence as a reference state, rather using mean-field assumptions about random gaslike "seas" of amino acids. We make a rigorous test of the method on a 2-dimensional lattice model, which previously identified flaws in the mean-field approaches (Thomas & Dill, 1995); it shows that the iterative method converges correctly to ratios of extracted energies that equal the corresponding ratios of the true underlying energies. We then apply the OPERA method to the protein recognition problem, as defined by Maiorov and Crippen (1992). We find that the predictive power of the extracted potentials does not always improve as the number of parameters used in them increases. The main point here is not to suggest that any one particular parameter set is fully adequate, or applies to all problems. Rather the purpose of this work is to identify a methodology for extracting energy-like information from protein structures. Nevertheless, many of the contact "energies" we extract have clear physical interpretations (see Appendix). Some of these are obvious, such as attractions between hydrophobic groups and between opposite charges, and repulsions between like charges. Others are less so. For example, "statistical potentials" generally find Pro-X (where X is any amino acid) to be unfavorable since Pro is generally found on protein surfaces (Thomas & Dill, in press; Kocher *et al.*, 1994). OPERA finds Pro to have the most favorable interactions with aromatic side chains (Phe, Tyr and Trp), which can be rationalized in terms of hydrophobic ring "stacking."

The OPERA method is readily generalizable. It can be applied to any system in which the distribution of states is known, but the underlying energies that cause this distribution are not. In the case of database-derived protein energetics, other types of interactions beyond pairwise interresidue terms can be treated using OPERA. The strength of the method is that it explicitly considers alternative conformations of each amino acid sequence; it is therefore only as good as the alternative conformations it uses. The gapless threading method used here is a simple and common way of generating alternative conformations. More sophisticated conformation-generating schemes, such as motif





threading and even protein folding algorithms, will allow even more accurate inferences about the energetics of protein folding.

Acknowledgments

P.D.T. was a Howard Hughes Medical Institute Predoctoral Fellow.

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). The protein data bank: a computer-based archival file for macromolecular structures. *J. Mol. Biol.* **112**, 535-542.

Bowie, J.U., Lüthy, R. & Eisenberg, D. (1991). A method to identify protein sequences that fold into a known three-dimensional structure. *Science* **253**, 164-170.

Bryant, S.H. & Lawrence, C.E. (1993). An empirical energy function for threading protein sequence through the folding motif. *Proteins* **16**, 92-112.

Casari, G. & Sippl, M.J. (1992). Structure-derived hydrophobic potential. Hydrophobic potential derived from X-ray structures of globular proteins is able to identify native folds. *J. Mol. Biol.* **224**, 725-732.

Flöckner, H., Braxenthaler, M., Lackner, P., Jaritz, M., Ortner, M. & Sippl, M.J. (1995). Progress in fold recognition. Proteins.

Godzik, A. & Skolnick, J. (1992). Sequence-structure matching in globular proteins: application to supersecondary and tertiary structure determination. *Proc. Natl. Acad. Sci. USA* **89**, 12098-12102.

Gunn, J.R., Monge, A., Friesner, R.A. & Marshall, C.H. (1994). Hierarchical algorithm for computer modeling of protein tertiary structure: Folding of Myoglobin to 6.2Å resolution. *J. Phys. Chem.* **98**, 702-711.

Hendlich, M., Lackner, P., Weitckus, S., Floeckner, H., Froschauer, R., Gottsbacher, K., Casari, G. & Sippl, M.J. (1990). Identification of native protein folds amongst a large number of incorrect models. The calculation of low energy conformations from potentials of mean force. *J. Mol. Biol.* **216**, 167-180.

Hobohm, U. & Sander, C. (1994). Enlarged representative set of protein structures. *Protein Science* **3**, 522-524.

Jones, D.T., Taylor, W.R. & Thornton, J.M. (1992) A new approach to protein fold recognition. *Nature* **358**, 86-89.

Kocher, J.P., Rooman, M.J. & Wodak, S.J. (1994). Factors influencing the ability of knowledge-based potentials to identify native sequence-structure matches. *J. Mol. Biol.* **235**, 1598-1613.

Kolaskar, A.S. & Prashanth, D. (1979). Empirical torsional potential functions from protein structure data. *Int. J. Peptide Protein Res.* **14**, 88-98.

Kolinski, A. & Skolnick, J. (1994). Monte Carlo simulations of protein folding. I. Lattice model and interaction scheme. *Proteins* **18**, 338-352.

Lawrence, C.E. & Bryant, S.H. (1991). Hydrophobic potentials from statistical analysis of protein structures. *Methods Enzymol.* **202**, 20-31.

Lüthy, R., Bowie, J.U. & Eisenberg, D. (1992). Assessment of protein models with three-dimensional profiles. *Nature* **356**, 83-85.

Miyazawa, S. & Jernigan, R.L. (1985). Estimation of effective interresidue contact energies from protein crystal structures: quasi-chemical approximation. *Macromolecules* **18**, 534-552.

Monge, A, Lathrop, E.J.P., Gunn, J.R., Shenkin, P.S. & Friesner, R.A. (1995). Computer modeling of protein folding: Conformational and energetic analysis of reduced and detailed protein models. *J. Mol. Biol.* **247**, 995-1012.

Nishikawa, K. & Matsuo, Y. (1993). Development of pseudoenergy potentials for assessing protein 3-D-1-D compatibility and detecting weak homologies. *Protein Eng.* **6**, 811-820.

Pellegrini, M., & Doniach, S. (1993). Computer simulation of antibody binding specificity. *Proteins* **15**, 436-444.

Pohl, F.M. (1971). Empirical protein energy maps. *Nature New Biology* **234**, 277-279.

Rashin, A.A., Ionif, M. & Honig, B. (1986). Internal cavities and buried waters in globular proteins. *Biochemistry* **25**, 3619-3625.

Rose, G.D., Geselowitz, A.R., Lesser, G.J., Lee, R.H. & Zehfus, M.H. (1985). Hydrophobicity of amino acid residues in globular proteins. *Science* **229**, 834-838.

Sippl M.J. & Weitckus, S. (1992). Detection of native-like models for amino acid sequences of unknown three-dimensional structure in a data base of known protein conformations. *Proteins* **13**, 258-271.

Sippl, M.J. (1990). Calculation of conformational ensembles from potentials of mean force. An approach to the knowledge-based prediction of local structures in globular proteins. *J. Mol. Biol.* **213**, 859-883.

Skolnick, J. & Kolinski, A. (1990). Simulations of the folding of a globular protein. *Science* **250**, 1121-1125.

Tanaka, S. & Scheraga, H.A. (1976). Medium- and long-range interaction parameters between amino acids for predicting three-dimensional structures of proteins. *Macromolecules* **9**, 945-950.

Thomas, P.D. & Dill, K.A. (in press). Statistical potentials extracted from known protein structures: How accurate are they? *J. Mol. Biol.*, in press.

Wilmanns, M. & Eisenberg, D. (1993). Three-dimensional profiles from residue-pair preferences: identification of sequences with beta/alpha-barrel fold. *Proc. Natl. Acad. Sci. USA* **90**, 1379-1383.

Wilson, C. & Doniach, S. (1989). A computer model to dynamically simulate protein folding: studies with crambin. *Proteins* **6**, 193-209.

Appendix: Interresidue interactions extracted by the OPERA method.

All of these C^{β}- C^{β} interaction potentials were extracted by the OPERA method using the training set and alternative structure set of Maiorov and Crippen (1992). The contact matrices use the Maiorov-Crippen (1992) non-binary definition of an interresidue contact. The iteratively refined two-residue-class potential performs no better than simply counting contacts between hydrophobic residues (Bryant & Amzel, 1990; Thomas & Dill, in press), so we do not list it here.

Table A1. Contact potential, three amino acid classes.

	VILMFWYA	EDRKGPSTHQN	C
VILMFWYA	-0.54	0.05	-0.52
EDRKGPSTHQN		0.13	0.00
C			-2.33

Table A2. Contact potential, five amino acid classes.

	VILMFWYA	GPSTHQN	C	ED	RK
VILMFWYA	-0.65	0.04	-0.68	0.32	-0.17
GPSTHQN		0.05	-0.26	0.12	0.17
C			-3.25	1.42	-0.09
ED				0.61	-0.43
RK					1.91

Table A3. Contact potential, ten amino acid classes.

	VILMF	HQN	C	ED	RK	A	G	WY	P	ST
VILMF	-1.84	-0.01	-1.82	0.79	-0.39	-0.77	0.53	-1.00	0.38	-0.29
HQN		0.42	-1.35	0.09	0.12	-0.20	0.10	-1.62	0.14	-0.20
C			-4.45	1.29	0.95	-0.30	-0.65	-0.52	0.10	-0.03
ED				1.30	-0.89	0.02	-0.24	0.56	2.26	0.10
RK					2.81	-0.14	0.22	-0.34	0.86	0.01
A						1.35	-0.39	-0.17	1.11	-0.34
G							-0.22	-0.01	0.94	0.39
WY								-0.19	-2.27	0.30
P									-0.65	0.53
ST										1.36

Table A4. Contact potential.

	C	M	F	I	L	V	W	Y	A	G	T	S	Q	N	E	D	H	R	K	P
C	-1.70	-1.23	-0.88	-0.48	-0.69	-0.94	-0.30	-0.96	-0.30	-0.42	-0.38	-0.20	-0.48	-0.32	0.04	0.55	-0.82	-0.48	0.00	0.07
M		0.36	-1.03	-0.41	-0.31	-0.94	-0.07	-1.10	0.05	0.00	0.06	-0.47	-0.54	0.31	0.02	1.07	-0.35	-0.43	0.55	-0.25
F			-0.01	-0.06	-1.02	-0.78	-0.39	-0.02	-0.05	0.21	-0.19	0.14	0.10	-0.02	0.19	0.20	-0.75	-0.22	-0.17	-0.43
I				-0.71	-1.04	-0.98	-0.39	-0.87	-0.04	0.40	-0.29	-0.13	-0.39	0.39	-0.20	0.04	-0.52	-0.08	-0.26	0.25
L					-1.14	-1.03	-0.97	-0.60	-0.57	-0.08	-0.39	-0.07	-0.18	-0.10	-0.05	0.50	-0.36	-0.10	0.10	0.09
V						-1.15	-0.60	-0.70	-0.60	-0.20	0.06	-0.31	-0.09	-0.24	-0.02	0.25	-0.35	-0.48	-0.08	-0.08
W							0.02	-0.99	-0.08	-0.14	0.07	-0.20	0.40	-0.08	0.32	0.24	-0.41	-0.78	-0.30	-0.44
Y								0.35	-0.37	-0.32	-0.23	0.25	-0.39	-0.74	0.22	0.11	-0.67	0.21	-0.20	-0.45
A									-0.08	-0.09	-0.22	-0.01	-0.11	-0.14	0.03	0.10	-0.15	0.07	0.00	0.41
G										0.04	0.13	-0.04	0.12	-0.18	0.40	-0.06	0.00	-0.15	0.10	0.40
T											0.20	0.05	-0.17	-0.27	0.15	-0.03	-0.27	-0.17	0.09	0.30
S												-0.13	0.40	0.37	0.30	-0.09	-0.59	0.01	0.10	0.44
Q													-0.08	-0.05	0.02	0.40	0.05	0.02	0.04	-0.21
N														-0.06	-0.25	-0.12	0.00	0.04	0.10	0.11
E															0.21	0.03	-0.53	-0.26	-0.09	0.33
D																0.00	-0.00	-0.15	-0.09	0.04
H																	0.14	-0.01	0.14	-0.22
R																		0.23	0.30	-0.02
K																			1.45	0.51
P																				0.20

Table A5. Distance-dependent potential having the greatest predictive power.

	<7Å	7-9Å	9-11Å
C-C	-3.17	-1.48	-0.43
C-M	-1.81	-0.96	-1.11
C-F	-1.46	-2.13	-1.38
C-I	-0.05	-1.23	-0.90
C-L	-0.98	-1.07	-0.94
C-V	-1.59	-1.19	-0.88
C-W	0.29	0.08	-0.95
C-Y	-1.65	-1.89	-0.97
C-A	-0.34	-1.34	-0.71
C-G	-0.44	-0.74	-0.43
C-T	-0.50	-0.70	-0.53
C-S	-0.10	-0.93	-0.56
C-Q	-0.17	-0.89	-0.50
C-N	-0.78	-1.14	-0.71
C-E	0.50	-1.23	0.17
C-D	0.62	0.49	-0.16
C-H	-1.64	-1.25	-1.51
C-R	-0.49	-0.64	-1.11
C-K	0.18	-0.81	0.04
C-P	0.36	-0.74	-0.84
M-M	-0.13	-0.35	-0.81
M-F	-1.78	-0.72	-0.71
M-I	-0.14	-0.31	-0.34
M-L	-0.50	-0.13	-0.90
M-V	-1.35	-1.36	-0.60
M-W	-0.42	0.25	-1.64
M-Y	-1.38	-0.81	-1.12
M-A	0.45	0.43	-0.18
M-G	0.88	-0.87	-0.37
M-T	0.53	-0.34	-0.11
M-S	-0.30	-0.92	0.01
M-Q	-0.27	-0.60	0.76
M-N	1.04	-0.84	-0.37
M-E	-0.13	0.88	-0.55
M-D	1.96	0.14	0.22
M-H	-1.03	-1.12	-1.45
M-R	-0.30	-0.60	-0.80
M-K	0.68	-0.20	0.25
M-P	0.48	-0.14	-0.50
F-F	-1.32	-2.13	-1.85
F-I	-1.02	-1.39	-1.44
F-L	-1.53	-1.17	-1.31
F-V	-1.23	-1.85	-1.77
F-W	-1.65	-0.80	-0.69
F-Y	-1.77	-1.20	-1.05
F-A	0.53	-0.73	-0.33
F-G	0.90	-0.25	0.31
F-T	-0.35	-0.40	-0.27
F-S	1.41	-0.15	-0.27
F-Q	0.03	0.35	-0.87
F-N	0.27	-0.56	-0.07
F-E	0.34	0.53	0.01
F-D	1.17	-0.14	-0.30
F-H	-1.50	-1.44	-0.95
F-R	-1.29	0.72	-0.20

F-K	0.53	-0.83	-0.21
F-P	-0.39	-0.84	0.22
I-I	-1.63	-0.68	-0.24
I-L	-1.65	-1.45	-1.06
I-V	-1.88	-0.67	-1.13
I-W	-0.90	-2.03	-1.33
I-Y	-1.42	-1.32	-1.16
I-A	-0.96	-0.84	-1.06
I-G	0.96	0.25	-0.18
I-T	0.08	-0.68	-0.42
I-S	-0.03	-0.10	-0.06
I-Q	-0.59	-0.51	0.22
I-N	1.97	-0.26	-0.44
I-E	-0.29	0.32	-0.70
I-D	0.39	0.09	-0.80
I-H	-1.02	-0.47	-1.68
I-R	-0.40	1.12	0.21
I-K	-0.61	-0.07	-0.30
I-P	1.93	-0.39	0.34
L-L	-1.96	-1.46	-0.96
L-V	-1.60	-1.35	-1.04
L-W	-0.93	-1.75	-1.34
L-Y	-0.69	-0.76	-0.63
L-A	-1.03	-0.29	-0.78
L-G	0.45	-0.35	-0.16
L-T	-0.28	-0.15	-0.22
L-S	0.19	-0.44	-0.73
L-Q	-0.10	0.36	-0.54
L-N	0.50	-0.43	-0.43
L-E	0.30	0.31	-0.61
L-D	2.13	0.37	-0.44
L-H	-0.69	-0.04	-0.42
L-R	0.23	-0.48	-0.91
L-K	0.04	0.08	-0.23
L-P	1.00	0.11	-0.09
V-V	-2.45	-0.57	-0.81
V-W	-1.19	-0.23	-1.56
V-Y	-1.57	0.00	-1.38
V-A	-0.85	-0.67	-0.84
V-G	-0.22	0.35	-0.24
V-T	0.48	-0.09	-0.65
V-S	-0.24	-0.16	-0.56
V-Q	0.22	-0.31	-0.81
V-N	-0.32	0.11	-0.65
V-E	0.17	0.23	-0.75
V-D	1.21	-0.28	-0.61
V-H	-0.77	0.46	-0.90
V-R	-0.63	-0.66	-0.67
V-K	0.06	-0.30	-0.07
V-P	0.89	-0.05	-0.19
W-W	0.15	-3.15	-1.26
W-Y	-1.62	-1.06	-1.45
W-A	0.87	-0.87	-0.88
W-G	-0.41	-0.20	0.33
W-T	0.93	-0.43	-0.52
W-S	0.49	-0.47	-0.16
W-Q	1.94	-0.75	-1.39
W-N	-1.52	1.14	0.55

W-E	1.31	0.23	-0.68
W-D	1.54	-1.02	-0.37
W-H	0.22	-0.82	-2.18
W-R	-1.12	0.03	-0.55
W-K	-0.69	0.06	-0.12
W-P	-0.29	0.58	1.41
Y-Y	-0.57	-0.38	-1.31
Y-A	-0.22	-0.56	-0.75
Y-G	-0.63	-0.35	-0.06
Y-T	-0.01	-0.67	-0.55
Y-S	0.59	-0.29	-0.21
Y-Q	-0.88	-0.96	-0.73
Y-N	-1.08	-0.44	-0.39
Y-E	0.90	0.10	-0.88
Y-D	0.30	-1.13	-0.14
Y-H	-1.10	-1.22	-0.69
Y-R	1.05	-0.86	-1.23
Y-K	-0.22	-0.23	-1.10
Y-P	-0.76	0.45	-0.55
A-A	0.10	0.19	-0.31
A-G	-0.31	0.06	0.23
A-T	-0.03	-0.23	-0.35
A-S	0.11	0.04	0.28
A-Q	-0.39	0.29	-0.47
A-N	-0.28	0.01	-0.10
A-E	0.01	0.30	0.18
A-D	0.17	0.59	0.38
A-H	0.21	0.22	-0.21
A-R	0.53	-0.12	0.44
A-K	-0.05	0.20	0.33
A-P	0.95	0.04	0.89
G-G	0.15	-0.15	0.41
G-T	0.49	0.08	0.25
G-S	-0.17	0.03	0.04
G-Q	0.16	0.00	-0.14
G-N	-0.06	0.12	0.17
G-E	1.18	0.12	0.12
G-D	-0.12	0.25	0.60
G-H	0.08	0.02	0.43
G-R	0.41	-0.43	0.52
G-K	0.24	-0.18	0.71
G-P	1.22	-0.24	0.01
T-T	0.37	1.16	-0.30
T-S	0.22	0.23	0.28
T-Q	-0.65	1.19	0.23
T-N	-0.52	0.12	0.12
T-E	0.41	0.95	-0.03
T-D	-0.14	1.08	-0.09
T-H	0.35	0.25	0.72
T-R	-0.31	-0.12	0.12
T-K	0.57	0.73	0.25
T-P	0.87	0.46	1.03
S-S	0.01	-0.35	0.04
S-Q	1.03	-0.19	0.35
S-N	0.71	0.76	0.64
S-E	0.09	2.05	0.51
S-D	-0.35	0.45	0.36
S-H	0.42	-0.52	-0.55

S-R	1.30	0.15	0.67
S-K	0.64	-0.27	0.18
S-P	2.02	-0.23	0.35
Q-Q	-0.46	0.27	-0.13
Q-N	0.15	-0.15	0.32
Q-E	0.92	0.50	0.26
Q-D	0.83	-0.26	0.88
Q-H	0.65	-0.25	0.13
Q-R	1.52	0.36	0.56
Q-K	0.19	1.32	-0.29
Q-P	-0.20	-0.20	0.30
N-N	-1.14	-0.46	0.44
N-E	-0.45	0.05	0.53
N-D	-0.24	1.16	0.28
N-H	0.97	-0.65	-0.10
N-R	-0.26	0.23	0.21
N-K	-0.14	1.68	0.43
N-P	0.31	-0.08	-0.46
E-E	0.57	0.94	1.40
E-D	1.70	0.63	1.39
E-H	-0.84	-0.71	0.14
E-R	-0.38	0.45	0.77
E-K	0.05	-0.19	0.13
E-P	2.21	-0.36	1.13
D-D	0.71	2.19	0.04
D-H	-0.24	-0.03	-0.22
D-R	0.18	-0.54	0.38
D-K	-0.38	0.65	0.53
D-P	1.93	1.24	0.52
H-H	1.33	-0.87	-1.22
H-R	-0.05	0.26	-0.39
H-K	0.67	-0.41	-0.36
H-P	1.24	-0.71	0.27
R-R	0.49	3.18	-0.35
R-K	0.65	0.42	0.29
R-P	0.42	-0.01	-0.32
K-K	2.73	0.35	0.67
K-P	1.70	0.01	0.03
P-P	1.55	0.96	0.24

Chapter 4

A Simple Protein Folding Algorithm using a Binary Code and Secondary Structure Constraints

Shaojian Sun, Paul D. Thomas and Ken A. Dill

This material originally appeared in *Protein Engineering*, vol. **8**, pp. 769-778 (1995).

The last author directed and supervised this work.

Abstract

We describe an algorithm to predict tertiary structures of small proteins. In contrast to most current folding algorithms, it uses very few energy parameters. Given the secondary structural elements in the sequence— α–helices and β–strands— the algorithm searches the remaining conformational space of a simplified real-space representation of chains to find a minimum energy of an exceedingly simple potential function. The potential is based only on a single type of favorable interaction between hydrophobic residues, an unfavorable excluded volume term for spatial overlaps, and for sheet proteins, an interstrand hydrogen bond interaction. Where appropriate, the known disulfide bonds are constrained by a square-law potential. Conformations are searched by a genetic algorithm. The model predicts reasonably well the known tertiary folds of 7 out of the 10 small proteins we consider. We draw two conclusions. First, for the proteins we tested, this exceedingly simple potential function is no worse than others having hundreds of energy parameters in finding the right general tertiary structures. Second, despite its simplicity, the potential function is not the weak link in this algorithm. Differences between our predicted structures and the correct targets can be ascribed to shortcomings in our search strategy. This potential function may be useful for testing other conformational search strategies.

Keywords

Binary hydrophobic interaction, genetic algorithm, protein folding.

Introduction

There are now several computer algorithms that aim to predict the folded structure of a protein from its amino acid sequence (Fasman, 1989; Merz & Le Grand, 1994). The physical approaches have not yet succeeded in predicting a broad range of protein structures from a single set of parameters. Artificial constraints are also usually required. One useful artificial constraint is to fix α–helices and β–strands as they are in the known native state, and then attempt to assemble them into the correct tertiary structure. If a conformational search algorithm cannot succeed on this restricted problem, it must surely fail when all the degrees of freedom are included. This is our approach here.

There have been earlier efforts of this type (see review by Maggiora *et al.*, 1991). Ptitsyn and Rashin (1975) represented the α–helices of myoglobin as cylinders and examined the degree of exposure of the hydrophobic and hydrophilic residues in various conformations. Janin and Chothia (1980) analyzed secondary structure packing geometry. Cohen *et al.* (1980, 1981, 1982) explored secondary structure packing by a combinatorial assembly approach. Chou *et al.* (1985, 1986) used an all-atom model with the ECEPP potential function to study secondary structure packing patterns, but explored conformations only in the vicinity of the starting structures. Recently Taylor (1991) has folded all-a-helical proteins by first predicting the secondary structure from multiply aligned sequences using pattern matching methods, then assembling the tertiary structures by a combinatorial process. Smith-Brown *et al.* (1993) folded several proteins based on their known secondary structures plus additional distance constraints among the secondary structural elements. Monge *et al.* (1994) have predicted structures of some 4-helix bundle proteins using a contact potential. Contact potentials are matrices containing 210 different energies, representing all the possible pairings of amino acids observed in a structural database (Miyazawa & Jernigan, 1985, 1993).

Many authors (Tanaka & Scheraga, 1976; Crippen & Viswanadhan, 1984, 1985; Maiorov and Crippen, 1992; Wilson and Doniach, 1989; Casari and Sippl, 1992; Sun, 1992, 1993, 1995; Bryant and Lawrence, 1993; Skolnick *et al.*, 1993) have developed empirical potentials based on the information of the available structural database, requiring many parameters. Here we find that a very simple semi-physical potential, in a constrained conformational search, works about as well as other methods and applies to a somewhat broader set of protein structures than the 4-helix bundles on which such algorithms often succeed. The search uses knowledge of the native secondary structures. The simple potential replaces the 210 or more parameters with a single hydrophobic interaction energy and a hydrogen bond term. Our premise is that hydrophobic interactions are important (Kauzmann, 1959; Dill, 1990). Simplified models suggest that the 3-dimensional structure of a native protein may be largely encoded within just the sequence in which the hydrophobic (H) and polar (P) amino acids are arranged (reviewed in Dill *et al.*, 1995). If so, a computer algorithm should be able to predict the general tertiary fold of a protein if it could explore conformational space sufficiently broadly with a simple potential function that prevents steric overlaps and maximizes the number of HH contacts.

Conformational search methods are not yet fast enough to find global optima of long chains in real-space representations using physical potential functions. Therefore, extra constraints, in the form of external knowledge or additional terms in the potential function, are currently needed to find these states. Here we assume that secondary structures are known. It is known, however, that even complete knowledge of secondary structure is not sufficient in itself to predict tertiary structures (Havel *et al.*, 1979; Momany *et al.*, 1975). Havel *et al.* (1979) showed that with the exact locations of the α–helices and β–strands, the average distance matrix error (DME) for structures generated by a distance geometry-based method is about 7Å, which is near the average DME of random structures, around 7.1Å. Nevertheless, knowledge of secondary structures puts severe constraints on

the conformational search, and makes it tractable. Given the correct information of secondary structures, is a simple potential based on hydrophobic interactions and hydrogen bonding sufficient to find the native-like chain folds?

Materials and Methods

The Chain Representation and Potential Function

We use a simplified geometric representation of a protein (Wilson & Doniach, 1989; Sun *et al.*, 1992; Sun, 1993, 1995). All backbone bond lengths and angles have their ideal values (Corey & Pauling, 1954). All the peptide bond dihedral angles are fixed in the *trans* (ω = 180°) conformation. A single virtual atom is used to represent each sidechain at the average position of the heavy atoms in the sidechain (the sidechain centroid). Since this position can vary for most amino acids depending on the sidechain rotamer, we fix the position to that of an average rotamer observed for that amino acid type in the PDB. The geometric variables which determine a protein conformation in this reduced representation are the backbone dihedral angles ϕ and ψ. The (ϕ, ψ) dihedral angles are constrained to be (-63.8°, -41.0°) and (-140.0°, 120.0°) for residues in the α–helical and β–strand conformations respectively where the secondary structure information is known in primary sequences. Conformational searching is performed over the remaining dihedral angles; i.e., those not in helices or strands in the known native structure.

Our potential function has two terms: a hydrophobic interaction and an excluded volume term: $E_{total} = E_{HH} + E_{ex}$. We classify the twenty amino acids into two groups.

The hydrophobic (H) residues are Ala, Cys, Ile, Leu, Met, Phe, Trp, Tyr, Val, and the hydrophilic (P) residues include Arg, Asn, Asp, Gln, Glu, Gly, His, Lys, Pro, Ser, Thr. This classification is based on the Miyazawa-Jernigan (1985) scale. We have not tried the many other similar classifications (Cornette *et al.*, 1987).

The hydrophobic interaction energy term is a short-ranged soft potential defined as:

$$E_{HH} = 1.0 \sum_{i} \sum_{j=2} \varepsilon_{ij} f(d_{ij})$$

$$f(d_{ij}) = \frac{1}{1 + \exp\left[\frac{(d_{ij} - d_0)}{d_t}\right]}$$

where ε_{ij}, the binary amino acid hydrophobic interaction coefficient, equals -1.0 if both i and j are hydrophobic residues, and equals zero otherwise. d_{ij} is the distance between the centers of sidechain centroids of i and j. The function $f(d_{ij})$ is an empirical distance dependence for the pairwise hydrophobic interactions. When $d_t = 0$, $f(d_{ij})$ is a step function (equal to one for distances less than d_0 and zero for greater separations), which is commonly used to evaluate non-bonded interactions in empirical potentials (following Miyazawa & Jernigan, 1985, for example) . Here we choose d_0 = 6.5Å, and d_t = 2.5Å. Thus the interaction function is sigmoidal rather than step-like, and perhaps more physical (Figure 1a). Beyond a cutoff distance of 12Å the hydrophobic interaction is set to zero.

The excluded volume term is represented by a soft sigmoidal potential:

$$E_{ex} = const \sum_{ij} \frac{1}{1 + \exp\left[\frac{(d_{ij} - d_{eff})}{d_w}\right]}$$

where d_{ij} is the distance between two C^{α} atoms or two sidechain centroids, d_w is 0.1Å, d_{eff} = 3.6Å for C^{α} atoms and 3.2Å for sidechain centroids (Figure 1b). The constant factor was chosen to be 5; this determines the hardness of the spheres in the excluded volume interaction. The model is not sensitive to this quantity, provided that it is considerably

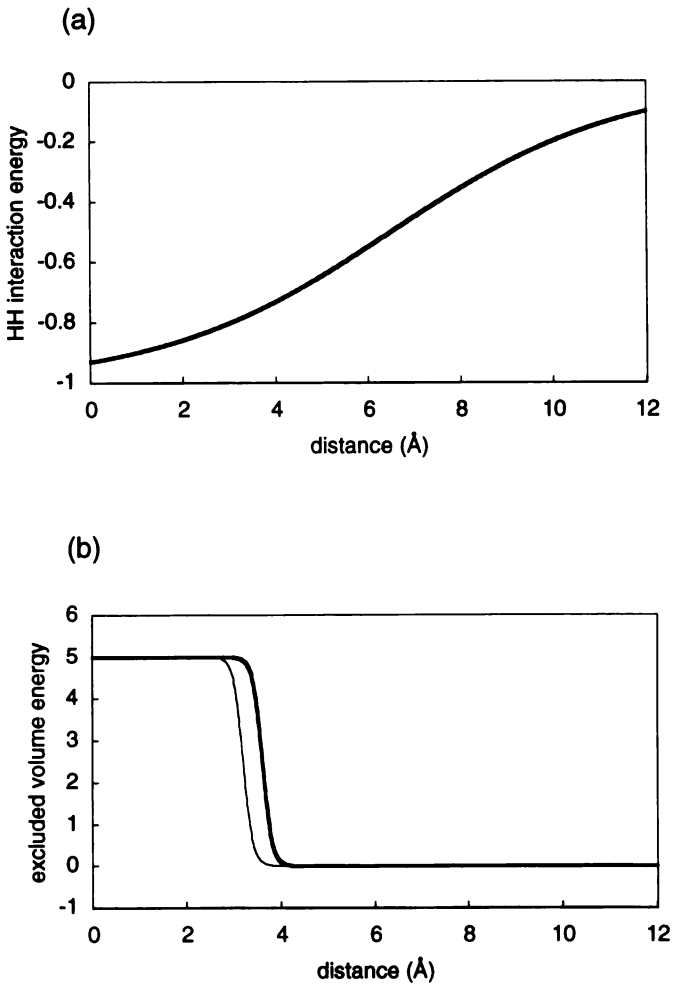


Figure 1. Interaction potential functions. a) Distance-dependent interaction between sidechain centroids of hydrophobic residues, and b) Excluded volume interactions between all sidechain centroids (lighter line) and between all α–carbons (heavier line).

larger than the interaction between a pair of hydrophobic residues. We found that the final structures usually have less than one excluded volume violation. The use of a plateau potential for the excluded volume interaction has the advantage that a given structure will not be rejected for a few excluded volume violations if it is favored by many hydrophobic interactions.

We also used an explicit step-function-like hydrogen bonding term in the protein backbone for β–sheet structures. The conditions for the hydrogen bonding are: 1) the distance between the hydrogen bonding atoms O and H is less than 2.5Å; 2) the angle between N—H—O is 120°-180°. An energy advantage of -1.0 units is given for each hydrogen bond that satisfies the above conditions. This is roughly equivalent to one hydrophobic contact. This term contributes only for sheet proteins since helices are already fixed as secondary structures. This term is therefore only relevant for three proteins we study here: crambin, the zinc finger motif and G protein.

To assess our predictions, we use both the root mean square error (RMS) and the distance matrix error (DME) to compare structures. They are defined as:

$$RMS = \left\{ \frac{1}{N} \sum_{i}^{N} (r_i - r_i^c)^2 \right\}^{1/2}$$

$$DME = \left\{ \frac{2}{N(N-1)} \sum_{ij}^{N} \left(r_{ij} - r_{ij}^c \right)^2 \right\}^{1/2}$$

where the superscript c indicates the "right answer" target conformation to which a structure is compared, either the crystallographic or the NMR native structure. Distance matrix error is usually a better measure of the overall similarity between two protein structures, although as one counterexample, the DME between a structure and its mirror image is zero, while their RMS difference can be large. Following the most common conventions, in this paper the RMS is calculated over all backbone atoms, and DME is over all C^{α} distances. We also count the total number of monomer-monomer contacts, which describes chain compactness: it is the number of pairs of residues whose C^{α} atoms are separated by less than 10.0Å. A smaller cut-off number could be used, say 6.5Å, which roughly equals the radius of the first shell of residue packing (Miyazawa & Jernigan, 1985), but the larger cut-off usefully reflects additional information about the chain compactness.

Conformational Searching

Following earlier treatments (Tuffrey *et al.*, 1991; Blommers *et al.*, 1992; Dandekar & Argos, 1992; Sun, 1993; Bowie & Eisenberg, 1994) we use a genetic algorithm (Holland, 1975; Goldberg, 1989) to search chain conformations to find the low energy states. The "genome" of a protein is a string of paired (ϕ, ψ) angles representing the conformation of each residue. A genetic algorithm run is performed in two steps, each of which contains a mutation operation, a crossover operation, and a selection operation.

The first step searches all of the available conformational space. The starting population is generated by fixing all (ϕ, ψ) angles of residues in the given secondary structure segments at their ideal values. The remaining, free (ϕ, ψ) angles are randomly selected from a dihedral dictionary of mono- and dipeptide conformations. This dictionary is the same one used by Sun (1993), and is compiled from known protein structures. It is edited to contain only (ϕ, ψ) pairs that satisfy Ramachandran map steric constraints. The mutation operation consists of replacing a single peptide unit of a protein with a different conformation randomly selected from the dihedral dictionary.

The second step is a more local search of conformational space. It starts with a homogeneous population: identical copies of the lowest energy structure found in the first step. The mutation operation now consists of perturbing the (ϕ, ψ) angles of a randomly chosen free residue, by a random angle between -5 and 5 degrees (if the perturbed (ϕ, ψ) angles are allowed within the Ramachandran map). We find this second step to be crucial in finding low energy states within the model, especially for structures having many secondary structure segments. It allows a conformation to adjust within a given area of conformational space, to probe the depth of the local energy minimum found in the first step.

For both steps, the number of mutation operations per chain (N) decreases as the search proceeds. We used an exponential function in which N depends on the number of generations over which the population has evolved n_{gen}: $N = 1 + Max \times \exp(-n_{gen}/n_{eff})$. We chose the values of Max and n_{eff} empirically to improve the conformational sampling; we selected values such that the GA search would not converge either prematurely (e.g. before the chain could become compact), or while trying too many simultaneous mutations (i.e. while $N > 1$). For most of the proteins in this paper, we set Max to 5 for both steps, and n_{eff} to 60 in the first step and 30 in the second step. The number of crossover sites per chain was a constant, set to 1. Of course, mutation and crossover sites are chosen only among residues which are not in secondary structures. The selection operation propagates only the 200 lowest-energy conformations into the next generation, from a total population of 200 parent conformations, 400 mutated offspring and 200 crossed offspring. We performed 100 separate genetic algorithm runs for each protein

Results

Using the algorithm described above, we search to find the lowest energy states of 10 small known proteins, according to our simple potential function. The model predicts 7 out of the 10 folds reasonably well. In addition to these 10 proteins of known structure, we predict the helix-packing topology of a putative 4-helix bundle designed by Kamtekar *et al.* (1993). For the proteins of known structure, Table 1 gives a summary of the energies for the Brookhaven Protein Data Bank (PDB) structure (as determined by x-ray crystallography or NMR spectroscopy) and the lowest-energy structures computed by the search strategy. It also gives the energy of the lowest-energy "model-native" conformation.

Table 1. Summary for computed proteins

Protein	$E_{c\alpha}$	E_{side}	E_{HH}	E_{HB}	E_{tot}	RMS(Å)	DME(Å)	cont	R_g(Å)
1ROP									
xtal	0.0	0.0	-36.9	0	-36.9	0.0	0.0	348	13.2
model-native	0.0	4.1	-38.8	0	-34.7	1.75	1.36	375	13.1
GA	0.0	0.7	-37.5	0	-36.8	1.68	1.65	341	13.2
1PPT									
xtal	0.0	0.0	-11.0	0	-11.0	0.0	0.0	180	10.7
model-native	0.0	0.7	-12.9	0	-12.2	2.60	1.08	201	10.6
GA	0.1	0.1	-15.1	0	-14.9	6.77	3.76	157	11.4
7ZNF									
NMR	0.0	0.0	-7.9	-2	-9.9	0.0	0.0	138	9.1
model-native	0.1	0.3	-6.9	0	-6.5	2.86	2.04	140	9.1
GA	0.0	0.1	-9.3	-1	-10.2	2.54	1.73	130	9.3
1HDD									
xtal	0.0	0.5	-25.0	0	-24.5	0.0	0.0	288	11.0
model-native	0.0	0.5	-27.1	0	-26.6	3.63	2.74	313	10.9
GA1	0.0	0.6	-30.6	0	-30.0	4.62	2.95	323	11.4
GA2	0.0	0.3	-30.2	0	-29.9	3.25	2.46	314	11.0
1R69									
xtal	0.0	0.0	-49.1	0	-49.1	0.0	0.0	407	10.1
model-native	0.5	1.0	-49.5	0	-48.0	2.37	2.06	501	9.7
GA	0.2	0.5	-52.8	0	-52.1	10.11	5.55	504	10.4
1CRN									
xtal	0.1	0.1	-33.8	-4	-37.6	0.0	0.0	289	9.7
model-native	0.8	4.3	-47.1	-2	-44.0	4.67	3.02	358	8.6
GA	0.6	0.8	-50.6	0	-49.2	3.00	2.56	355	8.6
GA, no S-S	0.5	0.4	-54.2	0	-53.3	9.56	4.87	413	8.1
1LE2									
xtal	0.0	0.3	-130.9	0	-130.6	0.0	0.0	923	18.2
model-native	1.4	11.6	-118.2	0	-105.2	6.73	5.61	1048	18.8
GA	0.3	2.3	-112.7	0	-110.1	10.74	5.66	1011	18.4
256B									
xtal	0.0	0.0	-98.1	0	-98.1	0.0	0.0	752	14.2
model-native	0.3	9.1	-113.3	0	-103.9	2.07	1.58	901	13.7
GA	0.2	4.4	-106.8	0	-102.2	4.34	2.43	876	13.9
2GB1									
NMR	0.0	0.0	-36.7	-13	-49.7	0.0	0.0	349	10.1
model-native	0.6	0.7	-34.6	-12	-45.3	7.14	6.02	320	12.3
GA	0.8	1.1	-44.4	-4	-46.5	6.47	4.72	347	11.0
1BBL									
NMR	0.1	0.0	-17.8	0	-17.7	0.0	0.0	171	9.4
model-native	0.2	0.2	-30.1	0	-29.7	3.33	2.89	280	7.9
GA	0.1	0.6	-36.4	0	-35.7	5.44	4.16	265	7.7

The model-native conformation is the closest we can come to reproducing the PDB structure within our simplified model. It is the lowest point of the local minimum near the experimentally determined conformation, using the same constraints as the model (i.e. ideal peptide geometry, average virtual-atom sidechain rotamers, fixed ideal secondary structure dihedrals, simple potential function). Model-native structures are found by means of a genetic algorithm search identical to the second step of the search outlined above, except that in addition to the simple potential described above, free dihedrals are restrained by a harmonic potential. Dihedral angles for secondary structures are fixed at their ideal values. Free dihedrals begin in the crystal or NMR structure conformation. The energetic penalty for non-native dihedrals is 0 for changes of less than 15 degrees in either ϕ or ψ, and 0.001 deg^{-2} for additional changes (a harmonic flat well potential). The model-native conformations enable us to assess how our many approximations may affect our predictions.

1. Repressor of primer (1ROP)

ROP is a small 4-helix bundle protein that is a dimer in the crystal structure. We compute the conformation of a monomer in isolation. Each monomer has 56 residues and forms a helix-turn-helix conformation (Banner *et al.*, 1987). This is the simplest conformational search, with only 12 angular degrees of freedom (residues 1-2, 29-31, 56). The computed structure is similar to the crystal structure (2.20Å RMS), and the energy of the two structures is approximately equal (Table 1). The helices pack with approximately the same angle as seen in the crystal structure. Figure 2 compares the conformations of the crystal and computed structures.

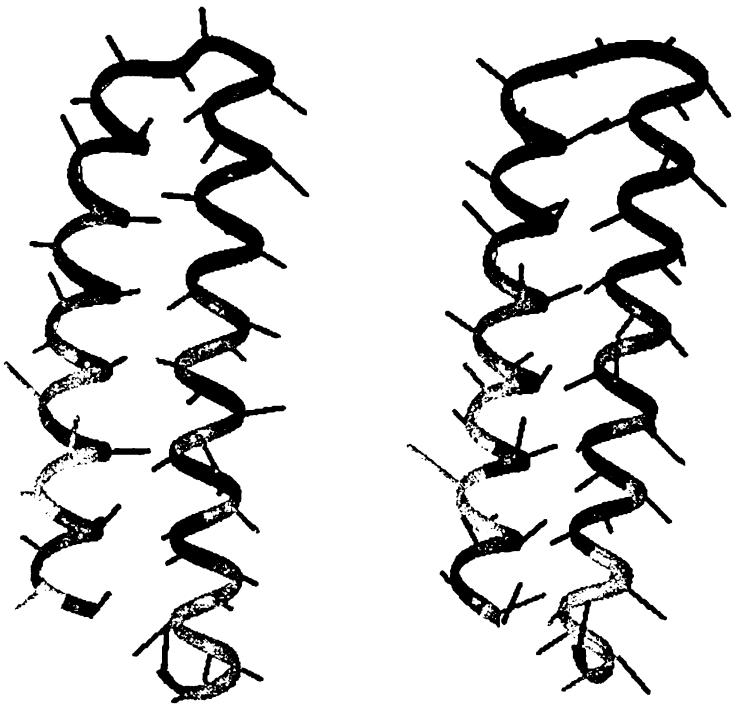


Figure 2. Structure comparison between the crystal structure ("native") and GA structure ("predicted") of ROP. H residues are in black; P residues are in gray. Protein backbones shown as ribbons; virtual bonds between α–carbons and sidechain centroids shown as sticks.

2. Avian pancreatic polypeptide (1PPT)

Avian pancreatic polypeptide is a 36-residue protein with a long helical segment near the C-terminal end and a polyproline helix at the N-terminus (Glover *et al.*, 1983). We have fixed residues 14-32 in the C-terminal helix; all other angles are degrees of freedom. The computed structure is less compact than both the crystal structure and the model-native structure (see Table 1 and Figure 3). The six N-terminal residues of the computed structure project away from the helix rather than packing against it, as in the crystal structure. These six residues are all polar residues (according to our classification

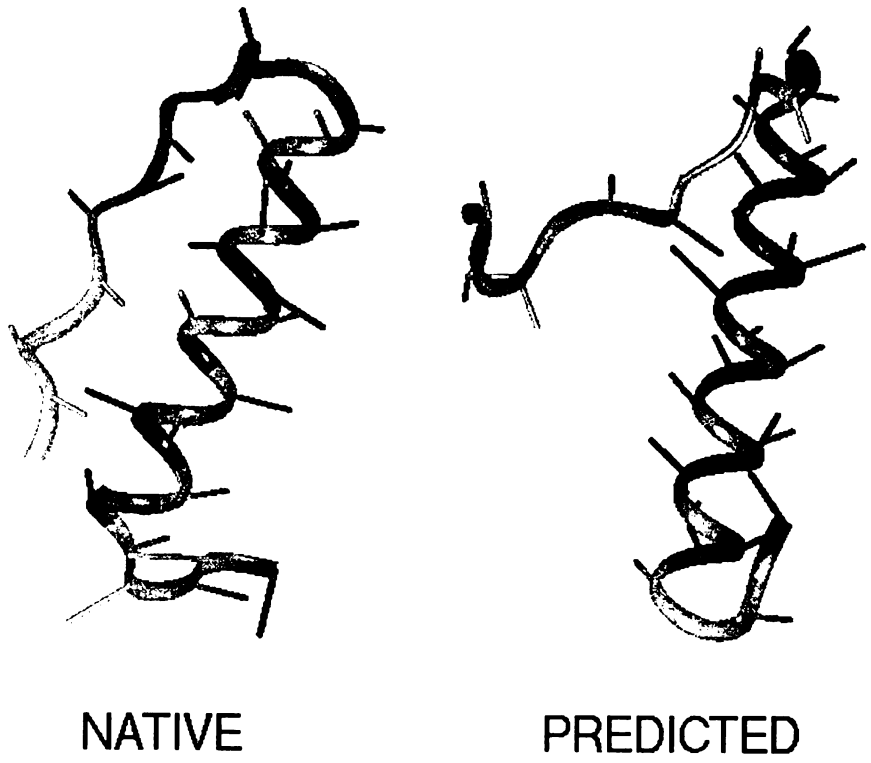


Figure 3. Structure comparison between the crystal structure ("native") and GA structure ("predicted") of Avian pancreatic polypeptide inhibitor.

above), so for our simple potential function there is no attraction between this segment and the rest of the molecule. Not including this segment, the structures match to 3.78Å RMS (residues 7-36). The computed structure is considerably lower in energy than both the crystal and model-native structures (by approximately 4 HH contacts), due primarily to the close packing of the four C-terminal residues against the helix. The four C-terminal residues are very flexible in the crystal structure; not including these residues, the model-native and computed structures are isoenergetic according to our simple potential function.

Figure 4. Structure comparison between the NMR structure ("native") and GA structure ("predicted") of the zinc finger motif.

3. Zinc finger (7ZNF)

According to its NMR structure (Kochoyan *et al.*, 1991) the consensus 30-residue zinc finger motif is an α–helix (residues 15-27) packed against a two-stranded antiparallel β–sheet (residues 2-4 and 10-12). We fix those angles, and allow all others to be degrees of freedom. Figure 4 shows one of the NMR structures, and the lowest energy computed structure. The computed structure is similar to the NMR structure (2.54Å RMS), and forms a hydrogen bond between the strands. This hydrogen bond is between residues 3 and 11, while the NMR structure pairs residues 3 and 12. In the native protein, the zinc ion is required for stability-- the ion is tetrahedrally coordinated by two cysteines in the

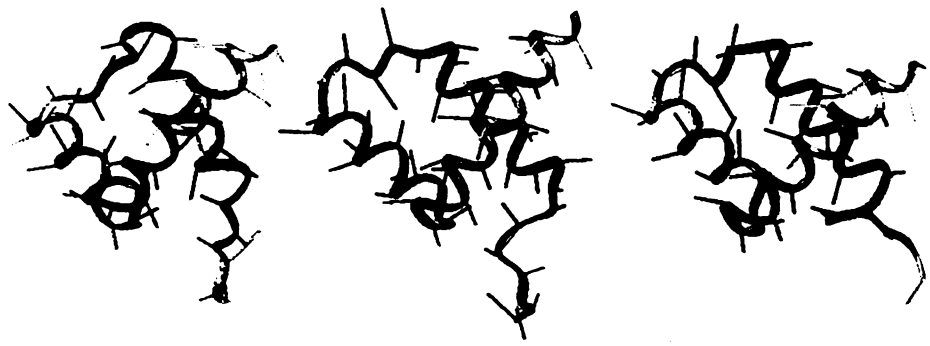


Figure 5. Structure comparison between the crystal structure ("native") and GA structures ("predicted") of Engrailed homeodomain.

sheet and two histidines in the helix. In our computed structure, hydrophobic interactions alone pack the helix against the sheet (primarily between Tyr 3, Cys 5, Tyr 7, Leu 18 and Ile 22); no zinc ion is present in the computation.

4. Engrailed homeodomain (1HDD)

The crystal structure of engrailed homeodomain (Kissinger *et al.*, 1990) has 57 ordered residues and contains three α–helices (residues 8-20, 26-36 and 40-56) packed against each other (Figure 5). For this protein, the search found two slightly different conformations of approximately equal energy (Figure 5). Both computed structures pack the three helices in a manner similar to the crystal structure, but differ primarily in the placement of the N-terminus. In the second predicted structure the N-terminus extends toward the third helix, as in the crystal structure. In the first predicted structure it extends toward the second helix instead. Not including the N-terminal segment (i.e. over residues 8-57), the structures match the crystal structure with an RMS of 3.38Å (Figure 5 middle)

Figure 6. Structure comparison between the crystal structure ("native") and GA structure ("predicted") of 434 repressor N-terminal domain.

and 2.63Å (Figure 5 right). The computed structures are lower in energy than both the crystal and the model-native structures (Table 1).

5. Amino-terminal domain of the 434 repressor (1R69)

Bacteriophage 434 Repressor contains 236 amino acids. It folds into two nearly equal-sized domains connected by a string of 40 amino acids. The crystal structure of the proteolytically cleaved amino-terminal domain (Mondragon *et al.*, 1989) has 63 ordered residues and is composed of five helical segments (residues 2-12, 17-24, 28-35, 45-51 and 56-61). This protein was the only one in our test set for which we found a conformation that is both lower in energy than the crystal structure and a completely different folding topology. Figure 6 shows that the computed structure is packed very differently than the

crystal structure, but nevertheless has good hydrophobic interactions. The computed structure is also lower in energy than the model-native structure. The computed structure, however, is less spherical than the crystal structure and has a larger radius of gyration (Table 1). Nevertheless, it has more contacts than the crystal structure, due to closer packing allowed by our crude excluded volume potential.

Analysis of the computed structure reveals that 7 hydrophobic sidechain centroids (Ile 2, Ile 11, Leu 13, Leu 15, Ala 18, Ala 51 and Trp 58) are quite exposed to solvent. By comparison, the model-native conformation has 8 exposed hydrophobic sidechain centroids (Ile 11, Leu 13, Leu 15, Ala 18, Ile 31, Phe 44, Val 56 and Trp 58), which is consistent with its smaller HH interaction energy. However, in the all-atom crystal structure of 1R69, only four hydrophobic sidechains are significantly solvent-exposed (Ile 11, Leu 13, Ala 18 and Val 56). Thus four hydrophobic residues which are not exposed in the crystal structure appear exposed in the model-native conformation, even though these structures differ by only 2.37Å RMS. We conclude that for this protein, the reduced-representation model (representing each sidechain as only a single fixed virtual atom) may be inadequate.

6. Crambin (1CRN)

Native crambin has 46 residues; Figure 7 shows the crystal structure (Teeter & Hendrickson, 1979). It has three disulfide bonds, between residues 3-40, 4-32, and 16-26. The secondary structure comprises two α–helical segments (residues 7-19 and 23-29) and a two-stranded antiparallel β– sheet (residues 1-4 and 33-35). In addition to the hydrophobic and hydrogen-bonding interactions, we included a harmonic potential to mimic the disulfide bonds. The spring constant was set to 0.33Å^{-2}, and the equilibrium distance between cysteine sidechain centroids was set to 2.7Å . Because the disulfide bond distance is considerably shorter than the excluded volume distance for two cysteine

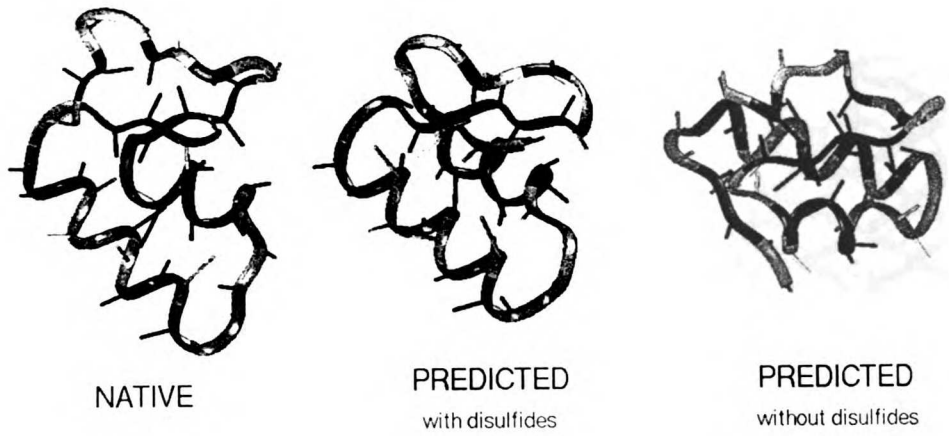


Figure 7. Structure comparison between the crystal structure ("native") and GA structures ("predicted") of crambin computed with and without the disulfide potential.

sidechain centroids (Figure 1b), no excluded volume penalty was given to overlapping cysteine sidechains.

Figure 7 compares the crystal and computed structures. While the structure computed with disulfides is significantly lower in energy than the crystal structure due to tighter packing (Table 1), the structures are similar (3.00Å RMS). They differ primarily in the conformation of the C-terminus. The antiparallel sheet hydrogen bonds do not form in the computed structure, although the strands come in close contact. When the disulfide interaction is not included, the search strategy finds even lower energy conformations, which do not resemble the crystal structure (Table 1, Figure 7). This is not surprising, since crambin has such a high content of hydrophobic residues, and our simple potential function is based primarily on maximizing contacts between hydrophobic residues. Table 1 shows that the lower-energy conformations have more contacts and a smaller radius of gyration.

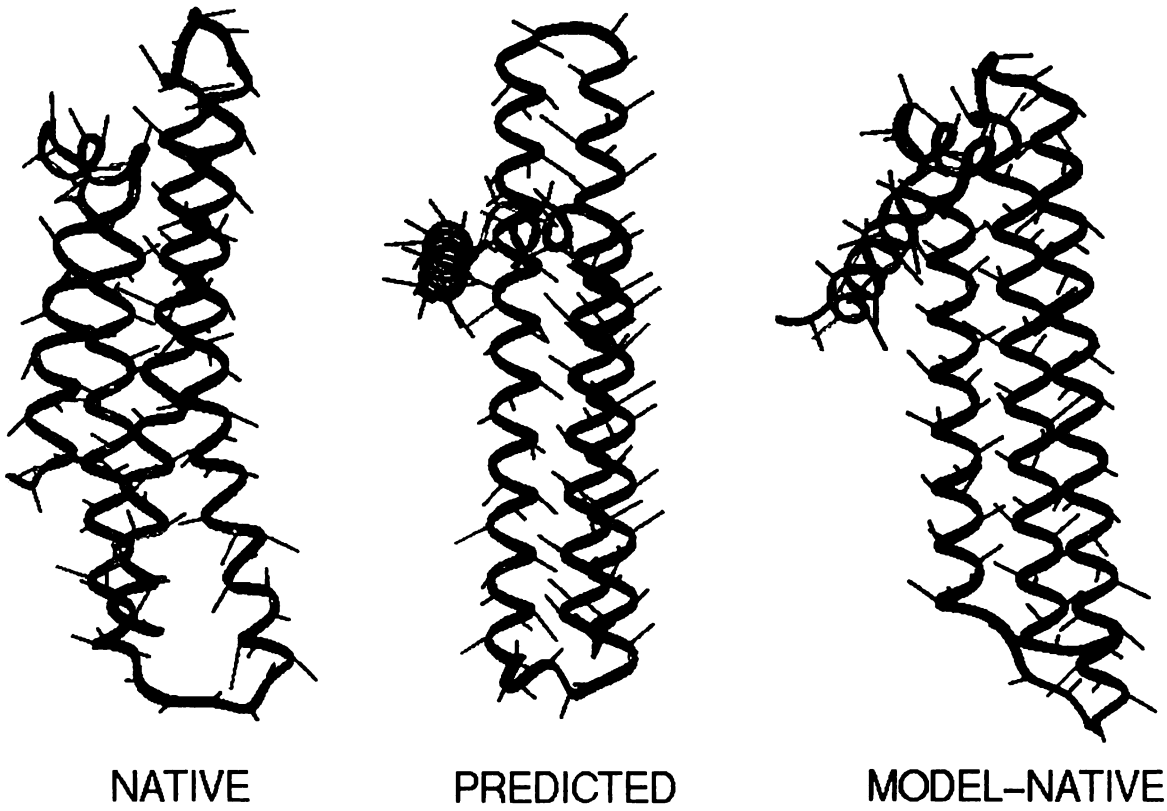


Figure 8. Structure comparison between the crystal structure ("native"), the GA structure ("predicted"), and the model-native structure of Apolipoprotein E.

7. Apolipoprotein E (1LE2)

Apolipoprotein E is a 4-helix bundle protein of 144 amino acids. Its structure comprises one short helical segment (residues 23-29) and 4 long helices (residues 3-20, 33-56, 68-103 and 108-142). The native crystal structure of Wilson *et al.* (1991) is shown in Figure 8. This protein posed some problems for our conformational search algorithm. There are five fixed secondary structure segments, four of which are very long, so manipulating them with mutations in (ϕ, ψ) space is difficult. A small change in (ϕ, ψ) space leads to a large change in relative positions of the ends of the helices. The search

strategy was unable to find any conformations of energy comparable to the crystal structure. The crystal structure is much lower in energy, by our simple potential function, than any of the searched conformations. Figure 8 shows the lowest energy computed structure, which is essentially a 3-helix bundle of the longest helices (the three C-terminal helices which we number 3, 4 and 5), in which the positions of the third and fourth helices are interchanged relative to the crystal structure. Interestingly, the model-native structure is also a 3-helix bundle of the C-terminal helices, but has the crystal structure topology for those three helices (Figure 8). The starting structure for the model-native search (ideal helices plus loop (ϕ, ψ) angles in the crystal conformation) is quite extended relative to the crystal structure; apparently the starting structure is outside the radius of convergence of the local conformational search. The failure of even this local conformational search to converge on the 4-helix bundle topology underscores the weakness of the conformational search when confronted with long elements of fixed secondary structure.

8. Cytochrome b_{562} (256B)

Cytochrome b_{562} is a 4-helix bundle protein (Matthews *et al.*, 1979) with 106 amino acids (Figure 9). We fixed only the four long helical segments (residues 3-19, 22-40, 55-80 and 84-105), leaving the other residues free (including a small helical fragment from residues 45-47). Figure 9 compares the computed structure to the crystal structure. The structures are similar (4.34Å RMS), although the computed structure is slightly more compact and lower in energy (Table 1). The computed structure correctly forms a small helix from residues 46 to 51, in approximately the same place as the crystal structure. The model-native structure is even lower in energy than the best computed structure and is close to the crystal structure (2.07Å RMS). Thus for this protein, even within our simple model, there is a structure of lower energy than our computed structure which is even more similar

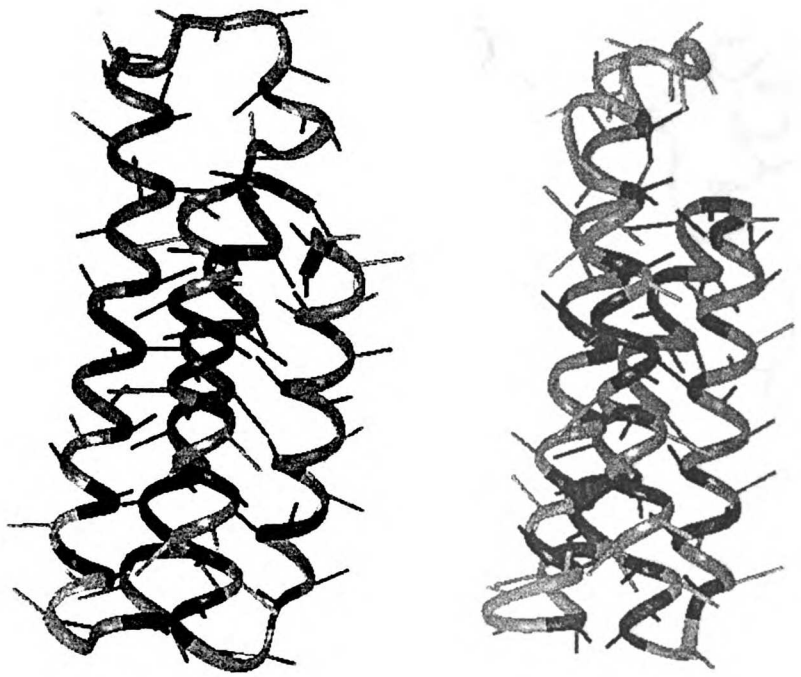


Figure 9. Structure comparison between the crystal structure ("native") and GA structure ("predicted") of cytochrome b_{562}.

to the crystal structure. A better conformational search strategy may be capable of finding such a structure using the present potential function.

9. B1 domain of protein G (2GB1)

The structure of the 56-residue immunoglobulin binding domain of protein G has been determined in solution by NMR spectroscopy (Gronenborn *et al.*, 1991). It has a 4-strand β–sheet (residues 1-7, 14-20, 42-46 and 51-55) packed against an α–helix (residues 23-35) (Figure 10). We included the hydrogen bonding interaction potential for this protein since it has b strands. Figure 10 shows one of the NMR structures, the lowest

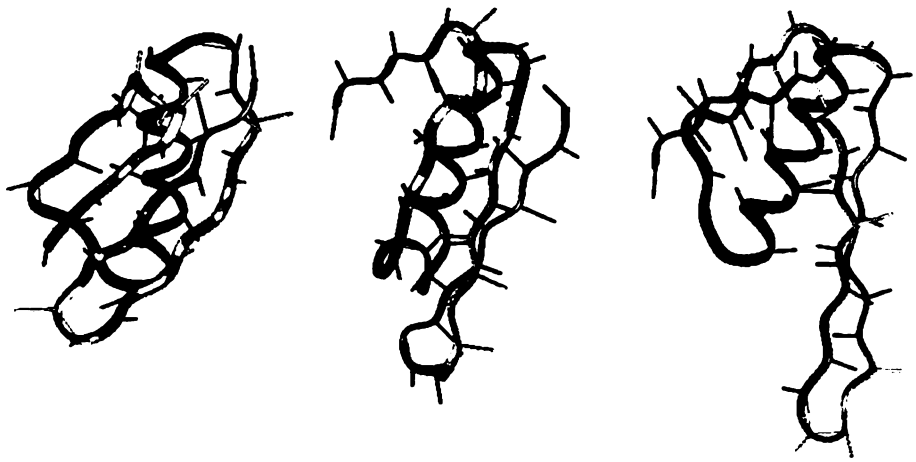


Figure 10. Structure comparison between the NMR structure ("native"), the GA structure ("predicted"), and the model-native structure of the B1 domain of Protein G.

energy computed structure, and the lowest energy model-native structure. The general folding pattern of the two structures is the same, with an approximate four-strand sheet packed against a helix. However, the strands are in different positions in the sheet. In the crystal structure, the strands are aligned in the order 2-1-4-3 (numbering the strands from N- to C-terminus), whereas in the computed structure the alignment is 1-2-3-4. Each β–hairpin has the wrong handedness. The problem here is not the simple potential function, but our simple representation of ideal secondary structures. The model-native structure shows clearly that the hydrogen-bonded ideal β–hairpins have a twist such that the strands tend to align 1-2-3-4 (Figure 10 right). The NMR structure has several strand dihedrals that vary by over 45 degrees from the ideal values (Tyr 3, Leu 5, Leu 7, Gly 14, Glu 42, Asp 46 and Phe 52), which allow the opposite twist. Because of the constraint of ideal sheet dihedrals, the model-native conformation is actually less similar to the NMR structure than is the computed structure (7.14Å RMS vs. 6.47); the two β-hairpins cannot form a

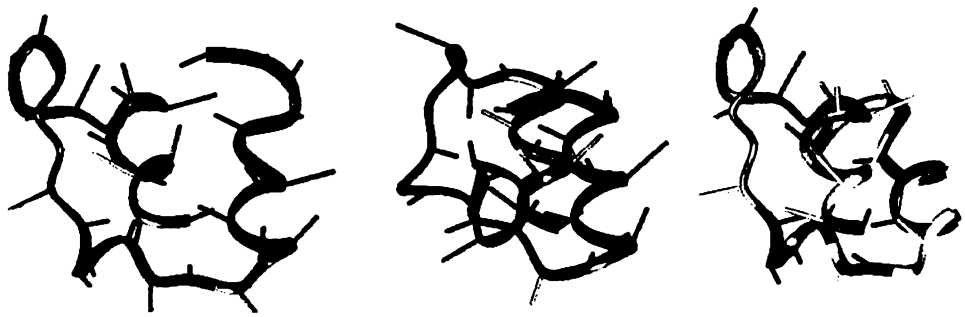
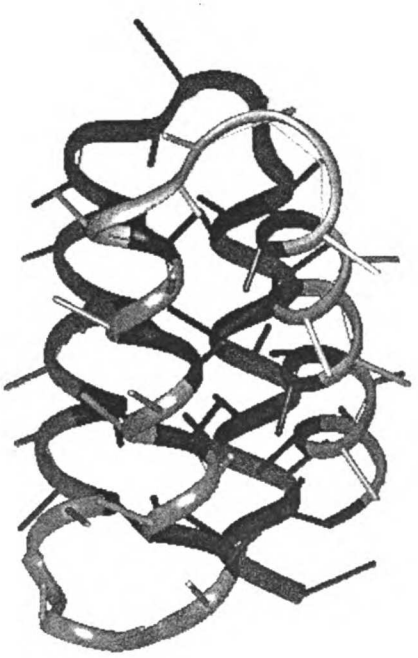


Figure 11. Structure comparison between the NMR structure ("native"), the GA structure ("predicted"), and the model-native structure of E3-Binding Protein.

single sheet. Thus, within the crude model, the search finds a way to form a 4-stranded sheet packed against the helix when the native topology cannot.

10. E3-binding domain (1BBL)

The NMR structure of a 51-residue peptide comprising the E3-binding domain of the E2 core of *E. coli* 2-oxoglutarate dehydrogenase multienzyme complex (Robien *et al.*, 1992) consists of two parallel α–helices (residues 14-22 and 41-48) connected by a segment containing a short extended strand (residues 25-26), a turn (residues 27-29), and a relatively long loop (residues 31-40). The 11 N-terminal and 3 C-terminal residues are disordered. We calculated the structure of the 39-residue ordered segment (residues 12-48), fixing only the α–helices. The computed structure is similar to the NMR structure (5.44Å RMS). There is a short strand (residues 23-25) and a turn (residues 28-30) in nearly the same positions as in the NMR structure. The two helices line up parallel to each other in roughly the correct orientation. The computed structure, however, is much more



PREDICTED

Figure 12. Predicted structure for the 4-helix bundle from Kamtekar *et al.* (1993).

compact and is much lower in energy than the NMR structure (Table 1). The calculated structure is even somewhat lower in energy than the model-native structure. The model-native structure has about the same DME compared to both the NMR structure and the computed GA structure (2.89Å to NMR vs. 2.97Å to GA). This result emphasizes the similarity between the calculated structure and the NMR structure, since the model-native structure is essentially a more compact version of the NMR structure (Figure 11).

11. Kamtekar *et al.* (1993) 4-helix bundle

For the 10 proteins above, we had a target native structure for comparison. Here we describe a test of a putative 4-helix bundle, for which the true native structure is not yet known. Kamtekar *et al.* (1993) genetically constructed several 4-helix bundles using a

binary code of just hydrophobic and polar monomers of different types. The structures are not yet known for the sequences that were obtained, but those structures have circular dichroism spectra consistent with 4-helix bundles. We studied one of their 76-residue sequences. We fixed the helical segments as designed by Kamtekar *et al.* (1993). Our computed structures show a 4-helix bundle shape with an overall twist among the packed helical segments. The computed structure is a bundle of antiparallel helices with a left-handed overall folding pattern among the helical segments. Figure 12 shows the lowest energy computed structure.

Conclusions

The main point of this paper is the suggestion that the hydrophobic, hydrogen-bonding and steric interactions that drive protein folding might ultimately be captured in exceedingly simple potential functions. We have described a simple computer algorithm that seeks the tertiary folds of small globular proteins based on knowledge of the sequence of hydrophobic and polar amino acids, the given native disulfide bonds, and the given α–helices and β–strands. The conformational space is explored by a genetic algorithm, of a simple potential function having very few parameters. Because our treatment assumes known secondary structure information, our potential function contains no secondary structure potential. The binary hydrophobic interaction and the excluded volume interaction, which constitute the current potential function, are not sufficient to be used for a full conformational search. Such searches tend to find structures that are more compact than native conformations, and with less secondary structure.

It could be argued that the level of success of this method is commensurate with that of many other folding algorithms -- neither considerably better nor considerably worse. It

tends toward the right general chain fold for 7 of our 10 test proteins. But we believe this algorithm has two general virtues. First, it is much simpler. There are essentially only three interaction energies: a hydrophobic interaction, a hydrogen bond, and an excluded volume interaction. We also include a disulfide bond term where appropriate. Some other potential functions use hundreds to thousands of energy parameters. Second, the potential function parameters are not derived from known protein structures (although the search method is directed by local conformational propensities taken from the PDB). The terms in the potential function were chosen to represent physical interactions explicitly and simply. We find that using the set of 210 database-derived contact energies of Miyazawa and Jernigan (1985) does not improve the results we describe here using only a simple HH interaction. We also find that using a step function for the HH interaction (with 6.5Å cutoff) does not perform as well as the distance dependent HH interaction described above.

Arguably, our simple potential function and reduced atomic representation are inadequate for only two of the 10 proteins we tested. The computed structure of 1PPT shows that some sort of additional interaction is required to bring the N-terminal extended polyproline-like helix against the C-terminal α–helix. The computed structure of 1R69 suggests that our simplified model, having only a single fixed virtual atom to represent each sidechain, may not always give a sufficiently detailed description of the protein. For the other proteins in our test set, the simple potential function favors the native topology despite our simplified protein representation.

In most cases, we were able to find structures of lower energy than the actual PDB structure, primarily because our crude excluded volume term allows structures to be more compact than the crystal structure (Table 1). Our potential function can be quite sensitive to small changes in structure, particularly changes in compactness. For example, compared to the NMR structure, the model-native structure has an RMS of only 3.3Å, yet it has less than 85% of the radius of gyration and has nearly twice the total interaction energy (Table 1).

For six of the proteins for which the computed structures were lower in energy than the PDB structures (1ROP, 7ZNF, 1HDD, 1CRN, 256B and 1BBL), the computed structure is quite similar to the PDB structure. In each of these cases the RMS error between the computed structure and the experimentally-determined structure ranges from roughly 1.7 to 5.4Å. For 7ZNF, 1HDD, 1CRN and 2GB1, the computed structure is also lower in energy than the model-native structure. Compared to the PDB structures, these computed structures have a smaller RMS and DME than the model-native structures have (Table 1). Thus in these cases, because of our approximations of ideal peptide and secondary structure geometry, a different combination of dihedrals than found in the PDB structures can actually reproduce the overall topology better.

For 2GB1, although its sheet topology differs from that of the NMR structure, the computed conformation is still similar to the NMR structure within the severe limits imposed on the conformational search. In this case, our approximation of fixing secondary structure dihedrals to canonical values prevents the conformational search from finding the native conformation. For 1LE2 the search strategy is unable to find any conformations having energies as low as the crystal structure; this protein apparently has too many long secondary structure elements for our GA search. For these two proteins, our results are clearly more limited by the search methods than by the potential function. We conclude that the present potential function may therefore be useful for testing and evaluating conformational search strategies. Perhaps with better search strategies, very simple hydrophobic and excluded volume interactions plus a simple hydrogen bonding potential will be sufficient to find the tertiary folds of proteins, at least at low resolution.

Acknowledgments

We thank the Pittsburgh Supercomputer Center for providing CPU time (grant No. DMB93005P) and NIH and ONR for funding. PDT is a Howard Hughes Medical Institute Predoctoral Fellow.

References

Banner, D.W., Kokkinidis, M. & Tsernoglou, D. (1987). Structure of the ColE1 rop protein at 1.7Å resolution. *J. Mol. Biol.* **196**, 657-675.

Blommers, M.J.J., Lucasius, C.B., Kateman, G. & Kaptein, R. (1992). Conformational analysis of a dinucleotide photodimer with the aid of the genetic algorithm. *Biopolymers* **32**, 45-52.

Bowie, J.U. & Eisenberg, D. (1994). An evolutionary approach to folding small alpha-helical proteins that uses sequence information and an empirical guiding fitness function. *Proc. Natl. Acad. Sci. USA* **91**, 4436-40.

Bryant, S.H. & Lawrence, C.E. (1993). An empirical energy function for threading protein sequence through the folding motif. *Proteins* **16**, 92-112.

Casari, G. & Sippl, M.J. (1992). Structure-derived hydrophobic potential -- hydrophobic potential derived from x-ray structures of globular proteins is able to identify native folds. *J. Mol. Biol.* **224**, 725-732.

Chothia, C. & Janin, J. (1982). Orthogonal packing of beta-pleated sheets in proteins. *Biochemistry* **21**, 3955-65.

Chou, K.C., Nemethy, G., Rumsey, S., Tuttle, R.W. & Scheraga, H.A. (1985). Interactions between an alpha-helix and a beta-sheet. Energetics of alpha/beta packing in proteins. *J. Mol. Biol.* **186**, 591-609.

Chou, K.C., Nemethy, G., Rumsey, S., Tuttle, R.W. & Scheraga, H.A. (1986). Interactions between two beta-sheets. Energetics of beta/beta packing in proteins. *J. Mol. Biol.* **188**, 641-649.

Cohen, F.E., Sternberg, M.J. & Taylor, W.R. (1980). Analysis and prediction of protein beta-sheet structures by a combinatorial approach. *Nature* **285**, 378-382.

Cohen, F.E., Sternberg, M.J. & Taylor, W.R. (1981). Analysis of the tertiary structure of protein beta-sheet sandwiches. *J. Mol. Biol.* **148**, 253-272.

Cohen, F.E., Sternberg, M.J. & Taylor, W.R. (1982). Analysis and prediction of the packing of alpha-helices against a beta-sheet in the tertiary structure of globular proteins. *J. Mol. Biol.* **156**, 821-862.

Cornette, J.L., Cease, K.B., Margalit, H., Spouge, J.L., Berzofsky, J.A. & DeLisi, C. (1987). Hydrophobicity scales and computational techniques for detecting amphipathic structures in proteins. *J. Mol. Biol.* **195**, 659-685.

Corey, R.B. & Pauling, L. (1953). Fundamental dimensions of polypeptide chains. *Proc. Roy. Soc. (London) B* **141**,10-20.

Crippen, G.M. & Viswanadhan, V.N. (1984). A potential function for conformational analysis of proteins. *Int. J. Peptide Protein Res.* **24**, 279-296.

Crippen, G.M. & Viswanadhan, V.N. (1985). Sidechain and backbone potential function for conformational analysis of proteins. *Int. J. Peptide Protein Res.* **25**, 487-509.

Dandekar, T. & Argos, P. (1992). Potential of genetic algorithms in protein folding and protein engineering simulations. *Protein Engineering* **5**, 637-645.

Dill, K.A. (1990). Dominant forces in protein folding. *Biochemistry* **29**, 7133-7155.

Dill K.A., Bromberg S., Yue K., Fiebig K., Yee D.P., Thomas P.D. & Chan H.S. (1995). Principles of protein folding: A perspective from simple exact models. *Protein Science* **4**, 561-602.

Fasman, G.D. (1989). The development of the prediction of protein structure. In *Prediction of protein structure and the principles of protein conformation* (Fasman, G.D., Ed.), pp. 193-301, Plenum Press, New York.

Glover, I., Haneef, I., Pitts, J., Wood, S., Moss, D., Tickle, I. & Blundell, T. (1983). Conformational flexibility in a small globular hormone. X-ray analysis of avian pancreatic polypeptide at 0.98-Ångstroms resolution. *Biopolymers* **22**, 293-304.

Goldberg, D. (1989). *Genetic Algorithms in Search, Optimization, and Machine Learning.* Addison-Wesley Publishing Company, Inc., Reading, Mass.

Gronenborn, A.M., Filpula, D.R., Essig, N.Z., Achari, A., Whitlow, M., Wingfield, P.T. & Clore, G.M. (1991). A novel, highly stable fold of the immunoglobulin binding domain of streptococcal protein G. *Science* **253**, 657-661.

Havel, T.F., Crippen, G.M. & Kuntz, I.D. (1979). Effects of distance constraints on macromolecular conformation. II. Simulation of experimental results and theoretical predictions. *Biopolymers* **18**, 73-81.

Holland, J.H. (1975). *Adaptation in natural and artificial systems: An introductory analysis with applications to biology, control, and artificial intelligence.* University of Michigan Press, Ann Arbor.

Janin J. & Chothia C. (1980). Packing of alpha-helices onto beta-pleated sheets and the anatomy of alpha/beta proteins. *J. Mol. Biol.* **143**, 95-128.

Kamtekar, S., Schiffer J.M., Xiong H., Babik J.M. & Hecht M.H. (1993). Protein design by binary patterning of polar and nonpolar amino acids. *Science* **262**, 1680-1685.

Kauzmann, W. (1959). Some factors in the interpretation of protein denaturation. *Adv. Protein Chem.* **14**, 1-63.

Kissinger, C.R., Sieker, L.C., Adman, E.T. & Jensen, L.H. (1990). Crystallization and preliminary X-ray diffraction studies of the engrailed homeodomain and of an engrailed

homeodomain-DNA complex. *Biochemical and Biophysical Research Communications* **171**, 257-259.

Kochoyan, M., Keutmann, H.T. & Weiss, M. (1991). Alternating zinc fingers in the human male associated protein ZFY: Refinement of the NMR structure of an even finger by selective deuterium labeling and implications for DNA recognition. *Biochemistry* **30**, 7063-7072.

Merz, K.M. & Le Grand, S.M., Eds. (1994). *The protein folding problem and tertiary structure prediction.* Birkhäuser, Boston.

Maggiora, G.M., Mao, B., Chou, K.C. & Narasimhan S.L. (1991). Theoretical and empirical approaches to protein structure prediction and analysis. *Methods of Biochemical Analysis* **35**, 1-86.

Maiorov, V.N. & Crippen, G.M. (1992). Contact potential that recognizes the correct folding of globular proteins. *J. Mol. Biol.* **227**, 876-888.

Mathews, F.S., Bethge, P.H. & Czerwinski, E.W. (1979). The structure of cytochrome b_{562} from Escherichia coli at 2.5Å resolution. *J. Biol. Chem.* **254**, 1699-1706.

Miyazawa, S. & Jernigan, R.L. (1985). Estimation of effective interresidue contact energies from protein crystal structures: quasi-chemical approximation. *Macromolecules* **18**, 534-552.

Miyazawa, S. & Jernigan, R.L. (1993). A new substitution matrix for protein sequence searches based on contact frequencies in protein structures. *Protein Engineering* **6**, 267-278.

Momany, F.A., McGuire, R.F., Burgess, A.W. & Scheraga, H.A. (1975). Energy Parameters in polypeptides VII. Geometric parameters, partial atomic charges, nonbonded interaction, hydrogen bond interactions, and intrinsic torsional potentials for natural occuring amino acids. *J. Phys. Chem.* **79**, 2361-2381.

Mondragon, A., Subbiah, S., Almo, S.C., Drottar, M. & Harrison, S.C. (1989). Structure of the amino-terminal domain of phage 434 repressor at 2.0Å resolution. *J. Mol. Biol.* **205**, 189-200.

Monge, A., Friesner, R.A. & Honig, B. (1994). An algorithm to generate low-resolution protein tertiary structures from knowledge of secondary structure. *Proc. Natl. Acad. Sci. USA* **91**, 5027-5029.

Robien, M.A., Clore, G.M., Omichinski, J.G., Perham, R.N., Appella, E., Sakaguchi, K. & Gronenborn, A.M. (1992). Three-dimensional solution structure of the E3-binding domain of the dihydrolipoamide succinyltransferase core from the 2-oxo-glutarate dehydrogenase multienzyme complex of *Escherichia coli*. *Biochemistry* **31**, 3463-3471.

Skolnick, J., Kolinski, A., Brooks, C.L. & Godzik, A. (1993). A method for predicting protein structure from sequence. *Current Biology* **3**, 414-423.

Smith-Brown, M.J., Kominos, D. & Levy, R.M. (1993). Global folding of proteins using a limited number of distance constraints. *Protein Engineering* **6**, 605-614.

Sun, S. (1993). Reduced representation model of protein structure prediction: statistical potential and genetic algorithms. *Protein Science* **2**, 762-785.

Sun, S. (1995). Reduced representation model of protein structure prediction: statistical potential and simulated annealing. *J. Theoretical Biol.* **172**, 13-32.

Sun, S., Luo, N., Ornstein, R. & Rein, R. (1992). Protein structure prediction based on statistical potential. *Biophys. J.* **62**, 104-106.

Tanaka, S. & Scheraga, H.A. (1976). Medium- and long-range interaction parameters between amino acids for predicting three-dimensional structures of proteins. *Macromolecules* **9**, 945-950.

Taylor, W.R. (1991). Towards protein tertiary fold prediction using distance and motif constraints. Protein Engineering 4, 853-870.

Teeter, M.M. & Hendrickson, W.A. (1979). Highly ordered crystals of the plant seed protein crambin. *J. Mol. Biol.* **127**, 219-23.

Wilson, C. & Doniach, S. (1989). A computer model to dynamically simulate protein folding: Studies with crambin. *Proteins* **6**, 193-209.

Wilson, C., Weisgraber, K.H., Wardell, M.R., Mahley, R.W. & Agard, D.A. (1991). Three-dimensional structure of the LDL receptor-binding domain of human apolipoprotein E. *Science* **252**, 1817-1822.