The protein primer (vol I)

Chapter 1. Introduction to protein research

This monograph provides an in-depth summary of protein physical chemistry appropriate to the end of 2001 in two volumes the first includes more general and often simpler material and the second adds less well established material necessary for extension of research on the construction and physiological applications. The first volume of this "protein primer" was begun reluctantly in 1997 to reexamine some critically important topics in protein chemistry that have drifted away from experimental confirmation. Our views on protein structure and function never very fashionable are now entirely ignored despite almost complete experimental confirmation but by 1997 most of the material had been published in books and papers much of it thirty years earlier. and it seemed pointless to beat that dead horse. However, new insights arising from the B-factor data in the Protein Databank have made protein physical chemistry once again exciting and provide the unique ways to utilized x-ray data quantitatively and the information about proteins necessary to support honest scientific study of the genome. Even the additions made in the last year make improvements in the subject matter of this Primer inadequate. Fortunately it now appears that I will have skilled help from Professor Chang-Hwei Chen of SUNY Albany. I have examined data so far only mesophiles although the numbers are in the thousands.. The extremophiles from the archaea , the immunoglobulins, the extracellular globulins, membrane proteins and various other special groups are yet to be examined in depth. Small samples of some of these suggests that variations in knot strength are the basis for most differences among these classes. So many proteins of many different kinds are built like enzymes to the extent that the pairing principle is clearly revealed but I don't know why.

The main themes of this Primer are the universal basis of folded stability and the fact that all enzymes including those with coenzymes are constructed in the same way strongly suggesting that their mechanisms are all modest versions of a single one. These topics are discussed at some depth in two places: Lumry, Chapter 1 in Protein-solvent interactions . Edited by Roger Gregory for Marcel Dekker , New York, and Lumry, Chapter 29 of Methods in enzymology, volume 256 (ed Johnson and Ackers). More concentrated discussions of important details are given in recent papers in Biophysical Chemistry. Those in December of 2002 will soon be accompanied by two more in that journal by the end of the summer 2003.

Publication on the WEB has become the way to go and particularly suitable for subjects like protein chemistry that are just beginning to emerge. Very few topics in this monograph are common knowledge, many are not even known to the general run of protein people and to our knowledge there are no texts suitable for graduate-student education in protein chemistry. Any parts of this monograph can be downloaded without cost or copyright restrictions although formally the articles are copyrighted in my name. Expansion and addition of chapters to include still newer material, to include more references will appear from time to time. We hope to answer questions via a chat file in additional WEB documents but my time is limited by the remaining writing tasls and old age and can be best used in more detailed papers on especially important topics least well known. The latter as illustrated in the previous paragraph generally will appear in conventional journals. For this monograph chapter titles and numbers as of 2003 will be retained but as updating becomes possible and necessary the chapters will be processes one by one. The header in each give the date changes.

Volume II is a collection of tools for protein work including the necessary mathematical theory, the nature of water and aqueous solutions, probability considerations in evolution and other essential special topics. It is unlikely to get to this site until the end of 2003.

The following topics are discussed in Volume I:

Proteins are not isotropic. They include a minimum of three substructures with very different properties and functions. Furthermore all enzymes and many other proteins, most mesophiles, consist of two semi-independent proteins each with its set of the three substructures. Attempts to solve the "protein-folding problem" or explain enzymic catalysis must be based on these subdivisions.

Thermal denaturation in dilute buffers does not produce anything approximating random-coil species. The normal product a species we call a bubble. is a soft, motile micelle-like species without much larger volume than the native species, Expansion of the latter to immersion of the polypeptide in bulk water occurs only in the low-density macrostate of pure water or in high concentrations of structure-breaking cosolvents like urea and hydrazine. The changes in standard enthalpy and entropy have different signs in the two steps although the heat-capacity changes are both positive. Model transfer processes into bulk water are not appropriate tools for understanding the bubble-forming process. The standard thermodynamic changes are reversed again except the heat capacity. Small model hydration data are applicable to the transfers of substrates and other ligands from bulk water to protein.

The key to understanding the thermodynamics of protein species has proved to be the zero value of the activation heat capacity. Its apparent non-zero values at intermediate mole fractions of structure breakers such as urea is due to the presence of both bubble and random-coil products: so that the two-state model of melting does not apply in that concentration range. The same artifactual behavior occurs at temperatures near 273K in pure water.

Stability of native species is a consequence of a small number of cooperative electrostatic clusters with unusual strength not yet explained. The larger parts of mesophilic proteins work against these clusters tending to destabilize native states. Expansion of bubble states to expose polypeptide to bulk water is prevented by positive free energy change as suggested by Kauzmann so at normal temperatures in the absence of structure-breakers the bubble state is the normal product. The "hydrophobic bond" defined as a cluster of aromatic and aliphatic side chains bound together by dispersion interactions have very little strength. Such sidechains gain importance as a consequence of the weak or non-existent permanent polarization .that allows versatility in packing so that their major importance is in reducing the effective dielectric constant in the regions of the strong hydrogen-bond clusters.

Structure breaking and structure making cosolvents alter activity coefficients but their direct attachment to protein is of secondary importance. Instead they act either by changing the ratio of the high-density to low-density macrostates of pure water or by destroying both at more effective concentrations. The latter effect is complete at about mole fraction 0.25 for hydrazine, hydrogen peroxide and urea. Because guanidinium chloride is an electrolyte, concentrations and results can differ. The cation seems to be important with calcium more effective than lithium and sodium scarcely effective at all. These are all Hofmeister matters but have lately become well understood. Structure makers like glycerol, ethylene glycol, ethanol, methanol and propanol use up all available water at mole fractions from 0.11 to 0.08.

Because amphiphiles larger than propanol tend to form micelles at low CMC values the situation quickly becomes complicated with higher mole fractions.

There is very little of small-molecule chemistry in enzymic processes. Enough good data on archae proteins may have now accumulated make it possible to examine their catalytic processes but we have not done so. The mesophilic proteins examined with Protein Databank (PDB) data have been found to be constructed to have the same features suggesting that they all use the same mechanism. That mechanism is based on mechanical activation by raising the potential energy of a pretransition state. This is a transient process dependent on collapsing the free volume and the enhancement is responsible for the low values of the apparent activation free energy and thus for the catalytic advantage.. Subsidiary conventional thermal activation taking the system from pretransition state to true transaction state for primary bond rearrangement is the apparent activation free energy. It is an add-on varying with construction of substrate but rarely more than about 20Kcal/mole with those substrate types for which the enzyme was "designed". The true total activation free energy is the difference beween the average reactants at constant temperature and that of the true transition state lying far above the amounts available from thermal activation that pure thermal activation as in small-molecule processes cannot satisfy the requirements for speed and substrate specificity of biology. Biology exists because evolution found a sufficiently efficient alternative and uses it not only for enzymic catalysis but for many other physiological functions. Muscle, ATPsynthetase and other protein motors are probably driven by the same transition production of potential energy though fluctuations in protein free volume. Because the mechanical mechanism acts only like a *nutcraker*, reaction pathways have been selected to consistent with that mechanism; inhibitors have to fit into the nutcracker with low enough mechanical requirement to be cracked. The force and work available are determined by the size of the enzyme and the adjustments of enthalpy and entropy available through rearrangements of free volumes. It is probable that nut-cracking is often not reversible so unlike conventional small-molecule rate processes equilibrium requirements such as microscopic reversibility do not obtain..

Proteins apparently tend to support physiology by manipulation of conformational enthalpy and entropy the quantitative characterization being the ratios of change in one to change in the other, known as the compensation temperature. Each substructure has its characteristic compensation temperature fixed at 354K for the small stabilizing substructures but varying for the larger structures within the range from near 500K down to 220K depending on ligation.. All of the smaller structures lose stability near 354K so that is the compensation temperature at which the activation free energy for melting of most mesophiles is zero. Increased stability up to about 373K seems to be supplied but infrequently by arrangements of disulfide groups but studies of the very heat-stable proteins from archae and hot springs may require modification of this possibility. Some of the latter melt as high as 403K but that is still low by comparison with the dragline spider silks.

Privalov and coworkers found that the standard heat-capacity changes in melting in dilute buffers can be normalized to a single value on division by the number of residues. Their collection of proteins included wild types of mesophiles. Later Murphy, Privalov and Gill found that the standard enthalpy and entropy changes in melting of the same proteins are normalized in that way. That complicated story is given in my paper on Parsimony in protein evolution (Biophysical chemistry, December 2002) with added detail in in this monograph. These depend on an extraordinarily high evolutionary selection Essentially all mesophiles are size variants of a single protein. This evolutionary achievement has been further exacerbated by the finding that only wild types satisfy the requirement. Interpretations of site-directed mutagenesis experiments thus jeopardized provide a fist goal in scientific studies of the genome and obviously an extraordinarily difficult one.

Since the temperature factors obtained routinely in diffraction studies of protein crystals are both precise and accurate, there can be little lattice disorder so the crystals must be hard. Considering the large amount of extramolecular water this is surprising but is a consequence of the strength and rigidity of the smaller substructures as illustrated by the values of the Youngs modulus reported by Morozov and Morozov.. Even though there is high hydration and few contacts between proteins, those contacts must by strong and stable. Hydration problems in crystals as in solution are not yet well understood primarily because the requirements of water have not been adequately taken into account. A major factor is the "non-freezing water" discovered by Kuntz. At low temperatures is strongly stabilizing native species against cold denaturation. The protein stabilizes the lower-density species of water thus preventing melting down to about 200K. As shown particularly elegantly by Timasheff and coworkers the protein-environment interface is qualitatively important as a mediator of folded stability and physiological function. Variations in interfacial free energy cause changes in free volume which is the major factor in determining protein activity coefficients in native and bubble states. Interfaces to water have intrinsically very high free energies that are very sensitive to the type and amount of cosolvents.

The expansion-constriction process of the large soft parts of proteins as modulated by environmental and functional states is a major factor in successful protein evolution but has been ignored because the geometric changes are smaller than the coordinate errors in diffraction studies. It is responsible for the circular-dichroism behavior in the peptide bands, fluorescence, proton exchange behavior, and so on. In ligand-free states the soft substructures oscillate to contracted states with periods of a few nanoseconds but the geometric changes are only a few tenths of angstroms detectable only in the temperature factors, CD, proton-exchange rates, etc.

There are several quite different kinds of mesophilic proteins of which only three are thus far clearly delineated. The free-energy surfaces of enzymes and many other mesophiles are strong displaying Arrhenius behavior in motions on those surfaces from one substate to another.. The immunoglobulins appear to be special kind of motile knot with low mean B factors and major changes in atom B factors depending on the details of their physiological processes. On the other hand myoglobin, hemoglobin and probably all the other myoglobin type respiratory proteins, (hemocyanins for example).have fragile free-energy surfaces. Although these show palindromic patterns in their B factors, the B factors are high and have a high average. Their conformational free-energy surfaces are fragile so the range of conformational fluctuations is large; definitive solution conformations may not exist despite the apparently simplicity of the diffraction studies. It is not surprising that Frauenfelder and coworkers have found that myoglobin and single-chain hemoglobins have many conformers . Whether there are other proteins with such variety due to their fragile surfaces is not yet known.

Limited utility of popular methods for studying proteins.

Enzymic catalysis is driven by transient cooperative fluctuations in the atom free volumes of the large substructures and reflects quantitatively residue exchanges in both major substructures Acceptable residue combinations in all enzymes regardless of size differ only in scale factors indexed by the number of residue, remarkable in itself but made more so by the selection of knot residues since the wild type of a protein satisfies linear indexing in this way but its mutants with the same residue number usually do not.(cf. Fig. @.) Residue differences generated in this way reveal a level of sophistication it may not be possible to describe in site-directed mutagenesis under practical limitations. Hit or miss residue substitutions not part of large well planned investigations are not likely to provide reliable information as to the critical involvement of the entire protein and further sophistication depends on finding very precise and very accurate methods for study of differences produced by mutation. Generally those now available are crude. so without improvement studies relating the mutagenesis to the genome seem pointless as a scientific undertaking.though not a financial one.

X-ray and neutron diffraction methods now and more so as precision improves are the major source of useful information although not in coordinate length and angle information in which errors exceed the important geometric changes in physiological function. Fortunately the B values in such studies have high precision and high accuracy for measuring the conformational changes and those changes are so small that the crystals remain isomorphous. Thus although protein diffraction studies now resemble the efforts of the sorcerer's apprentice, by fortunate accident the necessary information is accumulating at a great rate awaiting only the prince's kiss. As resolution improves, so do B values making all three moments of the scattering ellipsoids available for detailed description of the conformational factors in biology.although its successful extraction may be many years away.

NMR methods lack the precision found in the temperature factors from x-ray diffraction but make proton-exchange rates readily available. The substructures can be described in terms of those rates and residues but their coordinate changes, of order a few 0.1 Å, will probably continue to lie far below nmr errors.. Useful substitutes for the temperature factors from diffraction may be found but at present that avenue is not promising. The many secondary-level ones possible with this versatile instrumentation already provide much important information not tied to precise coordinate data.

It is not established that modeling of proteins for large-scale computing can produce reliable results or even that the results can be shown to be reliable. This is a very active area in protein study but the coordinate information from the Protein Databank on which it depends has much larger errors than the dependence of the potential-energy functions on distance and angles. The many approximations required to simplify the models to tractable solutions prevent precise comparison of results against experimental information. Important subtleties such as the expanded and contracted states are detectable only at resolutions much higher than those in general use in this kind of modeling. Predictions do not appear to have much reliability and the criteria for comparing results with experimental information cannot be precise. The situation at present resembles the attempts to produce accurate primary-bond quantum computations in the early years of the Mulliken group at the University of Chicago. Models for liquid water such as the STII models of Raman and Stillinger that missed the essential features of real water impeded research for many years Carloni et al have produced molecular-dynamics calculations giving the correct general qualitative description of structure and mechanism for the HIV-1 protease. This is the only qualitatively correct work of this kind so far published.

The great current popularity of site-directed mutagenesis rests on the textbook instructions available to anyone and without further instruction. The popularity is very much limited by three misconceptions. The first is that the translation of DNA information into proteins depends only on amino-acid residue sequence so that all members of a given protein family share some minimum set of conserved residues. Since most structural details of a protein are consequences of the construction of the small, hard substructures, its residues might be expected to be conserved. However, major exceptions have already appeared. Conservation is limited and different members of the same family to not have common sequences for these substructures. Instead it is the distribution of the free volume of the atoms that is conserved.

The second misconception is that the familiar secondary structures of Pauling, Corey and Branson play some special role in the construction and function of proteins. That does seem to be the case in fibers of which the dragline spider silks are stronger in tension than any other substance but it does not appear to be the case in ordinary mesophiles. However there is a major mystery to be solved since the smaller substructures have physical properties more like the spider silks than those of the larger substructures. That does not appear to require sheet or helix parts in the smaller substructures although both are found as participants in many. Structures of the spider silks have thus far defeated research attempts and the properties suggest that basic polypeptide chemistry is incomplete. The suggestion is that anti-parallel beta sheet and to a lesser extent the alpha helix has cooperative electron rearrangements giving properties something like graphite and thus like the carbon nanotubules. If arrays of regular polypeptide structures have cooperative shrinking and inductive electron displacements imparting covalency to the inter-chain hydrogen bonds, possible structures are easily drawn.

The third major misconception has already been mentioned. It is due to the high degree of cooperativity of residues in successful protein evolution. The high specificity in enzymic catalysis and in ligand binding in other physiological functions requires that all residues of a given protein have been individually selected for their contribution to the cooperativity of a protein. Then accurate assessment of the role of any residues requires simultaneous knowledge of roles of many and probably most the other residues. This is why, for example, change in a single residue makes a major change in quantitative specifity.

t

Confusing Biology with Chemistry

(quotation from Protein-solvent interactions , chap. 1. A new paragram for protein research.) The following quotation from the monograph on protein-solvent interactions describes the confusion between chemistry and biology still limiting research progress in protein chemistry. There is very little of the familiar chemistry learned with small molecules in biology but each year there is another desperate attempt to find some magic to make them one. Currently this effort is focused on "low-barrier hydrogen bonds" and proton tunneling, respectable phenomena but dependent in proteins on the mechanical basis of rate activation in biology for which they are an intended substitute. .

"This chapter addresses a serious problem frustrating progress toward understanding proteins and biological systems nearly as much now as in the earliest days of protein study. The problem arises from the confusion of chemistry with biology. The chemistry of the DNA double helix is that of its individual groups, all of which have been well studied by chemists and biochemists. The biological usefulness of the double helix lies in its structure, and this is an "invention" in which the intrinsic properties of these groups, i.e., their chemical and physical properties, have been utilized in evolution to produce a nonchemical result. To apply chemical knowledge to better understand biological mechanisms one must first discover the devices that make the mechanisms possible. The stable states of small molecules are determined by the intrinsic chemical properties of the molecules. Those of biological macromolecules individually and at higher and higher levels of cooperativity are intrinsically improbable, gaining their actual high probability only because of their coupling to the remainder of the biosphere. Chemistry deals with the cold earth and our attempt to breath life into cold molecules; biology deals with inventions not only unpredictable, but often so constructed as to subvert chemical expectations. The two disciplines require quite different orientations. Students of biology must be ever aware of the fact that chemistry is a poor guide to biology, a necessary foundation but not one extrapolatable to biological devices. Students of chemistry are adequately fortified by a general realization that in principle, often only in principle, chemical behavior can be traced back to the Schrõdinger equation and its solutions. To say that this is also true of biology is more than a minor sophistry .Designs for research on biological mechanisms that are based on logical progression from chemistry are unlikely to produce real progress.

The use of conformation changes in enzymic mechanisms is another such device. This is not the chemists' chemistry, but rather a construct found in the biosphere that makes selection, specificity, and rate control possible and adjustable by DNA modification. Once such a device is discovered, it is likely that we can copy it in abbreviated form, perhaps in some cases using classical chemical techniques. But the latter is not at all likely since one can neither understand nor duplicate something that is yet to be discovered. The problem is compounded by the fact that nature's solutions to many physiological problems depend on packaging several such devices in a single protein".

The mechanical biosphere

Eyring, Lumry and Spikes thought it most unlikely that the extraordinary rates and selectivity of reactions supporting biology are produces by thermal activation and non-covalent selection now seen fifty years later as even more crude and less promising. The most promising alternative then as now is mechanical activation of rates and vectorial adjustment in specificity and rates of substrate selection. The pros and cons of their arguments form the central theme of this volume. Starting from the estimates of atom free volume given as B factors in the Protein Databank these arguments progress to a wide variety of increasingly large protein assemblies with ever more complicated cooperative units. Obvious examples are the protein motors such as Boyer's ATP synthetase and muscle. And in view of the extraordinary success of the nutcracker mechanism for enzyme it will not be surprising to find the unique expansion-contraction device ubiquitous in protein systems has been found to solve other biological needs such as the immune systems and protein synthesis especially catalyis by RNA. The t-RNAs have enzyme construction suggesting that catalysis by RNA is also mechanical in mimicking that by protein enzymes. There appears to be a single thermodynamic principle underlying these successes. It has to be based on free energy but the criterion for success in natural selection is an every increasing efficiency in the use of free energy, the principle of free-energy complimentarity as discussed in the final chapters of this volume.