A model for the coupling of α -helix and tertiary contact formation

ANDREW C. HAUSRATH

Department of Biochemistry and Molecular Biophysics, University of Arizona, Tucson, Arizona 85721, USA

(RECEIVED April 19, 2006; FINAL REVISION April 19, 2006; ACCEPTED May 1, 2006)

Abstract

Peptides corresponding to excised α -helical segments of natural proteins can spontaneously form helices in solution. However, peptide helices are usually substantially less stable in solution than in the structural context of a folded protein, because of the additional interactions possible between helices in a protein. Such interactions can be thought of as coupling helix formation and tertiary contact formation. The relative energetic contributions of the two processes to the total energy of the folded state of a protein is a matter of current debate. To investigate this balance, an extended helix-coil model (XHC) that incorporates both effects has been constructed. The model treats helix formation with the Lifson-Roig formalism, which describes helix initiation and propagation through cooperative local interactions. The model postulates an additional parameter representing participation of a site in a tertiary contact. In the model, greater helix stability can be achieved through combinations of these short-range and long-range interactions. For instance, stronger tertiary contacts can compensate for helices with little intrinsic stability. By varying the strength of the nonlocal interactions, the model can exhibit behavior consistent with a variety of qualitative models describing the relative importance of secondary and tertiary structure. Moreover, the model is explicit in that it can be used to fit experimental data to individual peptide sequences, providing a means to quantify the two contributions on a common energetic basis.

Keywords: helix-coil transition; hydrophobic effect; binary pattern; secondary structure; tertiary interactions

The energetic contributions of individual amino acids to helix formation in peptides and proteins have been intensively studied by numerous methods. Helix-coil theory (Schellman 1955; Zimm and Bragg 1959; Lifson and Roig 1961) has been indispensable for the interpretation of this body of experimental data. The theory allows the combined influence of all interactions to be taken into account simultaneously, making their individual contributions to the experimental observables calculable from the theory (Qian and Schellman 1992). Therefore the theory can be used to resolve the contributions of individual interactions from experimental data. Knowl-

edge of the energetics of these interactions then endows the theory with predictive power.

The formalism introduced by Lifson and Roig (1961) (LR) to describe the helix-coil transition has proved to be particularly flexible and versatile and has been expanded and generalized in a variety of ways to include many interactions not included in the original formulation (Doig 2002). Progress in development of the theory has allowed quantification of the energetics of these varied interactions from experimental data. In turn, measurement of these energies has made it possible to estimate the helix content and other experimental observables for any peptide sequence (Munoz and Serrano 1994). As such, within its domain of applicability, helix-coil theory represents one of the chief success stories among theories of macromolecular conformation. In addition, it stands as an exemplar of productive interplay between experiment and theory.

Reprint requests to: Andrew Hausrath, Dept. of Biochemistry and Molecular Biophysics, University of Arizona, BSW 451, 1041 E. Lowell, Tucson, AZ 85721, USA; e-mail: hausrath@email.arizona.edu; fax: (520) 626-1925.

Article published online ahead of print. Article and publication date are at <http://www.proteinscience.org/cgi/doi/10.1110/ps.062292106>.

The challenge for future development is to make use of the existing body of knowledge about secondary structure formation in more complex situations where tertiary structure is also present. The model presented here represents a first step toward that goal. The basic approach is to modify the LR model so as to allow additional interactions between side chains distant in the sequence with the intent of maintaining the general structure of the theory and, in particular, its applicability to experimental situations.

Summary of the Lifson-Roig model

In the LR model, a given residue can be either in a helical (*h*) or coil (*c*) residue state, and so any molecular state of a polypeptide can be represented with a string of *h* and *c* symbols. A peptide with *N* residues has 2^N molecular states. Predictions of observable properties entail calculating a weighted average of that property over all of the molecular states. The statistical weight assigned to each molecular state is the product of the weights for the individual residues. The model explicitly considers only local interactions, and residue weights are assigned depending only on the state of the residue in question and its neighbors. The residue weight assigned to coil states is *u*, and the residue weight assigned to helix states is *v* or *w* depending on whether the residue has a coil neighbor (*v*) or not (*w*). For example, the molecular state represented by *ccchhhhhcchhhccc* would be assigned a weight of $uuuvwwwwvuuvvuuuu = u^8v^4w^5$ (Table 1).

Once the weights have been assigned for each molecular state, the model can be used to predict a variety of experimentally observable properties. In general, the calculation of any property is the weighted average of that property over all the possible states:

$$\langle A \rangle = \sum_{i=1}^N A(i)P(i) \quad (1)$$

where $P(i)$ is the probability of the state *i*, and $A(i)$ is the value of that property for state *i*. The probabilities $P(i)$ are the normalized state weights. The normalizing factor is the partition function *Z*, obtained by summing the weights for all the molecular states. Using the versatile transfer matrix formalism introduced in the LR model, this sum over states can be written in the compact form

$$Z_{LR}(u,v,w) = \begin{pmatrix} 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} w & v & 0 \\ 0 & 0 & u \\ v & v & u \end{pmatrix}^N \begin{pmatrix} 0 \\ 1 \\ 1 \end{pmatrix} = eW^Ne^+ \quad (2)$$

where *W* is the transfer matrix corresponding to Table 1, and *e* and *e*⁺ are the left and right end-vectors. Evaluation of the matrix product yields a partition function in the form of a polynomial in the variables *u*, *v*, and *w*.

Calculation of some types of ensemble averages can be accomplished by differentiation of a partition function *Z*. For example, the average helix content (the number of *w* residues) or coil content (the number of *u* residues) are obtained with

$$\langle n_w \rangle = \frac{w}{Z} \frac{\partial Z}{\partial w} \quad \text{and} \quad \langle n_u \rangle = \frac{u}{Z} \frac{\partial Z}{\partial u} \quad (3)$$

Other types of ensemble averages are accomplished by modifying the structure of the matrix product (Qian and Schellman 1992). Each matrix *W* corresponds to a single residue, and its entries comprise the possible states of that residue, only one of which can be used in a given molecular state. Site-specific ensemble averages can be calculated by using modified *W* matrices corresponding to the site of interest. The states of the molecule in which residue *k* is in a helix can be obtained by setting all but the *w* entry of the *k*th *W* matrix to zero. The modified matrix product will then sum only the states in which residue *k* has the residue state *w*.

$$P_w(k) = \frac{1}{Z_{LR}} eW^{k-1} \begin{pmatrix} w & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix} W^{N-k} e^+ \quad (4)$$

The LR formalism also is readily adapted to the description of heteropolymers. The differences between residues are accommodated by creating a different *W* matrix for each type of residue (Qian and Schellman 1992). A systematic survey of the effects of residue substitutions in host–guest peptide systems has enabled the determination of the appropriate LR model parameters for each amino acid (Chakrabartty et al. 1993, 1994; Doig et al. 1994; Doig and Baldwin 1995; Rohl and Baldwin 1997). Other experimental systems and approaches have provided independent measurements of the helical propensity scale (O’Neil and DeGrado 1990; Wojcik et al. 1990; Horovitz et al. 1992; Blaber et al. 1993; Blaber et al. 1994; Myers et al. 1997). These various scales can be used to make predictions of the various experimental observables obtainable from the theory for any peptide sequence. Inclusion of additional interactions has refined these predictions (Munoz and Serrano 1994; Doig 2002).

Table 1. Weight assignment table

Position <i>i</i> – 1	<i>i</i> + 1 <i>i</i>	<i>h</i> <i>h</i>	<i>c</i> <i>h</i>	<i>h</i> or <i>c</i> <i>c</i>
<i>h</i>	<i>h</i>	<i>w</i>	<i>v</i>	0
<i>h</i>	<i>c</i>	0	0	<i>u</i>
<i>c</i>	<i>h</i> or <i>c</i>	<i>v</i>	<i>v</i>	<i>u</i>

Note that the LR model parameters u , v , and w are statistical thermodynamic weights. Therefore they represent a resultant conformational tendency due to the combined influence of diverse physical effects. The understanding of these weights in terms of the contributions from particular physical interactions is naturally of great interest and includes contributions from backbone and solvent hydrogen bonds, side chain configurational entropy, solvation of the backbone, and solvent configurations. Experiments under conditions that modulate one or another of these effects can be used to dissect the relative contributions from these different physical interactions. The helical propensity was originally attributed largely to hydrogen bonding. As experimental data have accumulated and our understanding has evolved, the relative importance of particular interactions has been clarified, but even now, the origin of the helical propensity w is not universally agreed upon. Nevertheless, attributing the energies in the model to different physical effects has in no way undermined the validity of the model itself as an interpretive tool. The power of a statistical thermodynamic model is that it enables comparison with experiments without precise knowledge of the microscopic details of the experimental system.

Results

Inclusion of tertiary contacts in a homopolymer model

Helix-coil theory is a statistical mechanical theory of secondary structure formation. The simplifying assumption of the LR model is that any residue conformation can be assigned into one of two states, helix or coil. By making this two-state approximation, the matrix formalism can conveniently enumerate the possible states and perform ensemble averages over them. However the representation does not explicitly consider the spatial arrangements of the states, only the local secondary structure.

The HP lattice models of proteins constitute an analogous statistical mechanical model for tertiary contact formation (Lau and Dill 1989; Dill et al. 1995). To a great degree, the binary pattern of hydrophobic (H) and polar (P) residues in a protein sequence dictates its final folded form and characteristics (West and Hecht 1995; Cordes et al. 1996; Marshall and Mayo 2001). HP models capture this property in a solvable quantitative formalism, enumerating all possible states by making the simplifying assumption that individual residues lie on a lattice. Residues have two types, H or P, and weights are assigned to the states for each H–H contact, which is defined as a pair of H residues adjacent on the lattice. These discrete-state approximations make it possible to sum over the states and perform ensemble averages.

Both the helix-coil and HP models capture the essential character of an important aspect of the folding process and

are complementary in that each focuses on an aspect not explicitly considered by the other. The goal here is to devise a model which combines the characteristics of both.

The assumptions of the XHC model are

1. The conformation of any state can be described by a string of h 's and c 's, and the weights for a given conformation are assigned in terms of u , v , and w , as in helix-coil theory.
2. Residues can be of type H or P.
3. An H residue in an h state may form a tertiary contact with one other H residue in an h state, conferring an additional statistical weight x to that state. The allowed interactions are specified in a list of contact sets.

The task of enumerating the states consistent with 1–3 requires reconciliation of the helix-coil states and the tertiary contact arrangements. Assumption 3 states that tertiary contacts only can form between two residues that are both in a w state. Therefore, not all helix-coil states are compatible with all tertiary contact states. The strategy used here to obtain a partition function containing only the compatible terms proceeds in two steps. In the first step, a generating function is created that includes the allowed helix-coil states with all possible tertiary contacts present. In the second step, the terms that are compatible with the prescribed contact sets are selected from the generating function to form the partition function.

A generating function from which a partition function for the model can be obtained can be written using the LR matrix formalism:

$$G_{\text{hom}}(u, v, w, \{x_i\}) = e \left[\prod_{i=1}^N \begin{pmatrix} w(1+x_i) & v & 0 \\ 0 & 0 & u \\ v & v & u \end{pmatrix} \right] e^+ \quad (5)$$

where the tertiary contact parameter x_i is zero for P residues. This matrix product will be a polynomial in u , v , w , and the $\{x_i\}$, each of which represents “half” of a tertiary contact involving residue i . The factor $(1+x_i)$ creates an additional tertiary contact state from any residue with a w helix-coil state. Any term in the generating function polynomial will correspond to a helix-coil state with some number of potential half-contacts, but not all of these can correspond to physically reasonable states. For example, there are terms with only one x_i present, and in the XHC model, the x_i are assigned in pairs. Creation of a usable partition function then corresponds to selecting a subset of the terms in the generating function.

The variables representing the half-contact weights can be used to select the sets of states containing the desired tertiary contacts. Selection of terms from a generating function G in which both x_α and x_β are participating in a tertiary contact is accomplished by differentiation with respect to these quantities. Terms that do not contain

those variables contribute zero to the sum. Multiplying by the variables restores the original terms from the generating function. This strategy selects states in which both sites are participating in *some* tertiary contact, but these states may also include additional contacts. The states in which only residues α and β , and no others, are participating in a tertiary contact are obtained by setting the $\{x_i\}$ other than x_α and x_β to zero.

The situation becomes more complex when states with more than one tertiary contact (i.e., more than two sites) are considered. A state selection operator, which acts on a sum of states G to select the ones involving only the set of sites $\chi = \{x_\alpha\}$, will be denoted $SS[\chi]$.

$$SS[\chi]G = \prod_{x_\alpha \in \chi} x_\alpha \frac{\partial}{\partial x_\alpha} G|_{x_\gamma = 0, x_\gamma \notin \chi} \tag{6}$$

The state selection operators provide a flexible and systematic means to collect the desired terms for practical calculation purposes. This approach is applicable to interactions with arbitrary separations in the sequence, but the size of the matrix needed to compute the partition function does not increase as longer-range interactions are treated (Doig et al. 1994; Sun et al. 2000).

A partition function can be obtained as a sum over sets each specifying a specific collection of tertiary contacts. Each operator $SS[\chi]$ contributes to the partition function a sum over the terms in the generating function consistent with the contacts specified in the particular contact set χ_k . The strategy for obtaining the partition function is to apply the selection operator for each set to the generating function. The partition function is the sum of all of the sums obtained from the application of the individual operators. The states with no contacts, corresponding to the original LR model, are obtained by application of a selection operator which has an empty contact set.

$$\begin{aligned} Z_{\text{hom}}(u,v,w,x) &= \sum_k \prod_{x_\alpha \in \chi_k} x_\alpha \frac{\partial}{\partial x_\alpha} G_{\text{hom}}|_{x_\gamma = 0, x_\gamma \notin \chi_k} \\ &= \sum_k SS[\chi_k] G_{\text{hom}} \end{aligned} \tag{7}$$

For illustration purposes, calculations in this manuscript employ a 38-residue model with H residues only at positions 3, 7, 11, 14, 25, 28, 32, and 36. These sites correspond to the hydrophobic core of the designed peptide $\alpha\alpha$ (Fezoui et al. 1994, 1999). The sets of contacts between these sites used in these calculations are illustrated in Figure 5 of Xian et al. 2006 and are compiled in Table 2.

Properties of the homopolymer model

Summing over this ensemble of tertiary contact states and using the expression in Equation 7 to obtain the sum of

compatible helix-coil states, a partition function is obtained and so can be used for calculations of expectation values. This partition function is a polynomial in the variables u , v , w , and the $\{x_i\}$, which are variables kept in symbolic form for the purposes of state enumeration. However, in the following examples, once the expression for the partition function is obtained, all the $\{x_i\}$ are set to the same value, x . This corresponds to the simplifying assumption that all the hydrophobic contacts are energetically equivalent, and so the partition function will depend only on the variables u , v , w , and x .

The value of w is a measure of the intrinsic helical propensity of an individual residue. The larger the value of w , the more strongly that residue favors helix formation. A larger fraction of helical residues will result from increasing w . The model assumes productive tertiary contacts can only be formed between residues in a helical

Table 2. Tertiary contact sets used in example calculations

Contact set	Register	Pair 1		Pair 2		Pair 3		Pair 4	
1	0	3	36	7	32	11	28	14	25
2	0			7	32	11	28	14	25
3	0	3	36			11	28	14	25
4	0	3	36	7	32			14	25
5	0	3	36	7	32	11	28		
6	0					11	28	14	25
7	0			7	32	11	28		
8	0	3	36	7	32				
9	0	3	36					14	25
10	0	3	36			11	28		
11	0			7	32			14	25
12	0							14	25
13	0					11	28		
14	0			7	32				
15	0	3	36						
16 (LR model)	0								
17	1	7	36	11	32	14	28		
18	1			11	32	14	28		
19	1	7	36			14	28		
20	1	7	36	11	32				
21	1					14	28		
22	1			11	32				
23	1	7	36						
24	−1	3	32	7	28	11	25		
25	−1			7	28	11	25		
26	−1	3	32			11	25		
27	−1	3	32	7	28				
28	−1					11	25		
29	−1			7	28				
30	−1	3	32						
31	2	11	36	14	32				
32	2			14	32				
33	2	11	36						
34	−2	3	28	7	25				
35	−2			7	25				
36	−2	3	28						
37	3	14	36						
38	−3	3	25						

conformation. The effect of this assumption is that increasing the value of x also can drive helix formation. The number of tertiary contacts can be obtained from a partition function Z with the expression

$$\langle n_{\text{tertiary}} \rangle = \frac{x}{2Z} \frac{\partial Z}{\partial x} \quad (8)$$

The numerical factor 2 in the denominator reflects the fact that two half-contact weights are needed to form a single pairwise contact. Figure 1 displays numerical examples, which show that increased helical content is observed when either w or x is increased. To a lesser extent, the converse is also observed. For a given value of x , more tertiary contacts are formed when the helix-forming tendency is greater. However the maximum number of contacts attainable by the model (four in this example) cannot always be achieved by arbitrarily increasing the helical propensity w (Fig. 1B).

Figure 1 illustrates the basic property desired in the model, which is the interplay between the local interactions of secondary structure formation and the long-range interactions of tertiary contacts. Both contribute to helix stability and a deficiency in one can be made up for by improvements in the other.

Figure 1. Homopolymer and heteropolymer model comparison. Properties of the homopolymer model in which all sites have equal helical propensity and all tertiary contacts are identical, compared with the predicted properties for the $\alpha\alpha$ sequence. (A) Average helical content (number of w states) plotted vs. w . The five traces correspond to $x = 2$ (dotted line), $x = 5$ (short dashes), $x = 10$ (long dashes), $x = 15$ (dot-dash), and $x = 20$ (solid line). (B) Average number of tertiary contacts plotted vs. w . Traces correspond with those in A. (C) Average helical content plotted vs. x . The five traces correspond to $w = 0.2$ (dotted line), $w = 0.4$ (short dashes), $w = 0.6$ (long dashes), $w = 0.8$ (dot-dash), and $w = 1.0$ (solid line). The additional trace labeled $\alpha\alpha$ shows the predicted helical content for the helical hairpin molecule calculated with the heteropolymer model. (D) Average number of tertiary contacts plotted vs. x . The traces correspond to those in C with the addition of the predicted tertiary contacts for $\alpha\alpha$. In all cases, $u = 1$ and $v = 0.48$ (Chakrabarty et al. 1994).

Extension to heteropolymers

The general behavior of the model is apparent from study of the homopolymer case. However, to use the model to fit experimental data, it is necessary to consider the additional complications that arise from allowing the parameters to vary from site to site.

A generating function G_{HP} for a heteropolymer in which the helical parameters vary from site to site, and its corresponding partition function Z_{HP} are

$$G_{HP}(u, v, w, \{w_i\}, \{x_i\}) = e \left[\prod_{i=1}^N \begin{pmatrix} ww_i(1+x_i) & v & 0 \\ 0 & 0 & u \\ v & v & u \end{pmatrix} \right] e^+ = e \left[\prod_{i=1}^N X_i \right] e^+ \quad (9)$$

$$Z_{HP}(u, v, w, \{w_i\}, \{x_i\}) = \sum_k SS[\chi_k] G_{HP}$$

Here X_i represents the weight matrix for site i , which includes the x_i dependence. The weight for helical residue i is written as the product ww_i where w_i is the intrinsic helical propensity of a particular amino acid type at residue i and w is a variable tagging every helical residue, included so that the total helicity, fittable to circular dichroism data, can still be obtained with Equation 3. The value of w is set to 1 when not needed.

Assignment of the w_i allows the differences in helical propensity of the individual residues to exert their effects. Figure 1C shows the fraction of helical residues obtained for the sequence of $\alpha\alpha$ using the helical propensity scale obtained by Baldwin and coworkers (Chakrabarty et al. 1994).

The LR formalism also offers a convenient route to ensemble averages of site-specific properties. For example, the calculation of site-specific fractional helicity can be accomplished by modification of the weight matrix for residue k as described in the section on LR theory. This resulting generating function is a sum only over terms that contain w at this position. From this sum the terms consistent with the ensemble of tertiary contacts can be collected using SS operators to give $P_w(k)$, the probability of a w residue at site k .

$$G_{w_k} = e \left[\prod_{i=1}^{k-1} X_i \right] \begin{pmatrix} ww_k(1+x_k) & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix} \left[\prod_{i=k+1}^N X_i \right] e^+ \quad (10)$$

$$P_w(k) = \frac{1}{Z_{HP}} \sum_i SS[\chi_i] G_{w_k}$$

Calculations of the site-specific helicity in $\alpha\alpha$ are illustrated in Figure 2.

Amide hydrogen exchange offers an experimental route to analysis of the hydrogen bonding pattern of the

Figure 2. Predicted sitewise fractional helicity for the sequence of $\alpha\alpha$ by residue with dependence on x . The dotted, short-dashed, long-dashed, dot-dashed, and solid lines correspond to $x = 1.0, 2.0, 3.0, 5.0$, and 10.0 , respectively.

backbone and gives an indication of the local stability of a molecule (Clarke and Itzhaki 1998; Raschke and Marqusee 1998; Englander 2000). The hydrogen bonding pattern can be simulated with the model. Three residues N-terminal to a given site must be in a helical conformation to create a hydrogen bond. The hydrogen bond involving the amide proton at site k is formed if the site at $k - 2$ is in a w conformation, as the residues flanking site $k - 2$ must be either v or w for this to occur (Rohl et al. 1992). While such a local structure exists, exchange of the amide proton is assumed impossible, and the probability of such a local structure determines the protection factor at that site. Therefore the probability of site k being in hydrogen bond can be written $P_{Hbond}(k) = P_w(k - 2)$. A comparison of the observed hydrogen exchange pattern in $\alpha\alpha$ with calculations of site-specific helicity is used in the companion manuscript (Xian et al. 2006) to make an initial estimate of the value of the x parameter.

Discussion

The framework model (Kim and Baldwin 1982) and the hydrophobic collapse model (Kauzmann 1959; Tanford 1962; Jennings and Wright 1993) represent two canonical views of the protein folding process. The framework model assigns primary importance to the formation of secondary structure elements and suggests that, once formed, these can associate through formation of tertiary contacts. By contrast, the hydrophobic collapse model proposes that formation of tertiary contacts through the sequestering of hydrophobic side chains into a restricted volume can assist the subsequent formation of secondary structure by reducing the number of accessible conformational states. The folding processes of individual proteins vary and may correspond to different combinations of these two schemes (Baldwin and Rose 1999; Daggett and Fersht 2003).

Related considerations apply at equilibrium. The secondary and tertiary contacts must also be considered among the factors that maintain the folded structure of proteins. Again, the relative importance of the local contacts involved in secondary structure formation and the long-range contacts involved in tertiary structure formation is currently unclear. In part this is due to the fact that, while various experimental means exist to study the presence of secondary structure, the extent of hydrophobic contact formation is difficult to detect directly. But its signature may be discernible through its influence on experimentally accessible quantities. One motivation for this work is to provide a theory which will allow such inference.

Theoretical models have proved useful for understanding the individual processes in isolation. Helix-coil theory has been used to analyze the energetics of helix formation and the factors that stabilize helices. HP lattice models have provided a theoretical basis to understand the importance of the network of hydrophobic contacts that stabilize the tertiary structure of proteins. The formalism introduced here provides a quantitative model that draws on both helix-coil theory and the HP model to create a means to investigate the balance between the two effects. By treating both within a common framework without assuming the primacy of either, the relative contributions of the individual forces acting in concert can be determined from appropriately designed experiments.

Model parameters

The expressions above provide a means to obtain various predictions of the model such as the total fraction of helical residues and the probability of forming a particular hydrogen bond. These quantities are experimentally accessible by circular dichroism and amide hydrogen exchange respectively. Adjustment of the model parameters to optimally fit such data allows an estimate of their values. The model depends on both the helix-coil parameters $\{w_i\}$ and the tertiary contact parameters $\{x_i\}$. However, it is not possible to experimentally constrain all of these parameters independently. Use of the theory requires some simplifying assumptions, and the conclusions that can be drawn from using the model to analyze data are dependent on the validity of these assumptions.

The helix-coil parameters have been thoroughly investigated, and estimates of the intrinsic helix-forming capacity of individual residues have been made with a variety of experimental systems. While some differences exist between these independent determinations, the general trends are largely in agreement (Myers et al. 1997). The calculations used as examples in this article have made use of the scale presented by Chakrabarty et al. (1994). For the purpose of fitting experimental data

on a particular protein, a good strategy would be to select the scale obtained using the experimental system most closely resembling that protein. The optimal strategy would be to determine the local helix-coil parameters through measurements on peptides corresponding to isolated helices and then to use these parameters to fit data on an intact protein to obtain the tertiary contact parameters. In this case differences between the properties of a residue in an isolated peptide and in an intact protein can be attributed to interactions not present in the isolated peptide (Myers et al. 1997).

In LR helix-coil theory with no tertiary interactions, the weights u , v , and w are defined in terms of the $2N$ internal rotational angles. These angles are the degrees of freedom over which the sum of states is calculated. (Other degrees of freedom are formally assumed to have been taken into account by prior averaging.) Any choice of values for these angles defines an individual microstate, a particular spatial configuration of the entire chain. Within such a microstate, the distance between any two particular positions on the chain is fixed. The spatial distance between points on the chain is not an additional degree of freedom. It is a derived quantity dependent on the true degrees of the model. Formally, the process of carrying out the sum over all the microstates includes a sum over such derived intra-chain distances. Therefore the relative spatial position of regions of the chain, for instance, the distance between regions that might happen to be in helical conformations, is formally treated within helix-coil theory. Carrying out the sum over states identifies those regions of the chain that, in the entire ensemble of conformations, are more likely to be helical. In the very same process of summing over this ensemble, the degrees of freedom that define the relative spatial relationships of these regions have been averaged.

In the LR theory, no interactions take place between residues located in such helical regions. The XHC theory postulates the additional multiplicative factor $(1 + x_i)$ which creates two states from any pure helix-coil state, one of which (containing the x_i parameter) has the possibility of interacting with another region. The other (containing the multiplicative weight 1) is identical to the corresponding state in LR theory. The x_i parameter confers the additional weight due to tertiary contact formation to those states in which that interaction is taking place.

Physical basis of model parameters

In this model all tertiary contact weights $\{x_i\}$ are assumed identical. The statistical weight x assigned to a tertiary contact governs the behavior of the model, modulating the intrinsic helical tendencies with an imposed binary pattern of hydrophobic and polar residues. When x is

zero, the model reduces to helix-coil theory with no tertiary interactions. When x is large, tertiary contacts become dominant. The continuum of possibilities in between can mimic equilibrium properties of proteins that display a variety of folding mechanisms. A small value of x confers properties consistent with a “framework” folding mechanism, and a large x parameter results in properties more consistent with a “hydrophobic collapse” mechanism.

In the model, the x parameter exerts its effect at two sites simultaneously in the helix-coil equilibrium, but in turn, the cooperative neighbor interactions between adjacent residues allow this influence to be propagated along the chain. Therefore, the parameter, while acting non-locally between individual sites, does have a resulting influence on local helix formation. By this mechanism, the model achieves a coupling between secondary and tertiary structure.

The x parameter is a statistical thermodynamic quantity. It is the result of the combined influence of diverse physical effects, including hydrophobic interactions, rotational entropy of the side chains, van der Waals forces between groups in contact, etc. The assumption of an identical interaction parameter for all tertiary contacts is an oversimplification since the sites are not identical. The justification for employing this assumption is that it is the simplest starting point, which allows an estimate of the average strength of nonlocal interactions for a given structure. The formalism employed here can be adapted to treat site-specific x values when appropriate data become available, which will provide a more realistic picture. Moreover such values may prove to be more easily compared between systems.

Among the interactions contributing to the x parameter are the hydrophobic interactions. The hydrophobic effect is the primary energetic driving force for protein folding, and the pattern of hydrophobic residues is probably the dominant influence on the folded structure (West and Hecht 1995; Cordes et al. 1996; Marshall and Mayo 2001). In the XHC model, the nonlocal interactions are assumed to act through the hydrophobic sites. But the simple picture of pairwise attractive interactions between tertiary contact sites is misleading. The individual hydrophobic contacts in a folded protein are not primarily the result of an attraction between groups, but rather their association is a consequence of minimizing exposure to solvent. Therefore it should be considered a working assumption (which can be tested by experiment) that the largest contribution to the x parameter is the energy of transfer of a side chain from conditions where it is largely exposed to solvent to conditions where it is sequestered from solvent (Tanford 1962). The employment of pairwise contacts that form an interaction in the XHC model is a minimal requirement to ensure that there is a hydrophobic “core” to partition into,

and is a natural choice for the helical hairpin system used as an example here.

However, the pairwise interaction assumption is enforced only in the choice of contact sets as a device for the enumeration of states. The mechanism for selecting terms from the generating function could equally be used to select clusters of sites of any size, all of which are interacting simultaneously, in order to treat more complex molecules. The model does not explicitly include pairwise tertiary interactions but only sitewise participation in tertiary interactions. Interactions referred to as pairs are formally simultaneous sitewise interactions.

Independent of the microscopic physics that the model is intended to mimic, the strict statistical interpretation of the x parameter is as the equilibrium constant for participation in a tertiary contact. As the formation of such a contact between two sites also requires that both sites have a w conformation, the number of possible conformational states is reduced when the contact is formed. Therefore the x parameter is opposed by an entropic factor that reflects the loss of these states from the helix-coil equilibrium. This configurational entropy of the chain is treated within the helix-coil theory (see above) and does not enter into the tertiary contact parameter x .

The formulation of the partition function as a sum over sets of “allowed” contacts is intended to substitute for the lack of explicit consideration of the molecular geometry, which would be very difficult to treat exactly. Specification of which sets of contacts will be considered enables the partition function calculation to be tailored to particular protein folds and sequences. By avoiding certain combinations of residues in the allowed sets, it is possible to exclude contributions from contacts that result from impossible spatial configurations. For more complicated folds, it will be necessary to devise additional criteria on which to specify the contact sets, but the same types of analyses could be carried out.

Relationship to other models

The XHC model is a coupled equilibrium. It postulates an interdependence of the helix-coil equilibrium and the equilibrium of formation of tertiary contacts. Other models have been devised that explore the local and nonlocal driving forces specifying protein structure. For discussion purposes these can be divided into four categories, which are illustrated in Figure 3.

The simplest and most widely used model for interpretation of protein folding data is the two-state model (Becktel and Schellman 1987) (Fig. 3A). The utility and power of the two-state approximation for cooperatively folding proteins are well-appreciated. However, protein folding has aspects that are not addressed in this model, and many experimental studies are undertaken to obtain

Figure 3. Relationship of XHC model to other models. (A) Two-state model. (B) Multistate model. (C) Coupled two-state/multistate model in which the two-state potential is drawn as a double well and the compatible states of the multistate model are represented as horizontal lines in the individual wells. (D) Coupled dual-multistate model. Individual wells represent the states of one multi-equilibrium, and horizontal lines represent the compatible states of the other multi-equilibrium. The native (N) and unfolded (U) states for models C and D are ensembles of similar states.

information about the deviation of proteins from two-state behavior.

More complex multistate models (Fig. 3B) have the ability to mimic some of the properties not captured in a two-state model. One example is the helix-coil transition itself (Schellman 1955; Zimm and Bragg 1959; Lifson and Roig 1961). This model and its successors display a cooperative conformational transition driven by local interactions. Models that display similarly cooperative transitions but that are driven by nonlocal interactions also exist. HP lattice models can show a cooperative collapse transition (Chan and Dill 1991; Dill et al. 1995).

A different class of models considers two distinct processes to explicitly treat local and global effects separately. For example, the complex pattern of hydrogen exchange protection factors of globular proteins (Woodward and Hilton 1980; Englander and Kallenbach 1983) suggested the two-process model, in which both local and global equilibria contribute to the exchange process (Qian et al. 1994). This can be classified as a coupled model in which a two-state model corresponding to large-scale unfolding of the protein and a multistate model in which local fluctuations among the individual amide sites induce a large number of microscopic states both contribute to the exchange behavior (Fig. 3C).

While the two-process model explicitly considered states as controlled by both a global two-state equilibrium and a local multistate equilibrium, there is no communication between the exchangeable sites in the multistate equilibrium. Also in this category is the progenitor of the XHC model, a model for the formation of coiled-coil dimers developed by Qian (1994). The coiled-coil model pairs a two-state model for dimerization to a multistate model, the helix-coil transition of the individual chains. However, the coiled-coil model has the property that the sites in the multistate helix-coil equilibrium influence each other through cooperative local interactions. This helix-coil transition is further modulated by tertiary contacts formed during the dimerization reaction. The coiled-coil model first expressed a key idea central to the XHC model, which is that the influence of the tertiary contacts acts in a site-specific manner in this helix-coil equilibrium.

In the XHC model, two processes are treated, and each is a multistate equilibrium (Fig. 3D). For each set of tertiary contacts, many compatible helix-coil states exist. The structure of the problem is such that it is possible to carry out enumeration of the states by summing directly over the (relatively few) tertiary contact states, taking account of the cooperative local interactions using the LR formalism.

Predictions involving tertiary contacts

The primary purpose of the XHC model is to serve as a calculus of experimentally measurable quantities applicable to specific protein sequences. The example calculations above centered around quantities accessible to techniques for monitoring secondary structure formation. There are also quantities that are not easily accessible to experimental test that the model can be used to explore. The only presently available support for these conclusions lies in consistency with other predictions of the model that are testable. Several types of averages involving site-specific hydrophobic contacts fall in this category, and are presented in hopes that in the future it may become possible to obtain data that addresses them.

These can be obtained using a procedure similar to that used for site-specific helicity. The generating function below is a sum of states in which site k is participating in a hydrophobic contact. The particular terms from this generating function that are compatible with the chosen contact sets can be obtained using the SS operators. Dividing the resulting sum by the partition function gives $P_x(k)$, the fraction of states in which site k is participating in a tertiary contact.

$$G_{x_k} = e^{\left[\prod_{i=1}^{k-1} X_i \right]} \begin{pmatrix} ww_k(x_k) & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix} \left[\prod_{i=k+1}^N X_i \right] e^+ \quad (11)$$

$$P_x(k) = \frac{1}{Z_{HP}} \sum_i SS[\chi_i] G_{x_k}$$

To calculate the probability $P_x(\alpha|\beta)$ of states in which a contact between two residues α and β is formed, the appropriate states must be selected from the generating function. An efficient way to do so is to use the state selection operators corresponding to the contact sets that contain the particular interaction of interest. As the only way that individual terms can be included in the partition function is to be selected by one or another of the selection operators, using only those contact sets containing the desired contact provides all the desired states.

$$P_x(\alpha|\beta) = \frac{1}{Z_{HP}} \sum_{k,s,t,x_\alpha,x_\beta \in \chi_k} SS[\chi_k] G_{HP} \quad (12)$$

Structural specificity

Calculations of $P_x(k)$ show the probability that a given site participates in a contact is not necessarily equal to the same probability for its partner in the lowest energy state (Fig. 4).

This observation suggests that some nonnative contacts are significantly populated and, moreover, that the XHC model can be used to analyze the prevalence of nonnative tertiary contacts in the unfolded state ensemble. The general trend is that, for x values above 4 or 5, the presumed native contacts of the helical hairpin are predicted to become the most probable. However, a significant population of nonnative contacts persists even at substantially higher values of x . Therefore, the model simulation of the $\alpha\alpha$ helical hairpin displays “gemisch” behavior, in that the lowest energy state is not sufficiently lower in energy than other states to fully dominate the ensemble of conformations (Dill et al. 1995).

The specificity of the structure as predicted by the model can be more clearly seen in Figure 5. Larger values of x do drive the formation of a specific native structure. However the physical interactions represented by this parameter cannot be arbitrarily strong, and specificity must be achieved by working within the constraints imposed by the intrinsic properties of the amino acids. It is not effective to simply include sufficient numbers of interactions so as to accumulate a favorable energy. Both the helix-coil equilibrium and the arrangement of tertiary contacts balance favorable interactions against the unfavorable reduction in the number of configurational states upon formation of these interactions. The inclusion of too many tertiary interaction sites in the sequence could potentially increase the entropic cost of forming a subset of the possible interactions so as to overwhelm the interactions themselves. One route to a more specific structure could be to include only the minimum number of such sites.

The XHC model provides a quantitative basis to investigate the complex interplay of the influences of

both secondary and tertiary structure. In particular, it is able to take into account not only reciprocity but also opposition between the two. For example, forming a tertiary contact in a region of the sequence with little helical tendency could undermine the favorable contribution of the contact itself, as it would be entropically more costly to order such a region. But the result would be a more cooperative transition, in which breaking such a contact leads to loss of helicity. Given an estimate of tertiary contact parameter values, the model could be used to rationally devise sequences with a minimal number of tertiary interaction sites just sufficient to tip the balance toward helix formation, thereby achieving both stability and specificity. Doing so would mimic the design principle utilized by evolution in the case of natural proteins which are composed of helices intrinsically unstable in isolation and yet which in combination confer a highly specific structure.

Figure 4. Tertiary contact predictions. (A–D) Sitewise tertiary contact probabilities $P_x(k)$ as a function of x . Solid and dashed lines represent, respectively, residue 14 and 25 (A), 11 and 28 (B), 7 and 32 (C), and 3 and 36 (D). (E–H) Pairwise tertiary contact probabilities $P_x(\alpha|\beta)$ as a function of x . (E) Contact probability between 14 and 25 (solid line), 28 (dashed), 32 (dotted), and 36 (dot-dash). (F) Contact probability between 11 and 28 (solid), 25 (dashed), 32 (dotted), and 36 (dot-dash). (G) Contact probability between 7 and 32 (solid), 28 (dashed), 36 (dotted), and 25 (dot-dash). (H) Contact probability between 3 and 36 (solid), 32 (dashed), 28 (dotted), and 25 (dot-dash).

Figure 5. Relative tertiary contact probabilities as function of x . (A–F) In each grid, the probability of the 16 tertiary contacts between residues 3, 7, 11, and 14 (vertical axes) and residues 25, 28, 32, and 36 (horizontal axes) are shown. The lighter colors indicate higher probabilities. The positions along the diagonal correspond to the contacts present in the lowest energy tertiary contact state of the molecule, schematically represented in G.

Acknowledgments

I thank Megan McEvoy for the use of computer facilities and John Osterhout for helpful discussions and encouragement as well as for a critical reading of the manuscript. I also thank John Schellman for helpful suggestions on an early version of this work and for an influential introduction to helix-coil theory as a graduate student. This work was supported by the Department of Biochemistry and Molecular Biophysics, The University of Arizona, Tucson, Arizona.

References

- Baldwin, R.L. and Rose, G.D. 1999. Is protein folding hierarchic? I: Local structure and peptide folding. *Trends Biochem. Sci.* **24**: 26–33.
- Becktel, W.J. and Schellman, J.A. 1987. Protein stability curves. *Biopolymers* **26**: 1859–1877.
- Blaber, M., Zhang, X.J., and Matthews, B.W. 1993. Structural basis of amino-acid α -helix propensity. *Science* **260**: 1637–1640.
- Blaber, M., Zhang, X.J., Lindstrom, J.D., Pepiot, S.D., Baase, W.A., and Matthews, B.W. 1994. Determination of α -helix propensity within the context of a folded protein: Sites 44 and 131 in bacteriophage-T4 lysozyme. *J. Mol. Biol.* **235**: 600–624.
- Chakrabarty, A., Doig, A.J., and Baldwin, R.L. 1993. Helix capping propensities in peptides parallel those in proteins. *Proc. Natl. Acad. Sci.* **90**: 11332–11336.
- Chakrabarty, A., Kortemme, T., and Baldwin, R.L. 1994. Helix propensities of the amino-acids measured in alanine-based peptides without helix-stabilizing side-chain interactions. *Protein Sci.* **3**: 843–852.
- Chan, H.S. and Dill, K.A. 1991. Sequence space soup of proteins and copolymers. *J. Chem. Phys.* **95**: 3775–3787.
- Clarke, J. and Itzhaki, L.S. 1998. Hydrogen exchange and protein folding. *Curr. Opin. Struct. Biol.* **8**: 112–118.
- Cordes, M.H.J., Davidson, A.R., and Sauer, R.T. 1996. Sequence space, folding and protein design. *Curr. Opin. Struct. Biol.* **6**: 3–10.
- Daggett, V. and Fersht, A.R. 2003. Is there a unifying mechanism for protein folding? *Trends Biochem. Sci.* **28**: 18–25.
- Dill, K.A., Bromberg, S., Yue, K.Z., Fiebig, K.M., Yee, D.P., Thomas, P.D., and Chan, H.S. 1995. Principles of protein-folding: A perspective from simple exact models. *Protein Sci.* **4**: 561–602.
- Doig, A.J. 2002. Recent advances in helix-coil theory. *Biophys. Chem.* **101**: 281–293.
- Doig, A.J. and Baldwin, R.L. 1995. N- and C-capping preferences for all 20 amino-acids in α -helical peptides. *Protein Sci.* **4**: 1325–1336.
- Doig, A.J., Chakrabarty, A., Klingler, T.M., and Baldwin, R.L. 1994. Determination of free-energies of N-capping in α -helices by modification of the Lifson-Roig helix-coil theory to include N-capping and C-capping. *Biochemistry* **33**: 3396–3403.

- Englander, S.W. 2000. Protein folding intermediates and pathways studied by hydrogen exchange. *Annu. Rev. Biophys. Biomol. Struct.* **29**: 213–238.
- Englander, S.W. and Kallenbach, N.R. 1983. Hydrogen-exchange and structural dynamics of proteins and nucleic-acids. *Q. Rev. Biophys.* **16**: 521–655.
- Fezoui, Y., Weaver, D.L., and Osterhout, J.J. 1994. De-novo design and structural characterization of an α -helical hairpin peptide: A model system for the study of protein-folding intermediates. *Proc. Natl. Acad. Sci.* **91**: 3675–3679.
- Fezoui, Y., Braswell, E.H., Xian, W.J., and Osterhout, J.J. 1999. Dissection of the de novo designed peptide $\alpha\alpha$: Stability and properties of the intact molecule and its constituent helices. *Biochemistry* **38**: 2796–2804.
- Horovitz, A., Matthews, J.M., and Fersht, A.R. 1992. α -Helix stability in proteins 2: Factors that influence stability at an internal position. *J. Mol. Biol.* **227**: 560–568.
- Jennings, P.A. and Wright, P.E. 1993. Formation of a molten globule intermediate early in the kinetic folding pathway of apomyoglobin. *Science* **262**: 892–896.
- Kauzmann, W. 1959. Some factors in the interpretation of protein denaturation. *Adv. Protein Chem.* **14**: 1–63.
- Kim, P.S. and Baldwin, R.L. 1982. Specific intermediates in the folding reactions of small proteins and the mechanism of protein folding. *Annu. Rev. Biochem.* **51**: 459–489.
- Lau, K.F. and Dill, K.A. 1989. A lattice statistical-mechanics model of the conformational and sequence-spaces of proteins. *Macromolecules* **22**: 3986–3997.
- Lifson, S. and Roig, A. 1961. Theory of helix-coil transition in polypeptides. *J. Chem. Phys.* **34**: 1963–1974.
- Marshall, S.A. and Mayo, S.L. 2001. Achieving stability and conformational specificity in designed proteins via binary patterning. *J. Mol. Biol.* **305**: 619–631.
- Munoz, V. and Serrano, L. 1994. Elucidating the folding problem of helical peptides using empirical parameters. *Nat. Struct. Biol.* **1**: 399–409.
- Myers, J.K., Pace, C.N., and Scholtz, J.M. 1997. Helix propensities are identical in proteins and peptides. *Biochemistry* **36**: 10923–10929.
- O'Neil, K.T. and DeGrado, W.F. 1990. A thermodynamic scale for the helix-forming tendencies of the commonly occurring amino-acids. *Science* **250**: 646–651.
- Qian, H. 1994. A thermodynamic model for the helix-coil transition coupled to dimerization of short coiled-coil peptides. *Biophys. J.* **67**: 349–355.
- Qian, H. and Schellman, J.A. 1992. Helix-coil theories: A comparative-study for finite length polypeptides. *J. Phys. Chem.* **96**: 3987–3994.
- Qian, H., Mayo, S.L., and Morton, A. 1994. Protein hydrogen-exchange in denaturant: Quantitative-analysis by a two-process model. *Biochemistry* **33**: 8167–8171.
- Raschke, T.M. and Marqusee, S. 1998. Hydrogen exchange studies of protein structure. *Curr. Opin. Biotechnol.* **9**: 80–86.
- Rohl, C.A. and Baldwin, R.L. 1997. Comparison of NH exchange and circular dichroism as techniques for measuring the parameters of the helix-coil transition in peptides. *Biochemistry* **36**: 8435–8442.
- Rohl, C.A., Scholtz, J.M., York, E.J., Stewart, J.M., and Baldwin, R.L. 1992. Kinetics of amide proton-exchange in helical peptides of varying chain lengths: Interpretation by the Lifson-Roig equation. *Biochemistry* **31**: 1263–1269.
- Schellman, J.A. 1955. The stability of hydrogen-bonded peptide structures in aqueous solution. *C.R. Trav. Lab. Carlsberg [Chim.]* **29**: 230–259.
- Sun, J.K., Penel, S., and Doig, A.J. 2000. Determination of α -helix N1 energies after addition of N1, N2, and N3 preferences to helix/coil theory. *Protein Sci.* **9**: 750–754.
- Tanford, C. 1962. Contribution of hydrophobic interactions to stability of globular conformation of proteins. *J. Am. Chem. Soc.* **84**: 4240–4247.
- West, M.W. and Hecht, M.H. 1995. Binary patterning of polar and nonpolar amino-acids in the sequences and structures of native proteins. *Protein Sci.* **4**: 2032–2039.
- Wojcik, J., Altmann, K.H., and Scheraga, H.A. 1990. Helix coil stability-constants for the naturally-occurring amino-acids in water 24: Half-cystine parameters from random poly(hydroxybutylglutamine-co-s-methylthio-l-cysteine). *Biopolymers* **30**: 121–134.
- Woodward, C.K. and Hilton, B.D. 1980. Hydrogen isotope exchange kinetics of single protons in bovine pancreatic trypsin-inhibitor. *Biophys. J.* **32**: 561–575.
- Xian, W., Conolly, P.J., Oslin, M., Hausrath, A.C., and Osterhout, J.J. 2006. Fundamental interactions in protein folding: Measuring the energetic balance between helix formation and hydrophobic interactions. *Protein Sci.* (in press).
- Zimm, B.H. and Bragg, J.K. 1959. Theory of the phase transition between helix and random coil in polypeptide chains. *J. Chem. Phys.* **31**: 526–535.