Multiple Sequence Alignment as a Guideline for Protein Engineering Strategies

Alan R. Davidson

Summary

Many proteins lack the thermodynamic stability and/or solubility that is required for their use in a desired application. For this reason, it can be advantageous to improve these qualities through rational protein engineering. An effective means for achieving this goal is to use sequence alignment analysis to select amino acid substitutions that are likely to increase the thermodynamic stability or solubility of a protein. Advantages of using this approach are that generally only a small number of substitutions need to be tested, these substitutions are rarely debilitating to protein function, and knowledge of the three-dimensional structure of the protein of interest is not required. This chapter will describe approaches that have been used to exploit the information contained in sequence alignments for the engineering of improved protein properties.

Key Words: Protein solubility; mutagenesis; sequence alignment; protein engineering; protein stability.

1. Introduction

Although every protein present in living organisms has been honed by evolution to fold and function well in its natural environment, many of these same proteins stubbornly refuse to behave as desired by scientists who wish to exploit them for the advancement of their research projects. For this reason, much effort has been expended in developing generalized methods to alter the sequences of unco-operative proteins to make them more amenable to in vitro investigation. A central goal in these studies has been to overcome the two of the most common shortcomings of natural proteins: low thermodynamic stability and low solubility. In this chapter, the use of sequence alignment analysis to design amino acid substitutions that increase the stability or solubility of proteins is discussed. This technique has proven to be both simple and reliable.
Over the years, many different approaches have been used to rationally design stabilizing substitutions in a variety of proteins (1–3). These approaches often involve using the atomic resolution structure of the protein of interest to select substitutions aimed at introducing new favorable interactions or eliminating existing unfavorable ones. Comparison with the structures of thermophilic homologues can aid in such studies (4). However, structure-based approaches have the disadvantage of requiring the three-dimensional structure of the protein of interest to be known. A different approach to protein stabilization is to randomly mutagenize a protein and then perform multiple rounds of in vitro or in vivo selection to evolve the protein towards the desired properties (5). This methodology has also been successful, but requires an easily performed selective assay for the activity or stability of the protein of interest.

The design of stabilizing mutations through sequence alignment analysis provides a very advantageous alternative to other strategies. Sequence alignment-based methods involve comparing the sequence of a protein of interest with an alignment of homologous sequences. Positions in the protein of interest that are occupied by amino acids rarely seen in related sequences are substituted with the amino acid that is observed most often. Variations on this basic approach have been used to stabilize many different proteins including immunoglobulin domains (6,7), an SH3 domain (8), GroEL minichaperone (9), p53 (10), phytase (11,12), WW domains (13), and ankyrin repeats (14). Single mutations have been isolated in these studies that increase the melting temperature of the protein ($T_m$) by more than 15°C, and increase free energy of unfolding ($\Delta G_u$) by more than 1 kcal/mol. The effects of substitutions identified by this method are generally additive so that multiple mutants have been produce with increases of $T_m$ values of more than 30°C, and increases in $\Delta G_u$ of more than 3 kcal/mol (8). The success rate of the method as judged by whether predicted mutants are actually stabilizing ranges between 25 and 50% depending on the study. The advantages of sequence alignment-based methods are that they are simple to carry out, they do not require knowledge of the three-dimensional structure of the protein under investigation, and the resulting mutants generally maintain their biological function. Furthermore, the method has been shown to work with alignments ranging from thirteen to thousands of sequences.

Sequence alignment based methods can also aid in increasing the solubility of some proteins. For example, substitutions selected to stabilize a $\kappa$ domain intrabody also greatly increased the solubility of the domain in vivo in *Escherichia coli* (7). Alignment based methods specifically aimed at identifying substitutions to increase solubility have recently been successfully used (15–17), and these methods will also be discussed here.
Fig. 1. A portion of an SH3 domain sequence alignment. This small alignment is representative of the complete 266 member SH3 domain alignment used in a protein design project carried out in my laboratory (8). Conserved hydrophobic core positions are shaded; other highly conserved positions are boxed. Examination of the conserved positions is helpful in verifying the correctness of the alignment. It should be noted that gaps are seen between secondary structure elements, but not within in them, and that gap lengths can be quite variable. The PDB accession number is shown for each sequence.

2. Methods

2.1. Construction of a Sequence Alignment

The construction of a high-quality sequence alignment is a key step in any sequence alignment-based protein engineering strategy. Although there are many existing databases containing preassembled sequence alignments for most proteins (see Note 1), it is often best to construct an alignment from scratch that will be most suited for analysis of a particular protein of interest. The basic steps in constructing an alignment are outlined below. A more detailed description of the construction and analysis of a sequence alignment can be found in a review by Irving et al. (18). To illustrate some of the points made, an example of part of a large SH3 domain sequence alignment constructed in my laboratory is shown in Fig. 1.
2.1.1. Identification of Sequences Related to the Protein of Interest

2. Paste the amino acid sequence of your protein of interest into the box and submit using the default settings.
3. Press “Format” to see the results of your search after it has been completed.
4. Run multiple iterations of PSI-BLAST to maximize the chance that a diverse set of sequences related to your sequence of interest will be obtained (see Note 2).

2.1.2. Select Sequences to Use in the Alignment

1. Examine the pairwise alignments of your sequence with the sequences that have been identified in the PSI-BLAST search.
2. Select sequences of interest for sequence retrieval. This is done by putting a checkmark in the box at the beginning of the sequence entry. In choosing sequences to include in your alignment, avoid including too many obviously redundant sequences and sequences that are so distantly related to your sequence that alignment may be a problem. As a general rule, sequences with greater than 30% identity to your sequence of interest should be included.
3. After sequences have been selected, press “Get Selected Sequences”, then select Display “FASTA” and Send To “Text”. This should produce a window containing all of your selected sequence in FASTA format. These sequences should be saved as text.

2.1.3. Aligning the Selected Sequences

The most commonly used program for sequence alignment is Clustal (19). This program can be downloaded (ftp://ftp-igbmc.u-strasbg.fr/pub/ClustalX/) or run at a web site (http://www.ebi.ac.uk/clustalw/). Your sequences saved in FASTA format can be loaded directly into this program. There are many parameters that can be adjusted when using Clustal to create an alignment, and discussion of these is beyond the scope of this chapter. Detailed help menus are included with the program. If a three-dimensional structure of your protein or a closely related homologue is known, this structure can be used by the program as an aid in determining gap positions. Phylogenetic trees of your sequences can also be constructed as a means to evaluate the diversity of your alignment.

2.1.4. Refining the Alignment

To obtain reliable results from a sequence alignment, ensure that the sequences are aligned as well as possible. In many cases, sequence alignments must be manually adjusted after the use of an automated alignment program. A very good program to use for the viewing, editing, and analysis of sequence alignments is Jalview (http://www.jalview.org) (20). This program will automatically color conserve positions in the alignment, which is useful for detecting problem areas. Poorly aligned regions functions under the heading “Quali-
ators should be taken into consideration:
1. Sequence gaps (i.e., insertions or exons) are considered to be problematic. Gaps should generally be avoided.
2. The alignment should possess conserved residues that are predicted to respond to the hydrophobic core of the protein.
3. Sequences possessing unusual residue compositions may be examined closely to ensure that they are not homologous to the other sequences.
4. Sequences that are very difficult to align may be eliminated. It is best to err on the side of caution.
5. Highly redundant sequences should be removed if a large number of sequences that are nearly identical in frequency of occurrence of residues are found. In Jalview, in the “Calculate Redundancy” box, the level of redundancy to be removed is set at 90%. Sequences will be grouped as to their sequences display more than 90% identity. They may be eliminated if necessary.

2.2. Using the Sequence Alignment to Stabilize Amino Acid Substitutions

The basis of this method is to place the protein of interest to the frequency of occurrence in the sequence alignment. Gap positions within the protein of interest are assigned as what is observed in nature.

1. Calculate the frequency of occurrence of each amino acid at each gap position. These amino acids have been aligned with each other, and therefore positions near or within domains.
2. Compare the frequency of occurrence of each amino acid in the protein of interest in the consensus sequence. Position the gaps in the consensus sequence of the type of sequence domain so that the amino acid sequence does not break the consensus. In this way, we can place the gaps to maximize the frequency of occurrence of each amino acid.
problem areas. Poorly aligned regions can be detected in Clustal by selecting functions under the heading “Quality.” In improving an alignment, several factors should be taken into consideration:

1. Sequence gaps (i.e., insertions or deletions) should be minimized as much as possible. Gaps should generally lie between secondary structural elements.
2. The alignment should possess conserved hydrophobic positions at positions corresponding to the hydrophobic core of the protein. Polar residues should rarely be seen at these positions.
3. Sequences possessing unusual residues at highly conserved positions should be examined closely to ensure that they are not misaligned and that they are truly homologous to the other sequences throughout their length.
4. Sequences that are very difficult to align to the rest of the sequences should be eliminated. It is best to err on the side of caution in this respect.
5. Highly redundant sequences should be eliminated. If the alignment contains a large number of sequences that are very closely related, the calculations of the frequency of occurrence of residues at each position in the alignment will be skewed. In Jalview, in the “Calculate” menu, choose “Remove Redundancy.” The level of redundancy to be removed can be set at any level. For example, if it is set at 90%, sequences will be eliminated from the alignment so that no two sequences display more than 90% identity. Choosing the level of redundancy to eliminate may depend on the alignment being used. Many groups including my own routinely use 90% as the redundancy cutoff.

2.2. Using the Sequence Alignment to Design Stabilizing Amino Acid Substitutions

The basis of the design methodology is to compare the sequence of your protein of interest to the frequencies of occurrence of amino acids at each position in the sequence alignment. The goal of this comparison is to identify amino acids within the protein of interest that are relatively rare in comparison to what is observed in most homologues. The basic steps are as follows.

1. Calculate the frequency of occurrence of the most commonly observed amino acid at each position of the alignment (see Notes 3 and 4). Only positions that have been aligned with confidence should be considered in this analysis (e.g., positions near or within gaps may not be suitable).
2. Compare the frequency of occurrence of each residue at each position in your protein of interest to that of the most commonly occurring residue at each position. Positions in your protein that possess relatively rare residues compared with the consensus residue should be targeted for mutagenesis. An example of this type of sequence analysis is shown in Fig. 2, in which the sequence of the SH3 domain from the Abp1 protein (Abp1p) of yeast is compared to the consensus sequence derived from a 266 sequence SH3 domain alignment (8). In this case, we chose to mutate all positions at which the WT residue occurred with a relative frequency of 0.15 or less, where the relative frequency is the frequency of occur-
Fig. 2. Alignment of the Aβ1p SH3 domain sequence with the SH3 domain consensus sequence. Based on a 266-member sequence alignment, the most frequently observed amino acid residue at each position in the SH3 domain was determined and is given as the consensus sequence. The numbers below each sequence are the occurrence frequency (expressed as %) in the alignment observed for each residue in the Aβ1p SH3 domain and in the SH3 domain consensus sequence. The relative frequency of occurrence (Rel. Freq.) was obtained by dividing the Aβ1p SH3 domain residue occurrence with that of the consensus residue. The boxed positions had relative frequencies of less than 0.15 were selected for mutagenesis. Three substitutions (E7L, N23G, and V21K) led to significant stabilization of the Aβ1p SH3 domain (8).
rence of the residue in Abp1 divided by the frequency of occurrence of the residue that occurred most frequently at the given position, i.e., the consensus residue (see Note 5).

3. Use a standard polymerase chain reaction-based, site-directed mutagenesis protocol to replace rarely occurring residues in your protein of interest with the residue that occurs most commonly at that position in the alignment.

4. Purify the mutant proteins and assay their thermodynamic stability using spectroscopic or any other amenable techniques.

5. Combine single-site mutants with the largest stabilizing effects into one multiple mutant. Because the effects of mutants designed by sequence alignment are usually additive, the multiple mutant will likely be more stable than the singly substituted mutants (see Note 6).

2.3. Using Sequence Alignment Analysis to Design Amino Acid Substitutions to Increase Protein Solubility

In at least one case, the use of sequence alignment analysis to design stabilized mutants as described led to the isolation of proteins that were both more stable and more soluble (7). However, a different alignment-based approach has also been successfully used to design mutations specifically aimed at increasing protein solubility (15–17). This approach is carried out as follows.

1. Create an alignment as described in Subheading 2.1, containing the most closely related homologues to the protein of interest (a higher redundancy cutoff could be used to create this alignment).

2. Identify positions in the alignment at which your protein of interest possesses a hydrophobic residue and one or more (preferably more than one) closely related homologues possesses a polar residue. These positions are likely to lie on the surface of the protein structure as polar residues are rarely found in buried positions. The presence of a hydrophobic residue on the protein surface could lead to lowered solubility. An example of an alignment that could have been used to pinpoint solubilizing substitutions is shown in Fig. 3.

3. Substitute the targeted hydrophobic residue in your protein with a polar residue seen in a closely related sequence. Purify the mutant protein and assess protein solubility.

4. Combine solubilizing substitutions into one multiple mutant. If many potentially solubilizing mutants are identified, it may be more efficient to construct multiple mutants first. These substitutions are unlikely to be detrimental to protein structure or function because the substituted amino acids are seen in very closely related proteins.

3. Conclusions

Sequence alignments undoubtedly provide extremely useful information for the design of mutants that increase both the stability and solubility of proteins. The design of mutants using the methodology described is simple and does not require a known three-dimensional structure for a protein of interest. In addi-
Fig. 3. Identification of a solubility enhancing substitution in the Cystic Fibrosis Transmembrane Conductance Regulator (CFTR). Sequences of a short region of CFTR from the indicated species are shown. This region is within the first nucleotide-binding domain of CFTR (NBD1). A high degree of conservation is seen at most positions. However, position 429 is very hydrophobic in the human sequence (Phe), but Ser, a polar residue, is seen in the mouse and rat sequences. The F429S substitution was found to significantly increase the in vitro purified soluble yield of a human CFTR NBD1 construct (25).
tion, mutants selected using these techniques rarely cause an alteration of protein function.

4. Notes

1. Some databases containing alignments for various proteins or domains are Pfam (http://www.sanger.ac.uk/Software/Pfam/) (21), SMART (http://smart.ox.ac.uk/) (22), and HSSP (http://www.cmbi.kun.nl/gv/hssp/) (23). Because alignments in these databases are generated in a fully automated manner, care must be taken to refine them as described in Subheading 2.1.4.

2. The size and diversity of alignments that have been used for protein design have varied widely. The number of aligned sequences has ranged from 13 (11) to 4000 (14). At this point, the properties of an alignment needed to produce the best results are not known. As long as the sequences used are aligned well and are reasonably diverse, then it should be possible to identify stabilizing substitutions.

3. Unfortunately I do not know of any web-based servers that will automatically calculate the frequency of occurrence of all the amino acids at each position in an alignment. These calculations can be quite tedious for a large alignment. It is fairly straightforward to write a script in Perl (http://www.perl.com/) to perform this task. Another approach is to load the alignment into a Microsoft Excel worksheet. Once loaded, the “Autofilter” procedure (this command is found by selecting “Data,” then “Filter”) provides a very simple and powerful means to count the occurrences of any residue at any position of the alignment. Although it is rarely used for sequence alignment editing and display, Excel is also useful for these purposes.

4. A complication in calculating residue frequencies from any sequence alignment is in dealing with skew that may result from overrepresentation within an alignment of clusters of highly related sequences. Although using a redundancy threshold as described can help with this problem, further improvement in frequency calculation can be obtained by using a weighting algorithm to downweight the effect of overrepresented groups of sequences. A commonly used weighting algorithm that we have used in our studies is that of Henikoff and Henikoff (24). This procedure can be implemented through Perl scripts. If sequences to be included in an alignment are chosen carefully, weighting should not be essential for the success of the experiments describe in this chapter. Weighting protocols were not used in most of the studies referred to in this chapter.

5. The criteria used for the selection of positions to mutate has varied considerably from study to study. Thus, no strict rule can be stated on this issue. The most appropriate relative frequency cutoff will likely be different depending on the overall diversity of the alignment. The logical approach to this problem is to simply decide how many mutants can be reasonably made and then pick that number of positions displaying the lowest relative frequencies as compared with the consensus residue.

6. Some studies have taken the approach of first creating a “consensus” sequence of the protein of interest by substituting all the most favorable amino acids into it at one time (11, 14). However, because approximately half of the
substitutions designed by the sequence alignment based method are generally found to be destabilizing, it seems better to make single substitutions first to identify which ones are stabilizing.

Acknowledgments

Work pertaining to this chapter was funded by the Canadian Institutes of Health Research and by the National Cancer Institute of Canada. I thank Dr. Karen L. Maxwell for critical reading of the manuscript.

References


Introduction to Protein Structure

Second Edition

Carl Branden & John Tooze
Over a period of more than 3 billion years, a large variety of protein molecules has evolved to become the complex machinery of present-day cells and organisms. These molecules have evolved by random changes of genes by point mutations, exon shuffling, recombination and gene transfer between species, in combination with natural selection for those gene products that have conferred some functional advantage contributing to the survival of individual organisms.

Long before Darwin and Wallace proposed the theory of evolution and Mendel discovered the laws of genetics, plant and animal breeders had begun to interfere with the process of evolution in the species that gave rise to domesticated animals and cultivated plants. Considering their total lack of knowledge of both evolutionary theory and genetics, their achievements, brought about by forcing the pace of and subverting natural selection, were impressive albeit very gradual. With the advent of molecular genetics and in particular techniques for gene manipulation, we have now entered an era of genetic exploitation of organisms undreamed of only 50 years ago. We can now design genes to produce, in host organisms, novel gene products for the benefit of human beings; we are no longer restricted to selecting useful genes that arise by mutation. We are, however, only at the beginning of this new era, and so far we have only scratched the surface of the knowledge that is required for true engineering and design of protein molecules. We distinguish protein engineering, by which we mean mutating the gene of an existing protein in an attempt to alter its function in a predictable way, from protein design, which has the more ambitious goal of designing de novo a protein to fulfill a desired function.

Genome projects have now provided us with a description of the complete sequences of all the genes in more than a dozen organisms, and they will provide many more complete genome sequences within the next decade, including that of the human genome. These databases provide great opportunities for the analysis and exploitation of genes and their corresponding proteins. Central to reaping the intellectual and commercial benefits of this genetic information is the ability to find out the function of individual gene products. Almost all functional assignments to date have been based on sequence similarity to proteins of known function.

Knowledge of a protein's tertiary structure is a prerequisite for the proper understanding and engineering of its function. Unfortunately, in spite of recent significant technological advances, the experimental determination
of tertiary structure is still slow compared with the rate of accumulation of amino acid sequence data. This makes the folding problem, the successful prediction of a protein's tertiary structure from its amino acid sequence, central to rapid progress in post-genomic biology. We will, therefore, in this chapter first briefly describe implications of protein homology and methods for the prediction of secondary and tertiary structure before giving some examples of protein engineering and protein design.

**Homologous proteins have similar structure and function**

The term homology as used in a biological context is defined as similarity of structure, physiology, development and evolution of organisms based upon common genetic factors. The statement that two proteins are homologous therefore implies that their genes have evolved from a common ancestral gene.

Homologous proteins are mostly recognized by statistically significant similarities in their amino acid sequences. Usually, they also have similar functions although there are some known exceptions, where genes for ancient enzymes have been recruited at a later stage in evolution to produce proteins with quite different functions. An example is provided by one of the structural components in the eye lens that is homologous to the ancient glycolytic enzyme lactate dehydrogenase. Once a novel gene has been cloned and sequenced, a search for amino acid sequence similarity between the corresponding protein and other known protein sequences should be made. Usually, this is done by comparison with databases of known protein sequences using one of the standard sequence alignment computer programs.

Two proteins are considered to be homologous when they have identical amino acid residues in a significant number of sequential positions along the polypeptide chains. Using statistical methods based on comparisons of computer-generated random sequences, it is relatively straightforward to assess how many positions need to be identical for a statistically significant identity between two sequences. However, it is frequently found that two proteins with sequence identity below the level of statistical significance have similar functions and similar three-dimensional structures. In these cases, functionally important residues are identical and usually such residues form sequence patterns or motifs that can be used to identify other proteins that belong to the same functional family. Frequently, members of such families are also considered to be homologous, even though the identities are not statistically significant, only functionally significant. Databases for such families, based on identical or similar sequence motifs, are available on the World Wide Web (see pp. 393-394) and they are very useful for assigning function to a novel protein.

If significant amino acid sequence identity is found with a protein of known crystal structure, a three-dimensional model of the novel protein can be constructed, using computer modeling, on the basis of the sequence alignment and the known three-dimensional structure. This model can then serve as an excellent basis for identifying amino acid residues involved in the active site or in antigenic epitopes, and the model can be used for protein engineering, drug design, or immunological studies.

Since the sequence databases are large and growing exponentially, currently comprising more than 500,000 known protein sequences, the standard sequence alignment programs have been designed to provide a compromise between the speed and the accuracy of the search. As a result, they work well only when there is a reasonably high degree of sequence identity, usually of the order of 30% or more. Much more sensitive programs have been written that search for both identity and conserved structural properties and also for relatedness in different physical properties, but these inevitably require far more computing time. Carefully used, such programs can identify structural and functional similarity where the standard programs fail to do so.
Homologous proteins always contain a core region where the general folds of the polypeptide chains are very similar. This core region contains mainly the secondary structure elements that build up the interior of the protein; in other words, the scaffolds of homologous proteins have similar three-dimensional structures. Even distantly related proteins with low sequence identity have similar scaffold structures, although minor adjustments occur in the positions of the secondary structure elements to accommodate differences in the arrangements of the hydrophobic side chains in the interior of the protein. The greater the sequence identity, the more closely related are the scaffold structures (Figure 17.1). This has important implications for model building of homologous proteins; the more distantly related two proteins are, the more the scaffold must be adjusted to model the new structure.

Loop regions that connect the building blocks of scaffolds can vary considerably both in length and in structure. The problem of predicting the three-dimensional structure of a protein that is homologous to a protein of known three-dimensional structure is therefore mainly a question of predicting the structure of loop regions and side-chain conformations, after the scaffold has been adjusted. As mentioned in Chapter 2, loop regions do not have random structures, and their main-chain conformations cluster in sets of similar structures. The conformation of each set depends more on the number of amino acids in the loop and the type of secondary structure elements that it connects, whether they are α-α, β-β, α-β, or β-α connections, than on the actual amino acid sequences. Therefore it is possible to use a database of loop regions from proteins of known structure to obtain a preliminary model of the loops of an unknown structure. To model a protein structure, suitable main-chain loop conformations from this database are attached to the scaffold modeled to have a structure similar to that of the known homologous protein. Finally, the conformations of the side chains are predicted by energy refinement of the model, which minimizes the free energy of the protein by maximizing the interaction energies of the amino acids. Analysis of structures determined to high resolution has shown that only a few side-chain conformations frequently occur. These are called rotamers and model building of side chains employs databases of such rotamers.

An instructive example of the use of such procedures has been in modeling antigen-binding sites in immunoglobulins. These binding sites are built up from three hypervariable loop regions, CDR1–CDR3, from the variable domains of both the light and the heavy chains of immunoglobulins as described in Chapter 15. There is usually high sequence identity within the scaffolds of the variable domains in different immunoglobulin molecules. Consequently, the scaffold of variable domains of known three-dimensional structures can be used in modeling a new monoclonal antibody with a known amino acid sequence. However, the CDR regions of a new antibody are usually very different in sequence from those of any other known antibody, and their three-dimensional structures must be predicted. By comparing

Figure 17.1 The relation between the divergence of amino acid sequence and three-dimensional structure of the core region of homologous proteins. Known structures of 32 pairs of homologous proteins such as globins, serine proteinases, and immunoglobulin domains have been compared. The root mean square deviation of the main-chain atoms of the core regions is plotted as a function of amino acid homology (red dots). The curve represents the best fit of the dots to an exponential function. Pairs with high sequence homology are almost identical in three-dimensional structure, whereas deviations in atomic positions for pairs of low homology are of the order of 2 Å. (From C. Chothia and A. Lesk, EMBO J. 5: 823–826, 1986.)
known antibody structures and sequences, it has been shown that there is only a small repertoire of main-chain conformations for at least five of the six CDR regions and that the particular conformation adopted is determined by a few key conserved residues for each loop conformation. For example, three different conformations were found for the CDR3 regions of the light chains in nine known x-ray structures. More than 90% of the known sequences of light-chain CDR3 regions obey the sequence constraints of one or other of these three conformations. By using this repertoire of loop conformations, considerable success has been achieved in correctly predicting the structure of antigen-binding surfaces. An example of such a prediction compared with the actual structure, subsequently determined, is given in Figure 17.2.

**Knowledge of secondary structure is necessary for prediction of tertiary structure**

What can be done by predictive methods if the sequence search fails to reveal any homology with a protein of known tertiary structure? Is it possible to model a tertiary structure from the amino acid sequence alone? There are no methods available today to do this and obtain a model detailed enough to be of any use, for example, in drug design and protein engineering. This is, however, a very active area of research and quite promising results are being obtained; in some cases it is possible to predict correctly the type of protein, α, β, or α/β, and even to derive approximations to the correct fold.

Today's predictive methods rely on prediction of secondary structure: in other words, which amino acid residues are α-helical and which are in β strands. We have emphasized in Chapter 12 that secondary structure cannot in general be predicted with a high degree of confidence with the possible exceptions of transmembrane helices and α-helical coiled coils. This imposes a basic limitation on the prediction of tertiary structure. Once the correct secondary structure is known, we know enough about the rules for packing elements of secondary structure against each other (see Chapter 2 for helix packing) to derive a very limited number of possible stable globular folds. Consequently, secondary structure prediction lies at the heart of the prediction of tertiary structure from the amino acid sequence.

---

**Figure 17.2** An example of prediction of the conformations of three CDR regions of a monoclonal antibody (top row) compared with the unrefined x-ray structure (bottom row). L1 and L2 are CDR regions of the light chain, and H1 is from the heavy chain. The amino acid sequences of the loop regions were modeled by comparison with the sequences of loop regions selected from a database of known antibody structures. The three-dimensional structure of two of the loop regions, L1 and L2, were in good agreement with the preliminary x-ray structure, whereas H1 was not. However, during later refinement of the x-ray structure errors were found in the conformations of H1, and in the refined x-ray structure this loop was found to agree with the predicted conformations. In fact, all six loop conformations were correctly predicted in this case. (From C. Chothia et al., *Science* 233: 755–758, 1986.)
Unfortunately for predictive methods, secondary and tertiary structures are closely linked in the sense that global tertiary structure imposes local secondary structure at least in some regions of the polypeptide chain. The ability of a specific short sequence of amino acids to form an α helix, a β strand, or a loop region is dependent not only on the sequence of that region but also on its environment in the three-dimensional structure. For example, by analyzing all the known tertiary structures, it has been shown that peptide regions of up to five residues long with identical amino acid sequences are α-helical in one structure and a β strand or a loop in other structures. While this interdependence of secondary and tertiary structure complicates secondary structure predictions, it can, sometimes, be used to improve such predictions, by an iterative scheme in which a preliminary assignment of secondary structure is used to predict the type of domain structure, for example, a four-helix bundle or an α/β barrel. The structure type of the domain imposes additional constraints on possible secondary structure, which can be used to refine the secondary structure prediction.

**Prediction methods for secondary structure benefit from multiple alignment of homologous proteins**

Over 20 different methods have been proposed for predictions of secondary structure; they can be categorized in two broad classes. The empirical statistical methods use parameters obtained from analyses of known sequences and tertiary structures. All such methods are based on the assumption that the local sequence in a short region of the polypeptide chain determines local structure; as we have seen, this is not a universally valid assumption. The second group of methods is based on stereochemical criteria, such as compactness of form with a tightly packed hydrophobic core and a polar surface. Three frequently used methods are the empirical approaches of P.Y. Chou and G.D. Fasman and of J. Garnier, D.J. Osguthorpe and B. Robson (the GOR method), and third, the stereochemical method of V.I. Lim.

Although these three methods use quite different approaches to the problem, the accuracy of their secondary structure prediction is about the same. All three methods can be used to assign one of three states to each residue: α helix, β strand, or loop. Random assignment of these three states to residues in a polypeptide chain will give an average score of 33% correctly predicted states. The methods have been assessed in an analysis of single sequences of a large number of known x-ray structures comprising more than 10,000 residues. For the three-state definition of secondary structure, the overall accuracy of prediction was about 55%. Other objective assessments have given similar results.

However, when these predictive methods are used on a set of homologous proteins the predictive power is considerably higher. The underlying assumption is that secondary and tertiary structure has been more conserved during evolution than amino acid sequence; in other words only such changes have been retained during evolution that conserve the structure. Consequently, the pattern of residue changes within homologous proteins contains specific information about the structure. Conserved hydrophobic residues are usually in the interior of the protein with a high probability of belonging to helices or sheet strands. Insertions and deletions almost always occur in loop regions and not in the scaffold built up from helices and strands.

Several programs are now available that use multiple alignment of homologous proteins for prediction of secondary structure. One such program, called PHD, which was developed by Chris Sander and coworkers, EMBL, Heidelberg, has reached a mean accuracy of prediction of 72% for new structures. A large fraction of the remaining errors occur at the ends of α helices and β strands and, in addition, some errors occur because of occasional difficulties
in distinguishing between α helices and β strands. These latter errors can be corrected if the structural class, α, β, or α/β, can be deduced from a combination of physical studies, for example, circular dichroism spectra, and the general features of the secondary structure prediction. For example, if the prediction scheme assigns one or two short α helices among many β strands in a protein of the β class, there is a high probability that the regions of secondary structures are essentially correctly predicted but that they should all be β strands.

These predictive methods are very useful in many contexts; for example, in the design of novel polypeptides for the identification of possible antigenic epitopes, in the analysis of common motifs in sequences that direct proteins into specific organelles (for instance, mitochondria), and to provide starting models for tertiary structure predictions.

Many different amino acid sequences give similar three-dimensional structures

How many completely different amino acid sequences might give a similar three-dimensional structure for an average-sized domain of 150 amino acid residues? Simple combinatorial calculations show that there are a total of \(20^{150}\) or roughly \(10^{200}\) possible amino acid sequences for such a domain, given the 20 different amino acids in natural proteins. This number is much larger than the number of atoms in the known universe. A more laborious calculation shows that out of these \(10^{200}\) possible combinations we can extract about \(10^{28}\) members that have less than 20% amino acid sequence identity with each other and that therefore can be considered to have different sequences. In other words, there are \(10^{28}\) different ways of constructing a domain of 150 amino acids using the 20 standard amino acids as building blocks. We do not know how many of these can form a stable three-dimensional structure but, assuming say that one out of a billion \((10^9)\) can, we are left with \(10^{29}\) folded possible proteins. In the previous chapters we have seen that simple structural motifs arrange themselves into a limited number of topologically different domain structures. It has been estimated on reasonable grounds that there are about 1000 topologically different domain structures. Since there are \(10^{29}\) possible different sequences that might fold into \(10^3\) different structures, it follows that there are of the order of \(10^{26}\) different side chain arrangements with less than 20% amino acid sequence identity that can give similar polypeptide folds. Only a small fraction of these possible proteins will be found in nature.

For each of the 500 or so different domain structures that have so far been observed, we might at best know about a dozen of these different possible sequences. It is not trivial to recognize the general sequence patterns that are common to specific domain structures from such a limited knowledge base.

Prediction of protein structure from sequence is an unsolved problem

How to predict the three-dimensional structure of a protein from its amino acid sequence is the major unsolved problem in structural molecular biology. We would like to have a computer program that could simulate the action of the processes that operate in a test tube or a living cell when a polypeptide chain with a specific amino acid sequence folds into a precise three-dimensional structure. Why is this prediction of protein folding so difficult? The answer is usually formulated in terms of the complexity of the task of searching through all the possible conformations of a polypeptide chain to find those with low energy. It requires enormous amounts of computing time, in addition to the complication discussed in Chapter 6 that the energy difference between a stable folded molecule and its unfolded state is a small number containing large errors.
With the realization that there are only a limited number of stable folds and many unrelated sequences that have the same fold, biologically oriented computer scientists started to address what is called the inverse folding problem: namely, which sequence patterns are compatible with a specific fold? If this question can be answered, such patterns could be used to search through the genome sequence databases and extract those sequences that have a specific fold, such as the α/β barrel or the immunoglobulin fold.

However, given the large number of possible unrelated sequences for each fold and the limited number of known sequences, a variation of this problem has recently been addressed by a large number of groups; namely, which of the known folds, if any, is most compatible with a specific sequence? The methodology used is called threading because it involves threading a specific sequence through all known folds and, for each fold, estimating the probability that the sequence can have that fold. Considerable progress has recently been made in threading, and in blind tests several structures have been correctly predicted by different groups.

**Threading methods can assign amino acid sequences to known three-dimensional folds**

Threading methods, which are also called protein fold assignments or fold recognition, are a promising and rapidly evolving field of computational structural biology. The goal is to assign to each genome-derived protein sequence the protein fold to which it most closely corresponds, or to determine whether there is no known fold to which the sequence belongs. A further goal is to align the new sequence properly to the three-dimensional structure of the fold to which it belongs to provide a low-resolution model. In order to test different methods of threading, blind tests are arranged, called Critical Assessment of Structure Prediction (CASP), in which the participants are given sequences and invited to predict the fold and make an alignment before the structure is determined experimentally. We will briefly describe here the methods used by one of the more successful participants in these tests, the group of David Eisenberg at University of California, Los Angeles.

The first requirement for threading is to have a database of all the known different protein folds. Eisenberg has used his own library of about 800 folds, which represents a minimally redundant set of the more than 6000 structures deposited at the Protein Data Bank. Other groups use databases available on the World Wide Web, where the folds are hierarchically ordered according to structural and functional similarities, such as SCOP, designed by Alexey Murzin and Cyrus Chothia in Cambridge, UK.

For each fold one searches for the best alignment of the target sequence that would be compatible with the fold; the core should comprise hydrophobic residues and polar residues should be on the outside, predicted helical and strand regions should be aligned to corresponding secondary structure elements in the fold, and so on. In order to match a sequence alignment to a fold, Eisenberg developed a rapid method called the 3D profile method. The environment of each residue position in the known 3D structure is characterized on the basis of three properties: (1) the area of the side chain that is buried by other protein atoms, (2) the fraction of side chain area that is covered by polar atoms, and (3) the secondary structure, which is classified in three states: helix, sheet, and coil. The residue positions are rather arbitrarily divided into six classes by properties 1 and 2, which in combination with property 3 yields 18 environmental classes. This classification of environments enables a protein structure to be coded by a sequence in an 18-letter alphabet, in which each letter represents the environmental class of a residue position.

Each of the 20 different amino acids has different preferences for each of the 18 environmental classes; for instance a Leu has a high preference for being in a helical class with a high fraction of buried side chain area, whereas
an Asp has a very low preference for that position. Numerical values for these preferences, called 3D-1D scores, were derived from a set of well-refined high-resolution protein structures, together with sets of sequences similar to the sequences of the 3D structures. This produced a scoring table in which for each environmental class a numerical value of preference is associated with each of the 20 amino acids. This table is used to set up a 3D profile table of a protein structure, in which each residue position is assigned an environmental class with corresponding numerical values for preference for each type of amino acid. The essence of this method is that the three-dimensional structure is reduced to a one-dimensional array, which facilitates matching to a one-dimensional sequence.

A target amino acid sequence is aligned against this structure profile in such a way that the best possible match—the highest total score—is obtained, allowing gaps and insertions. Such an alignment is conceptually similar to alignment of two sequences and similar methods have been used. The match of a sequence to a 3D structure profile for a specific fold is expressed quantitatively by a value called the Z-score, which is the number of standard deviations above the mean alignment score for other sequences of similar length. A high Z-score means there is a high probability that the sequence has the corresponding fold.

The methods described here have subsequently been improved and extended by Eisenberg, but the principle remains essentially the same. Other groups use different methods to screen the sequence-structure alignments and different criteria to assess the matches. Manfred Sippl at the University of Salzburg, Austria, has developed a set of potentials to screen and assess the alignments, the essence of which is to maximize the number of hydrophobic interactions and to minimize the number of buried polar atoms that do not participate in hydrogen bonds. These and similar potentials are now used by many groups in their threading programs. Correct folds can be predicted with a reasonably high probability for small and medium-sized proteins. Correct alignment of the sequence to the selected fold is, however, less accurate.

Proteins can be made more stable by engineering

Protein engineering, via site-directed mutagenesis of DNA, can be used to answer very specific questions about protein stability, and the results of these studies are now being used to increase the stability of industrially important enzymes. To illustrate some of the factors of importance for protein stability that have been revealed by protein engineering studies, we have chosen the extensive work on the enzyme lysozyme from bacteriophage T4 that has been done by the group of Brian Mathews, University of Oregon, Eugene.

Lysozyme from bacteriophage T4 is a 164 amino acid polypeptide chain that folds into two domains (Figure 17.3). There are no disulfide bridges; the two cysteine residues in the amino acid sequence, Cys 54 and Cys 97, are far apart in the folded structure. The stability of both the wild-type and mutant proteins is expressed as the melting temperature, Tm, which is the temperature at which 50% of the enzyme is inactivated during reversible heat denaturation. For the wild-type T4 lysozyme the Tm is 41.9 °C.

We will discuss three different approaches to engineer a more thermostable protein than wild-type T4 lysozyme, namely (1) reducing the difference in entropy between folded and unfolded protein, which in practice means reducing the number of conformations in the unfolded state, (2) stabilizing the α helices, and (3) increasing the number of hydrophobic interactions in the interior core.

Disulfide bridges increase protein stability

The greater the number of unfolded conformations of a protein, the higher the entropic cost of folding that protein into its single native state (see Chapter 6). Reducing the number of unfolded conformations therefore increases
the stability of the native state. The most obvious way to decrease the number of unfolded conformations is to introduce a novel disulfide bond based on knowledge of the tertiary structure of the folded protein. The longer the loop between the cysteine residues, the more restricted is the unfolded polypeptide chain, giving more stabilization of the folded structure. To design such bridges is, however, not a simple task, since the geometry of an unstrained \( \text{CH}_2\text{S-S-CH}_2 \) bridge in proteins is confined to rather narrow conformational limits, and deviations from this geometry will introduce strains into the folded structure and hence reduce rather than increase its stability. It is, therefore, not sufficient to choose at random two residues close together in space to make such a bridge, rather the protein engineer must carefully select pairs of residues with main-chain conformations that fulfill the conditions needed for an unstrained disulfide bridge.

Mathews made a very careful comparison between the geometry of the 295 disulfide bridges in known X-ray structures and all possible pairs of amino acid residues close enough to each other in the refined T4 lysozyme structure to accommodate a disulfide bridge. This was followed by energy minimization of the most likely candidate disulfide bridges and an analysis of stabilizing interactions present in the wild-type structure that would be lost by mutation to a Cys residue. Such losses should be minimized. Three candidate disulfide bridges remained after this filtering, one of which, Cys 3-Cys 97, contained one of the cysteine residues (Cys 97) that is present in the wild type. The five amino acid residues—Ile 3, Ile 9, Thr 21, Thr 142, and Leu 164 (see Figure 17.3)—were mutated to Cys residues in separate experiments so that all single (3-97, 9-164, and 21-142) as well as combinations of double and triple disulfide bonds could be formed. In addition, the second Cys residue of the wild-type enzyme, Cys 54, was mutated to Thr to avoid the formation of incorrect disulfide bonds during folding.

The results of this careful design of novel disulfide bridges were very encouraging (Figure 17.4). All the mutants were more stable in their oxidized forms than wild-type protein. The longer the loop between the cysteine
residues of the mutants with single disulfide bonds, the larger was the effect on stability. Furthermore, the effects were additive so that the increase in Tm of 23 °C for the mutant with three disulfide bonds was approximately equal to that of the sum of the increases in Tm values for the three mutants with single disulfide bonds (4.8 °C + 6.4 °C + 11.0 °C = 22 °C). The effect on the stability of the protein from reducing the number of possible unfolded structures through introduction of disulfide bridges, the entropic effect, is even larger than these values show because the reduced forms of the mutants had a lower Tm than wild type, which indicates that favorable contacts in the folded structure had been lost by the mutations. These experiments show that engineered disulfide bridges can be combined together to enhance stability dramatically. Needless to say, knowledge of the three-dimensional structure of the protein is a prerequisite to engineer increased stability in this way.

**Glycine and proline have opposite effects on stability**

Glycine residues have more conformational freedom than any other amino acid, as discussed in Chapter 1. A glycine residue at a specific position in a protein has usually only one conformation in a folded structure but can have many different conformations in different unfolded structures of the same protein and thereby contribute to the diversity of unfolded conformations. Proline residues, on the other hand, have less conformational freedom in unfolded structures than any other residue since the proline side chain is fixed by an extra covalent bond to the main chain. Another way to decrease the number of possible unfolded structures of a protein, and hence stabilize the native structure, is, therefore, to mutate glycine residues to any other residue and to increase the number of proline residues. Such mutations can only be made at positions that neither change the conformation of the main chain in the folded structure nor introduce unfavorable, or cause the loss of favorable, contacts with neighboring side chains.
Both types of mutations have been made in T4 lysozyme. The chosen mutations were Gly 77–Ala, which caused an increase in $T_m$ of 1 °C, and Ala 82–Pro, which increased $T_m$ by 2 °C. The three-dimensional structures of these mutant enzymes were also determined: the Ala 82–Pro mutant had a structure essentially identical to the wild type except for the side chain of residue 82; this strongly indicates that the effect on $T_m$ of Ala 82–Pro is indeed due to entropy changes. Such effects are expected to be additive, so even though each mutation makes only a small contribution to increased stability, the combined effect of a number of such mutations should significantly increase a protein’s stability.

**Stabilizing the dipoles of α helices increases stability**

In Chapter 2 we described the α helix as a dipole with a positive charge at its N-terminus and a negative charge at the C-terminus. Negative ions, such as phosphate groups in coenzymes or substrates, are usually bound to the positive ends of such helical dipoles. The α helices that are not part of a binding site frequently have a negatively charged side chain at the N-terminus or a positively charged residue at the C-terminus that interacts with the dipole of the helix. Such dipole-compensating residues stabilize the helical forms of small synthetic peptides in solution. Do these helix-stabilizing residues also contribute to the overall stability of globular proteins? Of the 11 α helices of T4 lysozyme, 7 helices have negatively charged residues close to their N-terminals; two of the remaining four α helices were therefore chosen for engineering studies to answer this question (Figure 17.5).

Two different mutant proteins with single substitutions at the N-terminus of each of these helices, Ser 38–Asp and Asn 144–Asp, were made as well as the corresponding double mutant. The single mutants both showed an increase in $T_m$ of about 2 °C; the effects are additive since the double mutant had a $T_m$ about 4 °C higher than wild type. This corresponds to 1.6 kcal/mol of stabilization energy. From the x-ray structures of these mutants it is apparent that the stabilization is due to electrostatic interactions and not to specific hydrogen bonding between the substituted amino acid and the end of the helix. Alan Fersht in Cambridge, UK has shown, using a different system, the small bacterial ribonuclease, barnase, that a histidine residue at

![Figure 17.5 Diagram of the T4 lysozyme structure showing the locations of two mutations that stabilize the protein structure by providing electrostatic interactions with the dipoles of α helices. (Adapted from H. Nicholson et al., Nature 336: 651–656, 1988.)](image-url)
the C-terminus of a helix stabilizes the barnase structure by about 2.1 kcal/mol. Significant stabilization of α-helical structures might, therefore, be obtained by combining several such helix-stabilizing mutations.

**Mutants that fill cavities in hydrophobic cores do not stabilize T4 lysozyme**

We emphasized in Chapter 2 that burying the hydrophobic side chains in the interior of the molecule, thereby shielding them from contact with solvent, is a major determinant in the folding of proteins. The surface that is buried inside a folded protein contributes directly to the stabilization energy of the molecule. Studies of destabilizing mutants in barnase, where cavities have been engineered into the hydrophobic core of the wild-type enzyme by mutations such as Ile to Val or Phe to Leu show that the introduction of a cavity the size of one -CH₂- group destabilizes the enzyme by about 1 kcal/mol. By analogy it should be possible to stabilize a wild-type protein by making mutations that fill existing cavities in its hydrophobic core. Even though proteins have the atoms of their hydrophobic cores packed approximately as tight as atoms are packed in crystals of simple organic molecules, there are cavities in the cores of almost every protein.

T4 lysozyme has two such cavities in the hydrophobic core of its α-helical domain. From a careful analysis of the side chains that form the walls of the cavities and from building models of different possible mutations, it was found that the best mutations to make would be Leu 133–Phe for one cavity and Ala 129–Val for the other. These specific mutants were chosen because the new side chains were hydrophobic and large enough to fill the cavities without making too close contacts with surrounding atoms.

The two single mutants were constructed, purified, analyzed for stability, and crystallized. They were both less stable than wild type by 0.5 to 1.0 kcal/mol. The X-ray structures of the mutants provide a rational explanation for this disappointing result. It turns out that in order to fill the cavities, the new side chains in the mutants adopt energetically unfavorable conformations. This introduces strain in the structure, which obviously costs more energy than is gained by the new hydrophobic interactions. Even careful model building is obviously not sufficient to predict detailed structural and energetic effects of mutations in the hydrophobic core of proteins. Apparently, the observed core structure in T4 lysozyme, and probably in most proteins, reflects a compromise between the hydrophobic effect, which will tend to maximize the core-packing density, and the strain energy that would be incurred in eliminating all packing defects. Therefore, mutations designed to fill existing cavities may be effective in some cases, but they are not likely to provide a general route to substantial improvement in protein stability.

**Proteins can be engineered by combinatorial methods**

The ultimate goal of protein engineering is to design proteins to carry out predicted functions. However, we do not yet completely understand the rules governing protein folding and molecular recognition, making design of proteins difficult. Protein engineers have therefore invented **combinatorial methods**, in which libraries of related proteins are analyzed simultaneously. By sorting these libraries to select for a particular function, the small number of active proteins can be separated from millions of inactive variants. Combinatorial libraries have been used to increase the activity of enzymes, to improve the binding affinity and specificity of proteins, and even to identify novel peptide ligands. Additionally, researchers hope to use the structural and functional data obtained through library selection to improve their ability to engineer molecular interactions.

Combinatorial methods are often referred to as in vitro or directed evolution techniques. In nature, the random DNA mutations that lead to changes in protein sequences occur rarely and so evolution is usually a slow
Absorbance of GFP around 398 nm leads to a fluorescence emission peak at 509 nm with a quantum yield between 0.72–0.85, and hence the green colour observed in vivo and in vitro. The intensity ratio between the absorbance peaks is sensitive to factors such as pH, temperature, and ionic strength, suggesting the presence of two different forms of the chromophore. It has been shown that the two forms differ in their protonation states with the 398 nm band reflecting a protonated fluorophore whilst the 475 nm band reflects deprotonation of this group (Figure 10.46).

The unique fluorescence of GFP allows the expressed protein to be fused to other domains creating a chimeric protein with a fluorescent tag that can be measured and has been used in N- and C-terminal fusions to follow gene expression, protein–protein interactions, cell sorting pathways, and intracellular signalling.

**Circular dichroism**

Circularly polarized light travels through optically active media with different velocities due to different indices of refraction for right (dextro) and left (laevo) components (Figure 10.47). This is called optical rotation and the variation of optical rotation with wavelength is known as optical rotary dispersion (ORD). ORD is normally measured at a specific wavelength and temperature and is usually denoted by a parameter $\alpha$, the angle of rotation of polarized light.

The right and left circularly polarized components are also absorbed differentially at some wavelengths due to differences in extinction coefficients for the two polarized components (Figure 10.48). The addition of the left and right components yields $\varepsilon$ at every point from the relationship $\varepsilon = \varepsilon_L + \varepsilon_R$. When this light is
result in an ellipticity of 0.57°. In CD spectroscopy, the measured parameter is the rotation of polarized light expressed with the units of millidegrees. Modern CD instrumentation is capable of precision down to thousandths of a degree (1 millidegree). There are a number of ways of expressing the CD signal of a sample although frequently many incorrect forms are used in the literature. The observed CD signal (S) is expressed in millidegrees and is normally converted to $\Delta \varepsilon_m$ or $\Delta \varepsilon_{m\text{r}}$, where $\Delta \varepsilon_m$ is the molar CD extinction coefficient and $\Delta \varepsilon_{m\text{r}}$ is the mean residue CD extinction coefficient, respectively. In any CD experiment it is vital to know the protein concentration extremely accurately to avoid significant error. It can be shown that

$$\Delta \varepsilon_m = S/(32980CI) \tag{10.42}$$

where $C$ is the concentration in mol dm$^{-3}$ (or m) and $l$ is the path length in cm. The light path is usually a value between 0.1 and 1.0 cm. The units of $\Delta \varepsilon_m$ are therefore M$^{-1}$cm$^{-1}$ and the analogy with the molar extinction coefficient determined via absorbance measurement is clear. Alternatively, expressing the ellipticity as a mean residue CD extinction coefficient leads to

$$\Delta \varepsilon_{m\text{r}} = S_{m\text{r}}/(32980CI) \tag{10.43}$$

where the mean residue weight ($m_{\text{r}}$) is the molecular weight divided by the number of residues. In this instance $C$ is expressed in mg ml$^{-1}$. CD intensities are sometimes reported as molar ellipticity ($[\theta]_M$) or mean residue ellipticity ($[\theta]_{m\text{r}}$). These terms are calculated from the following equalities

$$[\theta]_M = S/(10CI) \tag{10.44}$$

where $C$ is again the concentration in mol dm$^{-3}$ or

$$[\theta]_{m\text{r}} = S_{m\text{r}}/(10CI) \tag{10.45}$$

Both $[\theta]_M$ and $[\theta]_{m\text{r}}$ have the units degrees cm$^2$ dmol$^{-1}$. $[\theta]$ and $\Delta \varepsilon$ may be inter-converted using the relationship

$$[\theta] = 3298\Delta \varepsilon \tag{10.46}$$

CD extinction coefficients are stronger for chiral absorbances (positive or negative) since they arise from rotation from circularly polarized light. The relationship between absorbance and ellipticity is not direct in an approach taken in this chapter. The values for ellipticity at a specific wavelength relationship to the absorbance values are different.

In the absence of any additional information, one approach is to use a combination of absorbance and ellipticity measurements ("pure" data) to determine the CD signal of homopolypeptide structure and may be used to give a general approach. In practice, CD spectroscopy can be already known from the sequence content of the protein and by using a known method for an unknown signal. A procedure for the determination of varying methods is a variation of these methods.

A variation of this method is to examine the ellipticity in magnetic fields (<77 K). CD spectroscopy in magnetic fields is reflected in circular dichroism spectrometry of proteins, but for magnetic fields are a powerful tool for the investigation of metal ions.

Vibrational Analysis

The vibrational analysis of a protein complex and its protein analysis spectra were measured to determine the degree of protein structure, with informative frequencies of 4000 and
CD extinction coefficients are the most logical unit since they are direct analogs of extinction coefficient in absorbance measurements and lead to values of $\Delta \varepsilon_{\text{mew}}$ in an approximate range $\pm 20$ whilst the corresponding values for $\Delta \varepsilon$ range from $\pm 3000$. Using the above relationships it will be apparent that values of $[\theta]_{\text{mew}}$ values are in the range $\pm 70000$.

In the determination of secondary structure content one approach is to assume that spectra are linear combinations of each contributing secondary structure type ("pure" $\alpha$ helix, "pure" $\beta$ strand, Figure 10.49) weighted by its relative abundance in the polypeptide conformation. Unfortunately the problem with this approach is that there are no standard reference CD spectra for "pure" secondary structure. More significantly synthetic homopolypeptides are poor models of secondary structure and most homopolymers do not form helices nor is there a good example of a "model" $\beta$ strand. An empirical approach involves determining the experimental CD spectra of proteins for which the structures are already known. Using knowledge of the structure the content of helix, turn and strand is defined accurately and by using a database of reference proteins these methods prove accurate and reliable when applied to unknown samples. A number of different mathematical procedures have been adopted using reference proteins of varying size and secondary structure content and all of these methods give similar results.

A variation of CD occurs when samples are placed in magnetic fields maintained at low temperatures ($\leq 77$ K). Under these conditions all molecules exhibit CD spectra and the technique is called magnetic circular dichroism (MCD). In most cases the spectra of proteins are too complex to interpret fully but for metalloproteins the MCD technique provides a powerful tool with which identify ligands and metal ions.

**Vibrational spectroscopy**

The vibrational spectra of proteins are extremely complex and lie at lower frequencies than electronic spectra within the infrared (IR) region. It is useful to divide the IR region into three sections, the near, middle and far IR regions with the most informative zone located at wavenumbers between 4000 and 600 cm$^{-1}$ (Table 10.12). Historically, IR

*Figure 10.49* CD spectra of proteins with diverse secondary structure content. Top: extensively $\alpha$ helical proteins; centre: proteins rich in $\beta$ strands; bottom: protein with $(\alpha + \beta)$ structure (reproduced with permission from Circular Dichroism - the conformational analysis of biomolecules, Fasman, G.D (ed) Kluwer Academic 1996.)