

## Protein Structural Families

This lecture started out to be a version of a recent review about protein structural families and their evolution, by Christine Orengo and Janet Thornton in *Annual Reviews of Biochemistry* vol. 74 pp. 867-900. But it made me realize that I have never talked about protein structure, and I have incorporated material from Brandén and Tooze, *Introduction to Protein Structure*. Probably Dr. Kahn covers a good deal of this in General Biochemistry, especially in the computer tutorials. I assume that everyone at least knows what  $\alpha$ -helices and  $\beta$ -sheets are.

As the number of three-dimensional structures of proteins determined by X-ray crystallography, and now also by nuclear magnetic resonance, has increased exponentially, two facts became evident: 1) more than 90% of protein structures are made up of several **domains**, smaller structures which have a structural identity within the whole structure of the protein, and which can fold up on their own to stable structures; 2) structural similarities among domains can be recognized across a wide range of functions, even when similarity of the primary sequence is very weak. Liver alcohol dehydrogenase, as I mentioned last time, is an example of both of these statements: the structure can be divided into the catalytic and coenzyme-binding domains – even though parts of the catalytic domain come both before and after the coenzyme-binding domain in the primary sequence – and the structure of the coenzyme-binding domain can be recognized in a number of NAD<sup>+</sup>-binding proteins and even proteins binding just adenine nucleotides. Domains have been fitted together in many different ways to produce multi-domain protein structures. Some multi-function enzymes such as fatty acid synthetases may be made up of separable subunits in some organisms, but domains of one polypeptide in other organisms. I mentioned that some aminoacyl-tRNA synthetases have two identical domains in one polypeptide, whereas others have them as separate subunits.

As those of you who have worked with structures in the Structural Computing Lab, room 202 in Lipman Hall, know, a complete protein structure showing all atoms, or even just the main peptide bond chain, is hard to analyze. Structures are greatly simplified by the ribbon diagrams introduced by Jane Richardson, in which  $\alpha$ -helices are shown as spiraling ribbons, strands of  $\beta$ -sheet as flat arrows with the head pointing toward the C-terminus of the chain. Other parts of the chain are string-like. This representation makes it much easier to recognize the similarity of structures, as in Fig. 2 from Orengo and Thornton (2005). An even more schematized representation is a **topology diagram**, which shows whether the  $\beta$ -sheet strands are parallel (C-terminal ends of strands at the same end of the sheet) or anti-parallel (C-terminal ends at opposite ends of the sheet), and how the strands are connected, which strand's head (C-terminal end) is connected to which other strand's tail (N-terminal end) -  $\beta$ -strands which are next to each other in the sheet need not be next to each other in the primary sequence. Examples are shown in Fig. 2.11 from Brandén & Tooze. If  $\beta$ -sheet domains in two proteins have the same connections between strands, they may be considered to belong to the same structural **family**, although sometimes it is difficult to tell whether they are evolutionarily related or evolved independently, just as mammalian serine proteases and bacterial serine proteases such as subtilisin probably evolved independently, and one may speak just of **fold groups** or just **folds**, which are defined purely by structural similarity, whereas saying that they belong to the same family suggests a common evolutionary origin.

Simple combinations of  $\alpha$ -helices and  $\beta$ -strands that recur are known as **motifs**. For instance, two  $\alpha$ -helices connected by a short loop and roughly at right angles are referred to as a helix-loop-helix motif; versions of this bind to DNA and bind calcium. Two antiparallel strands of  $\beta$ -sheet directly connected head to tail are called a **hairpin motif**. The structure in which

three successive antiparallel  $\beta$ -strands are connected head to tail, but the fourth in the peptide sequence is H-bonded in the sheet to the first, is called a **Greek key motif** (B&T fig. 2.15). Domains are built up from motifs.

Evolution of protein structures proceeds by duplication of coding sequences and separate evolution of the duplicated sequences and the structures they code for. Evolution of a single sequence, usually with retention of function, results in structures described as **orthologs**, while evolution of duplicated sequences results in greater divergence of structure and function, with such related structures known as **paralogues**. Structures which are similar but do not seem to have a common evolutionary origin are known as **analogues**. Examples are given in Orengo & Thornton Fig. 2.

Structures are classified by the CATH classification, at different levels of similarity (Orengo & Thornton Fig. 3); CATH stands for class, architecture, topology = fold, homologous superfamily. Three major classes are defined by what secondary structure predominates: mainly  $\alpha$ -helical, mixed  $\alpha$ -helical and  $\beta$ -sheet, and mainly  $\beta$ -sheet. Each of these classes contains a variety of architectures, 32 total; an architecture describes a general way of using secondary structure, for instance the  $\beta$ -barrel, in which  $\beta$ -sheet is wrapped around to form a cylinder. Green fluorescent protein, beloved by Dr. Ward, is one of the most perfect the  $\beta$ -barrels. Each architecture contains a variety of fold groups, 820 total, each fold group contains one or more evolutionarily homologous superfamilies, and each superfamily may contain a number of more closely related sequence families.

At first evolutionary relationships among proteins were sought by looking only at similarities of the primary sequence, but as more and more distant sequences were compared, often with multiple insertions and deletions, it became difficult to be sure whether they are related, even using computer algorithms to make the best matches. It became necessary to look also at three-dimensional structures, and these too can be automatically compared by computers, which allows recognition of more distantly related sequences. However, while two structures can be compared, full comparison of structures of many proteins at once is slow. If the structures are reduced to the secondary structures,  $\alpha$ -helix,  $\beta$ -sheet and 'random' which means everything that isn't the first two, the comparisons are much easier. I am not the computer person to say anything about how this is done; Allen Smith could do it much better.

Orengo and Thornton state that the most comprehensive survey of domain sequences contains about 7600 domain families, that account for up to 80% of sequences in completed genomes. Of course the remaining parts of genome sequences are likely to contain additional families, of proteins that do not fit into known families. Looking at structures cuts down on the number of domain families – if what sequence suggests are two families turn out to have similar structures, they comprise only one domain structural family. This cuts the number of known domains about in half, though of course the number of structures known is far less than the number of sequences.

The CATH structural classification of some 50,000 domains recognizes 1459 superfamilies, with about 820 different folds. A very small percentage of these folds – less than 10, less than 0.1% of the total – account for nearly 40% of the sequence families in the Protein Data Base and several superfamilies. Similarly, although 32 'architectures' have been defined – the next level of structure up – about 5 of these contain 30% of the fold groups and half the superfamilies. It seems that these structures are particularly well packed, but also able to have secondary structure units slide past each other, so that they can accommodate more change in individual residues without destabilization, which is an advantage for evolutionary

differentiation. And if they are particularly stable structures, it is all the more likely that the same folds have been arrived at multiple times during evolution.

A large majority, >70%, of the domain sequences in genomes which have been sequenced can be assigned to fewer than 2500 sequence families, and a small percentage of the families, <10%, comprise tens and even hundreds of sequences. At the other end of diversity, there are a lot of sequences which are assigned to some 45,000 families with less than 5 known members, typically specific to the organism or subkingdom – they were only invented once.

About 200 structure families can be found in all the kingdoms of life, including especially proteins involved in protein synthesis, but also some in metabolism and regulation. Newer families are those found in only one of the three main kingdoms of organism (archaea, prokaryotes, eukaryotes) or only in one lineage within these. Eukaryotes have produced mainly new  $\beta$ -folds – structures with primarily  $\beta$ -sheet – while prokaryotes have produced more  $\alpha$ -folds, and  $\alpha\beta$ -folds have appeared least often, though this may be because the stable structures had already appeared.

Since proteins usually are made up of more than one domain, and will comprise a *protein* family only if the domain composition is the same, there are some 50,000 different protein families, plus about 150,000 sequences which have been observed only once, i.e. in only one sequenced genome. Some of these are small proteins which just don't seem to have much secondary structure; examples might be the apolipoproteins, which seem to exist as patches on the surface of lipid globules and have no stable structure without them.

Less than 15% of the protein families – the members of each family are made up of the same combination of the same domains – are found in all three kingdoms, while between 50 and 75% are unique to the particular organism. This contrasts with the frequency of the families of *domain* structures which make them up: 40 - 60% are found in all 3 kingdoms, and 50% of the domains that have been noted come from only 200 domain structure families. It is easier to construct a new protein, by shuffling domains, than to create a new domain structure. Some 5000 domain families have been seen in the sequenced genomes; these could give rise to  $10^7$  two-domain protein families, or  $10^{11}$  three-domain families, but only 50,000 have been observed, or 200,000 if we count the 150,000 unique sequences that haven't found a family. A very small proportion of the domains, less than 1%, have been duplicated extensively. This is not unreasonable: there are domains which carry out a particular function, and carry it out over and over again in different proteins. Why reinvent the wheel? Brandén and Tooze give an example: chymotrypsin, and trypsin which is quite homologous, have about 245 amino acids in two domains (substrate peptides fit into a cleft between the 2 domains). This pair of domains is repeated over and over in other serine proteases, such as urokinase, Factor IX and plasmin, which however have additional domains attached, such as the epidermal growth factor and Kringle domains. The basic chymotrypsin pair of domains preserves the basic chemistry of the reaction catalyzed, while the added domains are responsible for the specificity of these proteases for particular substrates, or for other parts of their function such as binding to phospholipid membrane surfaces. Function may depend on two domains, as in serine proteases just mentioned or in binding ATP, but the two domains may or may not be consecutive in the sequence of the protein, either one may come first, and they may be associated with a number of other domains. However, very often when two domains are associated in a function, they are always in the same order, suggesting that the combination evolved only once.

Evolution then is believed to start with duplication of domains, followed by recombination of DNA to attach domains to new partners. Simultaneously, once the duplicated domains are not needed to carry out the same function, they can evolve independently by

mutational change of amino acids, and more grossly by introduction of additional amino acids or by deletion. Generally, however, the basic chemistry of the function is maintained; it is the specificity which is changed, either by amino acid changes or reassortment of domains. It is advantageous if domains catalyzing successive steps in a pathway can be attached together in a single protein, so that the intermediates in the pathway can be passed from one active site to another without wandering loose in the cytoplasm. Of course, the protein can be made up of separately synthesized subunits, but if the domains are fused into a single polypeptide they will be expressed completely coordinately. However, there is no tendency for successive reactions to be catalyzed by similar domains; different domain structures are needed for different chemical reactions to be catalyzed.

As genomes become more complex, the number of regulatory proteins increases faster than the number of total proteins, as the number of interactions increases more than linearly, and a rapidly increasing fraction of the functions are expressed only some of the time. The regulatory proteins come from a limited number of families, since they must all bind to DNA, and preserve a DNA-binding domain, while another domain binds to something else, small molecule or another protein. Transcription factors are less conserved than their target genes, suggesting that regulation evolves faster than structural genes.

In bacteria, nearly 100 of whose genomes have been sequenced, about 200 domain families are common to all species. About 27 common families have undergone no expansion – typically they are involved with protein synthesis, arrived at their function early in evolution and cannot improve on it. But another 30 families have expanded “linearly and dramatically” with expanding genome size. These are mostly associated with metabolism; I think of them as representing bacteria’ ability to metabolize a very wide range of substrates. I see a sort of “to those that have shall be given” principle: once a structural family is larger than most, the chances that new needs will be met by evolution from a structure of that family increase, and its fraction of total genes in the genome increases further. However, as the number of metabolic domains increases linearly with genome size, the number of regulatory domains increases more than linearly. Bacteria want to be able to multiply rapidly, and their rate of multiplication can be limited by how long it takes to duplicate their genome, so the proliferation of regulatory domains eventually limits the size of the genome, it doesn’t have room to take on yet more metabolic capabilities.