

## Post-translational Modification - The Peptide Chain

[Last time I talked about protein modification *during* protein synthesis, i.e. natural or artificially generated mutation - the mutation is carried out on the DNA, but we are interested in the effect on the protein - and incorporation of 'non-canonical' amino acids, with mention of selenocysteine. Today I want to talk about protein modification *after* synthesis, usually termed post-translational modification. I'm giving you the tables of contents of three whole volumes of *Methods in Enzymology* on the subject, as a guide to what *can* happen, because I can't talk about all of them. I would like to separate this into two major categories, modifications of the peptide chain and modifications of the side chains, with one sub-category that deserves separate recognition, interchange of disulfide bonds. Today I'll talk mainly about modifications of the peptide chain, because there are some interesting recent developments. In subsequent lectures I will talk about some of the many modifications of side chains.]

The point I want to make right now is there are a number of modifications of the N- and C-termini of proteins. Polypeptide chains are synthesized beginning with formylmethionine, but the formyl group always, and indeed the methionine usually, is removed in the mature protein. Quite often the mature protein has the amino terminus acetylated, usually if the amino acid is alanine or serine, sometimes glycine, methionine or aspartic acid; this is a pain if you are trying to determine the amino terminal sequence by Edman degradation. Amino-terminal glutamine easily cyclizes to form pyroglutamic acid, in which the  $\alpha$ -amino group replaces the separate nitrogen in amide linkage to the  $\gamma$  carboxyl. Sometimes this is an artifact of work-up of the protein, but usually it is real. Another modification of the amino terminus, as of lysine, histidine and arginine side chains, is methylation. In certain bacterial proteins, such as histidine decarboxylase of *Lactobacillus* and *Clostridium perfringens*, the amino terminal serine or threonine is deaminated to generate a pyruvoyl or  $\alpha$ -ketobutyryl terminus; this is used as an analog of pyridoxal 5'-phosphate in amino acid transamination and decarboxylation reactions. The enzyme is synthesized as an inactive precursor protein with a Ser-Ser sequence; the carboxyl of the first serine shifts to the  $\beta$ -hydroxyl and then is eliminated, taking the oxygen with it and leaving an  $\alpha$ -aminoacrylate residue; the double bond presumably shifts to the nitrogen, yielding an imino acid, and the =NH is easily replaced by oxygen from water, yielding the final keto acid.<sup>1</sup> The sequence before the split remains in the mature complex as a separate subunit (the complex is actually a hexamer with three each of these small and large subunits). The  $\alpha$ -amino group also reacts with sugars, as in glucuronylglycine in some fungal enzymes and non-enzymic glycation when blood sugar is high in diabetes.

There are a number of cases where the C-terminal carboxyl group of the polypeptide is converted to an amide. There aren't many other modifications of the C-terminus, though there is a removal of several amino acids associated with farnesylation or geranylgeranylation of a CySH side chain three residues in from the C-terminus, which I'll talk about in connection with that reaction.

There are also cases where an additional amino acid, most often arginine, is put on to a synthesized protein, usually at the amino terminus but in one case at the C-terminus (tyrosine added to tubulin), usually from the charged tRNA, but without ribosomes or genetic code; the result is a protein which has a seemingly normal sequence but includes an amino acid not found in the gene! This can also, as I mentioned in the introduction, be done artificially. A protein can be cleaved at a specific position by a specific protease, a single amino acid at the new C-terminus removed by a carboxypeptidase, and the protein incubated in an organic solvent with a protease stable under that condition and a high concentration of another amino acid. Under these conditions the proteolytic cleavage runs in reverse, because water is at a low concentration rather than 55 M, and the added amino acid is inserted and peptide bonds reformed.

Mainly I want to talk about things that happen to the peptide bond. Of course the simplest case is removal of the original N-terminal methionine; the next simplest case is processing of signal peptides.

This is a major topic which a whole lecture could be spent on, and you can find it in general biochemistry textbooks. The polypeptide as synthesized has an amino-terminal sequence rich in hydrophobic amino acids, which may bind to the signal recognition particle and be delivered to the membranes of the endoplasmic reticulum, where glycosylation takes place and the signal peptide is cleaved off. Or this preprotein goes to and into the cell membrane of the bacterial cell, or the outer membrane of the mitochondrion or chloroplast. In these cases the signal peptide is cleaved off in two stages, one on going through the outer membrane to the membrane space, one on going through the inner membrane; only when both signal sequences are removed does it fold to its final structure. Or, as with cytochrome  $c_1$ , the protein may go first to the matrix inside the inner membrane, then back to the intermembrane space. There is a specific signal peptidase, or several, which recognizes a specific sequence of the signal peptide to cleave it; by now computers are programmed to recognize this in a gene sequence. Proof of a signal peptide is obtained by comparing the amino acid sequence of the mature protein with that deduced from the gene; the signal sequence is the difference.

Another even vaster topic is the cleavage of a precursor form, called a proprotein, to yield the final active form. Even 30 years ago there was a very large book, *Proteases and Biological Control*, on this subject. The classic cases are the activation of chymotrypsinogen to chymotrypsin and the formation of peptide hormones such as insulin. Chymotrypsinogen is synthesized in the pancreas as the inactive zymogen chymotrypsin and secreted into the intestine in this form. The critical cleavage, by trypsin, is of the arg<sub>15</sub>-ile<sub>16</sub> peptide bond, freeing the  $\alpha$ -amino group of ile<sub>16</sub> to interact with the side chain carboxyl of asp-184; this allows small conformational changes which make the active site active in ways I shall discuss later. This form is called  $\pi$ -chymotrypsin; the amino terminal peptide is still attached to the rest of the protein by a disulfide bond of cys<sub>1</sub>. A further proteolytic cleavage removes the dipeptide ser<sub>14</sub>-arg<sub>15</sub>, yielding  $\delta$ -chymotrypsin, and others remove the dipeptide thr<sub>147</sub>-asn<sub>148</sub>, yielding the final product  $\alpha$ -chymotrypsin, the form usually studied.

Insulin is synthesized as a preproprotein 105 amino acids long. A 24-a.a. signal peptide is removed as it goes into the endoplasmic reticulum, yielding proinsulin. This folds and forms disulfide bonds between cysteines 7 and 67 and between 19 and 80. A trypsin-like protease cleaves this at arg<sub>31</sub> and arg<sub>60</sub>. The sequence 32-60 goes away, and arg<sub>31</sub> is trimmed off by a carboxypeptidase to yield mature insulin, with the A and B chains held together by the disulfide bonds. It will not fold up naturally if the disulfides are reduced, because folding information in the 32-60 sequence is no longer present. Peptide hormones and other bioactive peptides are generally formed similarly, by being cut out of an initially synthesized large protein; for instance, the precursor preproopiomelanocortin can yield, on different cleavages, corticotropin,  $\beta$ - and  $\gamma$ -lipotropin,  $\alpha$ -,  $\beta$ - and  $\gamma$ -MSH, enkephalin and endorphin, different peptides with different hormonal activities.

This is just cleavage of peptide bonds; can new peptide bonds be formed? Indeed they can; some cases are summarized by Cooper and Stevens<sup>2</sup>. In several cases, notably the RecA protein from *Mycobacterium tuberculosis*<sup>3</sup> and the catalytic subunit Tfp1p of the vacuolar H<sup>+</sup> ATPase of *Saccharomyces cerevisiae* (yeast)<sup>4</sup>, the protein is the same size as in other organisms, but the open reading frame in the gene is larger, and the homology is at the beginning and the end of the sequence. This could be due to RNA editing, but the only RNA observed is the size expected from the open reading frame. In the yeast case translation results in one copy each of the 69 kDa Tfp1p and of a 50 kDa spacer protein. Frame-shifting deletions in the spacer resulted in a truncated protein, the NH<sub>2</sub> terminal part of the mature protein + part of the spacer, while deletion of an entire codon still yielded the mature protein. Sequencing the appropriate peptide of the mature protein yielded a sequence running across the splice point, proving that there really is a peptide bond formed. In both cases splicing substitutes a cysteine from the amino terminus of the C-terminal domain for a cysteine of the spacer.

The well-known lectin (carbohydrate-binding protein) concanavalin A does even more<sup>5</sup>. It is synthesized as a 261-a.a. precursor which goes into the endoplasmic reticulum. Here it is clipped at

residues 119 and 130; the small peptide goes away, but the two-chain form is properly folded and binds carbohydrates. In the folded form the amino terminus apparently is very near residue 252; it displaces the amino group of residue 253, with formation of a new peptide bond. Subsequently the peptide 131-134 is trimmed off. Thus the mature protein has the original residue 135 as amino terminus and the original residue 119 as C-terminus. Concanavalin A made in *E. coli* with a bacterial signal peptide attached goes to the periplasmic space and does not rearrange, but a related protein which stays in the cytoplasm does rearrange in this way. All the changes occur at the carboxyl side of an asparagine; the residues following are not specified in the Cooper and Stevens paper. It is now known that the residue on the C-terminal side of the cleavage is always a cysteine, serine or threonine.

The most studied case is the DNA polymerases from the thermophilic archæa *Thermococcus littoralis* and *Pyrococcus* sp. GB-D, which Francine Perler at New England BioLabs was trying to clone. Homology of the gene to other DNA polymerases showed in *Thermococcus* two intervening sequences<sup>6</sup> (these are now termed inteins<sup>7</sup>, by analogy with introns; the pieces spliced together are called exteins), and expression of the gene in *E. coli* gave a mess. Removal of one of the inteins resulted in the expected 90 kDa polymerase and a 45 kDa protein representing the other intein. This has sequence similarity to 'homing' endonucleases which are coded by introns and install their DNA sequence into allelic genes that lack the introns, and in some cases endonuclease activity has been shown; it seems to help move the intein DNA sequence around in the genome of the organism.

Perler's group studied<sup>8</sup> the *Pyrococcus* intein by installing it in a chimeric gene for maltose-binding protein + paramyosin  $\Delta$ Sal; proteins containing the maltose-binding domain can be purified by adsorption on amylose. Letting *E. coli* synthesize protein overnight at 12° - 20° made lots of a three-part 132 kDa chimeric protein, MIP, as well as an apparently 180 kDa form they called MIP\*, and products of cleavage at either the N- or the C-terminus of the intein (M + IP, MI + P). Presumably MIP can accumulate because the temperature is so far below the growth temperature of the thermophile from which it came. Warming up the precursor MIP to 37° resulted in splicing, yielding mainly the spliced maltose binding protein-paramyosin (MP) + the intein (I), along with some IP. Splicing is fastest at pH 6, slower at higher pH. The slow migrating form, MIP\*, is a branched intermediate of the splicing reaction; it has two amino-terminal sequences, those of the maltose-binding protein and the intein. During a slow splicing reaction the amount of it increases, then decreases. At high pH (10) it goes back to the linear form MIP.

Splicing of MIP made in *E. coli* and purified by chromatography on amylose indicates that it doesn't require any other enzyme, and only the intein sequence, followed by serine, cysteine or in one case threonine, is necessary to get splicing. Presumably the intein folds to bring the splice points in close proximity, and presumably the extein structures have to allow this folding. The intein sequence begins with a cys or ser and ends, in the seven inteins known, with three hydrophobic amino acids, then his-asn. The released intein has been shown<sup>9</sup> by mass spectrometry to be released with a C-terminal succinimide formed from the asparagine. The branched intermediate is considered to involve an ester linkage to the OH or SH of the N-terminal ser or cys of the C-terminal extein, paramyosin (P) in this example. This paper found that incubating guanidine-denatured branched intermediate at pH 9 yielded M + IP, indicating that the more alkali-labile ester bond was from the amino-terminal extein to the C-terminal extein, rather than from the intein.

The now fairly accepted mechanism for splicing<sup>10,11</sup>, is shown in Fig. 2 of the handout. It involves (i) nucleophilic attack of the intein ser or cys on the C-terminal carboxyl of the amino terminal extein, an N-S or N-O shift as seen also in formation of the  $\alpha$ -keto group in histidine decarboxylase, and as observed with ordinary peptides under acid conditions, (ii) transesterification to the downstream ser or cys, forming the branched intermediate; (iii) attack of the NH<sub>2</sub> of the C-terminal asparagine of the intein on this carboxyl, yielding the succinimide C-terminus of the intein, (iv) O (or S) to N shift of the carboxyl back to the amino group of the downstream extein. This O-N shift must be rapid, since the MP product is alkali-stable. The histidine next to the asparagine seems to be involved in step iii but not i or

ii, since when it is replaced by other amino acids the branched intermediate is formed but splicing doesn't occur. The mechanism has also been investigated for the VMA intein of the vacuolar ATPase of *Saccharomyces cerevisiae*<sup>12</sup> *Sce* VMA intein for short.

Since Xu and Perler are at a company, they have developed a product, a protein purification system similar to those with a cleavable glutathione-S-transferase or maltose-binding protein but requiring no proteolytic enzyme for cleavage. Your protein is cloned into a vector with its C-terminus next to the amino-terminal cysteine of the *Sce* VMA intein, with a chitin-binding domain beyond the intein. The fusion protein is produced in *E. coli* and adsorbed onto a chitin column. The column is incubated overnight at 4° with dithiothreitol or  $\beta$ -mercaptoethanol, which plays the role of Cys<sup>455</sup>, the N-terminal of the second extein - it replaces the intein cysteine by transesterification. Your protein is thus released, with dithiothreitol in thioester linkage to the C-terminus. You could easily cleave this off with hydroxylamine, or replace it with [<sup>14</sup>C]-cysteine which would transesterify and then undergo S-N shift to form a peptide bond, thus stably radiolabeling the protein.

Some related reactions are mentioned by Shao & Kent. The 'sonic hedgehog' precursor protein - a name resulting from the imagination of *Drosophila* geneticists - which is important in patterning of embryonic structures, self-cleaves into two proteins at a cysteine residue. In this case the subsequent transesterification is not to another cysteine, but to the hydroxyl of a cholesterol, making one protein more hydrophobic<sup>13</sup>. So, even developmental molecular biologists need to know protein chemistry.

It remains to crystallize the splicing precursor protein to locate the various amino acids in space and determine whether only neighboring amino acids are needed to catalyze the reactions, or others more distant in space also act. If there are, they must be in the intein, since that is all that is needed; and the frequent appearance of self-cleavage reactions leaving an amino terminal Ser, Thr or CySH suggests that at least the N-O or N-S shift needs no unusual structure.

An even newer modification of an amino acid in a protein is inversion of an L-amino acid to a D-amino acid<sup>14,15</sup>. The funnel-web spider *Agelenopsis aperta* produces a venom containing peptide toxins which paralyze its prey by blocking voltage-sensitive Ca<sup>++</sup> channels. The toxins are synthesized as larger precursors containing also signal sequences for extracellular transport and acidic sequences cleaved off to produce the mature toxins. Two of these toxins, IVB and IBC, have identical sequences 48 a.a. long and the same disulfide bonds, but are separable on hplc. IVB is considerably more toxic. Protease cleavage at Glu<sup>42</sup> or CNBr cleavage at Met<sup>43</sup> yielded C-terminal hexa- or pentapeptides from the two toxins which were separable by hplc, while the peptides from the rest of the toxin were indistinguishable. Peptides with the C-terminal sequence, Gly-Leu-Ser-Phe-Ala, were synthesized with either L- or D-ser at position 46 (probably they tried other D-amino acids too, but the structure paper wasn't published yet) and coeluted with the peptides from the toxins, the L-ser peptide with that from IVC, the D-ser peptide with that from the more active IVB. They then synthesized the complete toxins with L- or D-ser at position 46, and showed that what folded correctly was identical to the natural toxin.

Crude venom converted IVC to IVB; this was better demonstrated when the venom was fractionated by gel filtration, separating the 30 kDa isomerase from an 86 kDa metalloprotease which otherwise degraded IVC. This points up one reason for such inversion: D-amino acid peptide bonds are not cleaved by the proteases that cleave natural L-amino acid peptide bonds, which the insect prey might use to detoxify IVC. But more importantly, the inversion allows the toxin to be a better fit to the Ca<sup>++</sup> channels it blocks, making it a more potent toxin; the repertory of protein conformation has been added to. I could give you more examples of similar but much smaller peptides with D-amino acids, but they are in the Kreil paper I am giving you.

In connection with what I just said about D-amino acid peptide bonds being stable to proteases, I mention that pharmaceutical companies are much interested in biologically active peptides with either D-amino acids or methylated nitrogens in the peptide bond, since these will be resistant to proteolytic cleavage, so that the pill can be taken by mouth, the peptide won't be chewed up in the stomach.

## References on polypeptide splicing and amino acid inversion

- <sup>1</sup>Recsei, P.A., Huynh, Q.K., and Snell, E.E., *Proc. Nat. Acad. Sci.* **80**:973-977 (1983)
- <sup>2</sup>Cooper, A.A., and Stevens, T.H., *BioEssays* **15**:667-674 (1993) Review.
- <sup>3</sup>Davis, E.O., Jenner, P.J., Brooks, P.C. Colston, M.J., and Sedgwick, S.G., *Cell* **71**:201-210 (1992) Splicing of *M. tuberculosis* RecA protein.
- <sup>4</sup>Cooper, A.A., Chen, Y., Lindorfer, M.A., and Stevens, T.H., *EMBO J.* **12**:2575-2583 (1993) Splicing of *S. cerevisiae* TFP1.
- <sup>5</sup>Bowles, D.J., and Pappin, D.J., *Trends Biochem. Sci.* **13**:60-64 (1988) Traffic and assembly of concanavalin A.
- <sup>6</sup>Perler, F.B., Comb, D.G., Jack, W.E., Moran, L.S., Qiang, B., Kucera, R.B., Benner, J., Slatko, B.E., Nwankwo, D.O., Hempstead, S.K., Carlow, C.K.S., and Jannasch, H., *Proc. Nat. Acad. Sci. USA* **89**:5577-5581 (1992) Intervening sequences in an Archaea DNA polymerase gene.
- <sup>7</sup>Perler, F.B., Davis, E.O., Dean, G.E., Gimble, F.S., Jack, W.E., Neff, N., Noren, C.J., Thorner, J., and Belfort, M., *Nucleic Acids Res.* **22**:1125-7 (1994) Nomenclature of protein splicing.
- <sup>8</sup>Xu, Ming-Qun, Southworth, M.W., Mersha, F.B., Hornstra, L.J., and Perler, F.B., *Cell* **75**:1371-1377 (1993) In vitro splicing of purified precursor and identification of branched intermediate.
- <sup>9</sup>Xu, Ming-Qun, Comb, D.G., Paulus, H., Noren, C.J., Shao, Y., and Perler, F.B., *EMBO J.* **13**:5517-5522 (1994) Analysis of the branched intermediate and its resolution by succinimide formation.
- <sup>10</sup>Xu, Ming-Qun, and Perler, F.B., *EMBO J.* **15**:5146-5153 (1996) The mechanism of protein splicing, as explicated by mutation of critical residues at the splice sites.
- <sup>11</sup>Shao, Y., and Kent., S.B.H., *Chemistry & Biology* **4**:187-194 (1997) Review.
- <sup>12</sup>Chong, S, Shao, Y., Paulus, H., Benner, J., Perler, F.B., and Xu, M.-Q., *J. Biol. Chem.* **271**:22159-22168 (1996) Similar study of splicing at the VMA intein of the vacuolar ATPase of *Saccharomyces cerevisiae*; development of a mesophilic *in vitro* splicing system and protein cleavage reaction.
- <sup>13</sup>Porter, J.A., Young, K.E. & Beachy, P.A., *Science* **274**:255-9 (1996).
- <sup>14</sup>Kreil, G., *Science* **266**:996-7 (1994) Commentary on the next paper.
- <sup>15</sup>Heck, S.D., Siok, C.J., Krapcho, K.J., Kelbaugh, P.R., Thadeio, P.F., Welch, M.J., Williams, R.D., Ganong, A.H., Kelly, M.E., Lanzetti, A.J., Gray, W.R., Phillips, D., Parks, T.N., Jackson, H., Ahlijanian, M.K., Saccomano, N.A., and Volkmann, R.A., *Science* **266**:1065-1068 (1994)