1

Introduction

1.1 Sources and roles of amino acids and peptides

More than 700 amino acids have been discovered in Nature and most of them are α-amino acids. Bacteria, fungi and algae and other plants provide nearly all these, which exist either in the free form or bound up into larger molecules (as constituents of peptides and proteins and other types of amide, and of alkylated and esterified structures).

The twenty amino acids (actually, nineteen α-amino acids and one α-imino acid) that are utilised in living cells for protein synthesis under the control of genes are in a special category since they are fundamental to all life forms as building blocks for peptides and proteins. However, the reasons why all the other natural amino acids are located where they are, are rarely known, although this is an area of much speculation. For example, some unusual amino acids are present in many seeds and are not needed by the mature plant. They deter predators through their toxic or otherwise unpleasant characteristics and in this way are thought to provide a defence strategy to improve the chances of survival for the seed and therefore help to ensure the survival of the plant species.

Peptides and proteins play a wide variety of roles in living organisms and display a range of properties (from the potent hormonal activity of some small peptides to the structural support and protection for the organism shown by insoluble proteins). Some of these roles are illustrated in this book.

1.2 Definitions

The term '*amino acids*' is generally understood to refer to the *aminoalkanoic acids*, $H_3N^+—(CR^1R^2)_n—CO_2^-$ with $n = 1$ for the series of α-amino acids, $n = 2$ for β-amino acids, etc. The term '*dehydro-amino acids*' specifically describes *2,3-unsaturated (or 'αβ-unsaturated')-2-aminoalkanoic acids*, $H_3N^+—(C{=}CR^1R^2)—CO_2^-$.

However, the term '*amino acids*' would include all structures carrying amine and acid functional groups, including simple aromatic compounds, e.g. anthranilic acid,

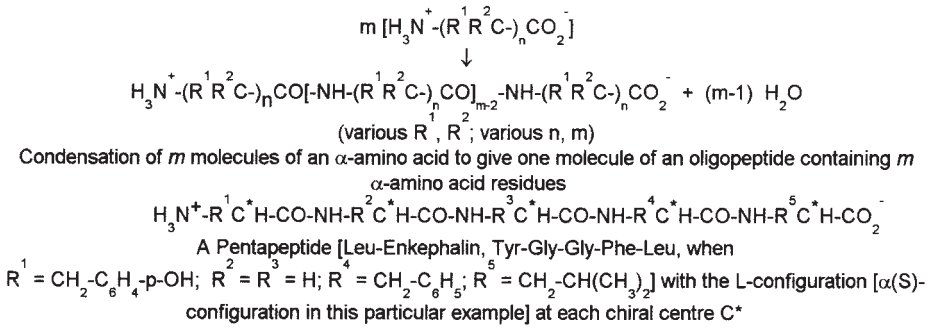


Figure 1.1. Peptides as condensation polymers of α-amino acids.

o-H_3N^+—C_6H_4—CO_2^-, and would also cover other types of acidic functional groups (such as phosphorus and sulphur oxy-acids, H_3N^+—$(R^1R^2C—)_nHPO_3^-$ and R_3N^+—$(R^1R^2C—)_nSO_3^-$, etc). The family of boron analogues $R_3N^{\bullet}BHR^1$—CO_2R^2 ($^{\bullet}$ denotes a dative bond) has recently been opened up through the synthesis of some examples (Sutton *et al.*, 1993); it would take only the substitution of the carboxy group in these 'organoboron amino acids' ($R = R^1 = R^2 = H$) by phosphorus or sulphur equivalents to obtain an amino acid that contains no carbon! However, unlike the amino acids containing sulphonic and phosphonic acid groupings, naturally occurring examples of organoboron-based amino acids are not known.

The term '*peptides*' has a more restricted meaning and is therefore a less ambiguous term, since it covers polymers formed by the condensation of the respective amino and carboxy groups of α, β, γ . . . -amino acids. For the structure with $m = 2$ in Figure 1.1 (i.e., for a dipeptide) up to values of $m \simeq 20$ (an eicosapeptide), the term '*oligopeptide*' is used and a prefix *di-*, *tri-*, *tetra-*, *penta-* (see Leu-enkephalin, a linear pentapeptide, in Figure 1.1), . . . *undeca-* (see cyclosporin A, a cyclic undecapeptide, in Figure 1.4 later), *dodeca-*, . . . etc. is used to indicate the number of *amino-acid residues* contained in the compound. *Homodetic* and *heterodetic* peptides are illustrated in Chapter 7.

Isopeptides are isomers in which amide bonds are present that involve the *side-chain amino group* of an αω-di-amino acid (e.g. lysine) or of a poly-amino acid and/or *the side-chain carboxy-group of an* α*-amino-di- or -poly-acid* (e.g. aspartic acid or glutamic acid). Glutathione (Chapter 8) is a simple example. Longer polymers are termed '*polypeptides*' or '*proteins*' and the term '*polypeptides*' is becoming the most commonly used general family name (though proteins remains the preferred term for particular examples of large polypeptides located in precise biological contexts). Nonetheless, the relationship between these terms is a little more contentious, since the change-over from polypeptide to protein needs definition. The figure 'roughly fifty amino acid residues' is widely accepted for this. Insulin (a polymer of fifty-one α-amino acids but consisting of two crosslinked oligopeptide

chains; see Figure 1.4 later) is on the borderline and has been referred to both as a *small protein* and as a *large polypeptide*.

Poly(α-amino acid)s is a better term for peptides formed by the self-condensation of one amino acid; natural examples exist, such as poly(D-glutamic acid), the protein coat of the anthrax spore (Hanby and Rydon, 1946). In early research in the textile industry, poly(α-amino acid)s showed promise as synthetic fibres, but the synthesis methodology required for the polymerisation of amino acids was complex and uneconomic.

Polymers of controlled structures made from *N*-alkyl-α-amino acids (Figure 1.1; —NR^n instead of —NH—, $R^1 = R^2 = H$; $n = 1$), i.e. $H_2{}^+NR^n—CH_2CO—[NR^n—CH_2—CO—]_mNR^n—CH_2—CO_2^-$, which are poly(*N*-alkylglycine)s of defined sequence (various R^n at chosen points along the chain), have been synthesised as *peptide mimetics* (see Chapter 9) and have been given the name *peptoids*. These can be viewed as peptides with side-chains shifted from carbon to nitrogen; they will therefore have a very different conformational flexibility (see Chapter 2) from that of peptides and will also be incapable of hydrogen bonding. This is a simple enough way of providing all the correct side-chains on a flexible chain of atoms, in order to mimic a biologically active peptide, but the mimic can avoid enzymic breakdown before it reaches the site in the body where it is needed.

Using the language of polymer chemistry, polypeptides made from two or more different α-amino acids are *copolymers* or irregular poly(amide)s, whereas poly(α-amino acid)s, $H—[NH—CR^1R^2—CO—]_mOH$, are *homopolymers* that could be described as members of the nylon[2] family.

Depsipeptides are near-relatives of peptides, with one or more *amide bonds replaced by ester bonds*; in other words, they are formed by condensing α-amino acids with α-hydroxy-acids in various proportions. There are several important natural examples of these, of defined sequence; for example the antibiotic valinomycin and the family of enniatin antibiotics. Structures of other examples of depsipeptides are given in Section 4.8.

Nomenclature for conformational features of peptide structure is covered in Chapter 2.

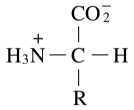
1.3 'Protein amino acids', alias 'the coded amino acids'

The twenty L-amino acids (actually, nineteen α-amino acids and one α-imino acid (Table 1.1)) which, in preparation for their role in protein synthesis, are joined *in vivo* through their carboxy group to tRNA to form α-aminoacyl-tRNAs, are organised by ribosomal action into specific sequences in accordance with the genetic code (Chapter 8).

'Coded amino acids' is a better name for these twenty amino acids, rather than 'protein amino acids' or 'primary protein amino acids' (the term 'coded amino acids' is increasingly used), because changes can occur to amino-acid residues after they have been laid in place in a polypeptide by ribosomal synthesis. Greenstein and

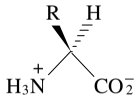
Table 1.1. *The twenty 'coded' amino acids (nineteen 'coded' L-α-amino acids, and one 'coded' L-α-imino acid): structures and definitions*[a]

Structure conventions for the L-α-amino acids are



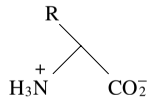
Fischer projection of an L-α-amino acid, requiring the carbon chain to be arranged vertically, with the carboxy group at the top

is equivalent to

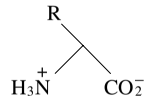


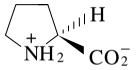
One of the commonly-used three-dimensional representations of an L-α-amino acid

which is equivalent to



Barrett representation of an L-α-amino acid

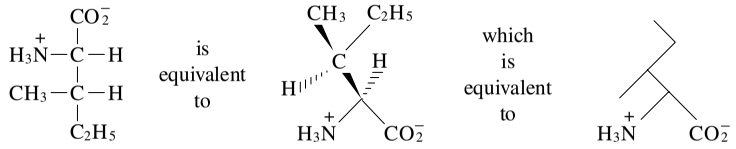
	Name of amino acid	Three-letter abbreviation	Single-letter abbreviation	Structures  *Amino acid side-chain, R =	Hydrophobicity High	Hydrophilicity High
One with no side-chain* (i.e. with a hydrogen atom)	Glycine	Gly	G	H		*
Four with saturated aliphatic side-chains* (hydrophobic side-chains)	Alanine	Ala	A	CH_3	*	
	Leucine	Leu	L	$CH_2CH(CH_3)_2$	*	
	Valine	Val	V	$CH(CH_3)_2$	*	
	Isoleucine	Ile	I	(S)-$CH(CH_3)C_2H_5$	*	

Ten with functionalised aliphatic side-chains* (mostly hydrophilic side-chains)	Arginine	Arg	R	$CH_2CH_2CH_2NHC(=NH)NH_2$																	*
	Aspartic acid	Asp	D	CH_2CO_2H																*	
	Asparagine	Asn	N	CH_2CONH_2																*	
	Glutamic acid	Glu	E	$CH_2CH_2CO_2H$																*	
	Glutamine	Gln	Q	$CH_2CH_2CONH_2$																*	
	Lysine	Lys	K	$CH_2CH_2CH_2CH_2NH_2$																	*
	Methionine	Met	M	$CH_2CH_2SCH_3$										*							
	Cysteine	Cys	C	CH_2SH										*							
	Serine	Ser	S	CH_2OH									*								
	Threonine	Thr	T	(R)-$CH(CH_3)OH$									*								
Four with aromatic or heteroaromatic side-chains* (most of these side-chains are hydrophobic)	Phenylalanine	Phe	F	$CH_2C_6H_5$								*									
	Tyrosine	Tyr	Y	CH_2-(p-OH-C_6H_4)													*				
	Histidine	His	H	CH_2-(imidazol-4-yl)															*		
	Tryptophan	Trp	W	CH_2-(indol-3-yl)											*						
The 'coded' α-imino acid	Proline	Pro	P														*				

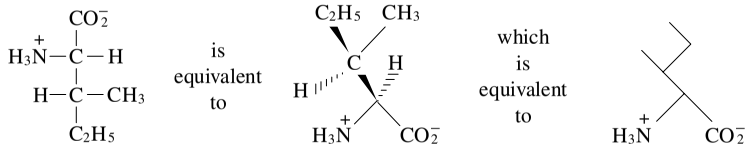
Notes:

1. The structure of each side-chain, R, is given for the 19 'coded α-amino acids', after each name. The full structure of the 'coded α-imino acid' proline is given. '*Three-letter*' and '*one-letter*' abbreviations are given for the 20. The *three-letter* abbreviation is the *first three letters of the name* for all twenty, *except* for asparagine (Asn), glutamine (Gln), isoleucine (Ile) and tryptophan (Trp). The *single-letter* abbreviated name is the first letter of their full name for *eleven* of them. Different letters are needed for the *other nine*, to avoid ambiguity: arginine (R), asparagine (N), aspartic acid (D), glutamic acid (E), glutamine (Q), lysine (K), phenylalanine (F), tryptophan (W) and tyrosine (Y).
2. All full names end in 'ine' except aspartic acid, glutamic acid and tryptophan. Adjectives are derived from the names by dropping the 'ine' or its equivalent ending and adding 'yl'; thus, alanyl, glutamyl, prolyl, tryptophyl, etc.
3. *Configurations.* The 'R/S' convention can easily be transferred to replace the Fischer 'D/L' system, while retaining the trivial names: L-enantiomers of all the coded amino acids are members of the S series except L-cysteine, which becomes R-cysteine through proper application of the R/S rules. Diastereoisomers (the isoleucine/allo-isoleucine and threonine/allothreonine pairs, 'allo' indicating inversion of the side-chain configuration of the coded amino acid) are less ambiguously named through the 'R/S' system, although the side-chain configuration can be indicated; for example, natural L-isoleucine is (2S,3S)-isoleucine:

Table 1.1. (*cont.*)



whereas L-alloisoleucine is (2S,3R)-isoleucine:



For the structures of natural L-*threonine* ((2S,3R)-threonine) and L-*allothreonine* ((2S,3S)-threonine), replace the side-chain ethyl group (C_2H_5) in isoleucine and alloisoleucine by OH.

4. *IUPAC–IUB nomenclature recommendations* (1983), reproduced in full in *Amino Acids, Peptides, and Proteins*, 1985, Vol. 16, The Royal Society of Chemistry, p. 387; and in *Eur.J.Biochem.*, 1984, **138**, 9, encourage the retention of trivial names for the common α-amino acids, but systematic names are relatively straightforward; thus, L-alanine is 2S-aminopropanoic acid and L-histidine is 2S-amino-3-(imidazol-4-yl)-propanoic acid (the name for the predominant tautomer).

5. 'Hydrophilic' and 'hydrophobic' are terms used to denote the relative water-attracting and water-repelling property, respectively, of the side-chain when the amino acid is condensed into a polypeptide (see Chapter 5). The term 'hydropathy index' may be used to place the amino acids in order of their 'hydrophilicity' (Kyte and Doolittle, 1985), and their relative positions are shown here on an arbitrary scale.

[a] Selenocysteine (i.e. cysteine with the sulphur atom replaced by a selenium atom) has been found in certain proteins, e.g. formate dehydrogenase, an enzyme from *Escherichia coli*, and it has very recently been shown to be placed there through normal ribosomal synthesis (Stadtman, 1996). Thus selenocysteine can now be accepted as the 'twenty-first coded amino acid'.

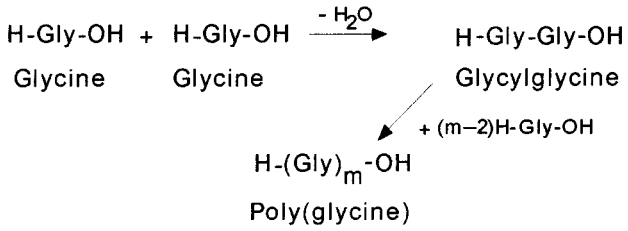


Figure 1.2. Polymerisation of glycine.

Winitz, in their 1961 book, listed 'the 26 protein amino acids', six of which were later found to be formed from among the other twenty 'protein amino acids' in the list of Greenstein and Winitz, after the protein had left the gene ('*post-translational* (sometimes called *post-ribosomal*) *modification*' or '*post-translational processing*'). Because of these changes made to the polypeptide after ribosomal synthesis, amino acids that are not capable of being incorporated into proteins by genes ('secondary protein amino acids', Table 1.2) can, nevertheless, be found in proteins.

1.4 Nomenclature for 'the protein amino acids', alias 'the coded amino acids'

The common amino acids are referred to through trivial names (for example, glycine would not be named either 2-aminoethanoic acid or amino-acetic acid in the amino acid and peptide literature). Table 1.1 summarises conventions and gives structures. The rarer natural amino acids are usually named as derivatives of the common amino acids, if they do not have their own trivial names related to their natural source (Table 1.2), but apart from these, there are occasional examples of the use of systematic names for natural amino acids.

1.5 Abbreviations for names of amino acids and the use of these abbreviations to give names to polypeptides

To keep names of amino acids and peptides to manageable proportions, there are agreed conventions for nomenclature (see the footnotes to Table 1.1). The simplest α-amino acid, glycine, would be depicted H—Gly—OH in the standard 'three-letter' system, the H— and —OH representing the 'H_2O' that is expelled when this amino acid undergoes condensation to form a peptide (Figure 1.2). The three-letter abbreviations therefore represent the 'amino-acid residues' that make up peptides and proteins.

So this '*three-letter system*' was introduced, more with the purpose of space-saving nomenclature for peptides than to simplify the names of the amino acids. A '*one-letter system*' (thus, glycine is G) is more widely used now for peptides (but is never used to refer to individual amino acids in other contexts) and is restricted to naming peptides synthesised from the coded amino acids (Figure 1.3).

Table 1.2. *Post-translational changes to proteins: the modified coded amino acids present in proteins, including crosslinking amino acids (secondary amino acids)*

Modifications to side-chain functional groups of coded amino acids

1. The aliphatic and aromatic coded amino acids may exist in αβ-dehydrogenated forms and the β-hydroxy-α-amino acids may undergo *post-translational dehydration*, so as to introduce αβ-dehydroamino acid residues, $-NH-(C=CR^1R^2)-CO-$, into polypeptides.

2. Side-chain OH, NH or NH_2 proton(s) may be substituted by *glycosyl*, *phosphate* or *sulphate*. These substituent groups are 'lost' during hydrolysis preceding analysis and during laboratory treatment of proteins by hydrolysis prior to chemical sequencing, which creates a problem that is usually solved through spectroscopic and other analytical techniques.

3. Side-chain NH_2 of lysine may be *methylated* or *acylated*: (N^{ε}-methylalanyl, N^{ε}-diaminopimelyl).

4. Side-chain NH_2 of glutamine may be *methylated*; giving N^5-methylglutamine, and the side-chain NH_2 of asparagine may be *glycosylated*.

5. Side-chain CH_2 may be *hydroxylated*, e.g. hydroxylysine, hydroxyprolines (trans-4-hydroxyproline in particular), or *carboxylated*, e.g. to give α-aminomalonic acid, β-carboxyaspartic acid, γ-carboxyglutamic acid, β-hydroxyaspartic acid, etc.

6. Side-chain aromatic or heteroaromatic moieties may be *hydroxylated*, *halogenated or* N-*methylated.*

7. The side-chain of arginine may be modified (e.g. to give ornithine (Orn), $R=CH_2CH_2CH_2NH_2$, or citrulline (Cit), $R=CH_2CH_2CH_2NHCONH_2$).

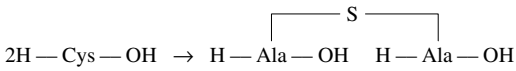
8. The side-chain of cysteine may be modified, as in 1 above, also selenocysteine (CH_2SeH instead of CH_2SH; see footnote a to Table 1.1), lanthionine (see 10 below).

9. The side-chain of methionine may be S-alkylated (see Table 1.3) or oxidised at S to give methionine sulphoxide.

10. Crosslinks in proteins may be formed by condensation between nearby side-chains.
    (a) From lysine: e.g. lysinoalanine as if from [lysine+serine$-H_2O$]

$$\rightarrow \quad \text{dehydroalanine} \quad \rightarrow \quad \begin{matrix} \text{H-Lys-OH} \\ | \\ \text{H-Ala-OH} \end{matrix}$$

    (b) From tyrosine: 3,3′-dityrosine, 3,3′,5′,3″-tertyrosine, etc.
    (c) From cysteine: oxidation of the thiol grouping ($HS-+-SH\rightarrow -S-S-$) to give the disulphide or to give cysteic acid (Cya): $-SH\rightarrow -SO_3H$ and alkylation leading to sulphide formation (e.g. alkylation as if by dehydroalanine to give lanthionine):



    (Further examples of crosslinking amino acids in peptides and proteins are given in Section 5.11.)

Nomenclature of post-translationally modified amino acids

Abbreviated names for close relatives of the 'coded amino acids' can be based on the 'three-letter' names when appropriate; thus, L-Pro after post-translational hydroxylation gives L-Hypro (trans-4-hydroxyproline, or (2S,4R)-hydroxyproline).

Current nomenclature recommendations (see footnote to Table 1.1) allow a number of abbreviations to be used for some non-coded amino acids possessing trivial names (some of which are used above and elsewhere in this book): Dopa, β-Ala, Glp, Sar, Cya, Hcy (homocysteine) and Hse (homoserine) are among the more common.

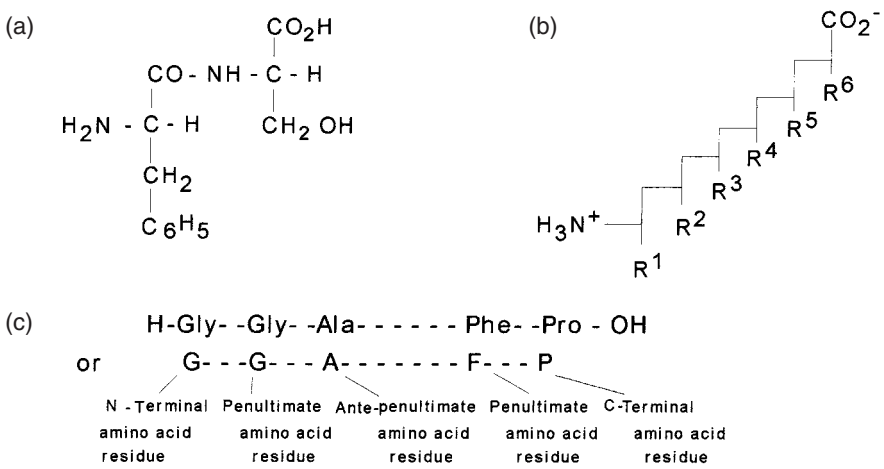


Figure 1.3. (a) The dipeptide L-phenylalanyl-L-serine in the Fischer depiction. (b) The schematic structure of a hexapeptide in the Fischer depiction, resulting in inefficient use of space. (c) The 'three-letter' and 'one-letter' conventions for a representative peptide, GGA---FP.

The 'three-letter system' has some advantages and has gradually been extended (Figure 1.4) to encompass several amino acids other than the coded amino acids. It is usually used to display schemes of laboratory peptide synthesis (Chapter 7) since it allows protecting groups and other structural details to be added, something that is very difficult and often confusing if attempted with the one-letter system.

The one-letter abbreviation (like its three-letter equivalent) represents 'an amino-acid residue' and the system allows the structure of a peptide or protein to be conveniently stated as a string of letters, written as a line of text, incorporating the long-used convention that the amino terminus (the '*N*-terminus') is to the LEFT and the carboxy terminus (the '*C*-terminus') is to the RIGHT. This convention originates in the Fischer projection formula for an L-α-amino acid or a peptide made up of L-α-amino acids; the L-configuration places the amino group to the left and the carboxy group to the right in a structural formula, as in Figure 1.3.

There are increasing numbers of violations of these rules; *N*-acetyl alanine, for example, being likely to be abbreviated Ac—Ala in the research literature or its correct abbreviation Ac—Ala—OH (but never Ac—A). This does not usually lead to ambiguity on the basis of the rule that peptide structures are written with the *N*-terminus to the left and the *C*-terminus to the right. Thus, Ac—Ala should still be correctly interpreted by a reader to mean CH_3—CO—NH—$CH(CH_3)$—CO_2H when this rule is kept in mind, since Ala—OAc (more correctly, H—Ala—OAc) would represent the 'mixed anhydride' NH_2—$CH(CH_3)$—CO—O—CO—CH_3 (there is a footnote about the term 'mixed anhydride' on p. 152).

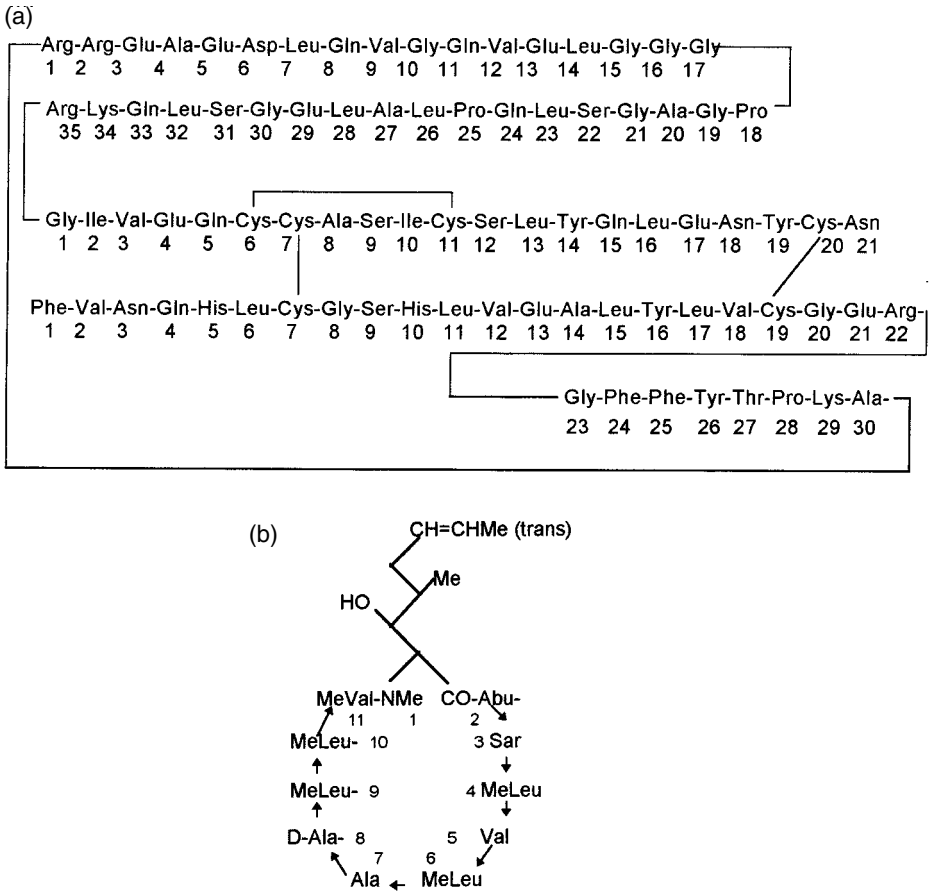


Figure 1.4. Post-translationally modified peptides: (a) Human proinsulin. (b) Cyclosporin A (Me is CH_3). As well as the post-translationally modified threonine derivative (residue 1, called 'MeBmt'), cyclosporin A contains one D-amino acid, four *N*-methyl-L-leucine residues, one 'non-natural' amino acid, Abu (butyrine, side-chain C_2H_5), Sar (sarcosine, *N*-methylglycine) and *N*-methyl-L-valine, but only two of the eleven residues are coded L-amino acids, valine and alanine.

Links through functional groups in side-chains of the amino-acid residues can be indicated in abbreviated structures of peptides (Figure 1.4). Cyclisation between the *C*- and *N*-termini to give a cyclic oligopeptide can also be shown in abbreviated structural formulae. Insulin (Figure 1.4) provides an example of the relatively common 'disulphide bridge' (there are three of these in the molecule), whereas cyclosporin A (a cyclic undecapeptide from *Trichoderma inflatum*, which is valuable for its immunosuppressant property that is exploited in organ-transplant surgery) is a product of post-translational cyclisation (Figure 1.4).