(1993) *Prot. Eng.* 6, 485–500
43 Brünger, A. T. and Nilges, M. (1993) *Q. Rev. Biophys.* 26, 49–125
44 Aszódi, A. and Taylor, W. R. (1994) *Biopolymers* 34, 489–506
45 Saitoh, S., Nakai, T. and Nishikawa, K. (1993) *Prot. Struct. Func. Genet.* 15, 191–204
46 Taylor, W. R. (1993) *Prot. Eng.* 6, 593–604
47 Havel, T. F. and Snow, M. E. (1991) *J. Mol. Biol.* 217, 1–7
48 Srinivasan, S., March, C. J. and Sudarsanam, S. (1993) *Prot. Sci.* 2, 277–289
49 Taylor, W. R., Jones, D. T. and Segal, A. W. (1993) *Prot. Sci.* 2, 1675–1685
50 Abagyan, R. A. (1993) *FEBS Lett.* 325, 17–22
51 Skolnick, J. and Kolinski, A. (1991) *J. Mol. Biol.* 221, 499–531
52 Hinds, D. A. and Levitt, M. (1992) *Proc. Natl Acad. Sci. USA* 89, 2536–2540
53 Baker, D., Chan, H. S. and Dill, K. A. (1993) *J. Chem. Phys.* 98, 9951–9962
54 Dill, K. A., Fiebig, K. M. and Chan, H. S. (1993) *Proc. Natl Acad. Sci. USA* 90, 1942–1946
55 Skolnick, J., Kolinski, A., Brooks, C. L., Godzik, A. and Rey, A. (1993) *Curr. Biol.* 3, 414–423
56 Godzik, A., Kolinski, A. and Skolnick, J. (1993) *J. Comput. Chem.* 14, 1194–1202
57 Rey, A. and Skolnick, J. (1993) *Prot. Struct. Funct. Genet.* 16, 8–28

Expanding the genetic lexicon: incorporating non-standard amino acids into proteins by ribosome-based synthesis

Steven A. Benner

Only 20 amino acids are normally incorporated into proteins synthesized in living cells, and this has limited the structural range of proteins that can be prepared. New methods that allow the incorporation of amino acids that are not normally encoded by natural genes are being developed: these include reassigning functions within the existing genetic code, and expanding the genetic code by constructing additional, non-natural codons. Used in conjunction with recent major advances in understanding protein structure–function relationships, these approaches should extend the range of *de novo* protein designs that are possible.

Proteins are synthesized in living organisms by a two-step process. First, a messenger RNA (mRNA) is synthesized when RNA polymerase copies the DNA in a gene, a process known as transcription. The mRNA is translated by a ribosome, a complex containing both protein and RNA that binds to the message and converts triplet 'words' in the RNA language, written using the four nucleic acid 'letters', adenine (A), uracil (U), guanine (G) or cytosine (C), into one of the 20 natural 'proteinogenic' amino acids in a polypeptide chain. Transfer RNA (tRNA) serves as an adaptor in this process. One end of the folded tRNA structure holds the amino acid, while the other presents three bases that are complementary to the triplet codon (the 'anticodon') to the mRNA. The ribosome then catalyses the synthesis of a peptide bond between the amino acid held by the tRNA molecule and the growing polypeptide chain.

With three-letter words, and only four letters to build them from, only 64 words (4^3) are possible in the genetic lexicon. This limits the number of types of amino acid that can be built into proteins by ribosome-based translation. In contemporary living organisms, this limitation is more severe than might be obvious at first glance. The genetic code is degenerate, meaning that most individual amino acids are encoded by more than one triplet codon. For example, six codons (UCA, UCG, UCU, UCC, AGC and AGU) all encode serine. In addition, three codons (UGA, UAA and UAG) encode 'stop'. When

S. A. Benner is at the Laboratory for Organic Chemistry, ETH, CH-8092 Zurich, Switzerland

all 64 of the possible codons are used up encoding 20 natural, or 'proteinogenic', amino acids. These 20 are only a small fraction of the thousands of amino acids that are conceivable, and proteins synthesized by ribosome-based translation of an mRNA can have only a limited number of structures.

Reconstruction of the genomes of ancient organisms shows that this limitation dates back at least 1.5×10^9 years to the point when the three primary kingdoms of life (archaebacteria, eubacteria and the eukaryotes) first diverged[1]. Indeed, much of contemporary biochemistry reflects the fact that only limited functionality can be encoded by a contemporary genome. For example, the absence of amino acids having side chains with appropriate redox potentials, or bearing an aldehyde group, or able to form good carbon anions with 'Umpolung' potential (a reversal of the normal patterns of nucleophilicity and electrophilicity within a molecule), creates a need for nicotinamide and flavin, pyridoxal, and thiamine cofactors, respectively. Alternatively, the presence of such cofactors in an RNA world emerging before ribosome-based synthesis of proteins may explain why the genetic code did not develop to include amino acids bearing such functionalities[1].

The need for more amino acids

The limited coding potential also places constraints on the mechanistic enzymologist. With the development of techniques for performing site-directed mutagenesis in enzymes[2], and the first synthetic genes designed to facilitate mutagenesis experiments[3], protein engineering became a promising tool for obtaining mechanistic insights into how enzymes work. Yet, early in this work, it became clear that the 20 proteinogenic amino acids simply do not have a sufficient range of functionality to address many of the most interesting mechanistic problems[2]. Many mechanistic questions begged for the substitution (for example) of tyrosine by fluorotyrosine, or other natural amino acids by other non-standard amino acids. Unfortunately, this substitution was not possible with the ribosome-based translation system in contemporary organisms.

The limitation has proven especially severe to those who wish to design *de novo* proteins that fold in solution and catalyse reactions[4]. A specific recent case is the design of a polypeptide that has catalytic activity as an oxaloacetate decarboxylase[5,6]. The design called for a polypeptide carrying an amino group with a low pK_a on a side chain. This was effected either by placing the terminal amine at the amino-terminal end of an α helix, or by placing the ϵ-amino group of a lysine in a positively charged environment, which would reduce its pK_a accordingly. While both of these goals were achieved by using just the 20 standard amino acids, they could have been achieved far more easily just by including a non-standard amino acid bearing an amino group with a lower pK_a (for example,

Getting more amino acids

Solid-phase synthesis

Solid-phase synthesis is, of course, one possible approach for preparing polypeptides containing any amino acid, including non-standard ones. However, the products of peptide synthesis have proven, in general, to be difficult to isolate in pure form when the polypeptide is longer than ~50 amino acids. While progress continues to be made, and longer polypeptides of sufficient purity can now be obtained in some laboratories[7], even these peptides are short by biological standards.

Rearranging the genetic code

In the 1970s, Hecht and his research group[8,9] began to implement a strategy for incorporating amino acids other than the standard ones into a protein using ribosome-based translation. In their strategy, one of the three 'stop' codons was recruited to play a coding role. A suppressor tRNA molecule, which bears a triplet anticodon complementary to the stop codon, was then recruited to act as an adaptor between the stop codon and a non-standard amino acid. Hecht then developed the chemistry for attaching a non-standard amino acid to the 3'-end of the suppressor tRNA.

In its design, this strategy was similar to that used by Miller *et al.*[10] in very early 'site-directed mutagenesis' experiments, where standard amino acids were incorporated by suppression of stop signals built into an mRNA by mutation. In addition, it is analogous to the incorporation of selenocysteine (the twenty-first proteinogenic amino acid) into polypeptides at a UGA codon[11,12], which occurs naturally in some organisms. Finally, it may be a model for the process by which codon signification drifts naturally over long periods of evolutionary time[13].

This approach was first put into practice in elegant work by Schultz and his group at Berkeley (CA, USA)[14,15], and by Chamberlin and co-workers at the University of California (Irvine)[16,17]. Schultz and coworkers have been particularly active in showing the scope and value of the new technology. They have incorporated isotopically labeled[18], and conformationally restricted amino acids into proteins[19], and used non-standard amino acids to study structure, stability[20], and mechanism[21] in enzymes. Brunner and his group at the ETH (Zürich, Switzerland) have used the technology to incorporate photoactive labeling groups into polypeptides[22]. The value, as a research tool, of incorporating non-standard amino acids into proteins by using ribosome-based translation has thus been thoroughly demonstrated.

The strategy does have limitations, however. Only two (at most) additional amino acids can be incorporated independently into proteins by this strategy: at least one of the three stop codons must be retained as a stop signal. Yet, some of the more interesting horizons involve incorporating two or more different non-standard amino acids built a protein (for example, a fluorescence-energy donor and a fluorescence-energy

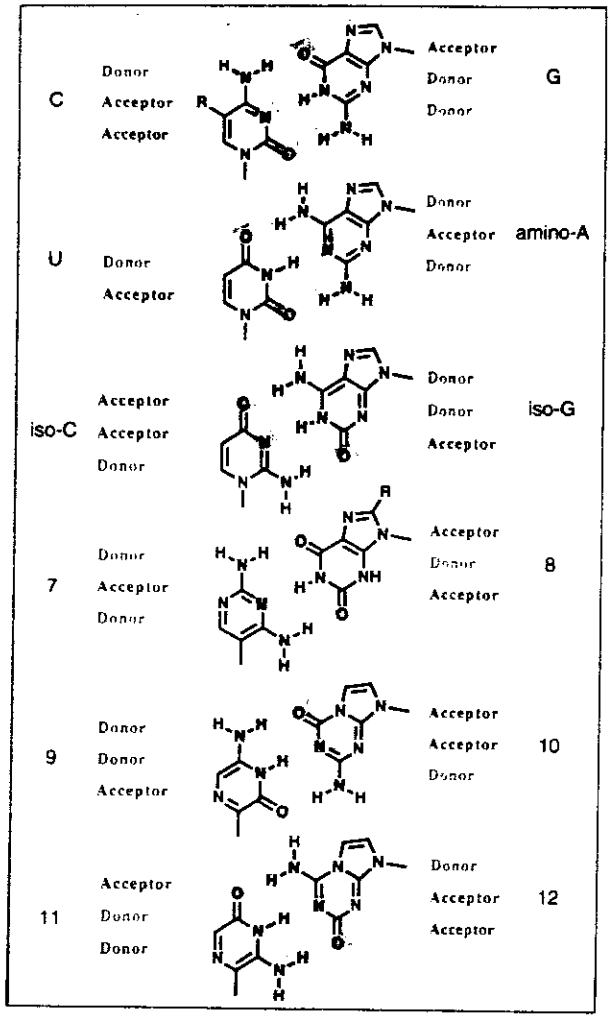

Figure 1

Twelve bases can form six independently replicating base pairs within the Watson–Crick geometry. All unnatural base pairs shown have been synthesized in our laboratory (S. Benner, unpublished).

efficiency of suppression of nonsense codons is not normally particularly high. Natural translation systems include release factors, i.e. proteins that bind to stop codons and encourage the disassociation of the ribosome from the message[23,24]; termination mediated by release factors competes with readthrough mediated by a suppressor tRNA, making efficient suppression of a nonsense codon difficult. Finally, when compared with the amounts of protein that can be produced in living cells, only small amounts of proteins can be prepared using *in vitro* translation systems. The problems that must be solved to use the Hecht–Schultz–Chamberlin strategy within living cells are herculean, although not necessarily insurmountable[25]. It would be necessary first to eliminate the triplet as a stop signal in all of the endogenous genes – removal of the release factor would, otherwise, almost certainly be lethal.

Expanding the genetic code

In Zürich, we approached the same problem from a different perspective. We began by analysing the structure of the Watson–Crick [illegible]

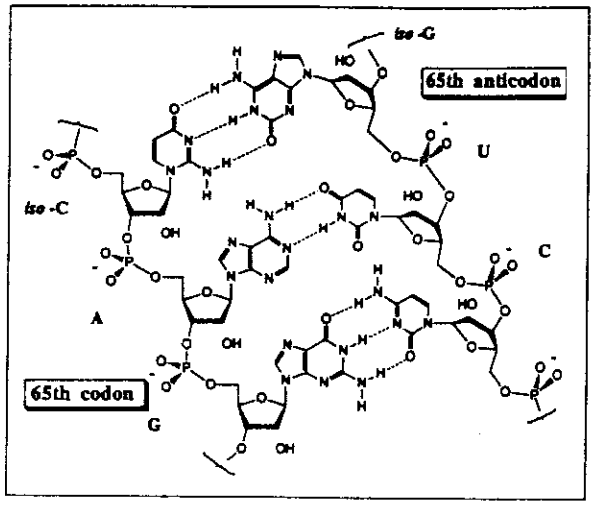

Figure 2

The 65th codon (incorporating the non-standard nucleoside iso-C) and its complementary anticodon (incorporating iso-dG) allow the incorporation of non-standard amino acids into proteins synthesized by translation, thereby expanding the genetic lexicon.

of complementarity by which it performs its duties as an element of a system for molecular recognition[26]. The first rule pairs a large heterocycle (a purine) with a small heterocycle (a pyrimidine) to achieve charge complementarity. The second rule pairs hydrogen-bond donors (-NH groups) with hydrogen-bond acceptors (an N or an O bearing an unshared pair of electrons). This analysis suggested that at least six base-pairing schemes were possible, given the Watson–Crick geometry (Fig. 1; Refs 27, 28). Because each of the base pairs is joined by a different hydrogen-bonding pattern, each should be independently replicable in the presence of all of the others.

Additional nucleoside bases that are independently replicable imply additional triplet codons (Fig. 2); 216 (6^3) and 512 (8^3) triplet codons are possible with six and eight independently pairing nucleoside bases, respectively. A full 1728 (12^3) triplet codons are possible if all 12 bases are available in the genetic alphabet. While some of these codons may be synonymous (due to wobble pairing) in natural translation systems, the inclusion of non-standard bases should certainly permit the design of additional codons that might be recognized by tRNA with the complementary non-standard bases in the anticodon loop, but not by standard tRNAs or release factors.

At the time of writing, nucleosides bearing all of the non-standard bases have been prepared (unpublished). For most of the nucleotides, the physical properties, especially those relevant to base pairing, complementarity and suitability for incorporation into an information-storage molecule, have been examined in detail. Enzyme-catalysed template-directed synthesis of oligonucleotides containing the non-standard bases has also been explored.

The value of additional base pairs for expanding the genetic lexicon was also appreciated by Bain and Chamberlin at the University of California (Irvine),

been set up to develop technology using the non-standard base pairs to encode non-standard amino acids in ribosome-based translation. The iso-C–iso-G base pair (Fig. 1) was chosen for the new codon–anticodon pair; the 65th codon was (iso-C)AG with the corresponding anticodon CU(iso-G) (Fig. 2). For comparison, the UAG codon, signifying 'stop' in the standard genetic code, was used as both reference and control in these experiments[29]. The non-standard amino acid chosen was iodotyrosine, in parallel with work done previously at Irvine.

To incorporate a non-standard amino acid by using a non-standard triplet codon containing a non-standard base pair, a tRNA containing a non-standard base in the anticodon loop (Fig. 3) was first prepared. This was then charged with iodotyrosine following chemistry similar to that developed by Hecht, Schultz, Bain and Chamberlin[4,8,9,14–21,29]. Next, a mRNA containing the new non-standard (iso-C)AG codon was prepared (Fig. 4). These components were then used successfully in an *in vitro* translation system[29]. When presented with a message containing the non-standard (iso-C)AG codon and a charged, non-standard, tRNA containing the non-standard anticodon CU(iso-G), ribosomes incorporated iodotyrosine with high (90%) efficiency (Table 1). This was considerably higher than the efficiency (~63%) with which the stop codon (UAG) was translated by a charged suppressor tRNA containing the CUA anticodon. As nearly every detail of the two systems was the same, the difference in yield could only be attributed to the difference in the first base of the codon.

This was the first time that an enzymic process had been observed where the non-standard base pair was accepted as readily as standard bases. With polymerases, non-standard bases are generally transcribed with lower efficiency than standard bases. In those instances where the effect has been studied in detail, the polymerase is able to sense incorporation of an unnatural base up to five bases beyond the site of incorporation[30].

There were further unexpected results. When the mRNA containing the UAG stop codon was incubated in the absence of charged suppressor tRNA, translation of the message was terminated, as expected. The polypeptide fell off the ribosome, and the mRNA disassociated. However, when a mRNA containing the non-standard (iso-C)AG codon was incubated without the charged non-standard tRNA, a new set of hydrophilic peptide products was isolated (Fig. 4; Table 1; Ref. 29). These proved to be the products of a frameshift mutation, where a ribosome skips over the non-standard iso-C in the mRNA and continues translating triplet codons in a new reading frame. This result underscores the fact that a nonsense codon (a codon that lacks a tRNA molecule to translate it) and a stop codon are functionally different. To stop translation and cause the ribosome to release the mRNA, a nonsense codon must evidently be recognized by release factor, not the nonsense codon, a frameshift mutation, not termination, follows.

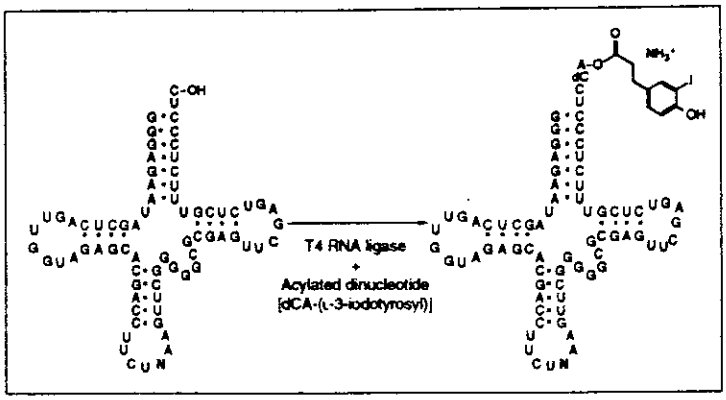

Figure 3

A truncated form of non-hypermodified tRNA derived from $tRNA^{Gly}_{CUN}$-dCA missing the last two bases, prepared by chemical synthesis, and charged with iodotyrosine at the 3'-end; N is either adenosine (the anticodon for the stop codon UAG) or 2'-deoxyisoguanosine (the anticodon for the non-standard codon isoC-AG).

a

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
	AUG	GGU	UUA	UAU	UUG	GCC	CUU	UUU	UAG	GGA	CUC	UAC	CUA	GGG	CUG	UUC	UAA	UGA
(i)	Met	Gly	Leu	Tyr	Leu	Gly	Leu	Phe	*End*									
(ii)	Met	Gly	Leu	Tyr	Leu	Gly	Leu	Phe	iTyr	Gly	Leu	Tyr	Leu	Gly	Leu	Phe	*End*	
(iii)	Met	Gly	Leu	Tyr	Leu	Gly	Leu	Phe		Arg	Asp	Cys	Thr	*End*				

b

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
	AUG	GGU	UUA	UAU	UUG	GCC	CUU	UUU	iCAG	GGA	CUC	UAC	CUA	GGG	CUG	UUC	UAA	UGA
(i)	Met	Gly	Leu	Tyr	Leu	Gly	Leu	Phe	*End*									
(ii)	Met	Gly	Leu	Tyr	Leu	Gly	Leu	Phe	iTyr	Gly	Leu	Tyr	Leu	Gly	Leu	Phe	*End*	
(iii)	Met	Gly	Leu	Tyr	Leu	Gly	Leu	Phe		Arg	Asp	Cys	Thr	*End*				

Figure 4

The mRNA molecules used to compare (**a**) rearranging and (**b**) expanding the genetic code with the 65th codon (Fig. 2) as alternative strategies for incorporating non-standard amino acids into translated peptides. Shown are the translation products (i, octamer), and (ii, hexadecamer) in the absence or presence, respectively, of suppression; and (iii, dodecamer) following a frameshift and successful termination at the next stop codon. iTyr is L-3-iodotyrosine. From Ref. 29.

These observations suggest that an expanded genetic lexicon could be implemented in an *in vivo* cell-culture system, as the non-standard codons will escape recognition by release factors. The efficiency of translation should be high. For this, however, further research must be done to identify the polymerases needed to replicate and transcribe, in a stable manner, duplex DNA containing non-standard base pairs. Furthermore, enzymes that phosphorylate non-standard nucleosides and charge tRNA molecules with non-standard amino acids must be obtained.

The next step: practical production systems

So where are the horizons? At present, the technology for incorporating non-standard amino acids into polypeptides by ribosome-based translation of codons formed from non-standard amino acids is expensive – far too expensive to be considered for the

Table 1. Translation of the UAG ('stop') and isoCAG ('non-standard') codons with different tRNA molecules[a]

	mRNA	tRNA[b]	Amino acid 'charge'[c]	Full-length products (%)	Termination products (%)	Frameshift products (%)
1	UAG	None	–[d]	4	96	0
2	UAG	CUA	None[d]	4	96	0
3	UAG	CUiG	None	3	97	0
4	UAG	CUA	Iodotyrosine	67	33	0
5	UAG	CUiG	Iodotyrosine	9	91	0
6	iCAG	None	–	3	25	71
7	iCAG	CUA	None	4	17	80
8	iCAG	CUiG	None	4	14	81
9	iCAG	CUA	Iodotyrosine	3	24	73
10	iCAG	CUiG	Iodotyrosine	91	8	1

[a] From Ref. 29.
[b] The codon (written 5' to 3') at the position designated in Fig. 2. Column 3 shows the anticodon (written 5' to 3') in the tRNA.
[c] The amino acid 'charge' is the amino acid carried at the 3'-end of the tRNA molecule. When none is present, no amino acid can be incorporated into the growing polypeptide chain. Note that readthrough is more efficient with the non-standard codon isoCAG than with the stop codon UAG.
[d] The symbol (–) signifies that no tRNA was present; (None) signifies that the tRNA present was not charged with an amino acid.

it is clear that this approach comes at a critical time in the development of our understanding of protein structure. It is now possible to predict the secondary structure of proteins starting from sequence data alone[31–34]. Structural predictions are being made and published for protein families before experimental structures become available[35,36]; comparison of several of the predicted structures with subsequently determined crystal structures has proven the predictions to be remarkably accurate[37,38].

In addition, the *de novo* design of polypeptides has progressed substantially over the past few years. In several cases, the conformation of designed peptides has been explored in solution by multidimensional nuclear magnetic resonance[5,39–42] and in the solid state by crystallography[43]. Small designed peptides, with remarkable catalytic power, have been known for over a decade[44]. Recently, however, the details of the catalytic mechanism have been elucidated in some detail[6]. With the tools now emerging to manage the conformation and design of proteins formed from the 20 natural proteinogenic amino acids, technology that allows access to the synthesis of proteins containing a greater variety of building blocks is likely to find wide use.

References

1 Benner, S. A., Ellington, A. D. and Tauer, A. (1989) *Proc. Natl Acad. Sci. USA* 86, 7054–7058
2 Knowles, J. R. (1987) *Science* 236, 1252–1258
3 Nambiar, K. P., Stackhouse, J., Stauffer, D. M., Kennedy, W. G., Eldridge, K. J. K. and Benner, S. A. (1984) *Science* 223, 1299–1301
4 Hecht, M. H., Richardson, J. S., Richardson, D. C. and Ogden, R. C. (1990) *Science* 249, 884–891
5 Johnsson, K., Allemann, R. K. and Benner, S. A. (1990) in *Molecular Mechanisms in Bioorganic Processes* (Bleasdale, C. and Golding, B. T., eds), pp. 166–187, Royal Society of Chemistry
6 Johnsson, K., Allemann, R. K., Widmer, H. and Benner, S. A. (1993)
7 Milton, R. C. D., Milton, S. C. F. and Kent, S. B. H. (1992) *Science* 256, 1445–1448
8 Hecht, S. M., Alford, B. L., Kuroda, Y. and Kitano, S. (1978) *J. Biol. Chem.* 253, 4517–4520
9 Roesser, R., Chorghade, M. S. and Hecht, S. M. (1986) *Biochemistry* 25, 6361–6365
10 Miller, J. H., Coulondre, C., Hofer, M., Schmeissner, U., Sommer, H. and Schmitz, A. (1979) *J. Mol. Biol.* 131, 191–222
11 Zinoni, F., Birkmann, A., Stadtman, T. C. and Bock, A. (1986) *Proc. Natl Acad. Sci. USA* 83, 4650–4654
12 Chambers, I., Frampton, J., Goldfarb, P., Affara, N., McBain, W. and Harrison, P. R. (1986) *EMBO J.* 5, 1221–1227
13 Osawa, S., Jukes, T. H., Watanabe, K. and Muto, A. (1992) *Microbiol. Rev.* 56, 229–264
14 Noren, C. J., Anthony-Cahill, S. J., Griffith, M. C. and Schultz, P. G. (1989) *Science* 244, 182–188
15 Robertson, S. A., Ellman, J. A. and Schultz, P. G. (1991) *J. Am. Chem. Soc.* 113, 2722–2729
16 Bain, J. D., Glabe, C. G., Dix, T. A., Chamberlin, A. R. and Diala, E. S. (1989) *J. Am. Chem. Soc.* 111, 8013–8014
17 Bain, J. D., Wacker, D. A., Kuo, E. E. and Chamberlin, A. R. (1991) *Tetrahedron* 14/15, 2389–2400
18 Ellman, J. A., Volkman, B. F., Mendel, D., Schultz, P. G. and Wemmer, D. E. (1992) *J. Am. Chem. Soc.* 114, 7959–7961
19 Mendel, D., Ellman, J. and Schultz, P. G. (1993) *J. Am. Chem. Soc.* 115, 4359–4360
20 Mendel, D., Ellman, J. A., Chang, Z. Y., Veenstra, D. L., Kollman, P. A. and Schultz, P. G. (1992) *Science* 256, 1798–1802
21 Chung, H. H., Benson, D. R. and Schultz, P. G. (1993) *Science* 259, 806–809
22 Baldini, G., Martoglio, B., Schachenmann, A., Zugliani, C. and Brunner, J. (1988) *Biochemistry* 27, 7951–7979
23 Konecki, D. S., Aune, K. C., Tate, W. and Caskey, C. T. (1977) *J. Biol. Chem.* 252, 4514–4520
24 Caskey, C. T. and Campbell, J. M. (1979) in *Nonsense Mutations and tRNA Suppressors* (Celis, J. E. and Smith, J. D., eds), pp. 81–96, Academic Press
25 Kast, P. and Hennecke, H. (1991) *J. Mol. Biol.* 222, 99–124
26 Benner, S. A. *et al.* (1987) *Cold Spring Harbor Symp. Quant. Biol.* 52, 53–63
27 Switzer, C. Y., Moroney, S. E. and Benner, S. A. (1989) *J. Am.*

28 Piccirilli, J. A., Krauch, T., Moroney, S. E. and Benner, S. A. (1990) *Nature* 343, 33–37
29 Bain, J. D., Chamberlin, A. R., Switzer, C. Y. and Benner, S. A. (1992) *Nature* 356, 537–539
30 Piccirilli, J. A., Moroney, S. E. and Benner, S. A. (1991) *Biochemistry* 30, 10350–10356
31 Benner, S. A. and Gerloff, D. (1991) *Adv. Enzyme Regul.* 31, 121–181
32 Benner, S. A., Cohen, M. A. and Gerloff, D. L. (1993) *J. Mol. Biol.* 229, 295–305
33 Gerloff, D. L., Jenny, T. F., Knecht, L. J., Gonnet, G. H. and Benner, S. A. (1993) *FEBS Lett.* 318, 118–124
34 Gerloff, D. L., Jenny, T. F., Knecht, L. J. and Benner, S. A. (1993) *Biochem. Biophys. Res. Commun.* 194, 560–565
35 Bazan, J. F. (1990) *Proc. Natl Acad. Sci. USA* 87, 6934–6938
36 Russell, R. B., Breed, J. and Barton, G. J. (1992) *FEBS Lett.* 304, 15–20
37 Knighton, D. R. *et al.* (1991) *Science* 253, 407–414
38 Thornton, J. M., Flores, T. P., Jones, D. T. and Swindells, M. B. (1991) *Nature* 354, 105–106
39 Ciesla, D. J., Gilbert, D. E. and Feigon, J. (1991) *J. Am. Chem. Soc.* 113, 3957–3961
40 Osterhout, J. J., Jr *et al.* (1992) *J. Am. Chem. Soc.* 114, 331–337
41 Zhou, N. E., Zhu, B. Y., Sykes, B. D. and Hodges, R. S. (1992) *J. Am. Chem. Soc.* 114, 4320–4326
42 Klaus, W. and Moser, R. (1992) *Prot. Eng.* 5, 333–341
43 Lovejoy, B., Choe, S., Cascio, D., McRorie, D. K., Degrado, W. F. and Eisenberg, D. (1993) *Science* 259, 1288–1293
44 Gutte, B., Daeumigen, M. and Wittschieber, E. (1979) *Nature* 281, 650–655

Design of protein structures: helix bundles and beyond

Chris Sander

The design of proteins or peptides with novel functions can be achieved either by modifying existing molecules or by inventing entirely new structures and sequences that are unknown in nature. Combinatorial-design strategies have led to the first *de novo* proteins, but these still lack some of the desired attributes. The most promising practical strategies for developing proteins with useful biological or chemical function combine theoretical design with experimental screening or selection systems.

Nature has evolved highly intricate and useful proteins over many millions of years, gradually optimizing protein function in response to selective pressure. When will humans be able to sidestep evolution and design novel proteins with desired structures and functions? Will the new proteins be redesigns of natural proteins, or *de novo* inventions with sequences not found in nature? The answers are not yet available, but the first steps towards them have been taken.

Re-engineering work has proven that the protein engineer has considerable latitude in modifying existing frameworks, not just in replacing surface- or active-site residues, but also in rearranging loop regions and replacing residues in the protein's interior. The rules of *de novo* design are already partially understood for simple structures, such as four-helix bundles, and the first stable proteins have been designed *de novo*. Simple functions, such as metal-binding sites, have been introduced into some designed proteins, while existing proteins have been functionally optimized by more intricate modification of the protein using, for instance, *in vitro* selection systems.

To illustrate the current capabilities and difficulties of structurally oriented protein design, this article discusses what has been achieved with a particularly simple class of protein fold – bundles of four α helices. These achievements include redesigning the topology of loop connections, redesigning the packing of the hydrophobic core, as well as creating *de novo* designs, i.e. proteins whose amino acid sequence is newly invented and not seen in nature.

A simple architecture: α-helix bundles

The architecture of the most commonly occurring type of four-helix bundle is particularly simple: each of the four helices is oriented antiparallel to its two nearest neighbors and parallel to its diagonal, more distant, neighbor. Residues on the interior helix-faces

C. Sander is at the European Molecular Biology Laboratory