Tetrahedron Vol. 47, No. 14/15, pp. 2543-2562, 1991
Printed in Great Britain

0040-4020/91 $3.00+.00

Protein Overproduction for Organic Chemists

STUART L. SCHREIBER* AND GREGORY L. VERDINE*
Department of Chemistry
Harvard University
Cambridge, Massachusetts 02138

(*Received in USA* 24 *September* 1990)

In this review we present the principles behind protein overexpression in bacteria, emphasizing how this biosynthetic system can be manipulated to generate large quantities of proteins for study. In addition to the classical (molecular biological) methods for constructing protein-overproducing bacterial, we discuss our recently developed (chemical/enzymatic) method, the Expression-Cassette Polymerase Chain Reaction (ECPCR). The chemical/enzymatic transformation of an unexpressable to an expressable gene afforded by ECPCR can routinely be carried out in the experimental organic chemistry laboratory, hence, ECPCR offers a convenient point of entry for chemists interested in macromolecular science.

Whereas the traditional approach of isolating proteins from natural sources often provided only limited quantities for structural and mechanistic studies, more recent efforts involving the use of "engineered" bacteria as protein biosynthetic factories have overcome this materials hurtle and thus transformed the study of proteins. The change in practice from classical isolation to engineered overexpression (overproduction) shifted the emphasis on expertise from protein biochemistry to molecular biology; this shift, however, had little consequence for chemists interested in entering macromolecular science, because it merely substituted one impenetrable discipline for another. It is thus ironic that an essentially chemical technique, the Polymerase Chain Reaction (PCR),[1-3] has quietly revolutionized the practice of molecular biology in the course of only four years. PCR involves the use of synthetic oligodeoxynucleotides (oligos) and a DNA-synthesizing enzyme (polymerase) to carry out the semisynthetic amplification of a target DNA sequence.[2] This technology forms the basis of our method for gene refabrication — the Expression-Cassette Polymerase Chain Reaction (ECPCR)[4] — which drastically reduces the level of required training in molecular biology and thus facilitates the entry of chemists into the field of protein science. Herein we briefly review molecular aspects of protein expression in bacteria, the design principles that govern engineering of protein-overproducing bacterial strains, and the use of ECPCR in protein dissection and overproduction. It is our intention in this forum to provide an entry-level discussion rather than a comprehensive treatment of these topics, and wherever possible direct the reader to more detailed accounts elsewhere.

THE HOST BACTERIUM FOR PROTEIN OVERPRODUCTION. The enteric bacterium *Escherichia coli* (*E. coli*)[5] is used to carry (host) and amplify foreign DNA sequences during recombinant DNA manipulations. In protein overproduction (see below), *E. coli* serves a dual role as host for a foreign gene sequence and expression surrogate for high-level biosynthesis of the encoded protein. Among the highly engineered *E. coli*

strains used in laboratory practice are those tailored genetically for specific use in protein overproduction; these are so commonly used that one can obtain suitable overproduction strains commercially. *E. coli* is used as DNA host and protein overproducer for several reasons: (1) the bacterium is easily grown using commercially available broths ("media"), it multiplies rapidly (doubles population once per 15-30 min), and requires little specialized equipment for growth (autoclave, incubator) and handling (centrifuge). (2) Foreign DNA (present in a plasmid vector, see below) can readily be introduced into *E. coli* and amplified *in vivo* by simply allowing the bacteria to multiply. (3) More is known about protein biosynthesis in *E. coli* than in any other organism.

ARCHITECTURE OF A HIGHLY EXPRESSED GENE IN *E. COLI*. First we shall examine the elements that direct protein biosynthesis in *E. coli*, so as to understand how this system may be exploited to "trick" bacteria into making large quantities of protein. The architecture of a highly expressed bacterial gene is shown in Figure 1. Flanking the DNA that specifies the amino acid sequence of the protein (coding sequence) are

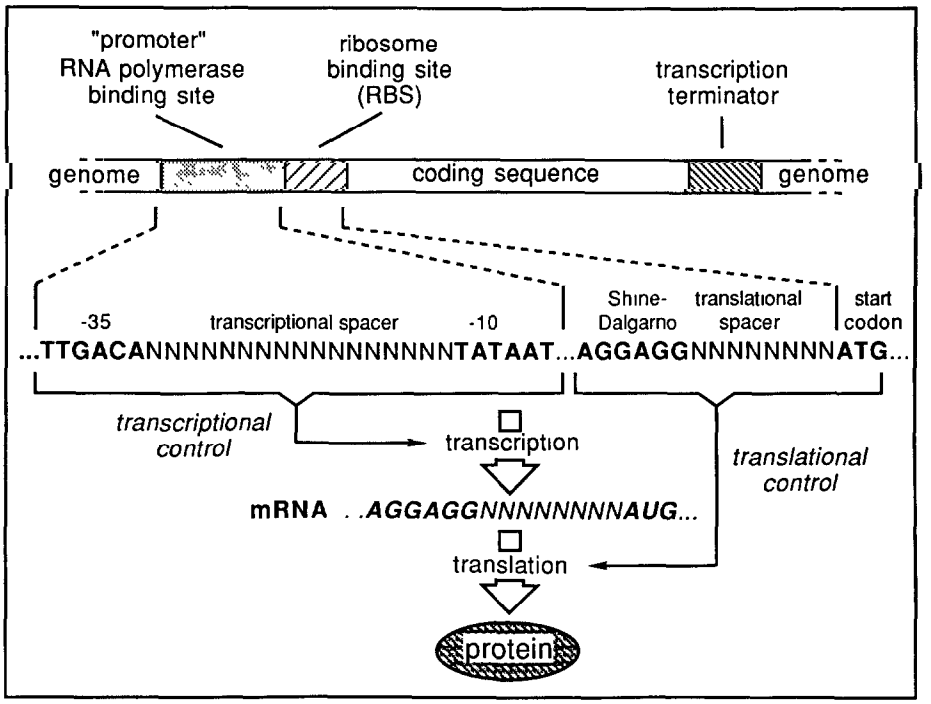

Figure 1 Architecture of a highly expressed gene in *E coli* The illustration at the top represents double-stranded DNA, with various control elements highlighted. Sequences of control elements are shown in the expansion below; by convention, the sequence of only the coding strand of double-stranded DNA is shown, in the 5' (left) to 3' (right) orientation. The sequence of the mRNA transcript is shown in italics N = any nucleotide

elements that control the initiation and termination of messenger RNA (mRNA) synthesis (promoter and transcription terminator, respectively), and the initiation of protein synthesis (ribosome binding site, or RBS). The promoter,[6] which functions as a binding site for RNA polymerase, is most active when it has the following

physical organization: a hexanucleotide sequence 5'-TTGACA-3' followed by a 17-nucleotide spacer and another hexanucleotide 5'-TATAAT-3'; the hexanucleotide elements are centered near positions -35 and -10, respectively (position +1 corresponds to the first nucleotide of the mRNA), and so are named the -35 and -10 regions. Between the promoter and the coding sequence is the RBS, which in mRNA controls the initiation of protein synthesis.[7] The RBS is nominally composed of two sub-elements: (i) the Shine-Dalgarno sequence, which appears to function by direct base-pairing with the 16*S* subunit of ribosomal RNA; and (ii) the translational spacer element, which positions the start codon (5'-ATG-3') on the ribosome relative to the anchored Shine-Dalgarno sequence. While general rules for designing a strong ribosome binding site have proven more elusive than those for promoters, some general principles apply: the consensus hexanucleotide Shine-Dalgarno sequence is more effective than shorter ones, and adenine/thymine-rich translational spacers of 4-8 nucleotides appear to be optimal.[7] Finally, downstream of the coding sequence is the transcription terminator, which aborts transcription by forming a stem-loop structure[8] in the mRNA. Transcription terminators can be viewed as ensuring economical usage of nucleotides (i.e., energy) in the cell by preventing mRNA synthesis beyond the end of the gene. Since many unrelated sequences can function as transcription terminators, they will not be further discussed here.

MOST NATIVE GENES ARE NOT SUITABLY EQUIPPED FOR OVEREXPRESSION IN *E. COLI*. Bacteria tightly regulate the biosynthesis of their proteins, and very few are required in large amounts within the cell (most are catalysts). As mentioned above, transcription and translation of bacterial (prokaryotic) genes are controlled by the promoter and ribosome binding site: even single-nucleotide changes in the -35, -10, and Shine-Dalgarno sequences from those shown in Figure 1 can lead to a 10^3-fold decrease in protein expression.[6,7] The regulatory region of each native bacterial gene is tuned by such changes to produce the appropriate levels of protein. The transcription of many genes is also controlled by the gating action of repressor proteins,[6] which bind DNA in or near the promoter and either: (i) block the access of RNA polymerase to the promoter, or (ii) stabilize duplex (base-paired) DNA structure in the -10 region and thereby prevent RNA polymerase from splitting open DNA in order to copy it.

The expression of genetic information in higher organisms (eukaryotes[9]) differs in several respects from that in bacteria (prokaryotes[9]), with important consequences for overexpression of eukaryotic genes in *E. coli*. One problematic aspect of most eukaryotic genes is that their coding sequence is non-contiguous: as shown in Figure 2, the stretches of coding sequence (exons) are interrupted by segments of non-coding sequence (introns).[10] The entire gene is transcribed — exons and introns — and the introns are subsequently excised in a process termed RNA splicing.[11] Introns in mRNA present a problem in bacterial overexpression because bacteria cannot splice mRNA. The eukaryotic gene, therefore, is useless for bacterial expression. However, if it were possible to take the mature mRNA, which does have contiguous coding sequence, and from that mRNA make a DNA replica, then the DNA copy could be introduced into bacteria and perhaps expressed. Indeed, technology has been developed to copy mRNA into DNA (called complementary DNA or cDNA).[12] A collection of cDNAs representing the total mRNA of a source is termed a cDNA library; the commercial availability of such libraries has largely eliminated the need for one to construct them. From a cDNA library, it is possible to isolate a pure cDNA encoding a protein of interest, using the techniques of molecular biology.[13] While the use of cDNA overcomes the problem of noncontiguity in eukaryotic genes, it still does not permit

high-level expression of the encoded protein in *E. coli*: cDNAs have no promoter and their (eukaryotic) translational initation sequences are not recognized by *E. coli*.

In summary, native bacterial genes come equipped with expression elements that can be recognized by *E. coli* but do not permit high-level expression. Eukaryotic cDNAs, on the other hand, are usually not expressed at all in *E. coli* owing to their lack of required sequence elements.

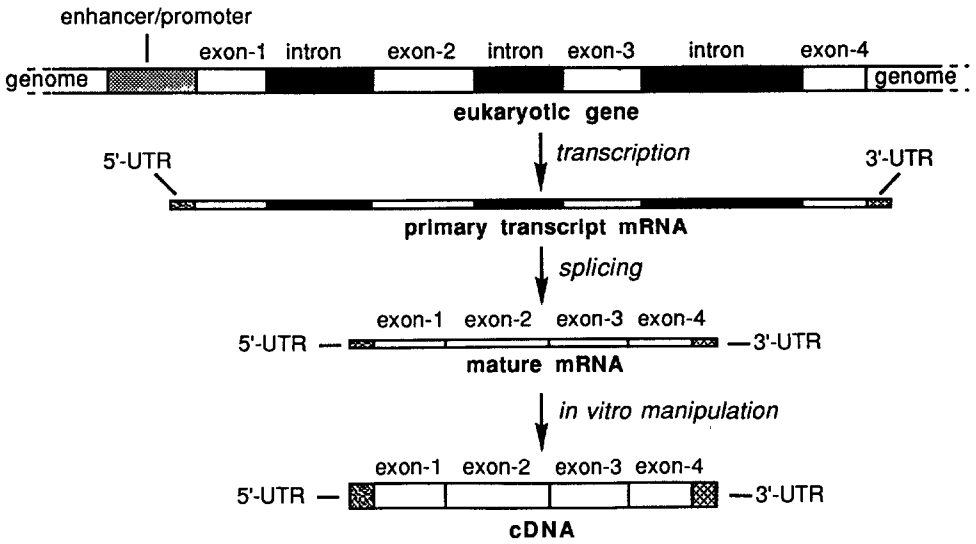

Figure 2. Physical organization of a typical eukaryotic gene, transcript and cDNA. Vertical lines within the coding sequences of the mature mRNA and cDNA denote exon junctions and not interruptions, the coding sequence is contiguous in mRNA and cDNA. The upstream regulatory region (enhancer/promoter) bears no resemblance to a prokaryotic control region. Tranlational control elements present in the enhancer/promoter are transcribed to become the 5'-untranslated region (5'-UTR) of the mRNA; noncoding sequence at the 3'-end (rightward, downstream) of the gene becomes the 3'-UTR. Transcription and splicing take place inside eukaryotic cells (*in vivo*), but synthesis of cDNA utilizes enzymatic reactions *in vitro*.

CONSTRUCTION OF AN OVERPRODUCER: STARTING MATERIALS AND PRODUCTS. The only sequence information present in a native gene (or cDNA) that is indispensible for overexpressing protein is the coding sequence. The molecular transformation required for constructing an overproducer is then apparent: formally, the coding sequence must be excised from its native context and then be inserted between a strong promoter, ribosome binding site (RBS) and transcription terminator.

In order for the newly refabricated gene to be stably propagated in bacteria, it must be present in a carrier DNA molecule called a plasmid or vector. Plasmids are circular, double-stranded DNA molecules having approximately 3,000~10,000 base-pairs (bp). Presented in Figure 3 is a graphical map of the vector pHN1+,[15] in which can be identified the following features common to virtually all plasmids:

(1) *An origin of DNA replication (ori).* The bacterial replication machinery copies plasmids with very high efficiency. Whereas the bacterium makes only one copy of its own genome, it can make up to

~500 copies of a plasmid. This *in vivo* DNA amplification system allows one to obtain large amounts of plasmid DNA by simply growing bacteria and harvesting the plasmid. The pUC origin of replication, present in pHN1+, is commonly used in protein overproduction because it directs intracellular plasmid synthesis to a high level (high *copy number*).

(2) *An antibiotic resistance gene.* The presence of this gene — usually the ampicillin-resistance gene β-lactamase (*bla*[16]) — allows selective killing of bacteria that lack the plasmid. Because the procedures for forcing bacteria to take up plasmids (transformation) are inefficient, there must be some way to weed out the large background of plasmid-less (untransformed) bacteria. Furthermore, bacteria tend to shed plasmids rapidly — especially overproducing plasmids — and continued antibiotic selection thus ensures the continued presence of the plasmid in the bacteria

(3) *Unique restriction sites.* These allow cleavage of the plasmid at specific sites using restriction enzymes, so that DNA can be inserted into the cleavage site. Unique sites are those that occur only once in the plasmid, and like many vectors pHN1+ has a cluster of sites called a multiple cloning site or polylinker.

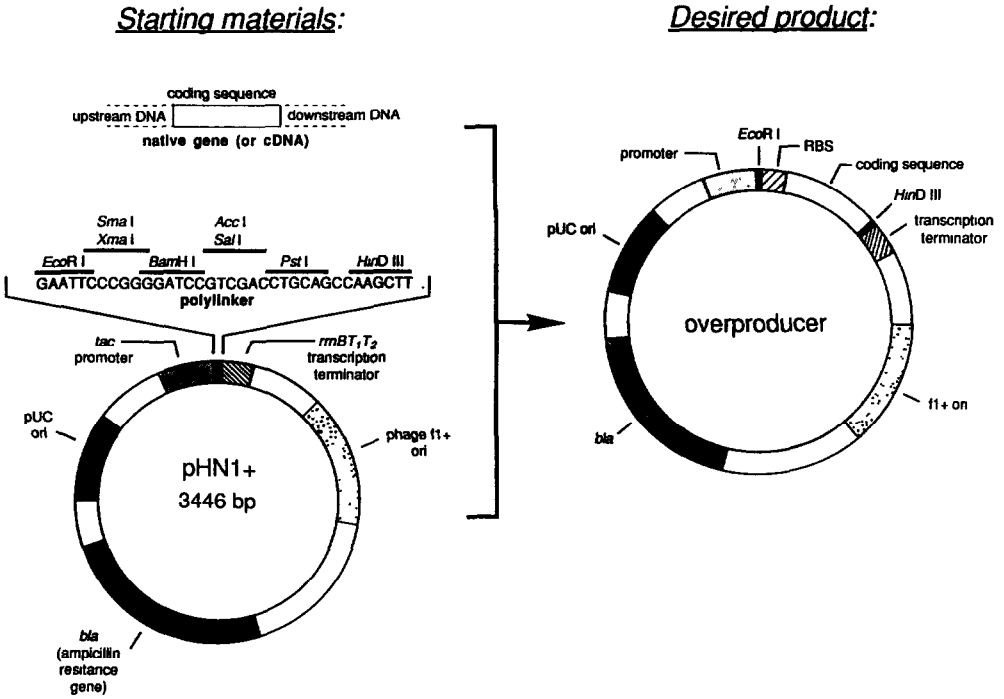

Figure 3. Starting materials and the desired product in construction of an overproducer. Circular figures are graphical maps of plasmid vectors; features denoted are explained in the text. The nucleotide sequence of the pHN1+ polylinker is shown in the expansion; bold bars indicate recognition sites for the restriction enzymes shown immediately above. The f1+ origin of replication (ori) is a bacteriophage (bacterial virus) replication origin that permits this vector to be packaged as single-stranded DNA inside a phage coat. Plasmids containing a phage origin are called "phagemids."

In addition to these nominal features found in all plasmids, pHN1+ has several additional elements that are found primarily in specialized expression vectors used for protein overproduction. Flanking the polylinker in pHN1+ are the strong, inducible (switchable, see below) *tac* promoter,[17] and *rrnB*T_1T_2 transcription terminator[18] (Figure 3). Constructing an overproducer using a vector such as pHN1+ requires the insertion of a RBS-equipped coding sequence into the polylinker (Figure 3). Given that the promoter and transcription terminator elements are contributed by the expression vector, *the fundamental problem of overproducer construction reduces to that of installing a RBS and restriction site at the upstream end of the gene and another restriction site at the downstream end.*[19] Once these installations are complete, the refabricated gene can be cut at the new restriction sites, the expression vector cut in the polylinker, and the gene covalently joined into the vector (this process of restriction digestion and enzymatic ligation is commonly called "cutting and pasting").

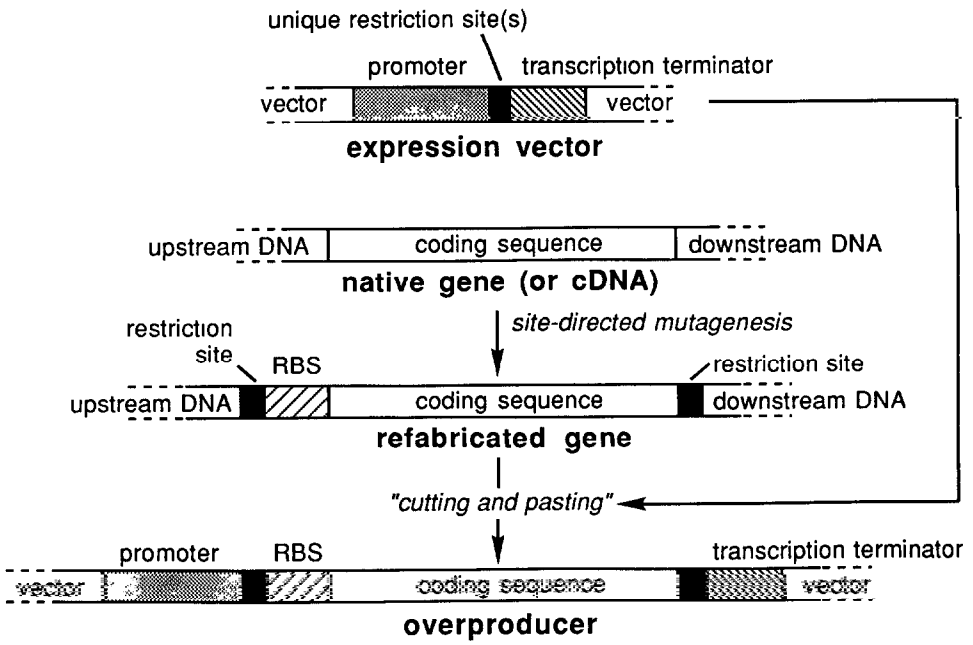

Figure 4. (a) Control region of a typical expression vector; (b) classical approaches to overproducer construction.

THE CLASSICAL APPROACHES TO OVERPRODUCER CONSTRUCTION. Before beginning a protein overproduction following the classical route, the native gene (or cDNA) must first be obtained in pure form, and its nucleotide must be known. Most often the gene is obtained as a plasmid "clone" that was isolated from a library. Should the clone have been isolated in another laboratory, it is necessary to either obtain the clone from the original cloner or to re-isolate the clone. Each of these options has its drawbacks: some investigators are understandably reluctant to give up their clones, and it can take an experienced co-worker months to re-isolate the clone from a library.[19]

Next, the installation of restriction sites and an RBS onto the gene of interest requires the use of molecular biological techniques like site-directed mutagenesis and nuclease BAL-31 deletion mutagenesis. In order to carry out site-directed mutagenesis the gene is required in single-stranded form; most often this is obtained by transferring the gene into the bacteriophage M13, which packages only one strand of its genome during viral assembly. After installing the RBS and restriction sites in several rounds of site-directed mutagenesis, the gene is ready for insertion into the expression vector. The BAL-31 deletion route involves cutting DNA outside but near the coding sequence and enzymatically "chewing back" to its ends. Restriction sites are then added by enzymatically attaching synthetic oligonucleotides (which contain the restriction site) onto the flush ends of the nibbled gene. Drawbacks to the BAL-31 technique include: (i) it affords little control over the precise location of the gene cut; and consequently (ii) only with great difficulty can an efficient RBS be installed onto the coding sequence.

Both of the above approaches require considerable expertise in molecular biology, and even for an experienced experimentalist obtaining the overproducer can often take months.

THE EXPRESSION-CASSETTE POLYMERASE CHAIN REACTION (ECPCR). Like classical routes, overproducer construction using ECPCR requires that the gene (or cDNA) sequence be known. However, unlike the classical methods, a pure clone is not required. A library can be used to donate the desired coding sequence, provided that the desired clone is represented in the library. This aspect of ECPCR alone can shave months off of the time required to obtain an overproducer.

ECPCR (Figure 5a) is an extension of the Polymerase Chain Reaction (PCR), a chemical/enzymatic DNA amplification procedure. PCR works as follows:[21] short synthetic oligodeoxynucleotides (primers) designed to bind different strands of a larger, double-stranded target sequence are synthesized. The target DNA is mixed with a large excess (~10^3-10^9-fold) of the primers, and the temperature is raised to thermally denature the double-stranded target. The temperature is lowered, and the primers bind their respective sites on the target. A thermostable DNA polymerase enzyme (usually *Thermus aquaticus* or *Taq* polymerase) and enzymatic DNA synthesis monomers (2'-deoxynucleoside 5'-triphosphates, or dNTP's) are present in the reaction mixture, so DNA synthesis proceeds from the 3'-ends of the primers, copying the target strand to which the primer is bound, until the end of the target is reached. This constitutes one cycle. Note that in this process the primer has been covalently incorporated into the copy. After three cycles, copies having both ends derived from primers generated, with target DNA in-between; these are exponentially amplified henceforth. PCR can be viewed as a *bona fide* synthetic chemical technique not unlike other enzymatic transformations employed in synthesis: an enzyme (polymerase) catalyzes a chemical conversion (polymerization) on synthetic primers.

In ECPCR, expression and restriction elements are designed into the synthetic primers (Figure 5b), and these become fused during amplification to the intervening coding sequence. The resulting expression cassette (Figure 5a) bears all of the sequence information necessary for cloning and expression in *E. coli*. These operations sidestep the need for training in molecular biology — the primers are generated by automated chemical synthesis, and the amplification step involves subjecting a mixture of commercially available reagents (polymerase, dNTP's, library) to an iterative thermal cycle. Using ECPCR, a coding sequence can be equipping with restriction sites and a new RBS in 24 h,[22] as compared to the many weeks of effort expended in molecular biological manipulations using the traditional approach.

Following synthesis of an expression cassette, it is then cut and pasted into an expression vector (e.g., pHN1+), and the resulting overproducer is transformed into *E. coli* to yield the overproducing bacterial strain. The cutting and pasting and bacterial transformation steps, common to the both traditional approach and ECPCR, are test-tube procedures in which we have found skilled organic experimentalists to excel.

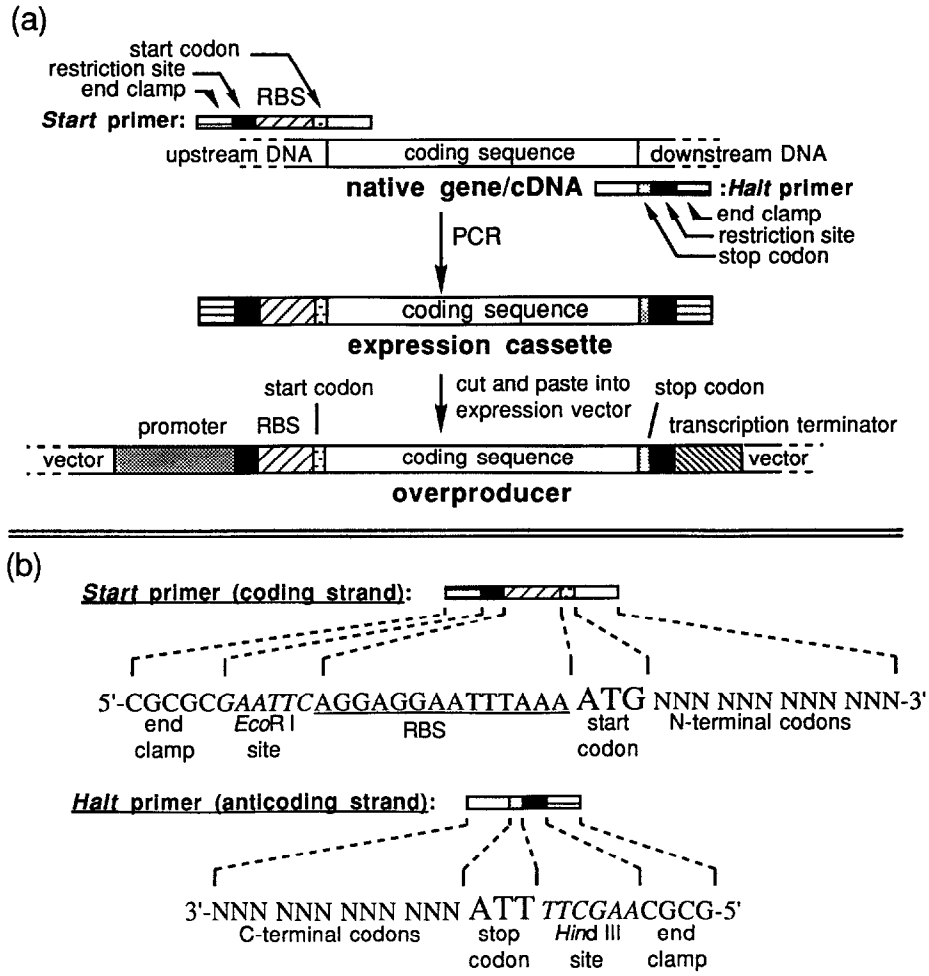

Figure 5 Overproducer construction using ECPCR. (a) Two primers (*start* and *halt*) containing expression information and restriction sites are are used in enzymatic amplification of the target coding sequence, thereby installing onto that coding sequence the primer-derived elements. The amplification product, an expression cassette, is then cut and pasted into an expression vector to yield the overproducer. (b) Sequences of typical *start* and *halt* primers; in this example, the two primers carry different restriction sites, which serves to differentiate the ends of the cassette after cutting. Start primers always correspond to the coding strand of the target; *halt* primers are anticoding, so the elements referred to as codons in the *halt* primers are technically anticodons. N- and C-terminal codons refer to those at the amino- and carboxyl-termini of the expression-cassette-encoded protein, respectively.

Assuming that one has access to an automated DNA synthesizer (oligos can also be purchased from custom supply houses), the only other equipment required for ECPCR are Pipetmen™ (μL-scale pipetting instruments) and three water baths.[23] All of the required materials, except in rare cases the gene or cDNA being refabricated, are commercially available. As opposed to more molecular biological methods, the gene donor DNA used in PCR need not be a pure clone; expression cassettes can be synthesized directly from a library provided that the desired gene is present in the library. This facet virtually eliminates the need to obtain a clone from a private source before proceeding with the overproduction, since libraries from numerous organisms and cell types are now commercially available. These considerations lead us to believe that ECPCR, which is an essentially chemical technique, may make protein overproduction accessible to those organic chemists who were previously prohibited by lack of training in molecular biology.[24]

APPLICATION OF ECPCR TO THE DOMAINAL ANALYSIS OF CD4. CD4 is a membrane-bound glycoprotein, expressed on the surface of helper T cells, that interacts with class II major histocompatibility (MHC) proteins and thereby augments the immune response to foreign antigens. In addition to its normal role in immune function, CD4 is also exploited as a specific cell-surface receptor by the human immunodeficiency virus (HIV).[25] Viral adhesion and hence infectivity is mediated by high-affinity binding of the HIV coat glycoprotein gp120 to the extracellular segment of CD4,[26] which suggests that CD4-based reagents may be able to competitively inhibit viral infectivity. Indeed, non-membrane-bound[27] and hybrid[28] CD4 derivatives have shown promise as antiviral chemotherapeutic agents. Despite the biological importance of CD4, high resolution structural information is not available, nor is detailed information regarding its interaction with gp120 or MHC class II molecules. The cDNA-derived amino acid sequence of CD4[29] suggests that it is composed of four extracellular domains of the immunoglobulin superfamily, a short membrane-spanning domain, and a single intracellular domain (Figure 6a). Several indirect lines of evidence suggest that that the two extracellular domains outermost from the membrane (domains 1 and 2) mediate interactions with gp120,[27f-h,30] with domain 1 being the primary locus of binding; domains 1 and 2 are also involved in binding MHC class II proteins.[31]

We reasoned that the interactions of CD4 with its receptors would best be understood by direct structural analysis, using truncated versions of CD4 that possess only the putative binding domains. In particular, we wished to use these CD4 fragment proteins to analyze the individual contribution of domains 1 and 2 to the overall CD4-gp120 binding energetics; this approach we have termed domainal analysis. Furthermore, we aimed to to generate quantities of these CD4 domain fragments sufficient to carry out X-ray crystallographic and NMR analyses. These goals could best be accomplished by the construction of *E. coli* strains engineered to overproduce CD4 domains 1, 2, and 1 + 2. It was the daunting prospect of investing ~1 yr in the construction of the requisite three overproducers using traditional methods that sparked the development of ECPCR, which enabled us to achieve overproduction and purification of the CD4 domain fragments[4] in a few months' time.

Shown in Figure 6b is a schematic structure of the CD4 cDNA.. The sequence encoding domains 1 and 2 is flanked on the upstream end by an N-terminal signal peptide[32] and on the downstream end by domain 3. Below the cDNA in Figure 6b. the strategy for ECPCR synthesis of CD4 domainal expression cassettes and their structures are presented. ECPCR primers were designed to copy the CD4 cDNA beginning at the

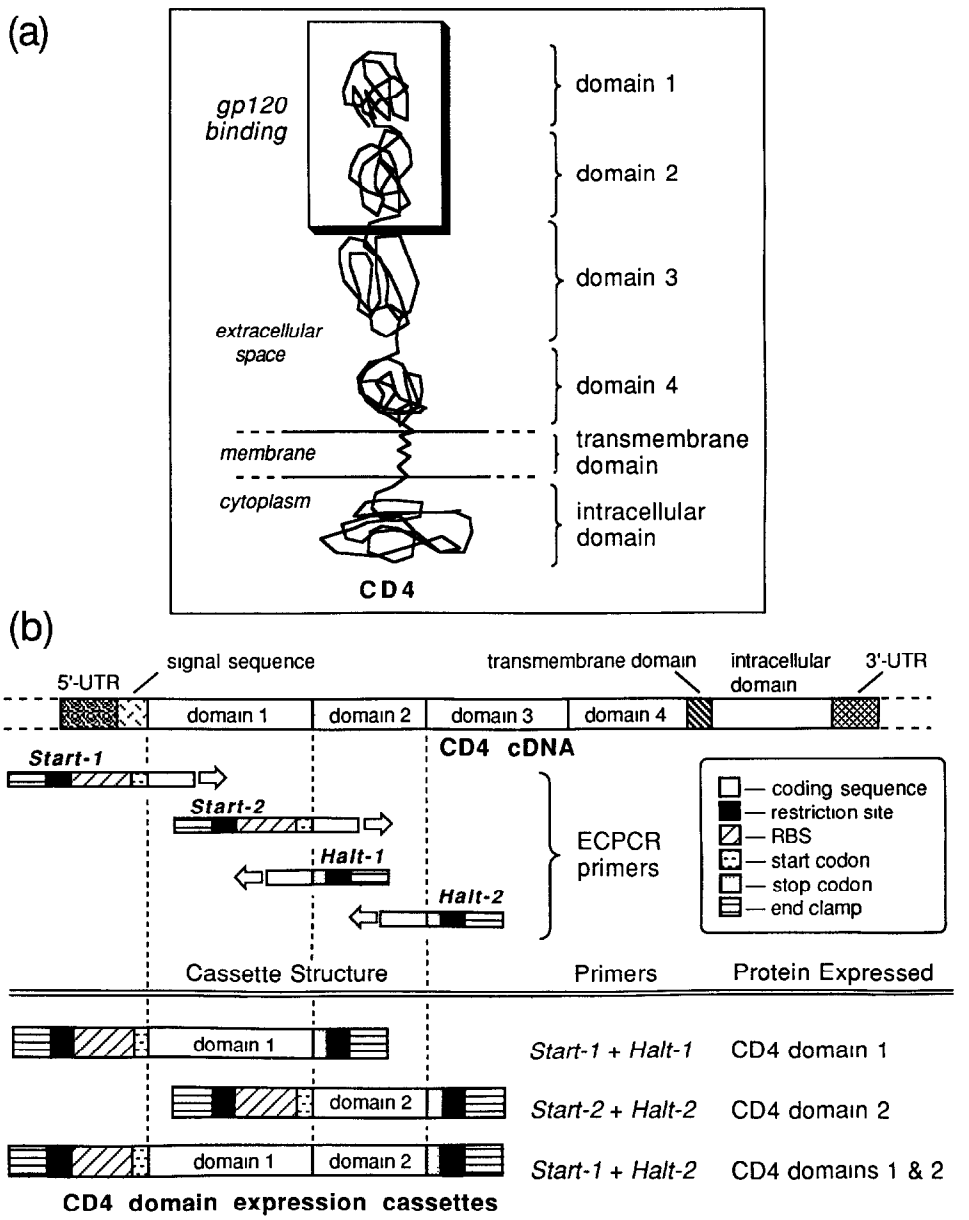

Figure 6 (a) Schematic structure of CD4, with relevant features denoted. (b) Top: structure of the CD4 cDNA; middle: ECPCR primers, with coding sequences aligned to correspond with identical sequences in cDNA; bottom: expression cassettes generated by ECPCR, primer combinations used to synthesize the various cassettes, and CD4 domain fragment encoded by the respective cassettes Drawings are not necessarily to scale

N-terminus of domains 1 or 2 (*Start* -1 and -2, respectively) and ending at the C-terminus of domains 1 or 2 (*Halt*-1 and -2, respectively). The primer combination of *Start*-1 and *Halt*-1 should thus amplify only CD4 sequence encoding domain 1, and the resulting expression cassette should yield an overproducer for CD4 domain 1; likewise the combination *Start*-2/*Halt*-2 was predicted to furnish domain 2 and *Start*-1/*Halt*-2 was predicted to furnish domain 1 + 2. Agarose gel electrophoretic analysis of the amplification reactions clearly demonstrated the synthesis of DNA having in each case the predicted molecular sizes (Figure 7A). In addition to using a pure cDNA clone as the CD4 coding sequence donor, we also demonstrated that double-stranded human cDNA libraries (Figure 7B) and even crude mouse single-stranded cDNA libraries[33] (Figure 7C) can be used as starting material.

The CD4 expression cassettes were cut and pasted into the expression vector pHN1+ and fully sequenced to confirm their predicted composition. The vector pHN1+ contains a switchable promoter that is ordinarily turned off in the cell but can be chemically induced (switched on) by the addition of isopropyl β-D-thiogalactopyranoside (IPTG) to the culture medium. Overproducing a protein imposes a severe metabolic burden on a bacterium, so it is important to be able to switch overproduction on and off at will. Ordinarily, the cells are grown in the "off" state until they reach a moderate density of rapidly reproducing cells (as measured by optical density on a UV/Vis spectrophotometer), and then the IPTG is added. After induction, cell division slows dramatically or stops, since so much energy has been diverted to protein overproduction.

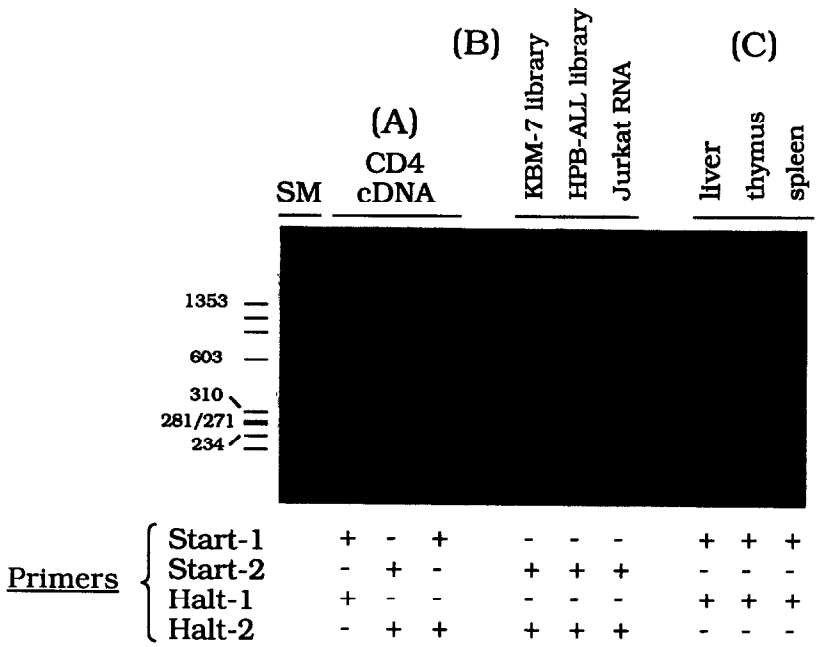

Figure 7. Agarose gel electrophoresis of ECPCR reaction products generated using different primer combinations (below photo) and different sources of the CD4 cDNA (above photo): SM: DNA size standard, marked in base-pairs (bp); (A), CD4 cDNA clone; (B) various human cDNA libaries; (C) various mouse cDNA libraries. The mouse libraries and the one marked "Jurkat RNA" were derived from a reverse transcription reaction of total RNA, and despite the crudeness of these preparations, yielded the desired expression cassettes.

Promoter switching also enables the analysis of proteins synthesized by induced and uninduced cells, and by comparison to observe induction-dependent synthesis of the desired protein. Shown in Figure 8 is a sodium dodecyl sulfate-polyacrylamide gel electrophoresis (SDS-PAGE) analysis of such an *induction test* for the CD4 domain fragments: CD4-d1 (domain 1 alone), CD4-d2 (domain 2 alone), and CD4-d12 (domains 1 and 2 together). Comparison of the leftmost lane in each panel (U: uninduced cells) with its nearest rightward neighbor (I: induced) reveals the presence of a new protein in the induced cells that is absent in uninduced cells, thus verifying induction-dependent synthesis of proteins having approximately the predicted size. Since the CD4 fragments do not undergo efficient folding inside *E. coli*, they form insoluble aggregates called "inclusion bodies." This insolubility can be advantageous: it permits the separation of the recombinant proteins from the majority of native *E. coli* proteins by mere centrifugation of lysed induced cells [Figure 8: compare Pt (pellet) lane with I lane]. Following purification by cation-exchange chromatography [Figure 8, Pu (pure) lane], the domain fragments were refolded *in vitro* in high yield (~80%) to generate proteins having a correctly folded structure.[34] Preliminary results of our domainal analysis studies have demonstrated that CD4-d1 and CD4-d12 are potent, competitive inhibitors of CD4-gp120 interactions;[34] further studies on the domainal analysis of CD4-gp120 and CD4-MHC class II molecules are in progress.

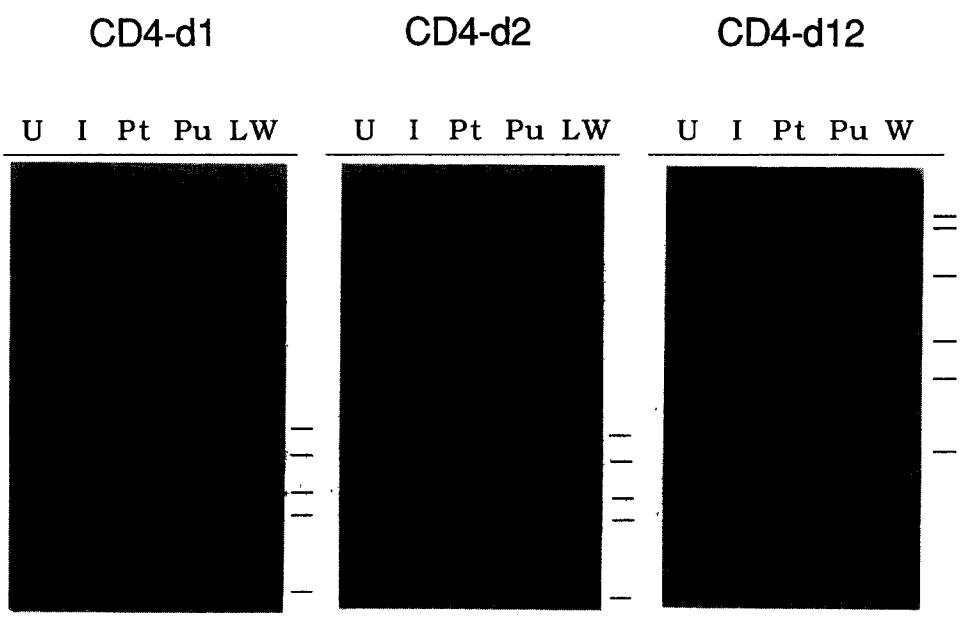

Figure 8. SDS-PAGE analysis of CD4 domain fragment overproduction and purification. U: uninduced whole cells; I, induced whole cells; Pt, lysed cell pellet; Pu, purified protein; LW, low-molecular-weight protein standards (from top: 16.9, 14.4, 10 7, 8 2, 6 2 kDa); W, high-molecular-weight protein standards (97.4, 66.2, 42.7, 31.0, 21.5, 14.4 kDa). The molecular weights of the CD4 fragments are as follows: CD4-d1, 12.2; CD4-d2, 8.3, CD4-d12, 19.7 kDa. CD4-d2 has a characteristic tendency to smear during electrophoresis.

USE OF ECPCR TO ADD NEW AMINO ACID RESIDUES TO THE ENDS OF A PROTEIN. CD4-d1, -d2, and -d12 all possess an additional methionine (Met) residue at their N-termini, encoded by the start (Met) codon, which was added by necessity. Other non-native coding sequences can be added to the primers, and this provides an avenue for the addition or replacement of amino acid residues to the overproduced protein, which can serve as useful tags or handles. In one such example, we have added a codon specifying a cysteine (Cys, C) between the native coding sequence and the stop codon in the *Halt*-2 primer to yield the primer *Halt*-2 Cys (Figure 9).[34] ECPCR of a CD4 cDNA using the primers *Halt*-2 Cys and *Start*-1 generated an expression cassette that encodes CD4 domains 1 and 2, with a new C-terminal Cys residue (CD4-d12C). The corresponding overproducing *E. coli* strain produced high levels of the CD4-d12C protein, which was purified and refolded in vitro. The refolded protein, a competitive inhibitor of CD4-gp120 interactions, has formed the basis for the construction of semisynthetic CD4 variants that may be capable of carrying out targeted modification of the surface of HIV.

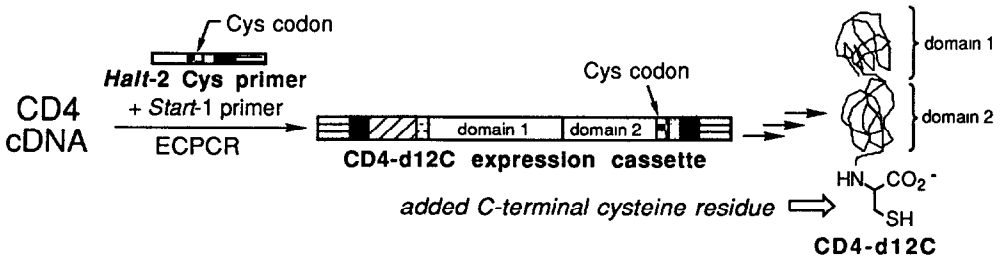

Figure 9 Use of ECPCR to add a C-terminal cysteine residue (checkerboard) to CD4 domain 1. See Figure 5 for legend of sequence elements.

OVERPRODUCTION OF THE HUMAN FK506-BINDING PROTEIN FKBP. The examples of protein overproduction described thus far involve dissection of a larger protein; however, ECPCR can also be used to overproduce entire native proteins, as exemplified in our overproduction of the human FK506-binding protein, FKBP. FKBP is the predominant receptor for the macrolide FK506,[35,36] a potent immunosuppressive drug recently shown to be highly effective in preventing organ transplant rejection in humans.[37] FKBP has a remarkable enzymatic activity: it catalyzes the interconversion of *cis* and *trans* rotamers about the prolyl-peptide amide bond[35,36] and has thus been termed a *rotamase* enzyme. FK506 is a sub-nanomolar, stoichiometric inhibitor of the FKBP rotamase activity; nonetheless, recent studies suggest that this rotamase inhibition is not *per se* the determining factor in the immunosuppressive activity of FK506.[38] In order to further understand the biological activity of FKBP and the FK506-FKBP complex at the molecular level, we recently undertook the isolation and cloning of a human FKBP cDNA.[39] The 108 amino acid protein encoded by the cDNA is processed *iv vivo* by cleavage of the N-terminal Met residue to yield the 11.8 kDa mature FKBP.

An expression cassette for FKBP (Figure 10) was generated in a manner analogous to the CD4 constructs, with two differences: (i) the start codons on the ECPCR *start* primer was already present in the cDNA, hence no amino acids were added during overproduction; and (ii) we decided to use as the *halt* primer one composed completely of native sequence information from the beginning of the 3'-UTR, with no restriction site and end-clamp. The downstream end of the expression cassette was therefore not digested before insertion

into pHN1+, but rather was inserted as a blunt end. Because of the particular sequences present at the ends of the expression cassette and vector, a unique *Hin*D III site was generated upon insertion of the cassette into the vector.[40]

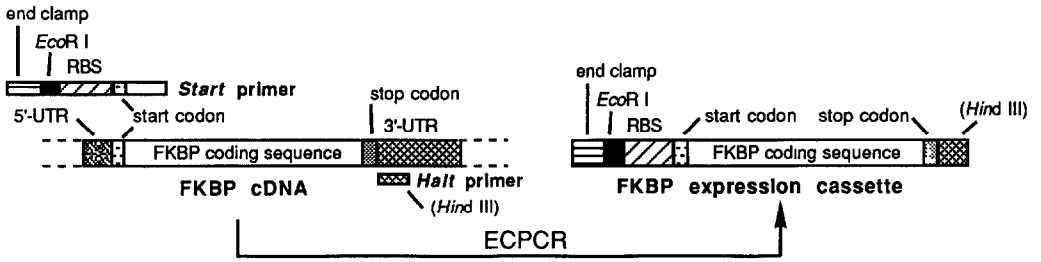

Figure 10. Construction of an FKBP expression cassette by ECPCR. Parentheses denote a partial *Hin*D III site (see text).

The human FKBP overproducer directs the synthesis of FKBP in *E. coli* at levels comprising ~5-10% of the total cellular protein. Since the N-terminal Met is excised in *E. coli*(as in human cells), the FKBP isolated from human and bacterial sources are identical. The ready availability of this protein in large quantities (~10 mg/L induced cells) has greatly facilitated structural and enzymological studies. Furthermore, the ability to generate isotopically labeled protein (^{15}N, ^{2}H) by growth of the overproducing strain on labeled nutrients has greatly facilitated the structural analysis of FKBP and FKBP/ligand complexes.[41]

OVERPRODUCER CONSTRUCTION USING "ATG" VECTORS. ECPCR relies on the use of expression vectors having a promoter and transcription terminator flanking unique restriciton sites. Another configuration of expression vector has also been used: one having all of the expression elements, including the RBS and start codon (5'-ATG-3'); such plasmids are accordingly termed "ATG" vectors (Figure 11). The fundamental problem of overproducer construction using an ATG vector[42] is somewhat different from that using a non-ATG vector (such as pHN1+), which has a promoter and transcription terminator but neither RBS nor start codon. In bacterial control regions, the spacing between the promoter and the RBS does not have very precise sequence or distance requirements,[5,6] hence in constructions using non-ATG vectors there is a great deal of flexibility in the sequence added upstream to the Shine-Dalgarno element (for example, one has a large choice of restriction sites to install at the upstream end). In contrast, the relationship of the start codon to the rest of the coding sequence is precisely controlled (three nucleotides comprise a codon, and the register in which the triplets are read is controlled by the start codon). Since ATG vectors already contain a start codon, the insertion of the coding sequence to the vector's start codon must be carefully controlled so as to maintain the proper register of the coding sequence (reading frame) and avoid the unnecessary addition of amino acids to the protein. Fortunately, two restriction enzymes with six-base recognition sites (which therefore cut infrequently) contain the ATG subsequence: *Nde* I, CATATG; *Nco* I, CCATGG. In order to construct the overproducer, either of these two restriction sites must be engineered into the beginning of the coding sequence of interest, with the start codon (ATG) in register. The *Nde* I or *Nco* I site in the coding sequence can then be fused precisely to the site in the

vector to generate a functional transcription-translation system. The *Nde* I or *Nco* I site, as well as another site downstream (*Bam*H I in Figure 11) can be introduced either by classical methods, or more expediently, by PCR.

Although ATG vectors have been used quite successfully in overproducer constructions, we have avoided their use for several reasons:

1. *Inability to control the sequence of the RBS.* Considerable evidence indicates that there is no universally optimal RBS.[7] In particular, the length and sequence of the translational spacer element appear to vary from one highly expressed gene to another. Since the coding information in the beginning of the gene varies from gene to gene, and this region can pair with the RBS to form translation-inhibiting structures, it is not surprising that the optimal RBS should vary from gene to gene. The most productive ATG vectors use an efficient RBS derived from the phage T7 gene 10 regulatory region,[42b] and such constructs often but not always yield high-level overproducers (S.L.S. and G.L.V., unpublished observations). Without the ability to permute the RBS, there is no straightforward way to increase expression.

2. *Limited choice of restriction sites.* If the coding sequence of interest contains an internal *Nco* I or *Nde* I restriction site, in addition to the start codon site, then one must go to great pains to cut selectively at the upstream site in the presence of the internal site. Restriction sites located near an end of a DNA molecule are cut much more slowly than those with long flanking sequences, so this can add another degree of difficulty when the upstream site has been generated by PCR. If the coding sequence contains an internal *Nco* I site, then the

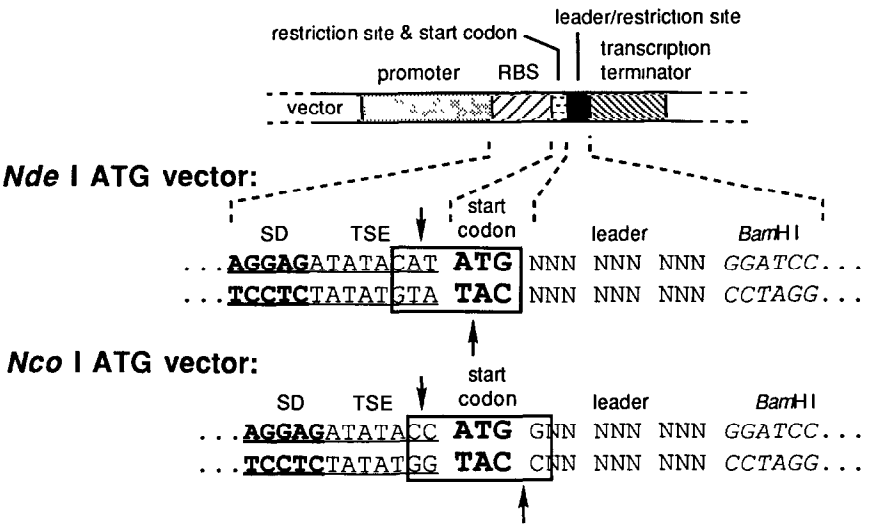

Figure 11. The control region of a typical ATG vector The RBS and start codon are already present in the vector, followed by a short leader sequence and a restriction site. The start codon is present within a restriction site, either *Nde* I or *Nco* I. Arrows denote internucleotide phosphodiester bonds cleaved by the restriction enzyme. Although the leader is shown in triplets, its precise sequence is unimportant, because it is removed during insertion of the coding sequence into the vector. N denotes any nucleotide; TSE denotes the translational spacer element.

problem can often be overcome by using an upstream *Nde* I site, and vice versa. Alternatively, it is possible to use another enzyme that leaves the same type of DNA end as *Nde* I or *Nco* I. The sites of phosphodiester cleavage by *Nco* I and *Nde* I are denoted by arrows in Figure 11. After *Nde* I cuts DNA, it leaves each product DNA strand with a 5'-TA overhang, which could be joined to any other DNA molecule having the same overhang. A survey of suppliers' catalogs reveals *Nde* I is the only such enzyme commercially available at present. It is generally true that restriction sites with two-base overhangs are difficult to ligate (paste), and this can be particularly troublesome for the beginner. While the four-base 5'-CATG overhang left *Nco* I can easily be ligated, the sequence of the *Nco* I site (5'-CCATGG) presents another problem: the ATG sub-sequence is followed by a G, and this G is the first base of the second codon. Thus, if one is limited to G at that position, only the amino acids valine, alanine, aspartic acid, glutamic acid, and glycine can be encoded. Another enzyme, *Afl* III, cleaves at the 5'-TCATGA sequence to leave a 5'-CATG overhang, and that site places an A in the second codon. An ATG vector constructed to possess an *Afl* III site could encode isoleucine, asparagine, lysine, serine, and arginine at the penultimate N-terminal position. Using ATG vectors containing either *Nco* I or *Afl* III, then, only ten of the twenty amino acids may be encoded at the second position. The use of *Nco* I (or *Afl* III) has one additional drawback: cytosines and guanine residues immediately preceeding the start codon have a deleterious effect on expression.[7]

Since ECPCR allows one install virtually any desired RBS, it can be used to produce precisely the same constructs as those created using ATG vectors. Significantly however, ECPCR can also be used to install RBS/start sequences different from the immutable ones found in ATG vectors, and thus it offers the valuable option of tuning expression. This flexibility of ECPCR lends itself to further developments of the method, some of which are described below.

ECPCR: FUTURE DIRECTIONS.

Selectable ECPCR.[43] As mentioned above, there is no universally optimal RBS; furthermore, simple rules for predicting which RBS will work best in conjunction with a given coding sequence have not yet emerged. Hence, there is an element of caprice in every overproducer construction, and the only strategy currently available to ensure high-level overexpression is to construct a number of overproducers having different RBS elements. Although ECPCR allows one to readily construct several overproducers, this approach is still sub-optimal in that a relatively small number of all possible RBS sequences can be screened. We have begun working on an alternative ECPCR system that allows the simultaneous screening of all possible RBS sequences. This system, *selectable ECPCR*, is based on the synthesis of *start* primers having highly degenerate RBS sequences, which are generated by incorporating mixtures of nucleotides during automated synthesis. ECPCR using this mixture of *start* primers generates a library of overproducers in which each clone has a different RBS. We have devised a system for rapidly selecting those overproducer clones that express the most protein. Selectable ECPCR should provide a general strategy for achieving maximal overproduction in each construction exercise.

ECPCR for preplasmic expression of proteins [43] Many proteins whose native environment is extracellular do not undergo correct folding in *E. coli*, but rather form insoluble aggregates of inactive protein. As noted above in the case of the CD4 fragments, this can sometimes be advantageous, because it can simplify

the purification. However, in order to obtain active protein, the purified, denatured protein must be refolded *in vitro*, and finding the proper conditions for this refolding operation can sometimes take months.

It has been demonstrated that many proteins that do not undergo cytoplasmic folding will fold properly if they are exported (secreted) to the periplasmic space (space between outer and inner cell wall) of the bacterium.[44] Found in the coding sequence of native *E. coli* proteins that are secreted into the periplasm is a characteristic N-terminal peptide motif termed a signal peptide or secretion signal, which is usually cleaved from the protein during its translocation across the membrane.[45] Periplasmic secretion of a protein can usually be engineered by addition of a signal peptide sequence to the N-terminus of the coding sequence. Recently, we have succeeded in adding a signal sequence to the N-terminal domain of ICAM-1,[46] a receptor for human rhinovirus[47] and the malaria parasite *Plasmodium falciparum*,[48] during overproduction by ECPCR.[43] The system, which we call "leapfrog ECPCR," promises to be valuable in structural studies on cell-surface-bound receptor molecules and soluble secreted proteins such as lymphokines and antibodies.

SUMMARY. The Expression-Cassette Polymerase Chain Reaction offers a new and rapid route to overproduced proteins in *E. coli*. This methodology can readily be brought into routine practice in organic chemistry laboratories and should thus facilitate the entry of chemists into studies involving native and redesigned proteins.

NOTE ADDED IN PROOF. The X-ray crystal structure of CD4 domains 1 and 2 was recently reported. The structure reveals significant interdomain interactions in the two immunoglobulin-like domains, as well as an atypical disulfide linkage in domain 2.[49]

ACKNOWLEDGEMENTS. We are indebted to our co-workers Dean Farmer, Huw Nash, Kurtis MacFerrin, Bob Standaert, and Mike Terranova for their contributions to this program. This research was supported by a grant from the National Institute of General Medical Sciences (S.L.S.), the Camille and Henry Dreyfus Foundation, Searle Scholars Program, Bristol-Myers Squibb, ICI Pharmaceuticals, and Hofmann-La Roche (G.L.V.).

EXPERIMENTAL PROCEDURES. Oligonucleotides were synthesized on an Applied Biosystems Model 381B automated synthesizer, using reagents purchased from the manufacturer. The PCR reactions shown in Figure 7 were carried out as described,[30] using three aluminum-block water baths (American Scientific Products Cat. #H2025-1A). All other PCR reactions were carried out using a programmable thermal cycler (MJ Research, Cambridge, MA, USA) as described.[38]

REFERENCES AND NOTES.

1. Saiki, R.K.; Scharf, S.; Faloona, F.; Mullis, K B.; Horn, G.T.; Erlich, H.A.; Arnheim, N. *Science* **1985**, *230*, 1350-1354.
2. For an entry-level discussion on PCR, see: Mullis, K.B. *Scientific American* **1990**, *262*, 56-65; Marx, J.L. *Science* **1988**, *240*, 1408-1410; Thornton, J.J. *Chem. Eng. News* **1989**, *67*, 18-30.
3. For a detailed discussion on principles and applications of PCR, see: Erlich, H A., Ed. *PCR Technology· Principles and Applications for DNA Amplification*; Stockton Press: New York, 1989.
4. MacFerrin, K.D.; Terranova, M.P.; Schreiber, S.L.; Verdine, G.L. *Proc Natl. Acad. Sci. USA* **1989**, *87*, 1937-1941.
5. For a comprehensive source of information on this bacterium, see: Neidhardt, F.C., Ed. *Escherichia coli and Salmonella typhimurium· Cellular and Molecular Biology*; American Society for Microbiology: Washington, 1987.
6. For a review on promoters and transcriptional control, see: McClure, W.R. *Annu. Rev Biochem.* **1985**, *54*, 171-204.
7. For a review on RBS's and translational control, see: Gold, L. *Annu Rev Biochem* **1988**, *57*, 199-233.
8 Palindromic RNA (and single-stranded DNA) sequences can form a double-stranded structure in which the single nucleic acid strand pairs with itself. The paired segments (*stem*) in such a structure are connected to each other by a *loop* of unpaired nucleotides, hence this motif is called a *stem-loop structure*. The base-pairing in stem-loop structures can be quite thermodynamically stable, and this property is responsible for their ability to abort transcription.
9. Prokaryotic cells lack a nucleus, as opposed to eukaryotic cells, which have a nucleus. This difference in cellular compartmentalization is accompanied by vast differences in gene expression.
10 Some lower eukaryotes (yeast, for example) have few intron-containing genes, while most genes from higher eukaryotes possess introns; we know no example of an intron-containing bacterial gene.
11. RNA splicing has received a great deal of attention in recent years. For a recent review on protein-mediated splicing, see: Padgett, R.A.; Grabowski, P.J.; Konarska, M.M.; Seiler, S.; Sharp, P.A. *Annu Rev. Biochem.* **1986**, *55*, 1119-1150. For a review on self-splicing, see: Cech, T.R. *Science* **1987**, *236*, 1532-1539.
12 Since the copying of mRNA into DNA can be viewed as the reverse of transcription, the enzyme that carries out this reaction has been dubbed *reverse transcriptase*. See ref. 12 for a discussion.
13. For an understandable, entry-level discussion of gene cloning and molecular biology, see: Watson, J.D.; Tooze, J ; Kurtz, D.T. *Recombinant DNA A Short Course*; Scientific American Books: New York, 1983.
14. To our knowledge, Ptashne was the first to generate an overproducer by replacing the upstream control elements of a gene. Backman, K.; Ptashne, M. *Cell* **1978**, *13*, 65-71. Gauarente, L ; Roberts, T.M.; Ptashne, M. *Science* **1980**, *209*, 1428-1430.
15. Plasmids are generally named beginning with a small "p" (plasmid) followed by a letter/number code (as in pHN1+ and pBR322). The letters often correspond to the initials of the person who constructed the plasmid (pHN1+ was constructed by Huw Nash, presently in the Department of Chemistry at the California Institute of Technology).
16. Bacterial genes are most often denoted by a three- or four-letter code in small, italicized letter, as in *bla* for β-lactamase and *sod* for superoxide dismutase gene. The *bla* gene is sometimes denoted on plasmid maps by the phenotypic designation Amp^r, although this is technically incorrect.
17. DeBoer, H.A.; Comstock, L J ; Vasser, M. *Proc Natl Acad Sci USA* **1983**, *80*, 21-25. Amman, E.; Brosius, J.; Ptashne, M. *Gene* **1983**, *25*, 167-178
18. Brosius, J ; Dull, T.J ; Sleeter, D.D., Noller, H F *J Mol Biol* **1981**, *148*, 107-127.

19. It should be noted that the choice of restriction sites is limited to either those that occur in the polylinker of the expression vector, or those that leave the same DNA ends as polylinker enzymes.
20. This is classically true but has changed recently: using PCR, genes (or cDNAs) encoding less than 1000 amino acids may be re-cloned in days. As discussed below, ECPCR allows one to take advantage of this rapid cloning asset of PCR.
21. In order to fully visualize the mechanics of PCR, one should work through several cycles on paper. Good graphic representations of this process can also be found in ref. 2.
22. This figure includes the time required to synthesize and purify the primers. The PCR reaction requires only ~2 h.
23 Alternatively, programmable thermal cycling devices capable of automating the PCR reaction are available for a modest price. A number of additional pieces of equipment are required in any overproduction exercise, regardless of the method used: gel electrophoresis apparati (which can be home-made), a cell disruptor, and equipment for chromatographic purification of proteins.
24. The rate of nucleotide misincorporation by *Taq* polymerase can be as high as 1/5000 (a typical gene has a size of ~1000 bp). In addition to such point mutations, a number of other minor artifacts can occur during PCR. It is therefore necessary to sequence the ECPCR-derived insert following an overproducer construction, to ensure its fidelity. This situation is not unique to overproducer constructions using ECPCR; the error during replication of bacteriophage M13 is also high, necessitating complete sequencing of most traditional constructs Numerous efforts are underway to discover thermostable DNA polymerases that are higher in fidelity than *Taq* polymerase
25. For reviews, see: Littmann, D.R. *Annu. Rev Immunol.* **1987**, *5*, 561-584; Robey, E.; Axel, R. *Cell* **1990**, *60*, 697-700.
26 For a review, see Sattentau, Q.J.; Weiss, R.A. *Cell* **1988**, *52*, 631-633.
27. (a) Smith, D.H., Bryn, R.A.; Marsters, S.A ; Gregory, T., Groopman, J E.; Capon, D.J. *Science* **1987**, *238*, 1704-1707. (b) Fischer, R.A.; Bertonis, J.M.; Meier, W.; Johnson, V.A.; Costopoulos, D.S., Liu, T.; Tizard, R.; Walker, B.D.; Hirsch, M.S., Schooley, R.T.; Flavell, R.A *Nature (London)* **1988**, *331*, 76-78. (c) Hussel, R.E.; Richardson, N.E.; Kowalski, M.; Brown, N.R.; Chang, H -C., Sicilliano, R.F.; Dorfman, T.; Walker, B.; Sodroski, J.; Reinhertz, E L. *Nature (London)* **1988**, *331*, 78-81. (d) Deen, K.C ; McDougal, J S.; Inacker, R.; Folena-Wasserman, G.; Arthos, J ; Rosenberg, J ; Maddon, P.J.; Axel, R *Nature (London)* **1988**, *331*, 82-84. (e) Traunecker, A , Luke, W.; Karjalainen, K. *Nature (London)* **1988**, *331*, 84-86. (f) Berger, E.A.; Fuerst, T.R.; Moss, B *Proc Natl Acad Sci USA* **1988**, *85*, 2357-2361. (g) Chao, B.H.; Costtopoulos, D.S., Curiel, T.; Bertonis, J.M ; Chisholm, P.; Williams, C ; Scholley, R.T.; Rosa, J. J., Fisher, R.A.; Maraganore, J.M. *J Biol. Chem* **1989**, *264*, 5812-5817. (h) Arthos, J.; Deen, K.C.; Chaikin, M.A , Fornwald, J.A.; Sathe, G.; Sattentau, Q.J.; Clapham, P.R.; Weiss, R.A.; MsDougal, J S.; Pietropaolo, C.; Axel, R ; Truneh, A ; Maddon, J.P , Sweet, R. *Cell* **1989**, *57*, 469-481.
28. Choudhary, V.K.; Mizukami, T.; Fuerst, T R.; FitzGerald, D.J.; Moss, B.; Pastan, I.; Berger, E.A. *Nature (London)* **1988**, *335*, 369-372. Capon, D.J ; Chamow, S.M.; Mordenti, J.; Marsters, S.A., Gregory, T , Mitsuya, H.; Bryn, R.A.; Lucas, C.; Wurm, F.M.; Groopman, J.E.; Broder, S.; Smith, D.H. *Nature (London)* **1989**, *337*, 525-531. Traunecker, A.; Schneider, J.; Kiefer, H.; Karjalainen, K. *Nature (London)* **1989**, *339*, 68-70.
29. Littman, D.R ; Maddon, P.J.; Axel, R. *Cell* **1988**, *55*, 541, and references cited therein.
30. Peterson, B.; Seed, B. *Cell* **1988**, *54*, 65-72. Landau, N R.; Wharton, M ; Littman, D.R. *Nature (London)* **1988**, *334*, 159-162. Richardson, N.E.; Brown, N R.; Hussey, R.E , Vaid, A.; Matthews, T.J.; Bolognesi, D.P.; Reinherz, E L. *Proc Natl. Acad Sci USA* **1988**, *85*, 6102-6106 Clayton, L.K.; Husey, R.E ; Steinbrich, R ; Ramachandran, H.; Husain, Y.; Reinherz, E.L. *Nature (London)* **1988**, *335*, 363-366.

31. Lamarre, D.; Capon, D J.; Karp, D.R.; Gregory, T.; Long, E.O.; Sekaly, R.-P. *EMBO J* **1989**, *8*, 3271-3277. Lamarre, D.; Ashkınazı, A., Fleury, S.; Smith, D.H.; Sekaly, R -P.; Capon, D J. *Science* **1989**, *245*, 743-746.

32. Many codıng sequences, both prokaryotic and eukaryotic, encode an N-terminal signal peptide, which can generally be ıdentified by comparisons wıth known signal sequences A sıgnal peptıde directs membrane translocatıon (secretıon) of a nascent polypeptıde, and is usually proteolytıcally cleaved after the translocatıon event.

33. The CD4 domains generated by ECPCR of the mouse lıbrary are actually mouse/human chimeras, wıth human sequence from the prımers, and mouse sequence ın between.

34. Terranova, M.P.; Farmer, DJ.; Schreiber, S.L.; Verdıne, G.L.; unpublıshed results.

35. Harding, M.W.; Galat, A.; Uehling, D.E.; Schreıber, S L. *Nature (London)* **1989**, *341*, 758-760.

36. Sıekierka, J J.; Hung, S.H.Y.; Poe, M.; Lın, C S ; Sıgal, N H. *Nature (London)* **1989**, *341*, 755-757.

37. Starzl, T.E.; Fung, J.; Ventkataramman, R.; Todo, S.; Demetrıs, A.J.; Jaın, A. *The Lancet* **1989**, 1000-1004.

38. Bierer, B.E.; Mattıla, P.S.; Standaert, R.F.; Herzenberg, L.A.; Burakoff, S.J.; Schreıber, S.L. *Proc. Natl Acad. Scı. USA* **1990**, ın press. Bıerer, B.E.; Somers, P.K.; Wandless, T.J.; Burakoff, S.J.; Schreıber, S L. *Science* **1990**, ın press. Schreiber, S.L. *Scıence* **1990**, ın press.

39. Standaert, R.F.; Galat, A.; Verdıne, G.L.; Schreıber, S.L *Nature (London)* **1990**, *346*, 671-674. See also: Makı, N.; Sekıguchı, F.; Nishımakı, J.; Mıwa, K.; Hayano, T.; Takahashi, N.; Suzukı, M. *Proc Natl Acad. Scı. USA* **1990**, *87*, 5440-5443.

40. The expressıon cassette ends ın an A resıdue, and the vector (after beıng cut by *Hın*D III and havıng the 5'-overhangs fılled in wıth dNTPs ın the presence of DNA polymerase I-Klenow) begıns wıth a AGCTT. When the two ends are joıned, the sequence at the fusıon sıte becomes AAGCTT, whıch ıs the recognıtıon sıte for *Hın*D III.

41. Mıchnıck, S.W.; Rosen, M.K.; Karplus, M.; Schreıber, S.L., unpublıshed results.

42 (a) Amman, E ; Brosıus, J. *Gene* **1985**, *40*, 183-190 (b) Studıer, F.W., Moffatt, B.A *J Mol. Bıol* **1986**, *189*, 113-130.

43. MacFerrın, K D., Schreıber, S L ; Verdıne, G.L.; unpublıshed results

44. For revıews, see Schein, C H. *Bıotechnology* **1989**, *7*, 1141-1148; Marston, F A.O. *Bıochem J.* **1986**, *240*, 1-12.

45. For a revıew, see· Gierasch, L M. *Bıochemıstry* **1989**, *28*, 923-930.

46. Dustın, M.L.; Staunton, D.E.; Springer, T.A. *Immunol Today* **1988**, *9*, 213-215.

47. Greve, J.M.; Davıs, G.; Meyer, A.M.; Forte, C.P ; Yost, S.C.; Marlor, C.W ; Kamarck, M.E.; McClelland, A. *Cell* **1989**, *56*, 839-847 Staunton, D E., Merluzzı, V J.; Rothlein, R., Barton, R.; Marlın, S.D.; Springer, T A. *Cell* **1989**, *56*, 849-853. Tomassını, J E.; Graham, D.; DeWitt, C M ; Lıneberger, D.W.; Rodkey, J A ; Colonno, R.J. *Proc Natl Acad Scı USA* **1989**, *86*, 4907-4911

48. Berendt, A.R ; Sımmons, D.L., Tansey, J , Newbold, C.I., Marsh, K. *Nature (London)* **1989**, *341*, 57-59.

49. Wang, J ; Yan, Y.; Garrett, T.P.J.; Lıu, J.; Rodgers, D.W ; Garlıck, R.L.; Tarr, G.E.; Husaın, Y.; Reınherz, E L.; Harrıson, S.C. *Nature* **1990**,*348* , 411-418. Ryu, S.-E.; Kwong, P.D.; Truneh, A , Porter, T.G., Arthos, J.; Rosenberg, M.; Daı, X.; Xuong, N.-h ; Axel, R.; Sweet, R.; Hendrickson, W.A. *Nature* **1990**, *348*, 419-426.