Total Gene Synthesis: Novel Single-Step and Convergent Strategies Applied to the Construction of a 779 Base Pair Bacteriorhodopsin Gene

Guo-Qiang Chen, Isaac Choi,
Banurekha Ramachandran, and J. Eric Gouaux*

Department of Biochemistry and Molecular Biology
The University of Chicago, 920 East 58th Street
Chicago, Illinois 60637

Received May 27, 1994

The chemical basis for the *de novo* design,[1] engineering,[2] and dissection of structure, function, and thermodynamic relationships of proteins[3] depends on rapid, efficient, and accurate methods to create and alter the amino acid sequence. Synthetic genes are critical for the design of new proteins and provide the basis for powerful mutagenesis strategies.[4] Existing methods for gene synthesis have many drawbacks, which include chemical synthesis of the entire duplex, multiple enzymatic reactions, purification of a large number of oligonucleotides (oligos), and length limitations.[5]

In this paper we describe two general strategies for the rapid synthesis of long genes or gene fragments using multiple overlapping oligos[6] and the polymerase chain reaction (PCR).[7] This general PCR gene synthesis (PGS) approach, one form of which is outlined in Figure 1, requires (i) chemical synthesis of less than 60% of the duplex, (ii) a single enzyme reaction, and (iii) purification of a minimal number of oligos and gives the desired gene as the only major product. We have also evaluated a convergent multistep synthetic scheme and have separately determined optimal conditions for Deep $Vent_R$(exo$^-$) and $Vent_R$ DNA polymerases (New England Biolabs, Inc.).[8]

The target gene encodes bacteriorhodopsin (bR), a polytopic, α-helical membrane protein from *Halobacterium halobium*.[9,10] As part of this laboratory's structural and thermodynamic study of membrane proteins, we designed a novel gene (see Figure 2) optimized for cassette mutagenesis and helix-swapping experiments. The wild-type gene[11] was used as the starting point for the introduction and removal of restriction sites[12] via amino acid-silent changes in the nucleotide sequence.[13] A retrosynthetic strategy[14] was employed in which the sites flanking the helices were introduced first and then additional, intervening sites were added. The resulting synthetic bR gene (sbR) has 33 unique restriction sites, spaced an average of 24 base pairs (bp) apart, which required the substitution of 76 nucleotides.

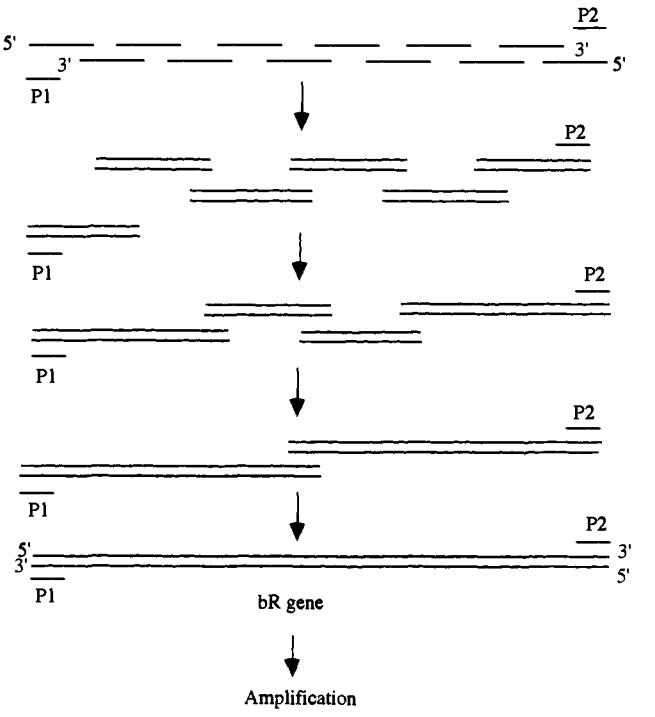

Figure 1. Illustration of one possible mechanism for the PGS process. The complementary ends of the long oligos create short regions of duplex DNA, thereby priming the elongation by DNA polymerase. Products of the elongation reactions serve as substrates for formation of longer duplexes, eventually resulting in the synthesis of full-length material, which is then amplified by the short oligos, P1 and P2.

The long oligos were designed according to the following criteria: formation of unique overlaps, approximately 20 bp in length, minimization of intramolecular base pairing, and the use of oligos between 70 and 100 nucleotides in length. The oligos were chemically synthesized and purified.[15] The lengths of the short oligos were selected to allow annealing at approximately 50 °C.

A systematic exploration of the reaction conditions indicated that the concentrations of long and short oligos and the annealing temperature were critical variables for the PGS reactions. Concentrations of deoxynucleotide triphosphates (dNTPs) and Mg^{2+} were chosen to maximize the fidelity of Taq DNA polymerase.[16] Conditions for the synthesis and amplification of the gene segments I, II, and III using three sets of four oligos (4-part)[17] were more permissive than those required for the entire gene using 12 oligos (12-part).[18] Optimization of the 12-part reactions compared to the 4-part reactions necessitated raising the annealing temperature and decreasing the concentration of both the long and the short oligos. Conditions for Deep $Vent_R$(exo$^-$) and $Vent_R$ DNA polymerases in the 12-part reaction were determined in an effort to increase the yield, purity, and fidelity of the product.[19] Utilization of $Vent_R$ DNA polymerase,

* Author to whom correspondence should be addressed.

(1) For example, see: Kamtekar, S.; Schiffer, J. M.; Xiong, H.; Babik, J. M.; Hecht, M. H. *Science* **1993**, *262*, 1680–1685.

(2) The following is a representative study: Kiefer, L. L.; Ippolito, J. A.; Fierke, C. A.; Christianson, D. W. *J. Am. Chem. Soc.* **1993**, *115*, 12581–12582.

(3) Matthews, B. W. *Annu. Rev. Biochem.* **1993**, *62*, 139–160.

(4) (a) Khorana, H. G. *Science* **1979**, *203*, 614–625. (b) For a recent example, see: Comess, K. M.; Shewchuk, L. M.; Ivanetich, K.; Walsh, C. T. *Biochemistry* **1994**, *33*, 4175–4186.

(5) (a) Engels, J. W.; Uhlmann, E. *Angew. Chem., Int. Ed. Engl.* **1989**, *28*, 716–734. (b) Hein, F.; Jansen, H. W.; Uhlmann, E. *Nucleosides Nucleotides* **1987**, *6*, 489–490. (c) Narang, S. A.; Dubuc, G.; Yao, F. L.; Michniewicz, J. J. *Biochem. Biophys. Res. Commun.* **1986**, *134*, 407–411. (d) Ye, Q.-Z.; Johnson, L.; Baragi, V. *Biochem. Biophys. Res. Commun.* **1992**, *186*, 143–149. (e) Jayaraman, K.; Puccini, C. J. *Biotechniques* **1992**, *12*, 392–398.

(6) Khorana, H. G. *Pure Appl. Chem.* **1968**, *17*, 349–381.

(7) Saiki, R. K.; Scharf, S.; Faloona, F.; Mullis, K. B.; Horn, G. T.; Erlich, H. A.; Arnheim, N. *Science* **1985**, *230*, 1350–1354.

(8) (a) Mattila, P.; Korpela, J.; Tenkanen, T.; Pitkänen, K. *Nucleic Acids Res.* **1991**, *19*, 4967–4973. (b) Kong, H.; Kucera, R. B.; Jack, W. E. *J. Biol. Chem.* **1993**, *268*, 1965–1975.

(9) Khorana, H. G. *J. Biol. Chem.* **1988**, *263*, 7439–7442.

(10) London, E.; Khorana, H. G. *J. Biol. Chem.* **1982**, *257*, 7003–7011.

(11) Dunn, R.; McCoy, J.; Simsek, M.; Majumdar, A.; Chang, S. H.; RajBhandary, U. L.; Khorana, H. G. *Proc. Natl. Acad. Sci. U.S.A.* **1981**, *78*, 6744–6748.

(12) The criteria for choosing the restriction enzymes were (i) production of sticky ends with 4 bp overhangs, (ii) effective recutting following overdigestion and ligation, (iii) minimal star activity, (iv) 6 bp recognition sites, (v) low cost, and (vi) high purity.

(13) Genetics Computer Group. *Program Manual for the GCG Package*, Version 7; April 1991; 575 Science Dr., Madison, WI 53711.

(14) Corey, E. J.; Wipke, W. T. *Science* **1969**, *166*, 178–192.

(15) The concentrations of the oligos were determined by absorbance measurements at 260 nm.

(16) (a) Tindall, K. R.; Kunkel, T. A. *Biochemistry* **1988**, *27*, 6008–6013. (b) Eckert, K. A.; Kunkel, T. A. *Nucleic Acids Res.* **1990**, *18*, 3739–3744.

(17) In general, for the 4-part reactions, 39.5 μL of distilled water was combined with 10 μL of 10× buffer (100 mM Tris-HCl, pH 8.3, 500 mM KCl, 15 mM $MgCl_2$, and 0.01% gelatin), 2 μL of a 10 mM solution of each of the dNTPs, 0.5 μL of Taq polymerase (5 units/μL), 1 μL of each short oligo (100 μM stock solution), and 10 μL of each long oligo (0.1 μM stock solution).

0002-7863/94/1516-8799$04.50/0

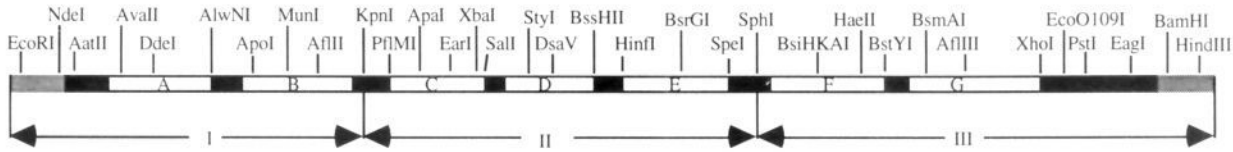

Figure 2. Illustration of the restriction sites, the transmembrane helical boundaries (A–G, unfilled regions), the intervening loops (filled regions), and the 5′ and 3′ flanking sequences (shaded regions) of the sbR gene. I, II, and III correspond to the sections synthesized separately using the 4-part PGS strategy.

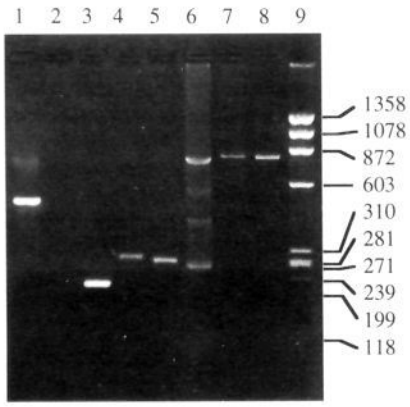

Figure 3. Analytical agarose gel (2%) of the 4-part and 12-part PCR results. Lanes 1 and 2 are the PCR positive and negative controls, respectively.[18] Lanes 3, 4, and 5 are the 4-part PGS reactions for fragments I, II, and III, respectively. The reagent concentrations were 1 μM short oligo, 0.01 μM long oligo, 200 μM dNTP, 1500 μM Mg^{2+}, and 2.5 U Taq DNA polymerase. The thermocycle program was 94 °C (1 min)/50 °C (2 min)/72 °C (3 min) for 25 cycles. Lanes 6, 7, and 8 are the 12-part PGS reactions for the full length sbR gene using Taq , Deep $Vent_R$(exo$^-$), and $Vent_R$ DNA polymerases, respectively. The reagent concentrations and the thermocycling programs are as follows: lane 6, 1 μM short oligo, 0.005 μM long oligo, 200 μM dNTP, 1500 μM Mg^{2+}, 2.5 U Taq, 94 °C (1 min)/50 °C (2 min)/72 °C (3 min), 31 cycles (this reaction mixture was concentrated 9-fold prior to loading on the gel);[19] lane 7, 0.1 μM short oligo, 0.005 μM long oligo, 200 μM dNTP, 2000 μM Mg^{2+}, 2 units Deep $Vent_R$(exo$^-$), 94 °C (1 min)/56 °C (2 min)/72 °C (3 min), 31 cycles; lane 8, 1 μM short oligo, 0.005 μM long oligo, 400 μM dNTP, 2000 μM Mg^{2+}, 2 units $Vent_R$, 95 °C (30 s)/56 °C (30 s)/72 °C (1 min), 31 cycles. Lane 9: Molecular weight markers, in bp.

which possesses 3′→5′ proofreading activity,[8] required an increase in the dNTP and short oligo concentrations and a decrease in the times for denaturation, annealing, and extension. Analysis of the crude 4-part and 12-part reactions is shown in Figure 3.

The desired 12-part PGS product was digested with either *Nde*I/*Bam*HI, *Nde*I/*Hind*III, or *Eco*RI/*Bam*HI and cloned into pUC19.[20] DNA sequencing[21] of both strands from seven clones (pGQ123B, I-IV) derived from the optimized conditions using Deep $Vent_R$(exo$^-$) DNA polymerase showed an error frequency of 0.24% (0.15% substitution, 0.07% deletion, and 0.02% insertion). After sequencing both strands of four independent clones obtained from a $Vent_R$ PGS reaction, we found a similar error frequency of 0.22% (0.13% substitution and 0.09% deletion). The correct, full-length sbR gene was obtained by excising the *Sty*I/*Bam*HI fragment from clone pGQ123BI and the *Nde*I/*Sty*I section from clone pGQ123BII; these fragments were ligated into pUC19 which had been digested with *Nde*I/*Bam*HI.[20] In a convergent approach, we separately synthesized fragments I, II, and III using three sets of four long oligos and six short oligos with Taq DNA polymerase.[22] The three fragments were sequentially inserted into a pUC19 cloning vector. Sequencing six independent colonies gave an average error frequency of 0.19%. We have confirmed the expression of the synthetic gene in *Escherichia coli* by Western blotting using both amino and carboxyl terminal monoclonal antibodies.[22]

The sequencing results indicate that the proofreading $Vent_R$ enzyme has minimal, if any, effect on the correctness of the final product. Instead, the intrinsic homogeneity of the chemically synthesized oligo templates probably has the greatest impact. Our results agree with estimates of errors in chemically synthesized oligos which are, on average, 0.15% per nucleotide.[5a] This value and our data indicate that the ideal length of a gene fragment prepared using the PGS approach is on the order of 400–600 bp.[5a,23]

In summary, we have determined optimal conditions for the single step, total synthesis of a gene using DNA polymerases with and without proofreading activity, we have evaluated linear and convergent synthetic strategies, and we have defined important factors for gene synthesis using the PGS approach.

Acknowledgment. We thank the Agarwal, Hochstrasser, Haselkorn, and Wool labs for many helpful suggestions and generous use of equipment and the Howard Hughes facility for oligo synthesis and DNA sequencing. The University of Chicago and the Kamen Fund are gratefully acknowledged for financial support. The gift of monoclonal antibodies from Dr. Koyama is sincerely appreciated.

(18) For typical 12-part PCR reactions using Deep $Vent_R$ DNA polymerase, 19 μL of distilled water was mixed with 10 μL of 10× buffer (200 mM Tris-HCl, pH 8.8, 100 mM KCl, 100 mM $(NH_4)_2SO_4$, 20 mM $MgSO_4$, 1% Triton X-100), 2 μL of each dNTP (10 mM stock solution), 1 μL of each short oligo (10 μM stock solution), 5 μL of each long oligo (0.1 μM stock), and 1 μL of Deep $Vent_R$ (2 units/μL). A tube without the long oligos was used as a negative control, and a positive control employed the template and primers provided by Perkin-Elmer Cetus. Following the addition of the reagents, the solution was quickly mixed and centrifuged, and one drop of mineral oil was added. Thermocycling was initiated immediately. After the end of the cycling, the reaction mixture was cooled to 0 °C and analyzed by gel electrophoresis on a 2% agarose gel.

(19) Although conditions for Taq DNA polymerase in the 12-part PCR reaction were not fully optimized, Taq generally gave a lower yield of the full-length product when compared to Deep $Vent_R$ DNA polymerase. One possible reason is that Taq is less thermostable than Deep $Vent_R$.

(20) Sambrook, J.; Fritsch, E. F.; Maniatis, T. *Molecular Cloning. A Laboratory Manual*, 2nd ed.; Cold Spring Harbor Laboratory Press: Plainview, NY, 1989; Vol. 1.

(21) (a) Sanger, F.; Nicklen, S.; Coulson, A. R. *Proc. Natl. Acad. Sci. U.S.A.* **1977**, *74*, 5463–5467. (b) DNA sequencing was carried out using the Taq DyeDeoxy Terminator Cycle Sequencing Kit supplied by Applied Biosystems and an Applied Biosystems automated sequencer.

(22) Yamaguchi, N.; Jinbo, Y.; Arai, M.; Koyama, K. *FEBS Lett.* **1993**, *324*, 287–292.

(23) (a) Box, G. E. P.; Hunter, W. G.; Hunter, J. S. *Statistics for Experimenters*. An *Introduction to Design, Data Analysis, and Model Building*; John Wiley and Sons: New York, 1978. (b) Given an average error in the chemically synthesized oligos of 0.15% per base, a Poisson distribution provides determination of the probability (p) of an oligonucleotide of length n being correct: n = 100 (p = 0.86); n = 200 (p = 0.74); n = 400 (p = 0.55); n = 800 (p = 0.30); and n = 1000 (p = 0.22). To obtain a correct sequence, at least two independent clones of a 400 bp gene and five clones of a 1000 bp gene should be sequenced. However, these estimations only account for the error in the template and do not include an estimation of the errors due to gene synthesis and amplification.