METHODS IN ENZYMOLOGY

SYNTHETIC GENE TECHNOLOGY:

APPLICATIONS TO ANCESTRAL GENE RECONSTRUCTION AND STRUCTURE-FUNCTION STUDIES OF RECEPTORS

Belinda S. W. Chang, Manija A. Kazmi and Thomas P. Sakmar

Running Title: Synthetic Genes

Please address correspondence concerning the manuscript to:

Thomas P. Sakmar, M.D.
Laboratory of Molecular Biology and Biochemistry,
Howard Hughes Medical Institute,
Rockefeller University,
1230 York Ave.
New York, New York 10021
Phone: 212-327-8288
FAX: 212-327-7904
E mail: sakmar@rockvax.rockefeller.edu

Author addresses:

Belinda S. W. Chang, Laboratory of Molecular Biology and Biochemistry, Rockefeller University, New York, New York 10021

Manija A. Kazmi, Laboratory of Molecular Biology and Biochemistry, Howard Hughes Medical Institute, Rockefeller University, New York, New York 10021

Thomas P. Sakmar, Laboratory of Molecular Biology and Biochemistry, Howard Hughes Medical Institute, Rockefeller University, New York, New York 10021

Introduction

The use of synthetic gene technology offers several major advantages for the study of G protein-coupled receptor (GPCR) structure and function,[1] and the incorporation of PCR into gene synthesis strategies means that genes of over 1 Kb can now be efficiently and economically synthesized in a matter of weeks.[2-4] In creating the gene of interest, the degeneracy of the genetic code can be used to yield nucleotide sequences that have useful properties such as large numbers of endonuclease restriction sites, optimized primer sites for the polymerase chain reaction (PCR) and sequencing, and desired levels of GC content and codon bias. Moreover, in certain cases, such as in studies of ancestral protein function, the easy and efficient creation of a gene *de novo* is critical.

Traditionally, studies of GPCRs have relied heavily on mutagenesis techniques to identify residues important for structure and function. While a number of methods are available for site-directed mutagenesis, the use of a properly designed synthetic gene offers many advantages, particularly where extensive mutagenesis is planned. The large number of restriction enzymes now available allows as many as forty unique restriction sites to be introduced for a synthetic gene of 1 Kb in length, in addition to the ability to incorporate as many optimized PCR primer sites as desired.

A novel and important application of synthetic gene technology in studies of GPCRs is for the recreation of ancestral receptors.[5] The superfamily of GPCRs consists of a large number of related seven-transmembrane proteins of highly varied function.[6]

Recreating 'fossil' receptors for functional studies in the laboratory can provide important insights into the constraints that shaped molecular structure and function in this diverse superfamily that are difficult to attain using more traditional mutagenesis methods. Except in cases where the ancestral receptor of interest is quite similar to extant sequences, synthetic gene technology is essential for this kind of study.

In addition, in studying receptor function via heterologous expression of mutant receptors, it is often useful to create chimeric receptors in which putative functional domains are exchanged. For example, the transfer of a cytoplasmic loop sequence from one pharmacological receptor subtype to another is one approach to study the specificity of ligand-dependent G protein activation. A synthetic gene for one receptor subtype can be engineered readily to produce a chimeric construct by exchanging a portion of a functionally different receptor subtype, or even of a reconstructed ancestral receptor. The use of synthetic genes allows domain exchanges without the potential limitations of naturally occurring restriction endonuclease cleavage sites.

In this chapter, a summary of methods and applications of synthetic gene technology is presented, with an emphasis on synthesizing ancestral genes. Algorithms for inferring ancestral sequences, and general considerations for designing synthetic genes are discussed in detail. Detailed laboratory procedures for the various steps in gene synthesis, from oligonucleotide preparation to cassette mutagenesis, are also given.

Applications of Synthetic Gene Technology

Many studies of GPCR function have relied on restriction fragment replacement (cassette mutagenesis) methods, which can be greatly facilitated by the use of a synthetic gene. Site-directed point mutations are easily introduced into cloned DNA by a variety of mismatch primer methods, some of which employ PCR. However, site-directed cassette mutagenesis can be successfully employed to introduce extensive alterations of the nucleotide sequence within a particular gene segment. Such an approach may be useful, for example, in structure-function studies of discrete receptor domains, such as the cytoplasmic loops of a GPCR.[7-9] A long stretch of amino acid residues easily can be replaced by a random sequence, by a homologous sequence from a related protein, or by a portion of an ancestral gene.

A new and promising approach to studies of gene function that also relies heavily on synthetic gene technology is for the study of ancestral genes. Here, the use of synthetic genes is not only useful and convenient, but in many cases where the gene has diverged significantly from existing molecules, absolutely essential. The only other option, which would be to start with an existing sequence, and use mutagenesis to incorporate all substitutions in the ancestral sequence, rapidly becomes infeasible for sequences that have even relatively small levels of divergence. Both of these approaches, and how they can incorporate synthetic gene technology, are discussed below.

Mutagenesis by Restriction Fragment Replacement (Cassette Mutagenesis)

Mutagenesis by restriction fragment replacement was first demonstrated in the naturally-occurring gene of bacteriorhodopsin.[10] This mutagenesis strategy was possible because of the fortuitous natural placement of unique restriction sites. Cassette mutagenesis in this case involved replacement of a restriction fragment by a synthetic duplex counterpart in order to introduce the desired codon alteration(s). However, the incorporation of large numbers of unique and evenly spaced restriction sites in a carefully designed synthetic gene increases the convenience and usefulness of this method. In addition, cloning is 'directional' and screening is not generally necessary for identification of a desired recombinant transformant.

Cassette mutagenesis is also of more general utility than most forms of site-directed mismatch primer mutagenesis because of the ease of producing defined mutations at multiple sites within a domain to yield deletions, extensive substitutions, domain swaps, or the construction of chimeric genes. In cases where initial mutagenesis experiments that target a particular amino acid did not prove informative, an alternative strategy is to prepare large numbers of mutations in a particular putative domain if an adequate functional screening method can be devised. Combinatorial cassette mutagenesis can be employed in such situations.[11-16] The general strategy of combinatorial cassette mutagenesis is to perform restriction fragment replacement with a set of synthetic duplexes that should provide codons for each of the 20 amino acids at one or more positions within the duplex.[17] This can be accomplished by synthesizing the noncoding (top) strand of the duplex with equal mixtures of all four bases in the first two positions of a codon, and with equal mixtures of guanine and cytosine at the third position. Inosine is inserted at each of the randomized base positions in the

bottom strand because it is able to pair with each of the four natural bases. The heterogeneous top strand oligonucleotides and the bottom strand oligonucleotide are annealed and the resulting duplex is ligated into an appropriate vector. Bacterial transformation essentially produces a library of mutants which can be cloned or studied batchwise depending upon particular circumstances.

Synthetic Ancestral Genes

A novel approach to the study of molecular structure and function that is complementary to mutagenesis methods is the study of ancestral gene function.[5,18,19] Often mutagenesis studies may be limited by the number of mutants that can reasonably be made and screened in the laboratory. *A priori* hypotheses that are critical in guiding the position and identity of amino acids to mutate may be difficult to formulate, especially for receptors about which little structural information is known. The lack of appropriate *a priori* hypotheses usually means one must be prepared to screen large numbers of mutants. In addition, random mutations may result in either non-functional receptors, or in receptors with wild-type function (in the particular assay used), for reasons which can be difficult to determine *a posteriori*, rendering the results uninformative in terms of receptor structure/function.

The superfamily of GPCRs is an ideal system in which to use ancestral genes to study molecular structure/function. In the molecular evolutionary radiation of these genes, selection has already screened out all the mutations that result in non-functional receptors. Moreover, by focussing attention on ancestral receptors where key functional

shifts are thought be have occurred (Fig. 1), important amino acid substitutions that form the basis of those functional shifts may be isolated. Essentially, reconstructing ancestral proteins allows the identification of mutations that caused major shifts in protein function, in the background in which these changes originally occurred. Reconstructing these critical mutations in the ancestral background avoids the problem of non-functional or improperly folded / expressed proteins, which may be the case if performed in extant molecules. Note that this approach also provides very specific *a priori* hypotheses about the nature of the amino acid substitutions, and the particular function affected.

Studies of ancestral gene function are made possible by recent advances in the statistical estimation of ancestral sequences, which are briefly summarized here. These methods are critical to the success of this approach. Although moving from the statistical inference of ancestral proteins to their actual synthesis in the laboratory is a major step, it can provide unique information about the context in which adaptive replacements may have occurred, and other structure / function information difficult to obtain using more traditional molecular methods.

Ancestral Gene Inference Methods

In an ideal situation, one would start from a well-supported molecular phylogeny, from which the ancestral protein sequence of interest would be unambiguously determined. However, given that a typical protein is composed of hundreds of amino acid sites, all of whose ancestral states must be inferred, it is

extremely rare that all sites can be reconstructed unambiguously. Moreover, often the most interesting evolutionary changes in biochemical function do not occur at extremely low levels of sequence divergence, where ancestral states are more easily inferred. This may be a problem particularly when divergent sequences are combined with varying rates of evolution.[20]

Our intention is not to provide a comprehensive review of phylogenetic methods, or methods for inferring ancestral states, which can be found elsewhere.[21,22] Instead, we highlight methodological considerations that we believe are most directly relevant when the goal is to proceed to reconstruct sequences in the lab. One obvious source of error in inferring ancestral states is lack of resolution of the tree itself, or the existence of a variety of plausible trees. If this problem cannot be overcome, it seems highly advisable to thoroughly explore the sensitivity of the inferred ancestral sequences to alternative resolutions of poorly supported nodes in the tree.[23,24] Even when one has great confidence in a single tree, there may still be ambiguity in inferences of ancestral states at particular internal nodes. Here we focus attention on ways to proceed in the face of such ambiguity.

Parsimony methods (implemented in programs such as PAUP* [25]) evaluate phylogenetic relationships and ancestral state assignments based on the amount of evolutionary change along the branches of the tree; specifically, trees or ancestral states that require the fewest changes are preferred.[21,26] Precisely what parsimony assumes about rates of change is still the subject of investigation,[22,27,28] but it is clear from simulation studies that it can fail under certain circumstances, such as when rates of change are highly unequal along different branches.[29,30] Although weighted

parsimony methods, and the use of step-matrices, can accommodate rather complex models of character change, it is difficult to correct for multiple substitutions at a site in an explicit model of evolution, or to take branch lengths into consideration. These properties of parsimony are problematic from the standpoint of reconstructing ancestral states.[23,31] Furthermore, in practice it is clear that equivocal assessments of ancestral states are common using parsimony. Even a small percentage of such ambiguities scattered along an entire sequence could yield a large number of permutations and combinations, which might then require the examination of a large number of proteins in the lab. For this reason it is important to consider alternative strategies to narrow down the possibilities.

Likelihood phylogenetic methods (implemented in programs such as PHYLIP,[32] MOLPHY,[33] PAML,[34] and NHML[35]) use as an optimality criterion a likelihood score, calculated according to a specified model of evolution.[36] This likelihood score represents the probability of observing the sequence data, given a particular tree topology and model of evolution, and is maximized in reconstructing phylogenetic relationships and ancestral sequences. In reconstructing ancestral states, likelihood methods offer several advantages over parsimony.[27,31,37] Likelihood methods not only use an explicit model of evolution, they also make use of additional information ignored by parsimony contained in branch lengths. An explicit model allows the incorporation of knowledge of the mechanisms and constraints acting on coding sequences, as well as the possibility of comparing the performance of different models, ultimately resulting in the development of more realistic models.[38] Finally, one of the frequently cited drawbacks of likelihood

methods, namely the computationally-intensive calculations required for most phylogenetic analyses, is not an issue for ancestral reconstructions since most of the computational burden is due to calculating the likelihood of different tree topologies, and not the calculations of the ancestral states themselves. Thus once a reliable tree has been obtained, in most cases it is entirely feasible to use likelihood methods to reconstruct ancestors.[27]

These properties of likelihood are especially important for ancestral state reconstruction, as sites which have ambiguous ancestral state assignments under parsimony can then be explored under different models in likelihood.[23,31] Likelihood methods not only offer the opportunity to assess the comparative fit of various models to the sequence data at hand, they also provide information about specific probabilities associated with particular ancestral reconstructions. This information can be extremely useful in narrowing down to one or a few reconstructions for the purpose of designing ancestral proteins for synthesis in the lab.

If differences in ancestral reconstructions depend on the choice of evolutionary model, then choosing a realistic model is critical. Not surprisingly, this has been the focus of recent developments in phylogenetic methods. Likelihood models describe molecular evolution at three different levels: nucleotide, amino acid, and codon. The simplest nucleotide model, Jukes-Cantor,[39] assumes equal base frequencies, and equal rates of transitions and transversions, which is clearly not realistic for most data sets. More complex nucleotide models incorporate parameters such as those allowing unequal base frequencies,[36]

transition / transversion bias,[40] among-site rate heterogeneity,[41] or nonstationary base composition.[35] However, even these models, and models such as the GTR model,[42] which allows unequal number of substitutions among all the different classes of nucleotides in the rate matrix, fails to take into account codon position or amino acid information.

Amino acid models tend to be even more parameter-rich, because they involve twenty states instead of only four, as with nucleotide data. The simplest of these models is the Poisson model, which assumes equal amino acid frequencies, and equal rates of substitution among all the amino acids.[43] Both of these assumptions are problematic, as amino acids are known to occur at very different frequencies, and rates of substitution among classes of amino acids can be highly variable due to functional and structural constraints at the protein level. Amino acid models have also been developed that incorporate parameters allowing unequal amino acid frequencies[44] and among-site rate heterogeneity,[41] in addition to a GTR model for amino acids, which allows for unequal numbers of substitutions in the rate matrix for all the different classes of amino acids.[34] However, it is not always necessary to estimate the substitution rate matrix from the sequence data at hand. Rate matrices have been calculated for a number of data sets, including those of Dayhoff[45,46] and Jones[47,48] for globular proteins, mitochondrial transmembrane proteins,[49] and rhodopsin proteins.[50] Use of rate matrices derived from other data sets (if appropriate) allows for the reduction in the number of parameters in the model of evolution. A significant advantage of amino acid models is that they avoid many of the problems associated with modeling evolution at the nucleotide

level, such as base compositional bias. However, in these models, all nucleotide-encoded information is lost, including those potentially relevant to the task of ancestral reconstruction.

Codon-based models of molecular evolution are among the most recent developments, and have the advantage of incorporating information on both nucleotide and amino acid levels. The original codon-based models assumed equal nonsynonymous to synonymous rate ratios among sites and lineages.[51,52] Subsequent models have allowed that ratio to vary across lineages or sites in the protein,[53,54] or even allowed the incorporation of unequal frequencies of different types of nonsynonymous substitutions based on the nature of the various amino acids.[55]

Given the diversity of models now available, choice of model for use in phylogenetic analysis and ancestral inference is critical. An inappropriate model of evolution can lead to inconsistency in the likelihood analysis, and convergence to an incorrect result.[21,28,56] Ancestral inference methods are particularly sensitive to model choice. The possibility of an incorrect result can be reduced by selecting a model of evolution that displays a good fit to the sequence data at hand.

Likelihood ratio tests can be used to compare two models of evolution that are nested with respect to one another, in order to determine whether the more complex models fits the sequence data significantly better than the simpler model.[36,57,58] For nested models, a more complex model (H_1) will contain all the parameters of the original model (H_0), as well as additional parameters. If the models are not nested, they cannot be directly compared using a likelihood ratio test,

and other methods, such as the generation of the distribution of the test statistic using Monte Carlo simulation, must be used.[38] For nested models, a more complex model (H_1), with additional parameters, should fit the data better than a simpler model (H_0), as judged by the likelihood score, or the natural logarithm of the likelihood, of each model (L_0, L_1). If H_0 is correct, this difference in fit to the data can be approximated by a χ^2 distribution, with degrees of freedom (df) equal to the difference in number of parameters between the two models:[59]

$$2(L_1-L_0) = \chi^2_{[df]}$$

However, if the observed difference is greater than the χ^2 critical value, then the simpler model (H_0) will be rejected, and the more complex model (H_1) will be the preferred model. In other words, in this case the more complex model fits the data even better than would be expected because of its additional parameters relative to the simpler model.

Although the exact details of the ancestral reconstruction methods such as model choice will differ according to the particular data set used, inference of an ancestral rod opsin gene is presented here as an example.[50] While different methods often infer the same ancestral reconstructions, occasionally parsimony methods may yield more than one most parsimonious reconstruction, as is illustrated here (Table I). Such ambiguity can be problematic if the aim is to reconstruct a protein in the lab. Although a few such ambiguous sites can be dealt with by synthesizing all possible combinations, this approach becomes intractable with even relatively few ambiguous sites. Parsimony methods alone offer no means of deciding among several most parsimonious reconstructions. In contrast, with likelihood methods it

is possible to compare the fit to the data of different models using likelihood ratio tests, if these are nested models. Furthermore, for each model information can be obtained about the relative probabilities of different amino acid reconstructions in the form of marginal posterior probabilities.[31] This information can provide a basis for choosing to create one or a few inferred sequences in the lab.

Table I shows a portion of the ancestral protein that gave rise to the rhodopsin family of genes inferred from a phylogeny of vertebrate rhodopsin and other visual pigment sequences using both parsimony and likelihood methods. For this data set, based on pairwise nested comparisons using likelihood ratio tests the HKY+Γ model (for nucleotides), and the GTR+Γ model (for amino acids) were chosen as the most appropriate for maximum likelihood ancestral reconstruction. Two simpler models, which show a significantly poorer fit to the data, JC and Poisson, are shown for comparison. In addition, maximum parsimony reconstructions are also given.

For many sites there is good correspondence between parsimony and likelihood inferences, and also among models of evolution utilizing different information (for example, nucleotides *versus* amino acids). However, this is not the case for all sites. In parsimony, ambiguity can result from different reconstructions generated from amino acids *versus* nucleotides (site 138), or it can result from more than one most parsimonious state assignment (sites 137, 139, 150). Choices can sometimes be made among these using likelihood methods, by choosing the model with the best fit to the data, and the reconstruction with the highest posterior probability under this model (site 137). However, this is not

always the case. Furthermore, strongly violating the assumptions of a model, or using a model that poorly fits the data (regardless of whether it is the best fit of the models compared) may result in spurious inferences. For example, site 139 is reconstructed as an Ile under the (in this case, oversimplified) Poisson amino acid model, whereas it is reconstructed as a Val under all other likelihood models. In addition, for a given model, the ancestral reconstruction with the highest marginal probability may not be correct, especially if that probability is not high, and other reconstructions have comparable posterior probabilities. Finally, different likelihood models which make use of different information, such as amino acids *versus* nucleotides, can yield different reconstructions (site 150), and in such cases it may be desirable to test both possibilities in the lab.

Gene Synthesis Methods

Synthetic Gene Design

In designing a synthetic gene, the ultimate goals are to facilitate subsequent mutagenesis and chimeric gene studies, while achieving high levels of *in vitro* expression of the designed gene. Ensuring ease and flexibility of genetic manipulation can be done in two ways. Incorporating large numbers of unique restriction sites is required for studies using replacement of restriction fragments, or cassette mutagenesis, whereas PCR-based mutagenesis methods are facilitated by the incorporation of appropriately optimized primer sites. Levels of protein expression can be optimized by

adjusting GC content and codon bias. These and other considerations are outlined briefly below.

The choice of restriction endonuclease sites to be considered in the design of a synthetic gene includes the following criteria: 1) reliability and availability, 2) high activity and freedom from any exonuclease activity, 3) a recognition sequence of five or more nucleotides, and 4) the generation of staggered rather than blunt ends. Because of the degeneracy of the universal genetic code, a very large number of potential nucleotide sequences can encode a given amino acid sequence. This potential variability in nucleotide sequence generates a large number of potential restriction maps.

The traditional approach for synthetic gene design was to begin with the native DNA sequence and restriction map, retain all potentially useful restriction sites, and then attempt to add new sites in the intervening sequences.[60] This approach, however, is not general. Manual approaches were used to reverse translate restriction endonuclease recognition sequences in order to consider the locations of all possible restriction sites in a particular amino acid sequence.[61] More recently, this general approach has been greatly facilitated by the availability of sequence analysis software packages such as LaserGene (DNAStar, Inc.) that allow the identification of all of the potential restriction sites within a putative gene. This can be accomplished by starting with the amino acid sequence and using a reverse translation algorithm to create a degenerate nucleotide sequence, from which potential restriction sites may be identified.

In order to reduce the number of restriction sites identified in the degenerate sequence, it is preferable to limit the file of restriction enzymes to those with recognition sequences (palindromic or interrupted palindromic) of at least 5 bases that generate cohesive ends of 2 or more nucleotides. Methylase-sensitive enzymes should be avoided, but in some cases nucleotides outside of the endonuclease recognition sequence can be altered to remove the methylase recognition sequence. Sites for enzymes generating blunt ends can be used if long gaps are present after all enzymes generating staggered ends are considered. However, blunt-end cutters should not be juxtaposed in the restriction map. Enzymes that generate identical cohesive overhangs should also not be juxtaposed. Unique restriction sites can be ensured by removing other potential sites from the sequence. Convenient cloning sites should be chosen for each end of the gene, as well as for gene construction via ligation of the long oligonucleotides should PCR methods fail. For example, a number of genes have been synthesized with an *Eco*R1 site at the 5'-end and a *Not*1 site at the 3'-end for ease of cloning.[1,2,61-63]

In order to optimize gene construction using PCR-based methods, and to ensure the success of PCR-based mutagenesis methods, and to facilitate later construction of chimeric genes by ‘swapping’ of PCR fragments, it is necessary to give careful consideration to the design of appropriately placed PCR primer sites. These primers should be designed with standard considerations in mind, such as minimizing hairpins, primer duplexes, mispriming, and optimizing the melting temperature. In addition, primers that will be used together should be designed so as to minimize

possible primer-dimers. All these can be done via many programs available for this purpose, such as Vector NTI (InforMax, Inc.).

After defining the nucleotide sequence that corresponds to the desired restriction map and optimized primer sites, a majority of the gene sequence may still remain undefined. In some cases, the natural sequence can be retained. However, other factors such as codon usage and GC content may also be considered in order to optimize expression levels and ease of molecular genetic manipulation.

Codon usage bias can be optimized for a particular expression system in order to achieve desired expression levels.[64] In expression systems such as those using *E. coli*, rare codons (*e.g.*, the AGA codon for Arg) have been shown to cause translational problems most likely due to limited tRNA availability, resulting in misincorporations, frameshifts (leading to truncated proteins), and overall reduced translational efficiency.[65-67] These rare codons, and over-represented codon pairs which also have been shown to slow translation,[68] should generally be avoided in designing synthetic genes. Optimizing codon usage frequencies can result in much higher expression levels. On the other hand, it may sometimes be appropriate to deliberately incorporate unpreferred codons in order to slow translation of signal sequences so that cellular membrane translocation systems are not saturated. This was the case for the expression in *E. coli* of the gene for the light-driven proton pump bacteriorhodopsin from *H. halobium*.[69]

The synthetic gene design process also allows for the reduction of GC content if desired. Stretches of four or more guanines or cytosines can be avoided where possible to minimize potential difficulties in oligonucleotide synthesis, PCR, and DNA

sequencing. For example, the GC content in dopamine receptor variants was reduced from 74.2% to 49.4% in the synthetic genes[62] In addition, some investigators have found it useful to place a mammalian translation initiation consensus sequence immediately preceding the initiation methionine codon.[60,61,70]

Finally, the user-defined DNA sequence should be translated to confirm the correct amino acid sequence. This is important because the amino acid sequence translated from a degenerate codon sequence will not always match the original. For example, with sixfold degenerate amino acids such as serine, four of the codons form TCN and two others AGY. The two types reduce to the single degenerate codon WSN. If this degenerate codon is translated, it will be assigned an unknown amino acid X, since WSN can expand to any of the following: TCN (Ser), ACN (Thr), AGY (Ser), AGR (Arg), TGY (Cys), TGA (Ter), or TGG (Trp).

Synthetic Gene Construction

Methods of synthetic gene construction rely on overlapping synthesized oligonucleotides of varying lengths. In earlier studies, these were extended using T7 DNA polymerase, and ligated together to form a complete gene.[71] The incorporation of PCR, and especially the use of heat stable enzymes such as *Pfu* that have additional proofreading functions and higher processivity than *Taq*, has not only made synthetic gene construction much faster, but also easier and more economical.[2-4,72-76] It is now entirely feasible to have a complete gene synthesized and expressed in a matter of weeks.

Because gene construction using PCR has rendered unnecessary purification steps to ensure full-length synthesized oligonucleotide template, and requires very little starting material, it has become feasible to synthesize genes using longer oligonucleotide fragments of up to 300 bases. Longer oligonucleotides offer several advantages in addition to rendering gene design more straightforward. Short oligonucleotides require on the order of 50-100 overlapping fragments, and it may become difficult to optimize them all for PCR. More importantly, regions of overlap are required to be about 20 basepairs in length, regardless of oligonucleotide length. Therefore fewer longer oligonucleotides are required to cover the entire synthetic gene, minimizing both the total amount of overlap required, and the total number of bases synthesized. This strategy is thus much more economical, and the oligonucleotides can be more quickly synthesized.

In order to demonstrate the procedure for synthetic gene construction, a recently-constructed ancestral rod opsin gene is presented here as an example (Fig. 2).[2] The entire gene (1114 bp) was constructed from five long synthetic oligonucleotides that were amplified and assembled using a stepwise PCR procedure (Fig. 3). Although it may be feasible to combine all the reactions in one PCR step,[4] a stepwise procedure was chosen primarily to facilitate troubleshooting should any of the PCRs fail. The five overlapping fragments were synthesized as single-stranded oligonucleotides on an Applied Biosystems oligonucleotide synthesizer. In the first round, each synthesized oligonucleotide was converted into a duplex and amplified in a PCR reaction containing flanking primers of 25-30 bp. In the second round of PCR, the resulting duplex PCR products were joined together pairwise (AB, BC, CD and DE) in separate

PCR reactions. The third round of PCR started the elongation of the gene through stepwise PCR reactions. For example, AB and BC were spliced together to give fragment ABC. In the fourth round fragments ABC and CD were spliced together to give ABCD. In the fifth and final round, ABCD was spliced together with DE to give the full-length gene ABCDE. The products of each round of PCR were separated on low-melt agarose gels and purified using a Qiaex II kit (Qiagen), or used directly in the next round of PCR. The final product was cloned directly into pCR-Blunt (Invitrogen), a vector specially designed for direct cloning of blunt-end PCR products generated using *Pfu* polymerase. Several recombinants were sequenced using flanking and internal sequencing primers.

Experimental Procedures

Oligonucleotide Synthesis

Automated oligonucleotide synthesis can be easily carried out on oligonucleotide synthesizers with commercially available solvents and reagents. The most commonly used chemistry involves the phosphite triester approach using protected β-cyanoethyl-phosphoramidite nucleosides. The fully protected 3'-terminal phosphoramidite of the oligonucleotide is coupled to a solid support such as control pore glass or polystyrene.[77] After protic acid treatment to remove the 5'-protecting group, the fully-protected incoming phosphoramidite is activated by tetrazole so that a phosphite triester bond is formed at high efficiency. The small amount of unreacted 5'-hydroxyl of the first

nucleoside is capped by a quantitative reaction with acetic anhydride in the presence of 1-methylimidazole. Finally, the newly formed internucleotide linkage is converted from a phosphite triester to a more stable phosphate triester by oxidation with iodine where water is the oxygen donor. The 5'-hydroxyl of the dinucleotide can now be deprotected with acid treatment to complete a cycle. The cycle is repeated until the full-length oligonucleotide is obtained. Thus, the oligonucleotide is elongated from 3' to 5'. Cleavage from the support and removal of phosphate and exocyclic amine protecting groups is achieved by treatment with concentrated ammonium hydroxide.

For synthesis of the ancestral rod opsin gene, automated oligonucleotide synthesis was performed on an Applied Biosystems model 392 DNA synthesizer. Phosphoramidite chemistry was employed using 40 pmole synthesis scales and standard cycle routines. Each synthetic oligonucleotide was automatically cleaved from the solid support after removal of the terminal 5'-hydroxyl protecting group. Each oligonucleotide solution was transferred into a screw top vial. After the addition of 2 ml of fresh concentrated ammonium hydroxide, the vial was tightly capped and heated at 55^{o} for at least 5 hrs. Each fully-deprotected oligonucleotide was dried by vacuum centrifugation in a polypropylene tube and the pellets were dissolved in 50 µl of TE (10 m*M* Tris-HCl, 1 m*M* EDTA, pH 7.4). An ultraviolet spectrum was measured from 310 nm to 210 nm after making the proper dilution in TE. The yield in total absorbance units at 260 nm was calculated. The crude oligonucleotides were subjected to PCR amplification as described below.

Stepwise PCR

Oligonucleotides corresponding to the overlapping fragments A-E (277, 300, 221, 308, and 159 bp, respectively) were synthesized (Fig. 2). Adjacent oligonucleotides had overlaps of 20-40 bp. A total of ten primers of 21-24 bp corresponding to the 5' and 3' region of each long oligo were synthesized. Primer pairs a1/a2, b1/b2, c1/c2, d1/d2, and e1/e2 flanked fragments A-E, respectively. These primer pairs and their corresponding crude oligonucleotide templates were used in a PCR to amplify the five gene fragments (Fig. 3). In the round I synthesis, 100 pmol of each template oligonucleotide was added to 50 µl PCR mixture (20 m*M* Tris-HCl, pH 8.0, 2 m*M* $MgCl_2$, 10 m*M* KCl, 6 m*M* $(NH_4)_2SO_4$, 0.1% Triton® X-100, 10 µg/ml BSA, 0.4 m*M* each dNTP, 2.5 U of *Pfu* polymerase and 1 µ*M* each flanking primer). The PCR program consisted of one denaturation step at 94° for 45 sec, followed by 25 cycles at 94° for 45 sec, 58° for 1 min, 72° for 2 min and a final incubation at 72° for 10 min. The fragments were separated on 2% NuSieve GTG agarose (FMC BioProducts) and purified using Qiaex II purification kits (Qiagen). These PCR purification conditions were used in each subsequent round of PCR.

For round II PCR, products A-E from round I were diluted 1:500 and 1 µl of each adjacent pair was used a template. Primer pairs a1/b2, b1/c2, c1/d2, and d1/e2 flanked fragments AB, BC, CD and DE, respectively. For round III, products AB and BC were diluted 1:500 and 1 µl of each was used as a template for primers a1 and c2. The resulting product ABC and product CD from round II were diluted 1:500 and 1 µl of

each was used as a template with primers a1 and d2 in round IV to generate fragment ABCD. In the fifth and final round of PCR, product ABCD and DE from round II were diluted 1:500 and 1 µl of each was used a template with primers a1 and e2 to amplify the entire gene.

Cloning of PCR product

The final PCR product (ABCDE) was cloned directly into pCR-Blunt (Invitrogen). In a 10 µl reaction, 1 µl of the purified product was mixed with 25 ng of vector and ligated under manufacturers guidelines. The ligated material was transformed into TOP-10 cells (Invitrogen) following standard protocols, and the recombinant transformants were plated onto LB plates containing 50 µg/ml kanamycin. Several recombinant clones were streak purified and plasmid DNA was purified from each using Qiagen mini prep kits. Each clone was sequenced with two internal and two flanking primers. Alternatively, the final PCR product may be digested with the restriction enzymes engineered at the 5'and 3' end of the gene and ligated into an expression vector using directional cloning. Sequence errors that are due to depurination during oligonucleotide synthesis, or due to misincorporations during PCR are expected to be randomly distributed among different clones, and can be easily corrected by combining error-free fragments by restriction cloning. Any remaining errors can be repaired using the QuikChange kit (Stratagene).

Cassette Mutagenesis

Site-directed mutagenesis of the synthetic gene was accomplished by synthesizing a pair of complementary oligonucleotides to form a duplex containing the desired codon alteration and the appropriate cohesive-terminal overhangs. After purification and annealing as previously described, the 5'-end non-phosphorylated synthetic duplex was ligated into the plasmid / gene DNA fragment linearized with the appropriated restriction endonucleases. Alternatively, for longer oligonucleotide replacements, the insert can be synthesized single-stranded, amplified using PCR, then cloned into the synthetic gene using the appropriate restriction sites.

Expression of Synthetic Genes

As discussed above, one of the advantages of the use of synthetic genes is that they can be easily transferred among a variety of vectors, and that codon usage can be optimized where relevant to achieve maximal levels of expression. Synthetic GPCR genes will generally be expressed in mammalian cells in tissue culture where pharmacological and cellular physiological effects can be correlated with structural changes introduced by mutation. In the case of the synthetic gene for bovine rhodopsin, large quantities of the opsin apoprotein can be produced in monkey kidney cells by transfection where transcription is under the control of the human adenovirus major-late promoter[78] or in stable cell lines.[79] The apoprotein in the plasma membrane can be regenerated with the chromophore 11-*cis*-retinal to form rhodopsin. The

recombinant rhodopsin can be solubilized with detergent treatment and purified using an affinity adsorption method.[8,78]

Conclusions

Because of their advantages for mutagenesis studies, synthetic GPCRs have been expressed in a variety of heterologous expression systems. For example, synthetic genes have been expressed in *E. coli*,[69] in monkey kidney cells in tissue culture,[60] in insect Sf9 cells,[80] and in yeast.[80-82] In visual pigment structure-function studies, synthetic receptor genes for the rhodopsin and for the human blue, green, and red cone pigment genes have been expressed in mammalian cells and purified from cell extracts after reconstitution with 11-*cis*-retinal chromophore.[63] Purified site-directed mutant pigments have been studied by a variety of biochemical[83-86] and biophysical techniques.[87-92] These studies have led to a greater understanding of the mechanism of wavelength regulation by visual pigments[84,93-98] and of the mechanism of rhodopsin-transducin interaction.[9,99]

However, despite advances in mutagenesis studies using synthetic gene technology, these studies are still limited by the requirements of some *a priori* knowledge of protein structure and function in order to decide which mutants to test. A new and promising approach that can shed light on this problem, and requires little prior information, is the study of ancestral genes. This approach has been made possible in recent years by advances in statistical methods in ancestral sequence

inference, and the obvious benefits of synthetic gene technology for bringing these ancestral genes into the laboratory are just beginning to be exploited.

In conclusion, gene synthesis should be considered when extensive long-term structure-function studies are planned, or when ancestral genes are the subject of study. The initial investment in gene design and oligonucleotide synthesis increases the ease and flexibility of later DNA manipulation. Improved economical automated DNA synthesis, PCR techniques, and the availability of a large number of quality restriction endonucleases have combined to make gene synthesis rapid and efficient for nearly all molecular biology laboratories.

Acknowledgements

We thank Michael Donoghue for discussions concerning ancestral inference, and Wing-Yee Fu for technical assistance on the design and synthesis of the Budgerigar UV pigment. Support for this work was provided in part by the Howard Hughes Medical Institute, the National Institutes of Health (DK54718, EY07138), and the Allene Reuss Memorial Trust.

TABLE I

RECONSTRUCTED ROD OPSIN ANCESTOR[a]

	Parsimony[c]		Likelihood[d]					
			Amino acids			Nucleotides[e]		
Site[b]	AA	BPS[e]	Poisson	GTR+Γ	Post. Prob's	JC	HKY+Γ	Post. Prob's
119	L	L	L	L	0.999	L	L	0.913
120	G	G	G	G	1.000	G	G	0.952
121	G	G	G	G	1.000	G	G	0.663
122	E	E	E	E	1.000	E	E	0.667
123	V	V	V	V	0.937	V	V	0.334
124	A	A	A	A	0.973	A	A	0.699
125	L	L	L	L	1.000	L	L	0.874
126	W	W	W	W	1.000	W	W	1.000
127	S	S	S	S	1.000	S	S	0.877
128	L	L	L	L	1.000	L	L	0.905
129	V	V	V	V	1.000	V	V	0.979
130	V	V	V	V	1.000	V	V	0.944
131	L	L	L	L	1.000	L	L	0.995
132	A	A	A	A	1.000	A	A	0.974
133	I	I	I	I	0.999	I	I	0.984

134	E	E	E	E	1.000	E	E	0.996
135	R	R	R	R	1.000	R	R	0.592
136	Y	Y	Y	Y	1.000	Y	Y	0.888
137	***I/V***	I	I	I	0.909	I	I	0.749
138	V	***G***	V	V	1.000	V	V	0.626
139	***I/V***	***I/V***	***I***	V	0.509	V	V	0.600
140	C	C	C	C	1.000	C	C	0.998
141	K	K	K	K	1.000	K	K	0.909
142	P	P	P	P	1.000	P	P	0.919
143	M	M	M	M	1.000	M	M	1.000
144	G	G	G	G	0.996	G	G	0.995
145	N	N	N	N	1.000	N	N	0.998
146	F	F	F	F	1.000	F	F	0.956
147	R	R	R	R	0.999	R	R	0.944
148	F	F	F	F	1.000	F	F	0.982
149	G	G	G	G	0.997	G	G	0.677
150	***S/D/G/N***	S	S	S	0.957	***D***	***D***	0.579
151	T	T	T	T	0.967	T	T	0.897
152	H	H	H	H	1.000	H	H	0.645

a[50]

[b]Amino acid reconstructions are for amino acid sites 119-152, numbered according to bovine rhodopsin.

[c]Parsimony reconstructions were done in PAUP*[25] using unweighted analyses for amino acids, and 2-to-1 Tv/Ts weighting for nucleotides. Discrepancies among reconstructions are highlighted in bold italics.

[d]Likelihood reconstructions were done in PAML[34] using the GTR+Γ model,[34] and the Poisson model[44] for amino acids, and the HKY85+Γ model,[100] and the Jukes-Cantor model[39] for nucleotides. Marginal posterior probabilities are given for each amino acid reconstruction for the best-fitting models, GTR+Γ and HKY85+Γ.

[e]Nucleotide reconstructions were translated to amino acids for the purposes of comparison with the amino acid reconstructions.

TABLE II

SYNTHETIC GPCR GENES

Gene	Length (bp)	Reference
Ancestral rod opsin	1114	Chang *et al.* in prep.
Budgerigar UV pigment	1153	Sakmar *et al.* in prep
Rhodopsin	1048	Ferretti *et al.* [60]
Red cone pigment	1130	Oprian *et al.* [63]
Green cone pigment	1130	Oprian *et al.* [63]
Blue cone pigment	1080	Oprian *et al.* [61]
D4-2 Dopamine receptor	1164	Chio *et al.* [101]
D4-2 Dopamine receptor	1170	Kazmi *et al.* [62]
D4-4 Dopamine receptor	1266	Kazmi *et al.* [62]
D4-7 Dopamine receptor	1410	Kazmi *et al.* [62]
Glucagon receptor	1472	Carruthers and Sakmar [1]
Galanin type 3 receptor	1104	Kolakowski *et al.* (GenBank AF042785)
Angiotensin receptor	1131	Noda *et al.* [102]
Calcitonin receptor	1493	Nussenzveig *et al.* [103]
Serotonin 5HT3 receptor	1906	D.S. Johnson (GenBank U59673)

Figure Legends

Fig. 1. Ancestral genes in studies of receptor structure and function. In the phylogeny depicted of a superfamily of related receptor genes, receptor A has evolved a function different from that of receptors B and C. Receptors with this new function are indicated by dashed lines. As an alternative to mutagenesis studies to discover the amino acid substitutions that underlie this shift in function, ancestral genes relevant to this functional transition can be synthesized and assayed instead (Ancestors #1, #2). This method offers the advantage of assaying mutations in the background in which they originally occurred, and avoids problems of misfolded or nonfunctional receptors, which is often the case in structure/function studies of extant sequences. See text for details.

Fig. 2. Gene synthesis by the stepwise PCR method. This method involves the synthesis of long overlapping oligonucleotides of 200-300 bps, followed by several PCRs to assemble the gene in a stepwise manner from the synthetic oligonucleotides. Once the complete gene has been obtained in this manner, it is then cloned into the appropriate expression vector. A schematic is presented for the synthesis of the rod ancestral opsin gene as an example. This gene is 1114 bps in length, and was designed with 29 unique restriction sites (*), and 10 PCR primer sites at the ends of each long oligonucleotide. The seven transmembrane segments are depicted as black rectangles and are labeled I-VII. Five gene fragments were prepared using an ABI oligonucleotide synthesizer: fragment A, *Eco*R1 to *Bgl*II (277 bases); fragment B, *Bgl*II to *Xba*I (300 bases);

fragment C, *Xba*I to *Pst*I (221 bases); fragment D, *Pst*I to *BstE*II (308 bases), and fragment E, *BstE*II to *Not*I (159 bases). Adjacent fragments were designed to overlap by at least 20 nucleotides. Detailed procedures for gene synthesis by this method are presented in the text.

Fig. 3 Synthesis of the rod ancestral opsin gene by stepwise PCR. (A) Assembly of the five long overlapping oligonucleotide fragments (A-E) is accomplished by five rounds of PCR. In the first round, synthesized oligonucleotide fragments are amplified using primers at each respective end. In the second round, adjacent pairs of oligonucleotide fragments are sewn together, and in subsequent rounds these joined fragments are concatenated until the entire gene is amplified. (B) Agarose gel showing the results of each PCR (3% agarose run in 1xTAE buffer and stained with ethidium bromide; ladders on either end are 0.5 ng φX-*Hae*III digest). Lanes A thru E are from Round I, AB thru DE from Round II, ABC from Round III, ABCD from Round IV, and ABCDE from Round V.

References

1 C. J. L. Carruthers and T. P. Sakmar, *Methods in Neurosci.* **25,** 322 (1995).

2 B. S. W. Chang and T. P. Sakmar, in prep.

3 C. Withers-Martinez, E. P. Carpenter, F. Hackett, B. Ely, M. Sajid, M. Grainger, and M. J. Blackman, *Protein Eng.* **12,** 1113 (1999).

4 W. P. C. Stemmer, A. Crameri, K. D. Ha, T. M. Brennan, and H. L. Heyneker, *Gene* **164,** 49 (1995).

5 B. S. W. Chang and M. J. Donoghue, *Trends Ecol. Evol.* **15,** 109 (2000).

6 W. C. Probst, L. A. Snyder, D. I. Schuster, J. Brosius, and S. C. Sealfon, *DNA Cell Biol.* **11,** 1 (1992).

7 O. Moro, J. Lameh, P. Högger, and W. Sadée, *J. Biol. Chem.* **268,** 22273 (1993).

8 R. R. Franke, T. P. Sakmar, R. M. Graham, and H. G. Khorana, *J. Biol. Chem.* **267,** 14767 (1992).

9 R. R. Franke, B. König, T. P. Sakmar, H. G. Khorana, and K. P. Hofmann, *Science* **250,** 123 (1990).

10 K.-M. Lo, S. S. Jones, N. R. Hackett, and H. G. Khorana, *Proc. Natl. Acad. Sci. U. S. A.* **91,** 2285 (1984).

11 K. Poindexter, R. Jerzy, and R. B. Gayle, *Nucleic Acids Res.* **19,** 1899 (1991).

12 W. C. Chan and T. Ferenci, *J. Bacteriol.* **175,** 858 (1993).

13 J. C. Hu, N. E. Newell, B. Tidor, and R. T. Sauer, *Protein Sci.* **2,** 1072 (1993).

14 L. M. Gregoret and R. T. Sauer, *Proc. Natl. Acad. Sci. U.S.A.* **90,** 4246 (1993).

15 A. P. Arkin and D. C. Youvan, *Proc. Natl. Acad. Sci. U.S.A.* **89,** 7811 (1992).

16 S. Delagrave, E. R. Goldman, and D. C. Youvan, *Protein Eng.* **6,** 327 (1993).

17 J. F. Reidhaar-Olson and R. T. Sauer, *Science* **241,** 53 (1988).

18 T. M. Jermann, J. G. Opitz, J. Stackhouse, and S. A. Benner, *Nature* **374,** 57 (1995).

19 U. M. Chandrasekharan, S. Sanker, M. J. Glynias, S. S. Karnik, and A. Husain, *Science* **271,** 502 (1996).

20 D. Schluter, *Nature* **377,** 108 (1995).

21 D. L. Swofford, G. J. Olsen, P. J. Waddell, and D. M. Hillis, *in* "Molecular Systematics" 2nd edn., (D. M. Hillis, C. Moritz, and B. K. Mable, eds.), p. 407. Sinauer, Sunderland, Massachusetts, 1996.

22 J. Felsenstein, *Annu. Rev. Genet.* **22,** 521 (1988).

23 J. Zhang and M. Nei, *J. Mol. Evol.* **44,** S139 (1997).

24 M. J. Donoghue and D. D. Ackerly, *Philos. Trans. R. Soc. London. Ser. B.* **351,** 1241 (1996).

25 D. L. Swofford, "PAUP*, Phylogenetic Analysis Using Parsimony (*and Other Methods)", Version 4.0. Sinauer, Sunderland, Massachusetts, 1999.

26 W. P. Maddison, *Syst. Zool.* **40,** 304 (1991).

27 P. O. Lewis, *in* "Molecular Systematics of Plants II: DNA Sequencing" , (P. S. Soltis, D. E. Soltis, and J. J. Doyle, eds.), p. 132. Kluwer, Boston, 1998.

28 Z. Yang, *J. Mol. Evol.* **42,** 294 (1996).

29 J. P. Huelsenbeck, *Syst. Biol.* **46,** 69 (1997).

30 J. Felsenstein, *Syst. Zool.* **27,** 401 (1978).

31 Z. Yang, S. Kumar, and M. Nei, *Genetics* **141,** 1641 (1995).

32 J. Felsenstein, "PHYLIP: Phylogeny Inference Package", Version 3.4. University of Washington, Seattle, WA, 1991.

33 J. Adachi and M. Hasegawa, "MOLPHY", Version 2.2. Institute of Statistical Mechanics, Tokyo, Japan, 1994.

34 Z. Yang, *Comput. Appl. Biosci.* **13,** 555 (1997).

35 N. Galtier and M. Gouy, *Mol. Biol. Evol.* **15,** 871 (1998).

36 J. Felsenstein, *J. Mol. Evol.* **17,** 368 (1981).

37 J. M. Koshi and R. A. Goldstein, *J. Mol. Evol.* **42,** 313 (1996).

38 N. Goldman, *J. Mol. Evol.* **36,** 345 (1993).

39 T. H. Jukes and C. R. Cantor, *in* "Mammalian Protein Metabolism" , (H. N. Munro, ed.), p. 21. Academic Press, New York, 1969.

40 M. Kimura, *J. Mol. Evol.* **16,** 111 (1980).

41 Z. Yang, *J. Mol. Evol.* **39,** 306 (1994).

42 Z. Yang, *J. Mol. Evol.* **39,** 105 (1994).

43 M. J. Bishop and A. E. Friday, *Proc. R. Soc. London. Ser. B* **226,** 271 (1985).

44 M. Hasegawa and M. Fujiwara, *Mol. Phylogenet. Evol.* **2,** 1 (1993).

45 M. O. Dayhoff, R. M. Schwartz, and B. C. Orcutt, *in* "Atlas of Protein Sequence and Structure", (M. O. Dayhoff, ed.), p. 345. National Biomedical Research Foundation, Washington, D. C., 1978.

46 H. Kishino, T. Miyata, and M. Hasegawa, *J. Mol. Evol.* **31,** 151 (1990).

47 D. T. Jones, W. R. Taylor, and J. M. Thornton, *Comp. Appl. Biosci.* **8,** 275 (1992).

48 Y. Cao, J. Adachi, T. Yano, and M. Hasegawa, *Mol. Biol. Evol.* **11,** 593 (1994).

49 J. Adachi and M. Hasegawa, *J. Mol. Evol.* **42,** 459 (1996).

50 B. S. Chang and M. J. Donoghue, in prep.

51 N. Goldman and Z. Yang, *Mol. Biol. Evol.* **11,** 725 (1994).

52 S. V. Muse and B. S. Gaut, *Mol. Biol. Evol.* **11,** 715 (1994).

53 Z. Yang, *Mol. Biol. Evol.* **15,** 568 (1998).

54 R. Nielsen and Z. Yang, *Genetics* **148,** 929 (1998).

55 Z. Yang, R. Nielsen, and M. Hasegawa, *Mol. Biol. Evol.* **15,** 1600 (1998).

56 J. P. Huelsenbeck, *Syst. Biol.* **47,** 519 (1998).

57 J. P. Huelsenbeck and B. Rannala, *Science* **276,** 227 (1997).

58 Z. Yang, N. Goldman, and A. Friday, *Mol. Biol. Evol.* **11,** 316 (1994).

59 W. C. Navidi, G. A. Churchill, and A. von Haeseler, *Mol. Biol. Evol.* **8,** 128 (1991).

60 L. Ferretti, S. S. Karnik, H. G. Khorana, M. Nassal, and D. D. Oprian, *Proc. Natl. Acad. Sci. U. S. A.* **83,** 599 (1986).

61 T. P. Sakmar and H. G. Khorana, *Nucleic Acids Res.* **16,** 6361 (1988).

62 M. A. Kazmi, L. A. Snyder, A. M. Cypess, S. G. Graber, and T. P. Sakmar, *Biochemistry* **XX,** XXX-XXX (2000).

63 D. D. Oprian, A. B. Asenjo, N. Lee, and S. L. Pelletier, *Biochemistry* **30,** 11367 (1991).

64 P. M. Sharp, E. Cowe, D. G. Higgins, D. C. Shiedls, K. H. Wolfe, and F. Wright, *Nucleic Acids Res.* **16,** 8207 (1988).

65 J. F. Kane, *Curr. Opin. Biotechnol.* **6,** 494 (1995).

66 M. D. Forman, R. F. Stack, P. S. Masters, C. R. Hauer, and S. M. Baxter, *Protein Sci.* **7,** 500 (1998).

67 A. Deana, R. Ehrlich, and C. Reiss, *Nucleic Acids Res.* **26,** 4778 (1998).

68 B. Irwin, J. D. Heck, and G. W. Hatfield, *J. Biol. Chem.* **270,** 22801 (1995).

69 S. S. Karnik, M. Nassal, T. Doi, E. Jay, V. Sgaramella, and H. G. Khorana, *J. Biol. Chem.* **262,** 9255 (1987).

70 M. Kozak, *Nucleic Acids Res.* **12,** 857 (1984).

71 D. D. Moore, *in* "Current Protocols in Molecular Biology" , (F. M. Ausubel, R. Brent, R. E. Kingston, D. D. Moore, J. A. Smith, J. G. Seidman, and K. Struhl, eds.), p. 8.2.8. Green Publishing Associates and Wiley-Interscience, New York, 1994.

72 R. W. Graham, T. Atkinson, D. G. Kilburn, R. C. Miller, and R. A. J. Warren, *Nucleic Acids Res.* **21,** 4923 (1993).

73 D. R. Casimiro, P. E. Wright, and H. J. Dyson, *Structure* **5,** 1407 (1997).

74 E. K. Jaffe, M. Volin, C. R. Bronson-Mullins, R. L. Dunbrack, J. Kervinen, J. Martins, J. F. Quinlan, M. H. Sazinsky, E. M. Steinhouse, and A. T. Yeung, *J. Biol. Chem.* **275,** 2619 (2000).

75 C. Prodromou and L. H. Pearl, *Protein Eng.* **5,** 827 (1992).

76 A. N. Vallejo, R. J. Pogulis, and L. R. Pease, *in* "PCR Primer: A Laboratory Manual", (C. D. G. Dieffenbach, ed.), p. 603. Cold Spring Harbor Laboratory Press, New York, 1995.

77 M. D. Matteucci and M. H. Caruther, *J. Am. Chem. Soc.* **103,** 3185 (1981).

78 D. D. Oprian, R. S. Molday, R. J. Kaufman, and H. G. Khorana, *Proc. Natl. Acad. Sci. U. S. A.* **84,** 8874 (1987).

79 P. J. Reeves, R. L. Thurmond, and H. G. Khorana, *Proc. Natl. Acad. Sci. U. S. A.* **93,** 11487 (1996).

80 M. S. Urdea, J. P. Merryweather, G. T. Mullenbach, D. Coit, U. Heberlein, P. Valenzuela, and P. J. Barr, *Proc. Natl. Acad. Sci. U. S. A.* **80,** 7461 (1983).

81 T. Tokunaga, S. Iwai, H. Gomi, K. Kodama, E. Ohtsuka, M. Ikehara, O. Chisaka, and K. Matsubara, *Gene* **39,** 117 (1985).

82 T. Tanaka, S. Kimura, and Y. Ota, *Nucleic Acids Res.* **15,** 3178 (1987).

83 G. B. Cohen, T. Yang, P. R. Robinson, and D. D. Oprian, *Biochemistry* **32,** 6111 (1993).

84 T. A. Zvyaga, K. C. Min, M. Beck, and T. P. Sakmar, *J. Biol. Chem.* **268,** 4661 (1993).

85 T. P. Sakmar, R. R. Franke, and H. G. Khorana, *Proc. Natl. Acad. Sci. U. S. A.* **86,** 8309 (1989).

86 K. Fahmy and T. P. Sakmar, *Biochemistry* **32,** 7229 (1993).

87 S. W. Lin, T. P. Sakmar, R. R. Franke, H. G. Khorana, and R. A. Mathies, *Biochemistry* **31,** 5105 (1992).

88 K. Fahmy, F. Jäger, M. Beck, T. A. Zvyaga, T. P. Sakmar, and F. Siebert, *Proc. Natl. Acad. Sci. U. S. A.* **90,** 10206 (1993).

89 J. F. Resek, Z. T. Farahbakhsh, W. L. Hubbell, and H. G. Khorana, *Biochemistry* **32,** 12025 (1993).

90 J. W. Lewis, I. Szundi, W. Y. Fu, T. P. Sakmar, and D. S. Kliger, *Biochemistry* **39,** 599 (2000).

91 O. P. Ernst, C. K. Meyer, E. P. Marin, P. Henklein, W. Y. Fu, T. P. Sakmar, and K. P. Hofmann, *J. Biol. Chem.* **275,** 1937 (2000).

92 M. Eilers, P. J. Reeves, W. Ying, H. G. Khorana, and S. O. Smith, *Proc. Natl. Acad. Sci. U. S. A.* **96,** 487 (1999).

93 T. Chan, M. Lee, and T. P. Sakmar, *J. Biol. Chem.* **267,** 9478 (1992).

94 Z. Wang, A. B. Asenjo, and D. D. Oprian, *Biochemistry* **32,** 2125 (1993).

95 E. A. Zhukovsky and D. D. Oprian, *Science* **246,** 928 (1989).

96 T. P. Sakmar, R. R. Franke, and H. G. Khorana, *Proc. Natl. Acad. Sci. U. S. A.* **88,** 3079 (1991).

97 S. W. Lin, G. G. Kochendoerfer, K. S. Carroll, D. Wang, R. A. Mathies, and T. P. Sakmar, *J. Biol. Chem.* **273,** 24583 (1998).

98 G. G. Kochendoerfer, S. W. Lin, T. P. Sakmar, and R. A. Mathies, *Trends Biochem. Sci.* **24,** 300 (1999).

99 K. Fahmy and T. P. Sakmar, *Biochemistry* **32,** 9165 (1993).

100 M. Hasegawa, H. Kishino, and T. Yano, *J. Mol. Evol.* **22,** 672 (1985).

101 C. L. Chio, R. F. Drong, D. T. Riley, G. S. Gill, J. L. Slightom, and R. M. Huff, *J. Biol. Chem.* **269,** 11813 (1994).

102 K. Noda, Y. Saad, A. Kinoshita, T. P. Boyle, R. M. Graham, A. Husain, and S. S. Karnik, *J. Biol. Chem.* **270,** 2284 (1995).

103 D. R. Nussenzveig, C. N. Thaw, and M. C. Gershengorn, *J. Biol. Chem.* **269,** 28123 (1994).