Nucleic Acids Research, 2002, Vol. 30, No. 10 ***e43***

DNAWorks: an automated method for designing oligonucleotides for PCR-based gene synthesis

David M. Hoover and Jacek Lubkowski*

Macromolecular Crystallography Laboratory, National Cancer Institute at Frederick, MD 21702, USA

Received November 27, 2001; Revised February 15, 2002; Accepted March 10, 2002

ABSTRACT

The availability of sequences of entire genomes has dramatically increased the number of protein targets, many of which will need to be overexpressed in cells other than the original source of DNA. Gene synthesis often provides a fast and economically efficient approach. The synthetic gene can be optimized for expression and constructed for easy mutational manipulation without regard to the parent genome. Yet design and construction of synthetic genes, especially those coding for large proteins, can be a slow, difficult and confusing process. We have written a computer program that automates the design of oligonucleotides for gene synthesis. Our program requires simple input information, i.e. amino acid sequence of the target protein and melting temperature (needed for the gene assembly) of synthetic oligonucleotides. The program outputs a series of oligonucleotide sequences with codons optimized for expression in an organism of choice. Those oligonucleotides are characterized by highly homogeneous melting temperatures and a minimized tendency for hairpin formation. With the help of this program and a two-step PCR method, we have successfully constructed numerous synthetic genes, ranging from 139 to 1042 bp. The approach presented here simplifies the production of proteins from a wide variety of organisms for genomics-based studies.

INTRODUCTION

In the post-genomic era, thousands of unknown proteins have become available for study. While in theory the structures and functions of many of these proteins may be determined by comparative analysis (1), in most cases, overexpression and purification of target proteins will be necessary (2,3). Although the use of naturally occurring genes might appear to be the quickest approach, many such genes will prove to be suboptimal for cloning and overexpression in heterologous systems like *Escherichia coli* or yeast. The potential problems include high G+C content, codon bias and complex intron/exon structures. An approach to overcoming the complications in cloning is gene synthesis. In this approach, the protein coding sequence can be directly optimized for the expression system of choice. Variants of this strategy include oligonucleotide ligation (4), the *FokI* method (5) and self-priming PCR (6). A particularly appealing method, due to its inherent simplicity, is assembly PCR (7). This involves generating overlapping oligonucleotides which, when assembled, form the template for the gene of interest. The oligonucleotides are then repetitively extended by PCR, to assemble the full-length gene in a single step.

While this method is simple in principle, in practice numerous complications can lead to errors in the synthesis. To reduce the possibility of errors during oligonucleotide synthesis, the oligonucleotides should be rather short, yet they must still be long enough to provide stable priming overlaps. Any deleterious secondary structures in the oligonucleotides also need to be avoided. Therefore, for large proteins with coding sequences of >300 nt, the process of designing these oligonucleotides is tedious and confusing. In the case of a single gene, the problem can be attacked by manual design, but for projects where high throughput is required (i.e. structural genomics) an automated strategy for synthetic gene design is needed.

In this report we describe a program, referred to as DNAWorks, which automates the process of oligonucleotide design for synthetic gene construction. As an input, the program requires an amino acid sequence of the target protein as well as any desired flanking sequences (for directional cloning). It then creates a set of oligonucleotide sequences (composing the gene of interest) that have been optimized to match the codon bias of the chosen host for expression and highly homogeneous melting temperatures of all overlapping oligonucleotide sections. Once synthesized, these oligonucleotides are combined and assembled in a two-step PCR protocol to form the synthetic gene. We have tested this protocol for 11 proteins encoded by genes ranging in lengths between 139 and 1042 nt, and have demonstrated its potential for high throughput gene synthesis.

MATERIALS AND METHODS

Materials

Oligonucleotides were purchased from either Life Technologies, Operon Technologies or Integrated DNA Technologies. They were synthesized on a 50 nmol scale and stock solutions were prepared at a concentration of ~1 mg/ml in water without additional purification. All restriction enzymes were obtained from New England Biolabs. Cloned *Pyrococcus furiosus* (*Pfu*) DNA polymerase was purchased from Stratagene. The Gateway cloning materials were from Invitrogen.

*To whom correspondence should be addressed. Tel: +1 301 846 5494; Fax: +1 301 846 7101; Email: jacek@ncifcrf.gov

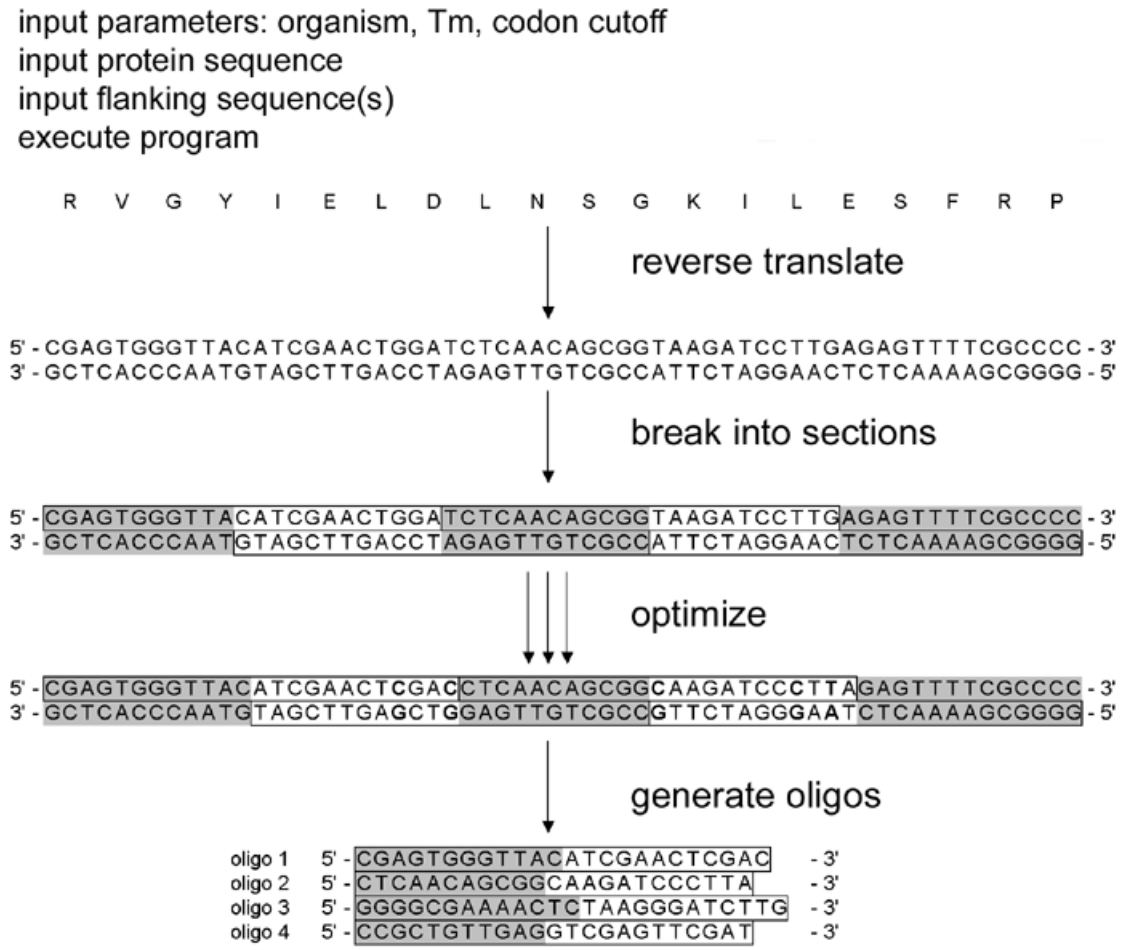

Figure 1. Outline of DNAWorks. Sections of equal melting temperature are delineated by shading, while oligonucleotides are outlined. During the optimization, silent mutations can be generated, as shown by nucleotides in bold, as well as shifts of section boundaries.

Target proteins

Amino acid sequences for the following proteins were obtained from GenBank: human CC-chemokine CCL27, also known as CTACK/ALP/ILC (hCCL27) (8–11), viral interleukin-8 (vIL-8) (12), lettuce mosaic virus protease (LPP) (13), potato virus A protease (PPP) (14), murine monocyte chemoattractant protein-5 (mMCP-5) (15), human monocyte chemoattractant protein-4 (hMCP-4) (16), a double mutant (T10C, S33C) of human lymphotactin (hLT) (17), human β-defensin-1 (hBD1) (18), -2 (hBD2) (19), -3 (hBD3) (20), and the chemokine receptor human CXCR4 (hCXCR4) (21).

Programming

The source code of DNAWorks was written in FORTRAN90 using Compaq Visual Fortran Professional Edition 6.1.0 (www.compaq.com/fortran/index.html). All executables were tested on a PC with a 650 MHz Pentium III processor running Windows 2000.

Program outline

DNAWorks was created with the idea of simplicity of use and moderate flexibility necessary to suit most common scenarios of designing the synthetic genes. The overall scheme of the input and output for the program is shown in Figure 1. To begin, several parameters used during the optimization of the resulting oligonucleotide sequences need to be defined. These parameters include melting temperature (currently limited to be within 58–70°C), the codon optimization scheme (explained below), the target genome [currently limited to *E.coli* class II genes, or those that are expressed at high levels during exponential growth, as determined by the factorial correspondence analysis (22)], and whether or not the program will check for and eliminate hairpins that may form within individual oligonucleotides. Codon frequencies, or the percent that a particular codon is used by the chosen organism to code for an amino acid, are read from an external text file that can be easily modified. Codon optimization represents the stringency in allowing only those codons equal to or greater than a threshold frequency to be used during optimization; for example, a value of 20 allows only codons of frequency 20% or higher to be used. However, to allow for optimization the program will always include codons with the highest and second highest frequencies. Thus, at a frequency value of 50, 18 of 20 amino acids (from *E.coli*) will be coded for by two codons each.

A protein sequence is input to DNAWorks, which then reverse-translates the protein sequence into a series of highest-frequency codons for the organism chosen. The sequence can

be entered either manually or as a file in FASTA or simple text format. Any desired flanking sequences can then be input and incorporated into the synthetic sequence. Flanking sequences for Gateway cloning (*attB1* and *attB2*) and short flanking sequences containing *Nde*I and *Bam*HI sites for cloning into pET vectors are presented automatically in the program as choices for the user. All sequences and changes in codon usage are updated both in the onscreen output and in the logfile.

Once the input information is entered and confirmed, the optimization algorithm is executed. At the beginning of its course, the back-translated initial sequence of the synthetic gene is divided into an odd number of contiguous sections, which are characterized by near-equal melting temperatures. Algorithms used to calculate the oligonucleotide melting temperatures are based on the nearest-neighbor model (23). A score is then assigned to each section on the basis of codon frequency, hairpin formation (within each oligonucleotide), and deviations from the desired melting temperature and size. The possibility of hairpin formation is determined by comparison of oligonucleotide sequences for individual sections and the sequences of reverse complements for adjacent sections. The number, length, composition (G+C content) and relative position of matches between those sequences is used for calculating the 'hairpin-formation' score. The scores associated with deviations of melting temperatures and sizes of individual sections beyond the input tolerances are calculated using a parabolic function. This approach restrains the values for those parameters from drifting too far from the numbers requested in the input. The overall score for a synthetic sequence is the sum of scores for all individual sections. In the ideal case, in which only the highest frequency codons are utilized, when the possibility of hairpin formation is eliminated, and all sections are characterized by requested (uniform) size and melting temperature, the score would be zero. Typically, there are multiple possibilities for the positions of sections within a sequence, due to the variability of the unpaired end lengths; therefore, all possible arrangements of sections within a gene sequence are tested, and the set of sections characterized by the lowest score is used as the initial sequence subjected to optimization. The results of the calculations as well as several diagnostic parameters are provided in the form of a logfile, created at the time of the program execution.

The coding (non-flanking) region of the synthetic sequence is then subjected to optimization. Because of the high number of possible sequences (assuming just two possible codons per amino acid residue, a protein consisting of 100 residues can be encoded by 2^{100} different genes), we employed a stochastic method of optimization (a variant of a simulated annealing algorithm), rather than a deterministic one like steepest gradient (24). Two primary benefits of this strategy are its robustness against premature termination as a result of entrapment in local minima, and time efficiency. Silent mutations of individual residues are created within the gene by randomly swapping codons with others from the available pool of codons for that residue. The choice of residues to be mutated is made based on their local section scores, and therefore residues residing within sections of high individual scores are more likely to be mutated than those residing within lower scoring sections. During each optimization step, the boundaries of sections are also redefined. The number of mutations as well as changes in the section sizes during each step diminish as the optimization process progresses (i.e. a parameter called 'temperature' in the simulated annealing protocol decreases). At the end of every step, the score is recalculated for each section, resulting in a new overall score for the entire sequence. If the overall score improves (decreases), the sequence is kept; otherwise, the choice between the new and the old sequences is controlled by the Boltzman distribution for a particular 'temperature' parameter, allowing worse scores at higher 'temperatures' while at low 'temperatures' only sequences with improved overall scores are accepted. When no further improvements of the overall score are obtained within a predefined number of trials (automatically determined by the program, depending on the size of the sequence), the output is generated and the program terminates. The initial and final synthetic gene sequences, a list of optimized oligonucleotide sequences, along with the scores for each section from both the initial and final sequence, are generated and written out.

Gene assembly and amplification

The assembly of the synthetic gene from component oligonucleotides was performed according to a previously described protocol (7); therefore, we are outlining only its general steps. Equal volumes of oligonucleotide solutions (each at a concentration of ~1 mg/ml) were mixed together and diluted with water to a final concentration of ~1 ng/µl for each oligonucleotide. The oligonucleotide mixture was diluted 5-fold with the PCR solution. The final concentrations of components were 0.2 ng/µl for each oligonucleotide, 20 mM for Tris–HCl (pH 8.8), 10 mM for KCl, 10 mM for $(NH_4)_2SO_4$, 6 mM for $MgSO_4$, 0.1% (v/v) for Triton X-100, 0.1 mg/ml for bovine serum albumin, 0.2 mM for each dNTP and 2.5 U for *Pfu* polymerase. The PCR protocol for gene assembly began with one 5 min denaturation step of 95°C, during which the polymerase was added to avoid any possible mispriming ('hot start' PCR). This step was followed by 25 cycles of a denaturation temperature 95°C for 30 s, a variable annealing temperature (dependent on the melting temperature chosen in the program) for 30 s and an extension temperature of 72°C for 1.5 min. The last step in this protocol was an incubation cycle at 72°C for 10 min. For gene amplification, 1 µl of the mixture resulting from the gene assembly reaction was used as the template, with the outermost oligonucleotides used as primers. The PCR protocol for gene amplification was essentially the same as gene assembly, except that the annealing temperature was raised to 62°C.

Cloning and sequencing

The synthetic gene fragments, purified by gel extraction, were either integrated into the vector pDONR201 using the Gateway cloning system (Invitrogen) or digested with appropriate restriction endonucleases and ligated into the vector pAED4 (25). The ligation products were transformed into DH5α *E.coli* cells and selected for on LB plates with 50 µg/ml ampicillin (when ligated into pAED4) or 35 µg/ml kanamycin (when integrated into pDONR201). The plasmids isolated were screened by either restriction digest analysis or PCR using primers complementary to vector sequences flanking the synthetic gene, and plasmids containing the gene of interest were sequenced in both the forward and reverse directions.

Table 1. Synthetic gene data

Gene	Gene length (bp)	Average T_m (min, max) (°C)	No. of oligonucleotides	Average oligonucleotide length (nt)
hBD1	139	60 (55, 62)	6	40
hBD2	154	62 (57, 64)	10	37
hBD3	166	52 (49, 54)	10	32
hMCP-4	300	56 (53, 66)	12	38
mMCP-5	330	65 (60, 69)	14	44
hCCL27	351	64 (59, 72)	16	42
HLT	360	64 (61, 68)	16	42
vIL-8	365	58 (55, 60)	18	36
LPP	756	56 (50, 59)	46	32
PPP	771	58 (51, 64)	40	38
hCXCR4	1042	67 (54, 72)	44	47

RESULTS

Program output and testing

Sequences of mature target proteins were obtained from GenBank and used as input for DNAWorks. For each protein, the program was executed with various combinations of melting temperature and codon frequency restraints in order to generate a synthetic sequence characterized by the lowest score. Whereas not all codons present in the resulting synthetic gene are of the highest frequency, due to the simultaneous optimization of other parameters, we found that only a small fraction of suboptimal frequency codons was utilized. The sets of calculated oligonucleotides were crosschecked using the program GCG (26) to verify the uniformity of melting temperatures across all oligonucleotides, the propensity to form hairpins, and the correctness of the sequences. The melting temperatures calculated by DNAWorks and GCG were found to differ by no more than 4°C. These differences are within the errors seen for nearest-neighbor predictive models (27). No stable hairpins were detected and all sequences correctly coded for the desired protein sequence. Table 1 shows the melting temperatures averaged over all overlaps as well as lengths of individual oligonucleotides for each of the tested genes.

Besides the clock frequency of the central processing unit, the execution time of DNAWorks is dependent on the length of the sequence to be optimized. The calculation time for a 200 bp sequence is ~5 min, while for 1000 bp it approaches ~15 min, when the program is executed on a PC with a 650 MHz Pentium III processor. Consequently, for an average protein of approximately 200–300 amino acids, the set of oligonucleotides with overlap melting temperatures between 58 and 65°C can be generated within 1–2 h using a typical personal computer.

Gene synthesis

The procedure for synthesizing genes is relatively fast and streamlined, with the most laborious step being the solubilization of lyophilized oligonucleotides. Although the initial step of gene assembly may result in a spectrum of incorrect products, in all cases tested by us the final gene amplification reaction gave rise to a dominant single band of the correct size as shown by gel electrophoresis (Fig. 2). This dramatic improvement is likely due to the fact that only a very small population of the assembly products contains the outermost sequences in a productive arrangement, and that this population is dominated by the correct assembly. The synthetic genes were cloned either directly into the expression vector pAED4 using *Nde*I and *Bam*HI sites, or indirectly through the Gateway donor vector pDONR201 (for sequencing) and then into pAED4 as an *Nde*I/*Bam*HI fragment. With the aid of the Gateway cloning system, ~90–95% of the screened transformants cloned contained the correct size insert and could be subjected to sequencing. Thus, using the Gateway system, the gene can be made, cloned into an expression vector, and subjected to sequencing within 2 days. Otherwise, the gene must be digested and then ligated before being transformed, adding an additional day for ligation and transformation. Once cloned, between 60 and 90% of the screened transformants cloned contained the correct size insert and could be subjected to sequencing.

All screened plasmids were sequenced in both the forward and reverse direction, and an error in gene synthesis was verified only if it occurred in both directions. This eliminated any false negatives due to inaccurate sequencing runs. All errors seen were either single nucleotide deletion, single nucleotide insertion or mismatch mutations. As shown in Table 2, deletion and mismatch mutations dominated, with only one insertion mutation found in hLT. The overall ratio of correct sequences to total sequences is ~2/3. The overall rate of error incorporation is 1.8 errors per 1000 bases sequenced, with the worst case being mMCP-5 with five errors per 1000 bases sequenced.

While correct sequences for most of the synthetic genes were found after sequencing enough clones, two genes required additional work to obtain the final product. Two clones of LPP were sequenced, and deletions were found on one end of the first gene and on the other end of the second gene. The presence of a unique restriction site between the two deletions allowed the correct halves to be digested and combined to form a correct complete gene. Three clones of hLT were sequenced and errors were found in each of the plasmids. Because the quality of the sequencing chromatograms was much better for the plasmid with a single insertion, correction of this clone by site-directed PCR using the original oligonucleotides was attempted. The resulting plasmids still contained the original insertion, therefore new oligonucleotides were synthesized and used for site-directed PCR. This corrected the insertion and resulted in the final synthetic gene. However, it should be pointed out that only a small number of clones (two for LPP, three for hLT) were sequenced; screening a larger number of clones may have uncovered an error-free sequence.

DISCUSSION

A large-scale, high-throughput means of generating numerous proteins depends dramatically on robust, reliable and routine methods of cloning genes and expressing proteins. An approach utilizing synthetic genes presents an attractive option. The strategy described in this report, utilizing the appropriate computer software, minimizes the effort involved in designing oligonucleotides used for PCR-based gene synthesis. In many cases, the time required between identification

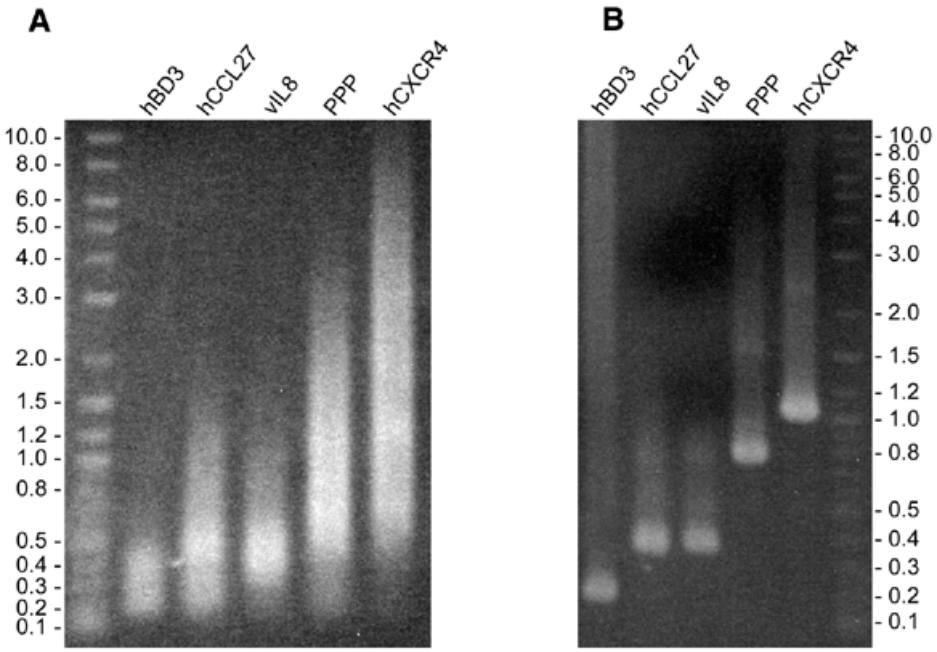

Figure 2. (**A**) Agarose gel electrophoresis of the gene assembly and (**B**) gene amplification products for selected genes hBD3, hCCL27, vIL-8, PPP and hCXCR4. The PCR protocol for gene amplification is as described in Materials and Methods. The annealing temperatures during the gene amplification reactions were as follows: hBD3, 54°C; hCCL27, 58°C; vIL-8, 55°C; PPP, 54°C; hCXCR4, 58°C.

Table 2. Error analysis data

Gene	Correct/ total[a]	No. of deletions	No. of insertions	No. of mismatches	Errors/ clone[b]	Total bp sequenced[c]	Errors/kb[d]
hBD1	3/5	2	0	1	0.6	695	4.3
hBD2	3/3	0	0	0	0	462	0
hBD3	8/8	0	0	0	0	1328	0
hMCP-4	3/5	2	0	2	0.8	1500	2.7
mMCP-5	1/3	1	0	4	1.67	990	5.1
hCCL27	1/2	0	0	1	0.5	702	1.4
hLT(mut)	0/3[e]	1	1	1	1	1080	2.8
vIL-8	3/5	2	0	2	0.8	1825	2.2
LPP	0/2[e]	2	0	0	1	1512	1.3
PPP	1/1	0	0	0	0	771	0
hCXCR4	1/1	0	0	0	0	1140	0
Average	63%				0.58		1.8

[a]The ratio of synthetic gene clones verified as correct by sequencing to the total number of clones sequenced.
[b]The total number of errors (single nucleotide deletion, insertion or mismatch mutations) divided by the total number of clones sequenced.
[c]The sum of base pairs sequenced for a set of clones.
[d]The total number of errors divided by the sum of base pairs sequenced.
[e]The correct gene was generated using additional methods. See text for details.

of a protein sequence and obtaining an expression product can be as short as 1 week. Unfortunately, protein expression is strongly dependent on post-translational events. Thus, although the synthetic gene is optimized for expression, the yield of any particular protein may vary considerably.

The frequency of nucleotide errors is largely dependent on the quality of the oligonucleotides, rather than the fidelity of the polymerase. In published results of *Pfu* fidelity assays (28), it was found that 0.42% of transformants were *lacI*$^{-}$ after amplifying the *lacI* gene within a 1.9 kb *Eco*RI fragment.

Assuming that within this fragment there are 349 sensitive nucleotides in the *lacI* gene which, when mutated, would cause the *lacI* gene product to become non-functional, it can be estimated that under standard operating conditions, *Pfu*-mediated PCR will minimally give rise to approximately 0.012 errors per kilobase of cloned PCR product [(0.0042 × 1000)/349]. Because errors could occur outside the sensitive sites without affecting *lacI* activity, a better estimate of total errors (both silent and destructive mutations) would be approximately 5.4 times higher, 0.065 errors per kilobase of cloned PCR product. The overall error rate seen from the data presented in Table 2 is 1.8 errors per kilobase of sequenced synthetic gene product, 32 times higher than that estimated from PCR-mediated errors alone.

Theoretically, the likelihood of incorporating errors into oligonucleotides increases with the size of the oligonucleotides, and therefore the lengths of synthetic oligonucleotides were kept to a minimum. However, there was little correlation between the average size of the oligonucleotides and the frequency of error generation. Also, there is little correlation with the length of the gene and the number of errors. As shown with the hCXCR4 and PPP genes, longer genes did not contain any errors, while shorter genes, such as mMCP-5 and hLT, contained multiple errors per clone. Thus, we find that random sequence errors introduced during oligonucleotide synthesis depend mostly on the quality of the synthetic source, and not systematically from either the length of the oligonucleotides or the number of oligonucleotides needed for the synthetic gene.

Because most of the synthetic gene is constructed from two overlapping oligonucleotide chains, it is highly unlikely that the same error will occur in complementary oligonucleotides. Thus, any error that does arise in a single oligonucleotide has at most a 50% chance of being incorporated into the synthetic gene. Interestingly, in our experiments insertion errors occurred with the lowest frequency. Deletion and mismatch errors seemed to be dominant and arose in most of the transformants. This is probably due to the technical process of oligonucleotide synthesis. Most of the errors could be overcome by either screening a larger number of transformants or correcting the mistakes by site-directed mutagenesis performed on the original 'faulty' genes. However, for longer genes it might be best to correct multiple deletion errors by resynthesis of the oligonucleotides.

Mispriming (oligonucleotides priming at unintended sites) could arise in specific cases. In the genes investigated here, no mispriming was found. The use of longer oligonucleotides, at the expense of increasing errors, should minimize mispriming, allowing full gene synthesis in a single step. Based on our experience, we suggest that the oligonucleotide overlap melting temperatures should be set to at least 58°C and can be raised to as much as 70°C to minimize mispriming. Combining longer oligonucleotide overlaps and 'touchdown' PCR (29) can help eliminate mispriming. However, proteins containing multiple repeats of amino acid sequences would produce highly similar DNA sequences, increasing the chance of mispriming. In such cases, it would be beneficial to screen for possible mispriming before synthesizing the oligonucleotides. At present, this feature is not available in DNAWorks, but will be incorporated into future versions. However, possible mispriming can be analyzed with several other available software tools. Although the assembly protocol works relatively well for small genes (shorter than ~500 bp), in the case of longer genes some problems begin to arise. In such cases, PCR mispriming could become more prevalent, as reported in other studies (30,31). These problems can be overcome by dividing the gene synthesis into regions of 200–300 nt, and then amplifying the gene from combined purified gene fragments with the outer primers.

Overall, once the synthetic gene is designed using DNAWorks and the oligonucleotides are synthesized, it should take 3–4 days to clone the gene and submit for sequencing. Based on the results described here, from six initial transformants at least one of these clones will have the correct sequence. Several features, currently not implemented in DNAWorks, will likely increase the utility of this program. Among these are the ability to insert restriction sites within the sequence to facilitate restriction analysis and mutagenesis work on a particular gene, and monitoring long-range repeats to minimize mispriming in large genes (currently mispriming can be avoided by breaking the gene assembly step into smaller steps). The availability of a wider range of other multiple cloning system flanking sequences would give the user flexibility in cloning the synthetic gene directly from the PCR product. These and other extensions will be incorporated into future versions of DNAWorks.

The executable version of DNAWorks for Windows can be downloaded from the website http://mcl1.ncifcrf.gov/lubkowski.html.

ACKNOWLEDGEMENTS

We would like to acknowledge Dr David Waugh, Dr Howard Peters and Rachel Kapust for their support and suggestions. This research was sponsored in part by the Intramural AIDS Targeted Antiviral Program of the Office of the Director, National Institutes of Health (J.L.).

REFERENCES

1. Bork,P., Dandekar,T., Diaz-Lazcoz,Y., Eisenhaber,F., Huynen,M. and Yuan,Y. (1998) Predicting function: from genes to genomes and back. *J. Mol. Biol.*, **283**, 707–725.
2. Baxter,S.M. and Fetrow,J.S. (2001) Sequence- and structure-based protein function prediction from genomic information. *Curr. Opin. Drug Discov. Dev.*, **4**, 291–295.
3. Gerlt,J.A. and Babbitt,P.C. (2000) Can sequence determine function? *Genome Biol.*, **1**, 1–10.
4. Heyneker,H.L., Shine,J., Goodman,H.M., Boyer,H.W., Rosenberg,J., Dickerson,R.E., Narang,S.A., Itakura,K., Lin,S. and Riggs,A.D. (1976) Synthetic lac operator DNA is functional *in vivo*. *Nature*, **263**, 748–752.
5. Mandecki,W. and Bolling,T.J. (1988) FokI method of gene synthesis. *Gene*, **68**, 101–107.
6. Dillon,P.J. and Rosen,C.A. (1990) A rapid method for the construction of synthetic genes using the polymerase chain reaction. *Biotechniques*, **9**, 298, 300.
7. Stemmer,W.P., Crameri,A., Ha,K.D., Brennan,T.M. and Heyneker,H.L. (1995) Single-step assembly of a gene and entire plasmid from large numbers of oligodeoxyribonucleotides. *Gene*, **164**, 49–53.
8. Morales,J., Homey,B., Vicari,A.P., Hudak,S., Oldham,E., Hedrick,J., Orozco,R., Copeland,N.G., Jenkins,N.A., McEvoy,L.M. *et al.* (1999) CTACK, a skin-associated chemokine that preferentially attracts skin-homing memory T cells. *Proc. Natl Acad. Sci. USA*, **96**, 14470–14475.
9. Broxmeyer,H.E., Kim,C.H., Cooper,S.H., Hangoc,G., Hromas,R. and Pelus,L.M. (1999) Effects of CC, CXC, C, and CX3C chemokines on proliferation of myeloid progenitor cells, and insights into SDF-1-induced chemotaxis of progenitors. *Ann. N. Y. Acad. Sci.*, **872**, 142–162.

10. Baird,J.W., Nibbs,R.J., Komai-Koma,M., Connolly,J.A., Ottersbach,K., Clark-Lewis,I., Liew,F.Y. and Graham,G.J. (1999) ESkine, a novel beta-chemokine, is differentially spliced to produce secretable and nuclear targeted isoforms. *J. Biol. Chem.*, **274**, 33496–33503.
11. Ishikawa-Mochizuki,I., Kitaura,M., Baba,M., Nakayama,T., Izawa,D., Imai,T., Yamada,H., Hieshima,K., Suzuki,R., Nomiyama,H. *et al.* (1999) Molecular cloning of a novel CC chemokine, interleukin-11 receptor alpha-locus chemokine (ILC), which is located on chromosome 9p13 and a potential homologue of a CC chemokine encoded by molluscum contagiosum virus. *FEBS Lett.*, **460**, 544–548.
12. Parcells,M.S., Lin,S.F., Dienglewicz,R.L., Majerciak,V., Robinson,D.R., Chen,H.C., Wu,Z., Dubyak,G.R., Brunovskis,P., Hunt,H.D. *et al.* (2001) Marek's disease virus (MDV) encodes an interleukin-8 homolog (vIL-8): characterization of the vIL-8 protein and a vIL-8 deletion mutant MDV. *J. Virol.*, **75**, 5159–5173.
13. Revers,F., Yang,S.J., Walter,J., Souche,S., Lot,H., Le Gall,O., Candresse,T. and Dunez,J. (1997) Comparison of the complete nucleotide sequences of two isolates of lettuce mosaic virus differing in their biological properties. *Virus Res.*, **47**, 167–177.
14. Puurand,U., Makinen,K., Paulin,L. and Saarma,M. (1994) The nucleotide sequence of potato virus A genomic RNA and its sequence similarities with other potyviruses. *J. Gen. Virol.*, **75**, 457–461.
15. Jia,G.Q., Gonzalo,J.A., Lloyd,C., Kremer,L., Lu,L., Martinez,A., Wershil,B.K. and Gutierrez-Ramos,J.C. (1996) Distinct expression and function of the novel mouse chemokine monocyte chemotactic protein-5 in lung allergic inflammation. *J. Exp. Med.*, **184**, 1939–1951.
16. Garcia-Zepeda,E.A., Combadiere,C., Rothenberg,M.E., Sarafi,M.N., Lavigne,F., Hamid,Q., Murphy,P.M. and Luster,A.D. (1996) Human monocyte chemoattractant protein (MCP)-4 is a novel CC chemokine with activities on monocytes, eosinophils, and basophils induced in allergic and nonallergic inflammation that signals through the CC chemokine receptors (CCR)-2 and -3. *J. Immunol.*, **157**, 5613–5626.
17. Kennedy,J., Kelner,G.S., Kleyensteuber,S., Schall,T.J., Weiss,M.C., Yssel,H., Schneider,P.V., Cocks,B.G., Bacon,K.B. and Zlotnik,A. (1995) Molecular cloning and functional characterization of human lymphotactin. *J. Immunol.*, **155**, 203–209.
18. Bensch,K.W., Raida,M., Magert,H.J., Schulz-Knappe,P. and Forssmann,W.G. (1995) hBD-1: a novel beta-defensin from human plasma. *FEBS Lett.*, **368**, 331–335.
19. Harder,J., Bartels,J., Christophers,E. and Schröder,J.M. (1997) A peptide antibiotic from human skin. *Nature*, **387**, 861.
20. Harder,J., Bartels,J., Christophers,E. and Schröder,J.M. (2001) Isolation and characterization of human beta-defensin-3, a novel human inducible peptide antibiotic. *J. Biol. Chem.*, **276**, 5707–5713.
21. Federsppiel,B., Melhado,I.G., Duncan,A.M., Delaney,A., Schappert,K., Clark-Lewis,I. and Jirik,F.R. (1993) Molecular cloning of the cDNA and chromosomal localization of the gene for a putative seven-transmembrane segment (7-TMS) receptor isolated from human spleen. *Genomics*, **16**, 707–712.
22. Medigue,C., Rouxel,T., Vigier,P., Henaut,A. and Danchin,A. (1991) Evidence for horizontal gene transfer in *Escherichia coli* speciation. *J. Mol. Biol.*, **222**, 851–856.
23. SantaLucia,J.,Jr (1998) A unified view of polymer, dumbbell, and oligonucleotide DNA nearest-neighbor thermodynamics. *Proc. Natl Acad. Sci. USA*, **95**, 1460–1465.
24. Press,W.H., Teukolsky,S.A., Vetterling,W.T. and Flannery,B.P. (1986) *Numerical Recipes in FORTRAN*. Press Syndicate of the University of Cambridge, Cambridge, UK, pp. 387–436.
25. Studier,F.W., Rosenberg,A.H., Dunn,J.J. and Dubendorff,J.W. (1990) Use of T7 RNA polymerase to direct expression of cloned genes. *Methods Enzymol.*, **185**, 60–89.
26. Womble,D.D. (2000) GCG: the Wisconsin Package of sequence analysis programs. *Methods Mol. Biol.*, **132**, 3–22.
27. SantaLucia,J.,Jr, Allawi,H.T. and Seneviratne,P.A. (1996) Improved nearest-neighbor parameters for predicting DNA duplex stability. *Biochemistry*, **35**, 3555–3562.
28. Cline,J., Braman,J.C. and Hogrefe,H.H. (1996) PCR fidelity of *Pfu* DNA polymerase and other thermostable DNA polymerases. *Nucleic Acids Res.*, **24**, 3546–3551.
29. Don,R.H., Cox,P.T., Wainwright,B.J., Baker,K. and Mattick,J.S. (1991) 'Touchdown' PCR to circumvent spurious priming during gene amplification. *Nucleic Acids Res.*, **19**, 4008.
30. Baedeker,M. and Schulz,G.E. (1999) Overexpression of a designed 2.2 kb gene of eukaryotic phenylalanine ammonia-lyase in *Escherichia coli*. *FEBS Lett.*, **457**, 57–60.
31. Withers-Martinez,C., Carpenter,E.P., Hackett,F., Ely,B., Sajid,M., Grainger,M. and Blackman,M.J. (1999) PCR-based gene synthesis as an efficient approach for expression of the A+T-rich malaria genome. *Protein Eng.*, **12**, 1113–1120.