Efficient in silico designing of oligonucleotides for artificial gene synthesis

Abhishek D. Garg
Faculty of Biological Sciences, University of Leeds, Leeds LS2 9JT, United Kingdom

Method Article

Keywords: oligonucleotides, artificial gene synthesis, secondary structures, in silico, DNA

Posted Date: January 22nd, 2009

DOI: https://doi.org/10.1038/nprot.2009.15

Abstract

Introduction

‘Artificial/Chemical Gene-Synthesis’ has become an important contemporary technique, which has been refined several folds owing to application of Polymerase Chain Reaction \(PCR) \(Engels, 2005). This procedure has been used in past for artificially synthesizing a number of genes \(Engels, 2005). The principle of Artificial Gene-Synthesis has been elucidated in Figure 1 \(Casimiro et al, 1997; Engels, 2005; Engels & Uhlmann, 1989; Tsuchiya et al, 2006). Technically speaking, this procedure involves two phases i.e. upstream-phase and downstream-phase. In the upstream phase, the oligonucleotides \(whose sequence is derived from target gene to be synthesized) are designed manually or in silico while in the downstream phase, these designed oligonucleotides are chemically synthesized and assembled together to give stable ds-DNA duplexes/target gene. The peculiarity of these oligonucleotides is that, each forward \(5’→3’) oligonucleotide overlaps with the corresponding reverse \(3’→5’) oligonucleotide in around 20 terminal nucleotide-bases, thus allowing formation of an annealed-product. Current protocol focuses on the upstream-phase since it’s a success-defining phase for artificial gene-synthesis. In the current protocol, various state-of-art bioinformatics tools have been brought together to provide a comprehensive-approach towards in silico designing of oligonucleotides. Further, for the first time, DNA/mRNA secondary-structure formation analysis has been included in such a protocol.

Reagents

None

Equipment

Computer: A PC loaded with MS Windows OS \(2000/Me/XP/Vista). Here, ‘Mozilla Firefox v.3.0.5’ is specially recommended for browsing related to softwares mentioned below. Computer Softwares: 1. DNA mfold v.3.2 URL: "http://frontend.bioinfo.rpi.edu/applications/mfold/cgi-bin/dna-form1.cgi":http://www.chromdb.org/rnai/vector_info.html 2. FoldRNA URL: "http://www.softberry.com/berry.phtml?topic=foldrna&group=programs&subgroup=rnastruct":http://www.softberry.com/berry.phtml?topic=foldrna&group=programs&subgroup=rnastruct 3. Gene Design β2.0 URL: "http://baderlab.bme.jhu.edu/gd/":http://baderlab.bme.jhu.edu/gd/ 4. MB Advanced DNA Analysis v.6.83 URL: "http://www.molbiosoft.de":http://www.molbiosoft.de 5. Tandem Repeats Finder \(TRF) v.3.21 URL: "http://tandem.bu.edu/trf/trf.download.html":http://tandem.bu.edu/trf/trf.download.html

Procedure

Overall methodology for in silico designing of the oligonucleotides required for artificial gene synthesis has been divided into various task levels \(I to VI) such that each task consists of various respective steps. Task I: Generation of ‘Preliminary DNA Sequence’ The peptide/protein sequence of interest \(which ought to be more than 20 amino acids long) should be first back-translated to derive a ‘Preliminary DNA Sequence’. The peptide/protein sequence should be back-translated using the software, ‘Gene Design β2.0’ \(Richardson et al, 2006), which is accessible from the following link: "http://baderlab.bme.jhu.edu/gd/":http://baderlab.bme.jhu.edu/gd/ 1. For back-translating using the above software, enter the peptide/protein sequence in RAW format into the ‘Reverse Translate’ module present in Gene Design software. 2. Now the software would ask for selecting a specific codon-usage table \(genetic-code). Selection of this table would depend upon the organism from which your protein/peptide originates; for e.g. if organism of origin is E. coli, then select the ‘E. coli Radio-button’ present in the software. After these procedures run the module by hitting the ‘Reverse Translate’ button \(Keep everything else as default). 3. After the software has processed the peptide sequence \(which probably takes a second), suitable back-translated DNA sequence would be given as output. Consider this to be our ‘Preliminary DNA Sequence’, which is now ready as input for Task II. Task II: Codon Juggling and Generation of ‘Optimized Trial DNA Sequence’ After back/reverse-translation, one might observe presence of certain repetitive residues in the ‘preliminary DNA Sequence’. In long-run this may become a problem since these repetitive sequences may cause the DNA sequence to form secondary structures like hair-pin loops \(due to self-annealing) or wrong dimer products \(due to inter- or intra-molecular interactions) during later annealing cycles. Therefore, such DNA repeats may interfere with gene-construction \(Engels & Uhlmann, 1989). In fact on the level of mRNA, certain inverted-repetitive sequences may lead to formation of RNA-secondary loop structures, which may hamper the translation of proteins \(Looman & et al., 1986). In order to avoid these repetitive DNA sequences from causing problems in gene-constructions, the phenomenon of codon-degeneracy must be utilized to change our ‘preliminary DNA sequence’ into ‘Trial DNA Sequence’ such that the ‘trial DNA sequence’ codes for the same peptide as the preliminary sequence but both differ in their sequence especially at the third position of various codons \(Engels & Uhlmann, 1989; Griffiths et al, 1999). This strategy may prove to be useful in getting rid of these repetitive sequences while still not altering the peptide coded by the sequence. These sequences are called ‘Trial DNA Sequences’ since these are still under test for repeats, restrictions sites, secondary structures etc. 4. Now, suitable ‘Trial DNA sequences’ could be constructed from ‘Preliminary DNA sequence’ using the ‘Codon Juggling/Juggle’ module present in the software, ‘Gene Design β2.0’ \(Richardson et al, 2006). Here, feed the above ‘preliminary DNA sequence’ \(produced at the end of Task I) as input in RAW format to the ‘Codon juggling’ module; subsequently select your respective target organism from the pull-down menu to keep up with the genetic-code, and hit the ‘Next Step: Results’ button. This software will now utilize four different algorithms to produce four corresponding sequences \(namely - Optimized, Most Different, Next Most Optimized and Random DNA sequences) as output, which have nucleotide sequences most different from original sequence but all encode the same peptide/protein \(Richardson et al, 2006). 5. From the fours sequences produced by the software, select two ‘trial DNA sequences’ on the basis of their identity \(which should be as small as possible) with the original ‘preliminary DNA sequence’. 6. Now enter both of these above trial DNA sequences in RAW format into the ‘Oligo Design’ module present in the Gene Design β2.0 software, one-by-one. Keep all other parameters as default, click on the ‘Design Oligos’ button. Since one of the above two ‘trial DNA sequences' is supposed to be synthesized later by chemical synthesis, during which they would be broken up into oligonucleotides \(Tsuchiya et al, 2006), a preliminary analysis for Oligonucleotide is required to be run on both of the above sequences for synthesis-related optimization \(e.g. insertion of silent restriction site). This can be done using the 'Oligo Design' module present in the Gene Design software. 7. During this preliminary analysis, the software might ask for adding an extra unique restriction site in the 'trial DNA sequences' for designing optimal oligonucleotides. Here, a module called 'Silent Site Insertion' present in the Gene Design software may well be utilized. If the software doesn't ask for this procedure then directly proceed to Step 11. This module allows insertion of unique restriction site within a sequence without altering what the codons in that position are coding. This strategy could be used for inserting a silent restriction site in the sequence by giving the query sequence \(which in this case are the trial DNA sequences) as input and selecting the unique restriction site. The software requires presence of at least 2 restriction sites within the DNA sequence of interest in order to generate optimal oligonucleotides. 8. The software suitably guides the user in selecting the site; henceforth follow the instructions of the software \(under the 'Silent Site Insertion' module). The unique restriction site might be selected ad libitum but remember that the restriction enzyme being selected shouldn't exhibit 'star activity'. This characteristic of restriction enzyme might be confirmed from the following list: "http://catalog.takara-bio.co.jp/en/product/basic_info.asp?unitid=U100005230":http://catalog.takara-bio.co.jp/en/product/basic_info.asp?unitid=U100005230 . Restriction sites of enzymes exhibiting star activity shouldn't be selected. 9. The software would now insert the relevant restriction site in the sequence giving us our 'modified trial' DNA sequence. The software at this point will show this phase – 'You can take this sequence to another module now', along with the modified DNA sequence and the details of the restriction sites inserted. At this stage, take the sequence back to the 'Oligo Design' module. 10. Now, once you have returned to the 'Oligo Design' module you would observe that your modified DNA sequence has been automatically inserted into the query area. Now, keeping all the settings as default, click on the 'Design Oligos' button. 11. The software will now give the oligonucleotides for the respective 'Trial DNA sequences' and their overlapping sequences as output along with other statistics of interest. Note these down or save this page using the 'Save As' option. 12. After the above exercise, both of the 'Optimized trial DNA sequences' would be preliminarily found to be producing acceptable oligonucleotides. This test must be followed up by the test to confirm whether both of these sequences are coding for the same peptide or not. This can be done by back/reverse-translating these sequences using the strategy used in Task I. After the above step, both sequences should be found to be coding for the same peptide/protein as desired. These sequences are now ready to be taken for 'Repeat Analysis' so as to select one sequence over another for chemical/artificial gene synthesis. Task III: Repeat Analysis on the 'Optimized Trail DNA Sequences' As mentioned previously, the repeats present in DNA sequences may hamper overall success of the present experiment at various levels. Thus, a 'Repeat Identification & Quantification Analysis' must be performed on the two 'Optimized Trial DNA Sequences' \(Engels & Uhlmann, 1989) derived at the end of Task II. This analysis is also supposed to decide the choice of DNA sequence out of the above two, to be used for further analysis and experimentation. DNA repeat analysis on the two 'optimized trial DNA sequences' may be done using the software, 'Tandem Repeats Finder v.3.21' \(Benson, 1999), which is designed to locate and display tandem DNA repeats in a sequence. This software may be downloaded from the following link: "http://tandem.bu.edu/trf/trf.download.html":http://tandem.bu.edu/trf/trf.download.html . The original ‘preliminary DNA sequence’ \(Derived at the end of Task I) should also be included in this analysis for comparing the extent of success achieved due to codon juggling \(that was done in Task II). 13. The above three sequences \(i.e. two optimized trial and one preliminary, DNA sequences) should be fed in FASTA format to the ‘Tandem Repeat Finder software’, separately; followed by running it by clicking the “Start Search” button in the upper tray. 14. After running the program, an output consisting of two files will be produced such that one file \(i.e. Repeat Table) gives data about location, size, copy number, and nucleotide content of repeats while second file gives data about alignment of the repeats present in each sequence with a consensus sequence of that repeat. These files could be assessed by clicking on the “View Report” button in the upper tray. 15. Now, based upon the data present in the above two tables and comparison with the data for the original ‘preliminary DNA sequence’; one ‘optimized trial DNA sequence’ should be selected out of the two, as the ‘final DNA sequence’. Main parameters responsible for deciding a particular sequence over another are: \(1) Good Percent Match score and \(2) Low Entropy \(S). The DNA sequence having a higher entropy \(calculated based upon the percent composition) shouldn’t be selected; because, as per laws of thermodynamics \(Moore et al, 2005; Zuker, 2000), such a sequence with higher entropy has better probability of posing stable repeat-mediated problems than others with lower entropy \(since high value of entropy i.e. S or ΔS, means less Gibbs free energy i.e. G or ΔG value; and lesser the value of Gibbs free energy, better the spontaneity of an event, hence the repeats in DNA sequence with high entropy value would have better spontaneity towards posing problems like formation of rapid and stable secondary structures). Therefore, the sequence with highest percent match score and lowest entropy must be selected as the ‘Final DNA Sequence’ at the end of this Task. Task IV: DNA and mRNA Secondary Structure Analysis on ‘Final DNA Sequence’ At the end of Task III, ‘final DNA sequence’ would be retrieved. This sequence now is further needed to be analysed for presence/absence of DNA and mRNA secondary structures \(Engels & Uhlmann, 1989) so that necessary codon juggling \(if required) could be further performed to abolish these secondary structures from the sequence. 16. DNA secondary structure analysis could be performed using the software, ‘DNA mfold v.3.2’ \(Zuker, 2003) accessible at "http://frontend.bioinfo.rpi.edu/applications/mfold/cgi-bin/dna-form1.cgi":http://frontend.bioinfo.rpi.edu/applications/mfold/cgi-bin/dna-form1.cgi . This software predicts pattern of DNA folding based upon energy predictions. Here, ‘Preliminary DNA sequence’ \(generated at the end of Task I) should also be included \(for the sake of comparison) along with an ‘External Control’ DNA Sequence. This external control DNA sequence should originate from the same organism \(from which the peptide/protein originated) and should have a comparable number of amino acids. This will serve the purpose of being a natural sequence so as to allow comparison between our results \(for synthetic sequence) and this ‘natural sequence’. Such sequence might be selected ad libitum through any genomic database e.g. NCBI Gene Database. For example, if the peptide/protein for which the synthetic gene is being produced is 40 amino acids long and originates from E. coli than a suitable ‘External Control’ DNA Sequence might be the sequence of E. coli K12 gene \(NCBI GeneID: 2847682) coding for 41 amino acids. 17. The DNA mfold software is customizable for a number of chemical parameters. Here, present analysis should be performed under two conditions i.e. at Temperature = 37^{O}C,

\[Na+] = 1M \(viz. Condition I - to get an estimate of secondary structures in physiological conditions) and at Temperature = 62^{O}C, \[Na+] = 300 mM, \[Mg++] = 1.75 mM \(viz. Condition II - to get an estimate of secondary structures at annealing conditions, to be used in later PCR cycles). Other parameters should be used as default. 18. Once the above parameters have been set, the respective sequences \(i.e. Final DNA Sequence, Preliminary DNA Sequence and 'External Control' DNA Sequence) should be entered \(one-by-one) into the software in the FASTA format. Data should be gathered for each of the sequence in both Conditions I and II. After entering the sequence, run the software by hitting the button 'Fold DNA'. 19. After running the software with above parameters for the three sequences, the output would be retrieved in the form of several energy statistics as well as structural diagrams. Here, the statistics to focus upon are, 'No. of structures formed' and 'ΔG of each structure'. Structural Diagrams are produced for each type of structure formed by the respective sequences. Out of the presented hypothetical structures formed by these sequences only those structures would be expected to form more often \(than others did), which have the lowest Gibbs free energy values \(ΔG). Thus, structures with high Gibbs free energy \(i.e. tending towards 0 Kcal/mol) would be preferable since they would have lesser probability of forming secondary structures. For example in Condition II, a structure with ΔG = -22.17 Kcal/mol would have more probability of forming rapid and stable secondary structures than the one with ΔG = -0.77 Kcal/mol. For the sake of comparison it is worth mentioning here that the DNA Sequence of E. coli K12 gene has a ΔG value in range of -0.29 to 0.65 Kcal/mol. 20. In this case, it should be observable in the output for both Conditions I as well as II that, the 'final DNA sequence' forms structures of higher Gibbs free energy \(i.e. tending towards 0 Kcal/mol) when compared to the 'preliminary DNA sequence', such that their values are considerably closer to those of the 'external control' DNA sequence; meaning thereby that this 'final DNA sequence' would have a very low probability of forming stable secondary structures in physiological or PCR conditions \(Zuker, 2000). Retrieval of results where 'final DNA sequence' has ΔG values tending towards 0 Kcal/mol would mean that our 'codon juggling' worked well and that oligonucleotides for this sequence wouldn't form spontaneous stable secondary structures. However, if this is not found to be the case in the output then, Task II should be repeated again-and-again on the 'final DNA sequence' of interest unless and until the problem is fixed. In almost all cases though, the protocol should work and the ΔG-based stability of the 'final DNA sequence' should be easily observable at this point. 21. Similarly, mRNA secondary structure analysis should also be done on the sequences using the software called 'FoldRNA' accessible from the following webpage link: "http://www.softberry.com/berry.phtml?topic=foldrna&group=programs&subgroup=rnastruct":http://www.softberry.com/berry.phtml?topic=foldrna&group=programs&subgroup=rnastruct . This software predicts RNA/mRNA secondary structures through energy minimization. This type of analysis would further ensure the robustness of 'final DNA sequence'. 22. Here, the above three sequences \(final, preliminary and external control) should be entered into the software-query field in RAW format, one after the other. The 'Run Java Viewer' option must be enabled. The software should then be initiated by clicking on the 'Process' button. 23. The output would be retrieved from the software on a new page along with the respective structure in a separate Java Window. Here, the statistics to focus on are, 'No. of Structures' and their 'ΔG values'. Again, the 'final DNA sequence' should pass this test and should be expected to give a good mRNA structure \(with high Gibbs free energy and less secondary structures). If the 'final DNA sequence' doesn't produce a stable mRNA then it has to undergo the same changes as recommended in Step 20. Here, the ΔG value of mRNA secondary structure formed by ‘final DNA sequence’ may not tend towards 0 Kcal/mol as much as it did for DNA secondary structures but it would be considerably less than that for the ‘preliminary DNA sequence’. Secondary mRNA structures, may not have a ΔG value tending heavily towards 0 Kcal/mol since in contrast to DNA, mRNA often utilizes such secondary structures for functional regulation \(Murray et al, 2003). Hence, their complete absence might not be possible. 24. At the end of this task, it would be expected that our ‘final DNA sequence’ could now be finalized without any further codon-modification and this DNA sequence could now be taken into the next Task. Task V: Modification of Final DNA Sequence Now as our ‘final DNA sequence’ is ready at the end of Task IV, we may now modify it in certain ways \(without disturbing the sequence of peptide/protein that it codes for) so as to add certain sequence and modify certain others, so that the gene resulting from the artificial synthesis of this DNA duplex could later be inserted properly and transcribed/translated in the expression system \(e.g. pET). At the end of Task IV, the DNA sequence that we would get would actually produce a blunt ended DNA duplex. Blunt ended DNA duplexes/genes may be inserted easily into certain vectors in presence of proper ligation system however certain vectors might not be very compatible to such insertions. Thus, in latter situations we would be required to introduce some restriction sites at the end of our DNA duplex so as to allow its proper insertion into various vectors. 25. Firstly, the ‘final DNA sequence’ should be checked properly for the presence of appropriate start and stop codons. In case of either’s absence, they should be added manually to make our sequence translationally autonomous. 26. Before adding restriction sites at ends of our DNA, it should be ascertained that the restriction sites that we are using for insertion aren’t present inside the synthetic gene’s DNA sequence. For this purpose, feed the above sequence in RAW format into the software, ‘MB Advanced DNA Analysis v.6.83’ \(Simakov, 2006) downloadable from "http://www.molbiosoft.de":http://www.molbiosoft.de and digest the sequence with all the 711 restriction enzymes available in the software. 27. The output of the software should be self-explanatory and should help in selection of unique restriction sites absent in the ‘final DNA sequence’ but present in the target vector. Here, the restriction site to be inserted at the ends of the DNA sequence should be selected based upon: \(1) the type of vector being used, \(2) the restriction sites within the vector which are to be utilized for insertion and \(3) types of restriction sites present within the final DNA sequence-based duplex/gene. Also note that the restriction enzymes which exhibit ‘star activity’ should be avoided. 28. Now, the user can manually insert the respective restriction site sequences at the 5’ and 3’ end of the ‘final DNA sequence’. Here, while inserting the sequences manually, take special care regarding the exact 5’ and 3’ orientations of the ‘final DNA sequence’ and the restriction sites. 29. Lastly, as later cycles of gene synthesis involve PCR procedures, wherein there exists a risk of losing some nucleotides at the 5’ end of the amplicons, it’s advisable to add 2 or 3 nucleotides at the end of the final DNA sequence which could be cleaved off later by restriction enzymes \(Engels & Uhlmann, 1989). Thus now, manually add 3 nucleotides at each end of the above sequence e.g. ATC. The sequence thus retrieved is our synthetic gene’s ‘template DNA sequence’, and it is broadly depicted in Figure 2. This sequence could now be taken to the next task. Here, remember again that the miscellaneous restriction site being inserted at the ends of the sequence should be unique such that it should be absent both in our target gene as well as the corresponding vector. Task VI: Designing of the Oligonucleotides for later Chemical/Artificial Gene

Synthesis As discussed in the Background section, the current protocol is an upstream process leading up to the actual in vitro artificial synthesis of the gene encoding for our target peptide/protein. The in vitro downstream methodology to be used for synthesizing our gene construct would be the ‘SPR method of artificial gene synthesis’ \(Tsuchiya et al, 2006). This procedure involves; \(1) Designing of two oligonucleotides of equal length having a ~20 b.p. overlap; \(2) Synthesizing these oligonucleotides using Automated DNA Synthesizer utilizing phosphoaramdite \(or any other suitable) method on solid phase/surface and \(3) Hybridizing these oligos and extending them using PCR \(See Figure 1) \(Engels & Uhlmann, 1989; Tsuchiya et al, 2006). Here, the step 2 and 3 heavily depend upon successful execution of step 1 i.e. proper in silico designing of the two oligonucleotides, which is exactly the focus of the present Task. 30. Here, the designing of oligonucleotides for our gene of interest would be done using the software, ‘Gene Design β2.0’ \(Richardson et al, 2006). For designing the oligonucleotides, feed the ‘Oligo Design’ module within the ‘Gene Design β2.0’ software with the ‘template DNA sequence’ \(derived at the end of Task V) in RAW format and set following parameters in the software – Target Oligo-Length: 100 b.p., Overlap Melting Temperature: 62oC and Chunk overlap: 20 b.p. 31. After the parameters have been set and the sequence entered in the query field, run the software by clicking on the button, ‘Design Oligos’. 32. In its output, the software would now show a series of overlapping oligonucleotides \(depending upon how long the template DNA sequence is) in both Forward \(5’-3’) and Reverse \(3’-5’) orientations, such that each oligonucleotide overlaps with another by 20 b.p. The software would also give the exact value of Tm which applies to these oligonucleotides, which in practice wouldn’t be very far away from 62^{O}C. Once the respective oligonucleotides have been designed at the end of Task VI, these oligonucleotides could than be synthesized using Automated DNA/Gene Synthesizer utilizing the phosphoaramdite method on solid phase/surface \(Engels & Uhlmann, 1989). Later, once the oligonucleotides have been synthesized, they could be mixed, hybridized and the complete gene could be synthesized via PCR reaction \(Casimiro et al, 1997; Kodumal et al, 2004; Tsuchiya et al, 2006). Following the synthesis of the gene, it can be inserted into the respective vectors by following the manufacturer’s protocols for that particular vector \(Kodumal et al, 2004).

Timing

1-2 days

Critical Steps

1. In the steps like Step 2 in the above protocol, please select the organism of origin correctly or else this might lead to selection of wrong genetic-code, which would thereby spoil the later 'reverse translation' results. 2. As apparent in the above protocol, there are many instances were the user has to select a certain restriction enzyme site. Please remember that don't select a restriction enzyme site, whose recognizing enzyme exhibits 'star activity'. Because, enzymes exhibiting star activity lead to unspecific cleavages during certain in vitro conditions. 3. In steps like Step 20, it is very crucial to remember that only those DNA sequences are good for which the G value is high or tending towards 0 kcal/mol since

higher is this value lesser is the probability of those DNA sequences forming spontaneous structures in physiological or PCR conditions.

Troubleshooting

Though every effort has been made to use softwares or bioinformatics resources having permanent links yet there exists a fair possibility that over a long period of time these links might go inactive due to server changes, domain changes etc. In such conditions, do not panic\! These softwares have peer-reviewed publications behind them and hence they need to be online always thus if you don't find a software functional at that link, just type the name of that program/software in Google and in author's experience the latest link would be found without much problems.

Anticipated Results

Main problems that may arise during downstream phase of artificial gene synthesis are mostly attributable to faulty designing. Such problems might include: \(1) wrong or no annealing of the oligonucleotides and \(2) hampering of oligonucleotide annealing due to repeat-mediated DNA and/or mRNA secondary structure formations \(Engels & Uhlmann, 1989; Richardson et al, 2006). Previously researchers use to manually design oligonucleotides, which was an extremely cumbersome and error-prone procedure \(Richardson et al, 2006). However, advent of bioinformatics created a great demand for in silico methodologies for designing of these oligonucleotides. Though many types of software have been designed for this purpose \(Richardson et al, 2006) yet the problem has been that the overall designing procedure is so complex and demanding that a single-platform based analysis cannot be considered comprehensive enough. In the present protocol, an attempt has been made to address these problems. Present protocol is better than most other existing in silico protocols for oligonucleotide designing since along with certain other things, this protocol uniquely incorporates the testing for mechanics of repeat-mediated DNA and mRNA secondary structure formations following codon juggling. Previous to this, codon juggling or modifications were performed and it was assumed that it would be enough to avoid secondary structure formation on all nucleic acids levels \(Hoover & Lubkowski, 2002; Jayaraj et al, 2005; Rouillard et al, 2004). However, the current protocol allows users to confirm the extent of stable secondary structures formed by either DNA or mRNA of target DNA sequence and if required, repeat the codon modifications to reach acceptable levels. Further, in the present protocol, the ‘Gene Design β2.0’ software has been extensively used since it has better and more flexible options than certain other previous softwares such as Gene2Oligo, DNAWorks and GeMS \(Richardson et al, 2006). It is envisaged that, this protocol would ensure effective and easy in silico designing of oligonucleotides to be used for later artificial/chemical gene synthesis. The best anticipated result coming out of this protocol is that, the final sequence that the user derives after applying this protocol would be very different from their original back-translated DNA sequence and would not form stable secondary structures. For the sake of example of how efficient this protocol is, kindly have a look at Figure 3. Here, three scenarios \(A, B and C) have been presented which consist of secondary structures formed by 3 highly similar DNA sequences \

(in a 'before' and 'after' orientation). In scenario 'A', the initial DNA sequence \(upper) was not modified much and hence the resulting sequence \(below) still forms stable secondary structures however in scenario 'B' and 'C', the initial DNA sequences \(upper) were modified extensively using above protocol and hence their resulting sequences \(below) are making less stable secondary structures. This example demonstrates in nutshell, the overall anticipated results after following this protocol.

References

Benson G \(1999) Tandem repeats finder: a program to analyze DNA sequences. Nucleic Acids Res 27: 573-580 Casimiro DR, Wright PE, Dyson HJ \(1997) PCR-based gene synthesis and protein NMR spectroscopy. Structure 5: 1407-12 Engels JW \(2005) Gene synthesis on microchips. Angew Chem Int Ed Engl 44: 7166-9 Engels JW, Uhlmann E \(1989) Gene Synthesis \(New Synthetic Methods 77). Angew Chem Int Ed Engl 28: 716-734 Griffiths JFA, Miller JH, Suzuki DT, Lewontin RC, Gelbart WM \(1999) Introduction to Genetic Analysis, 7th edn. New York: W. H. Freeman & Co. Hoover DM, Lubkowski J \(2002) DNAWorks: an automated method for designing oligonucleotides for PCR-based gene synthesis. Nucleic Acids Res 30: e43 Jayaraj S, Reid R, Santi DV \(2005) GeMS: an advanced software package for designing synthetic genes. Nucleic Acids Res 33: 3011-6 Kodumal SJ, Patel KG, Reid R, Menzella HG, Welch M, Santi DV \(2004) Total synthesis of long DNA sequences: synthesis of a contiguous 32-kb polyketide synthase gene cluster. Proc Natl Acad Sci U S A 101: 15573-8 Looman AC, et al. \(1986) Secondary structure as primary determinant of the efficiency of ribosomal binding sites in Escherichia coli. Nucleic Acids Res 14: 5481-5497 Moore JW, Stanistski CL, Jurs PC \(2005) Chemistry, the Molecular Science. USA: Brooks Cole Murray RK, Granner DK, Mayes PA, Rodwell VW \(2003) Harper's Illustrated Biochemistry. New York: McGraw-Hill Medical Richardson SM, Wheelan SJ, Yarrington RM, Boeke JD \(2006) GeneDesign: Rapid, automated design of multikilobase synthetic. Genome Res 16: DOI - 10.1101/gr.4431306 Rouillard JM, Lee W, Truan G, Gao X, Zhou X, Gulari E \(2004) Gene2Oligo: oligonucleotide design for in vitro gene synthesis. Nucleic Acids Res 32: W176-80 Simakov O \(2006) MB Advanced DNA Analysis v.6.83: URL: "http://www.molbiosoft.de":http://www.molbiosoft.de Tsuchiya Y, Morioka K, Shirai J, Yoshida K, Inumaru S \(2006) Comparison of artificial synthesis methods of gene. Nucleic Acids Symp Series No. 50: 275-276 Zuker M \(2000) Calculating nucleic acid secondary structure. Curr Opin Struct Biol 10: 303-310 Zuker M \(2003) Mfold web server for nucleic acid folding and hybridization prediction. Nucleic Acids Res 31: 3406-3415

Acknowledgements

The author would like to acknowledge the financial support received from, Department for International Development \(DFID, UK), University of Leeds \(Leeds, UK) and Commonwealth Scholarship Commission \(CSC, UK) in the form of the ‘Commonwealth Shared Scholarship Scheme 2007-08’, which enabled him to complete his studies at University of Leeds \(Leeds, UK). Further the author would also like to acknowledge the guidance and critical review provided by Prof. Dr. John Colyer \(Professor of Biotechnology, Institute of Membrane and Systems Biology, University of Leeds, UK). This protocol is

derived from an assignment submitted to Prof. Colyer as a part of Masters Coursework and the assignment was reviewed thoroughly and awarded with very high marks by Prof. Colyer, inspiring me to write this protocol.

Figures

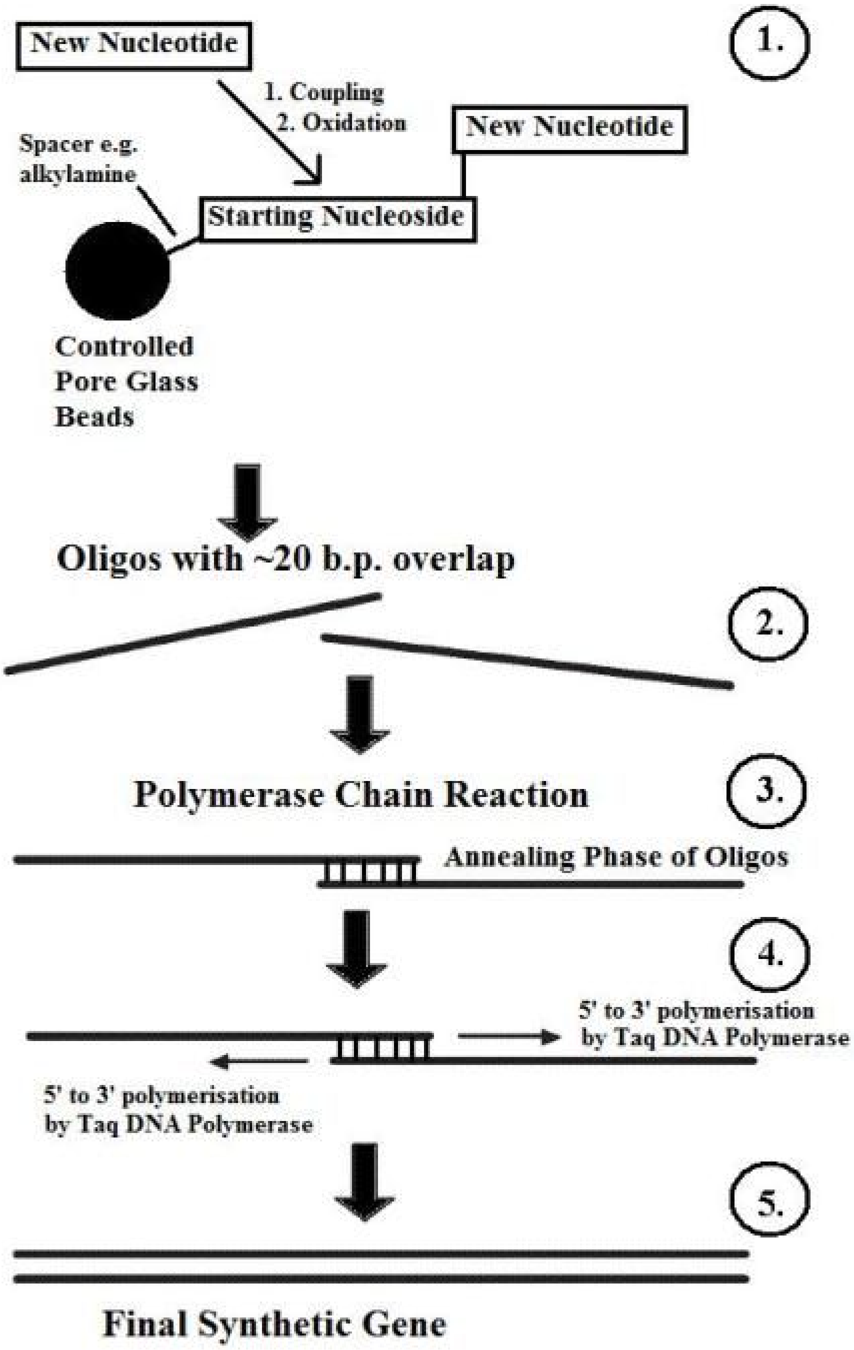
1.
New Nucleotide
1. Coupling
2. Oxidation
New Nucleotide
Spacer e.g. alkylamine
Starting Nucleoside
Controlled Pore Glass Beads
Oligos with ~20 b.p. overlap
2.
Polymerase Chain Reaction
3.
Annealing Phase of Oligos
4.
5' to 3' polymerisation by Taq DNA Polymerase
5' to 3' polymerisation by Taq DNA Polymerase
5.
Final Synthetic Gene

Figure 1

Principle of artificial gene synthesis. Artificial Gene Synthesis, principally involves stepwise coupling of individual nucleotides together to yield oligonucleotides, which are further arranged/annealed amongst themselves into double-stranded (ds) DNA-duplexes based upon overlapping base complementarity. This complementarity-based hydrogen bonding is followed by chemical-based or chemoenzymatic-based ligation of phosphoric esters resulting in stable ds-DNA duplexes or our final target gene. This is further elucidated in following steps: (1) After the oligonucleotides (of both 5&#x2019;&#x2192;3&#x2019; and 3&#x2019;&#x2192;5&#x2019; orientations) for the target gene have been designed (manually or in silico), the oligonucleotides are then synthesized via various methods like trimester, phosphate, phosphoramidite and H-phosphonate. Oligonucleotides of both orientation are synthesized by polycondensation of corresponding nucleoside phosphates. In the present figure, the phosphoramidite method has been depicted; (2) Once the oligonucleotides of both orientations have been synthesized, they are mixed together; (3) Owing to around 20 b.p. overlapping complementarity, these corresponding oligonucleotides anneal together; (4) The annealed product thus formed than acts as a template for PCR and (5) PCR-mediated polymerisation helps in production of the final DNA duplex or artificially synthesized gene. Here, steps 2 to 6 depict the SPR method of Artificial Gene Synthesis (Engels &amp; Uhlmann, 1989; Tsuchiya et al, 2006).

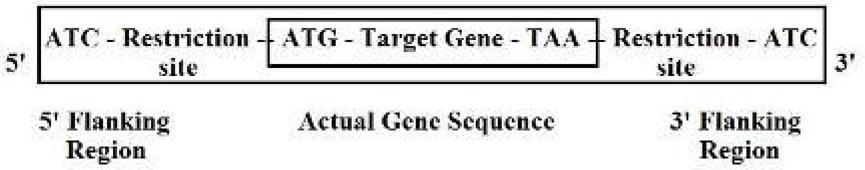

Figure 2

Broad depiction of the in silico designed &#x2018;template DNA&#x2019;. Here, the actual target gene sequence is flanked on both sides by corresponding restriction sites as well as 3 extra nucleotides to safeguard against PCR-mediated terminal nucleotide loss. The start codon (ATG) and stop codon (TAA) mark the ends of the actual target gene sequence.

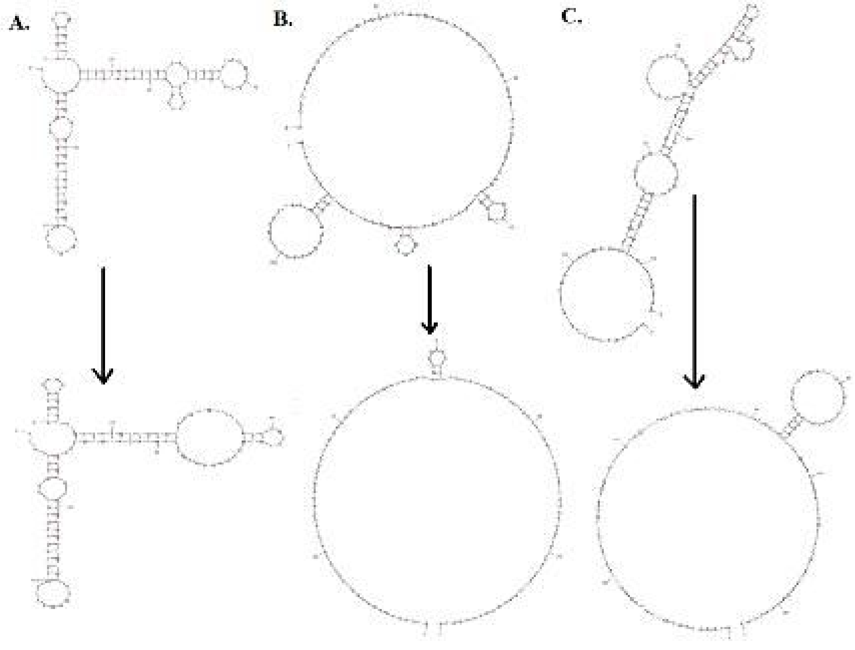

Figure 3

Efficiency of Present Procotol based on DNA secondary structure formation. (A)Here, the DNA Sequence derived from back-translation (upper) wasn't modified according to the present protocol and hence the resulting DNA sequence (below) still retained extensive stable secondary structures. (B and C) Here, the two DNA sequences (upper) were extensively modified using the present protocol and hence the resulting respective DNA sequences (below) were found to form less extensive stable structures.