Project Report

Codon Optimization

Shubhra Gupta
CBS 521
May 5, 2003
Computational Bioscience
Arizona State University
Shubhra.Gupta@asu.edu

Advisor
Dr. Lokesh Joshi
Plant and Molecular Biology Department
Arizona State University
ljoshi@asu.edu

Contents

Abstract

In this report we will see what we have done in the project till now and what future work we are going to do. We have developed the interface for codon optimization and it is working fine with Apache server. For future work we are writing a working program in C, which can convert sequence of nucleotides of one species into the sequence of nucleotides of another desired species by replacing codons using codon frequency table [29, 30] and also in future we shall see how this software works.

1. Introduction

Codon optimization is a technique recently used by many scientists to improve the protein expression in living organism by increasing the translational efficiency of gene of interest [1-4, 6-13, 15-18, 20-28]. Optimizing codon for the custom gene design is the best way to increase the functionality of gene. In this report we will explain this process in detail and also look at some available software tools for this purpose. We are also developing a web-based tool for this purpose. The necessary background material of biology is given in the following subsections.

1.1 DNA Pathway

DNA is the basic hereditary material in all cells and contains all the information necessary to make proteins. In normal DNA, the bases form pairs: A to T and G to C. These are called complementary pair. Two complementary chains that are arranged in an anti-parallel manner form a duplex of DNA.
RNA is a polymer that contains ribose rather than deoxyribose sugars. The normal base composition is made up of guanine, adenine, cytosine, and uracil.
When a cell receives a signal saying that a certain protein is needed, the code for producing protein is made. The DNA double helix unwinds and one strand of the helix becomes a template for producing the protein-coding template. This template is a single strand of opposite bases (from DNA) and is called RNA (Ribonucleic Acid).
Bases that are floating in the cell join up with opposite bases. Uracil takes the place of Thymine. This template is called mRNA (messenger RNA), because it serves as a code messenger between DNA and protein. The process of creating an mRNA from DNA is called **transcription.**
Whatever information is stored in mRNA that is used to make protein. Outside the nucleus, the proteins are built based upon the code in the RNA. This process is called **translation.**
All proteins are composed of one or more linear unbranched polymers. These polymers are constructed of monomers. These monomers are called **amino acids.** There are 20 different amino acids found in most proteins. Instead of these 20 amino acids there are 2 more amino acids, which involve in start and stop the protein chain.
Each amino acid is made of three nucleotides called **triplet** or **codon**. Typically 100-300 amino acids are linked together in chains. The linkage between the amino acids is called the **peptide bond**.

All codons are used with the same frequency. Some codons are commonly used while others are not. Codons that are “rare” usually correlate with a reduced intracellular level of aminoacyl-tRNAs. Aminoacyl-tRNAs are the biochemical building blocks used by ribosomes to match a codon with the correct amino acid during protein synthesis (translation). If a particular aminoacyl-tRNA is deficient, then use of the corresponding codon will slow the rate of protein synthesis and can decrease the yield of a desired protein product.

It has been experimentally documented that genes, which use “rare” codons, tend to be less well expressed (at the point of protein translation) than genes, which use "frequent" codons. The relative frequency of use for each codon can vary significantly between species, although certain codons are infrequently used across species.

Like, custom gene synthesis was used to produce different versions of the human erythropoietin (EPO). Gene all of whom encoded the same protein but used different DNA sequence, biased towards either yeast or human codon usage. So optimization of codons is useful in this regards.
When we talk about codon, protein automatically comes with that. So one has to know, what is protein and how it works?

1.2 Proteins

Approximately 50% of the dry weight of living matter is protein. Protein is not just the simple storage or structural material. The types and functions of proteins are as varied as the functions of life itself. All the thousands of catalysts that make the chemical reactions of living matter possible are proteins. These catalysts are **enzymes**.
Protein is also responsible for the movement of living organisms. Muscles are largely composed of precisely ordered protein molecules. Proteins are responsible for transporting many materials through the circulatory system. For example, Hemoglobin, which transports O2 and CO2 in the blood. The clotting of the blood requires the interaction of a number of different proteins.
Antibodies, those extraordinary molecules that can recognize and inactive virtually any foreign substance, are proteins.
Milk (Beta-lactoglobulin) is the major example of protein.
Keratin is the major protein in hair and nails.
Proteins always contain carbon, hydrogen, oxygen and nitrogen atoms and usually sulfur atom as well. All proteins composed of one or more liner unbranched polymers. The monomers of which these polymers are constructed are called **amino acids**.

1.3 Four Levels of Protein Structure

-Primary structure refers to the **"linear"** sequence of amino acids.
-Secondary structure is called **“local”.** The most common secondary structure in proteins is the **alpha (a) helix** and the **beta (b) helix** (sometime called b-pleated sheet).
-Tertiary structure is called **"global".** This is the three-dimensional folding of a single polypeptide chain.
-Quaternary structure involves the association of two or more polypeptide chain into a multi-subunit structure.
So one can understand how protein is related to codon optimization.

The rest of the paper is organized as follows. Section 2 gives an introduction to codon optimization (CO). Section 3 discusses biology behind codon optimization and section 4 talks about the objective, importance and advantages of CO. Section 5 and 6 deals about the work done before and the available software for this purposes. Resources used by us are given in section 7 and they are further described in section 8 giving also the algorithm for CO. Next section gives the work completed by us and also discuss about the future work. Some examples of codon optimization are given in section 10.

2. Introduction to Codon Optimization

2.1 Codon Optimization

Codon optimization is a technique to maximize the protein expression in living organism by increasing the translational efficiency of gene of interest by transforming DNA sequence of nucleotides of one species into DNA sequence of nucleotides of another species. Like, plant sequence to human sequence, human sequence to bacteria or yeast sequences, etc.

2.2 Working of Interface for Codon Optimization

Given a DNA sequence of nucleotides, breaks that sequence into triplet (codons) and replace triplets with new one, generated with a given frequency distribution. In this process amino acid will be same, but codon of low frequency of an amino acid will replace with codon of high frequency, according to desired species frequency distribution table. For example, suppose in one species amino acid Leucine (codon) UUA has low frequency 7.5 and in desired species Leucine (codon) CUG has high frequency 39.8, so UUA will replace with CUG.

2.3 The Output

The output would be a new sequence of nucleotides of given species having the changes of some triplets.

In this way one can get a DNA sequence of nucleotides of another species with same amino acids but some low frequency codon replace with high frequency codon. So codon optimization involves replacing wild type DNA sequences with more highly expressed species sequences (without changing the antigen). Studies have indicated a direct correlation with expression levels and immunogenicity.

3. Biology Behind Codon Optimization

Each organism has its preferred choice of nucleotide usage to encode any particular amino acid, called as codon usage. Humans and plants vary in their codon preferences for translating mRNA into proteins. Plants, in general, prefer G and C rich codons as compared to mammals' they rich in A and T codon. Like Allege has 60% - 70% G and C

rich codon. Different plants have different percentage of G and C rich codon. Because of this difference in codons, we need codon optimization. This is only one region.

Transformation of heterologous genes with low G and C content in plants often leads to very low yield. The Dictyostelium genome has a higher AT content than the human [38].

By codon optimization process, we can remove stop codon from sequences. Translation of messenger RNA occurs on ribosomes. This process takes place in three phases, 1. Initiation 2. Elongation 3. Termination. In elongation phase, a tRNA molecule capable of base pairing with the next mRNA codon arrives at an adjacent site on the ribosome. In this way, the message that was originally transcribed into the mRNA molecule is translated into a defined sequence of amino acids in the polypeptide. Elongation of the polypeptide continues until the end of the message. End comes with stop codons for which no tRNA molecule exits. Because of this occurrence, the ribosome splits into subunits and another round of protein synthesis reassembled. To stop this splitting, we need to remove stop codons. Codon usage contributes to ribosome stabilization.
An "open reading frame" (ORG) is a string of codons with no **stop codon**.

4. Why Codon Optimization?

4.1 Objectives

Use high frequency codons to maximize or increase efficiency of functional unit (protein expression level). Purpose is to increase the protein expression levels to ensure efficient production for research and clinical trials. It will help in finding disease, drug discovery, etc.

4.2 Importance

Codon optimization has importance in DNA vaccination for HIV [14,19]. Few human trials have indicated the promise of DNA based immunity for HIV infection. It has Importance in numerous animal tests. To remove stop codons, to clone, in custom design of synthetic genes, to improve the functionality of genes, to increase protein expression level and for lower production costs, in drug development and it have numerous other applications.

4.3 Advantages of Codon Optimization

- Match codon frequencies in target and host organisms to ensure proper folding.
- Bias GC content to increase mRNA stability or reduce secondary structures.
- Minimize tandem repeat codons or base runs that may impair gene construction or expression.
-Customize transcriptional and translational control regions.
- Insert or remove protein trafficking sequences.
- Remove/add post translation modification sites in encoded protein (e.g. glycosylation sites)
- Add, remove or shuffle protein domains.
- Insert or delete restriction sites.
- Modify ribosome binding sites and mRNA degradation sites.
- Adjust translational rates to allow the various domains of the protein to fold properly.

- Reduce or eliminate problem secondary structures within the mRNA

5. Earlier Work

There is a web-based software of codon optimization process, which is free and which can convert human sequence to yeast sequence, yeast sequence to bacteria sequence, etc. But this software is not able to transform nucleotide sequence of given species into nucleotide sequence of desired species.
For example, yeast sequence to plant sequence or one plant sequence to another plant sequence.
Lancia G. and Gambotto A. of Molecular Medicine Institute (MMI) developed that software. The name of this tool is Codon Optimization Algorithm [36].

6. Some Available Softwares

Entelechon is the syntheticgenes company. This company has tool Codon Usage Table analysis. This tool can analyze codon usage tables and lists the relative usage of codons for each amino acid. This company has some other java enabled browser products like, Gene synthesis, Protein backtranslation, Sequence inversion, DNA reader, etc. These people are working on gene synthesis, sequencing and other fields [31].

GENEMAKER is a codon optimization service. This service provides software for codon optimization, which name is Blue Heron. These people claim that their software Blue Heron can improve 100- to 500-fold in protein expression under codon optimized sequence and give perfect accuracy including genes 10,000 bps. This service provides some other software also. Minimum price of those software's are started from $5,000 - $15,000 [34].

Aptagen is another company, which provides tool for codon optimization. Their software name is Gene Forge. These people claim that their software can increase protein expression up to 10-fold and also this software can insert or delete restriction sites. Price of this tool is US$9 per bp [35].

DNA Builder Software is designs oligonucleotides for gene synthesis [39].

General Codon Usage Analysis (GCUA) is a program. This program is designed to perform various tasks that are of use for evaluating codon usage in a set of genes. This software can calculate percentage usage of each amino acid in a protein sequence. It can also calculate the Number (N) of times a particular codon in a gene or set of genes. Limit of this software is 15,000 bp [32].

MCLAB is The Molecular Cloning Laboratories. Mclab offers expertise in codon maximization, recombinant protein expression and purification of protein [37].

7. Resources for Codon Optimization

1. The following interface to run the process of codon optimization has been made. This interface is a web base tool. This tool has two ends, front end and back end. In the front end of this tool we have interface to run the codon optimization

process and back end is supported by CGI & C program. Figure 1.1 is some part of that interface.

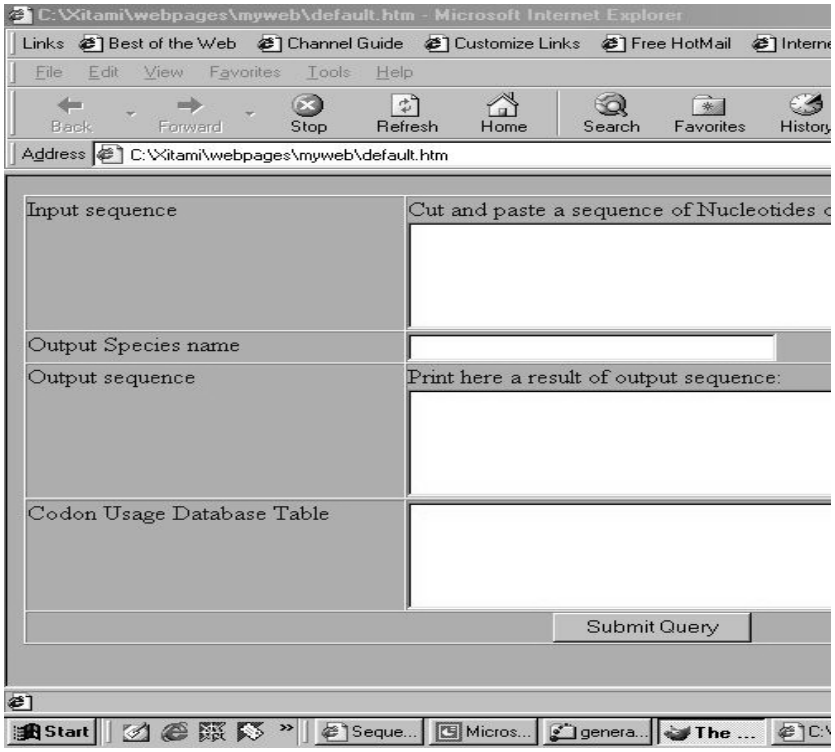

Figure 1.1 (Front end)

UUU 35.7	UCU 11.9	UAU 4.0	UGU 4.0
UUC 7.9	UCC 19.8	UAC 7.9	UGC 0.0
UUA 15.9	UCA 7.9	UAA 4.0	UGA 0.0
UUG 35.7	UCG 15.9	UAG 0.0	UGG 19.8
CUU 19.8	CCU 11.9	CAU 23.8	CGU 4.0
CUC 19.8	CCC 11.9	CAC 4.0	CGC 4.0
CUA 19.8	CCA 0.0	CAA 35.7	CGA 19.8
CUG 35.7	CCG 11.9	CAG 11.9	CGG 0.0
AUU 43.7	ACU 27.8	AAU 31.7	AGU 7.9
AUC 19.8	ACC 4.0	AAC 35.7	AGC 19.8
AUA 11.9	ACA 31.7	AAA 39.7	AGA 7.9
AUG 31.7	ACG 4.0	AAG 23.8	AGG 4.0
GUU 11.9	GCU 19.8	GAU 11.9	GGU 11.9
GUC 15.9	GCC 11.9	GAC 23.8	GGC 0.0
GUA 0.0	GCA 27.8	GAA 23.8	GGA 23.8
GUG 11.9	GCG 4.0	GAG 15.9	GGG 15.9

Table 1.1
Codon Usage table with frequencies
[Triplet] [Frequency per thousand]

2. In this the codon usage table from the web site http://www.kazusa.or.jp/codon to select the codon frequency table of given species is used. A typical table is Table 1.1.

3. Apache web server is used to run the interface for codon optimization. Apache web server is Linux based and it can support CGI programming. This web server is free and it is believe that 60% of Internet is on Apache web sever. And more than that one can use some C library for CGI programming (like cgic an ANSI C library for CGI programming) which is friendlier on Linux server.

8. CGI Programming and Algorithm for Tool

CGI (Common Gateway Interface) is a script file, which stored on remote web server and executed on a web server in response to request from user. Mainly it is interface between user and web server.
It can handle two types of languages

- Compiled language C, C ++
- Interpreted language perl, JCL (job control language).

CGI is a standard for interfacing external applications with information servers, such as HTTP or Web servers. A plain HTML document that the Web daemon retrieves is static, which means it exists in a constant state, a text file that doesn't change. A CGI program, on the other hand, is executed in real-time, so that it can output dynamic information.

CGI script can handle two types of HTML form, Get and Post.
If form has METHOD="GET" in its FORM tag, CGI program will receive the encoded form input in the environment variable QUERY_STRING.
If form has METHOD="POST" in its FORM tag, CGI program will receive the encoded form input on stdin.

To run codon optimization process using interface on web server we need a compiling language. I am using C language, to execute the process of codon optimization.
Which follows below algorithm.

Input: Given a DNA sequence of nucleotides of a species
Output: A sequence of nucleotides of desired species

Step1: Enter a sequence of nucleotides

Step2: Based on desired species name select codon frequency table of that species

Step3: Parse the amino acids from codon frequency table

Step4: Select first highest, second highest and third highest frequency codons from an amino acid.

Step5: According to these frequencies, replace the codon of low frequency in input sequence with high frequency codon to keep an amino acid same

Step6: print the nucleotide sequence of desired species

9. Work Completed and Future work

9.1 Work Done

- Developed interface using HTML language
- Downloaded and configured web server Xitami, which can support CGI programming. working fine with perl program
- Now writing C program
    - C program which can read and write the data from web page

9.2 Future Work

-Complete the C program and see how this tool will work

10. Some Examples on Codon Optimization

Codons of Arginine and leucine amino acids in eukaryotic gene products can be especially problematic when expressed in *E. coli,* because they can exhibit mismatched codon usage.
For example, mammalian genes frequently use the AGG codon for Arginine.
This is an extremely rare codon used by *E. coli* and correlates with low levels of its corresponding tRNA. Hence, the mammalian protein translates inefficiently in *E. coli* and produces a low protein yield [5].
Examples of mismatched codon use that hamper protein production in heterologous expression systems include genes from yeast and plants. Matching codon usage may increase yield as much as 5-10 fold.

Another example, the amino acid glutamine is encoded by two codons "GAA" and "GAG". The relative usage of these codons varies dramatically between the prokaryotic organism E. coli and humans as follows

	H. Sapiens	***E. coli***
GAA	42%	68%
GAG	58%	32%

When designing a gene to be expressed in human cells, bias toward use of "GAG" to code for glutamine would be preferred; when expression is intended in E. coli, use of "GAA" would be preferred.
Gene synthesis allows for complete control over codon usage, a different DNA sequence can be designed and manufactured that has been uniquely optimized for expression of any desired amino acid sequence for expression in any host cell.

11. Acknowledgement

The author would like to thank Prof. Lokesh Joshi and Prof. Renaut for their help.

12. References

[1] Baev D, Lil XW, Edgerton M, Genetically engineered human salivary histatin genes are functional in Candida albicans: development of a new system for studying histatin candidacidal activity, MICROBIOL-SGM **147** (2001), pp. 3323-3334.

[2] Cid-Arregui A, Juarez V, zur Hausen H, A synthetic E7 gene of human papillomavirus type 16 that yields enhanced expression of the protein in mammalian cells and is useful for DNA immunization studies, J VIROL **77** (2003), pp. 4928-4937.

[3] Deml L, Bojak A, Steck S, et al., Multiple effects of codon usage optimization on expression and immunogenicity of DNA candidate vaccines encoding the human immunodeficiency virus type 1 Gag protein, J VIROL **75** (2001), pp. 10991-11001.

[4] Deng TL, Bacterial expression and purification of biologically active mouse c-Fos proteins by selective codon optimization, FEBS LETT **409** (1997), pp. 269-272.

[5] Eric M. Slimko, Henry A. Lester**,** Codon optimization of Caenorhabditis elegans GluCl ion channel genes for mammalian cells dramatically improves expression levels, JOURNAL OF NEUROSCIENCE METHODS **00** (2003) pp. 1-7.

[6] Fitch DHA, Strausbaugh ID, Low codon bias and high-rates of synonymous substitution in drosophila-hydei and drosophila-melanogaster histone genes, MOL BIOL EVOL **10** (1993), pp. 397-413.

[7] Fuller M, Anson DSA, Helper plasmids for production of HIV-1-derived vectors, HUM GENE THER **12** (2001), pp. 2081-2093.

[8] Hale RS, Thompson G, Codon optimization of the gene encoding a domain from human type 1 neurofibromin protein results in a threefold improvement in expression level in Escherichia coli, PROTEIN EXPRES PURIF **12** (1998), pp. 185-188.

[9] Hamdan FF, Mousa A, Ribeiro P, Codon optimization improves heterologous expression of a Schistosoma mansoni cDNA in HEK293 cells, PARASITOL RES **88** (2002), pp. 583-586.

[10] Holler TP, Foltin SK, Ye QZ, et al., Hiv-1 integrase expressed in escherichia-coli from a synthetic gene, GENE **136** (1993), pp. 323-328.

[11] Horvath H, Huang JT, Wong O, et al., The production of recombinant proteins in transgenic barley grains, P NATL ACAD SCI USA **97** (2000), pp. 1914-1919.

[12] Kim CH, Oh Y, Lee TH, Codon optimization for high-level expression of human erythropoietin (EPO) in mammalian cells, GENE **199** (1997), pp. 293-301.

[13] Kotula l, Curtis PJ, Evaluation of foreign gene codon optimization in yeast – expression of a mouse ig kappa-chain, BIO-TECHNOL **9** (1991), pp. 1386-1389.

[14] Ludwig Deml, Alexandra Bojak, Stephanie Steck, Marcus Graf, Jens Wild, Reinhold Schirmbeck, Hans Wolf and Ralf Wagner, Multiple Effects of Codon Usage Optimization on Expression and Immunogenicity of DNA Candidate Vaccines Encoding the Human Immunodeficiency Virus Type 1 Gag Protein, AMERICAN SOCIETY FOR MICROBIOLOGY J VIROL **75** (2001) pp. 10991–11001.

[15] Massaer M, Mazzu P, Haumont M, et al., High-level expression in mammalian cells of recombinant house dust mite allergen ProDer p 1 with optimized codon usage, INT ARCH ALLERGY IMM **125** (2001), pp. 32-43.

[16] Meetei AR, Rao MRS, Hyperexpression of rat spermatidal protein TP2 in Escherichia coli by codon optimization and engineering the vector-encoded 5 ' UTR, PROTEIN EXPRES PURIF **13** (1998), pp. 184-190.

[17] Nagata T, Uchijima M, Yoshida A, et al., Codon optimization effect on translational efficiency of DNA vaccine in mammalian cells: Analysis of plasmid DNA encoding a CTL epitope derived from microorganisms, BIOCHEM BIOPH RES CO **261** (1999), pp. 445-451.

[18] Narum DL, Kumar S, Rogers WO, et al., Codon optimization of gene fragments encoding Plasmodium falciparum merzoite proteins enhances DNA vaccine protein expression and immunogenicity in mice, INFECT IMMUN **69** (2001), pp. 7250-7253.

[19] Richard H. L., Michael J. P., Combinatorial Optimization in Rapidly Mutating Drug-Resistant Viruses, JOURNAL OF COMBINATORIAL OPTIMIZATION **3** (1999) pp. 301–320.

[20] Slimko EM, Lester HA, Codon optimization of Caenorhabditis elegans GluCl ion channel genes for mammalian cells dramatically improves expression levels, J NEUROSCI METH **124** (2003), pp. 75-81.

[21] Sinclair G, Choy FYM, Synonymous codon usage bias and the expression of human glucocerebrosidase in the methylotrophic yeast, Pichia pastoris, PROTEIN EXPRES PURIF **26** (2002), pp. 96-105.

[22] Takenaka Y, Haga N, Harumoto T, et al., Transformation of Paramecium caudatum with a novel expression vector harboring codon-optimized GFP gene, GENE **284** (2002), pp. 233-240.

[23] Te'o VSJ, Cziferszky AE, Bergquist PL, et al., Codon optimization of xylanase gene xynB from the thermophilic bacterium Dictyoglomus thermophilum for expression in the filamentous fungus Trichoderma reesei, FEMS MICROBIOL LETT **190** (2000), pp. 13-19.

[24] Valencik ML, McDonald JA, Codon optimization markedly improves doxycycline regulated gene expression in the mouse heart, TRANSGENIC RES **10** (2001), pp. 269-275.

[25] Vervoort EB, van Ravestein A, van Peij NNME, et al., Optimizing heterologous expression in Dictyostelium: importance of 5 ' codon adaptation, NUCLEIC ACIDS RES **28** (2000), pp. 2069-2074.

[26] Wells KD, Foster JA, Moore K, et al., Codon optimization, genetic insulation, and an rtTA reporter improve performance of the tetracycline switch, TRANSGENIC RES **8** (1999), pp. 371-381.

[27] Woo JH, Liu YY, Mathias A, et al., Gene optimization is necessary to express a bivalent anti-human anti-T cell immunotoxin in Pichia pastoris, PROTEIN EXPRES PURIF **25** (2002), pp. 270-282.

[28] Wu L, Barry MA, Fusion protein vectors to increase protein production and evaluate the immunogenicity of genetic vaccines, MOL THER **2** (2000), pp. 288-297.

[29] http://www.kazusa.or.jp/en/
[30] http://www.kazusa.or.jp/codon/
[31] http://www.entelechon.com/eng/cutanalysis.html
[32] http://bioinf.may.ie/gcua/index.html
[33] http://www-leibniz.imag.fr/GDR-INFOGENOMES/JC-bulmer.html
[34] http://www.blueheronbio.com/genemaker/codon.html
[35] http://www.aptagen.com/codon-optimization.htm
[36] http://www.pitt.edu/~agamb/
[37] http://www.mclab.com/html/codonMaximization.html
[38] http://www.fwn.rug.nl/becel/research.html
[39] http://cbi.swmed.edu/computation/cbu/DNABuilder.html