PRIMERDB: A SYNTHETIC DATABASE FOR PRIMER/OLIGONUCLEOTIDE HYBRIDIZATION AND EFFICIENCY PREDICTION

Randeep Singh, Sunil Kumar and Lalit Gupta
Philips Research Asia - Bangalore (PRA-B)
{randeep.singh,sunil.jaglan,lalit.gupta}@philips.com

ABSTRACT
Primer/oligonucleotide design is the most important step for any array or PCR based assay. There are several parameters that affect the outcome of the whole assay. One of the most important parameter that affects the hybridization efficiency of the probe with the target sequence is its free energy. In the present work, we have designed primers for genes associated with 30 monogenic disorders using Primer3. The database were created with free energy parameters and rating of the sequences was obtained through modified Net-Primer where the secondary structure parameter prediction was made more robust. In addition to these parameters, we have calculated various other sequence features including the fraction of A, T, G and C, position of dimers and trimers. The PrimerDB database will be made public and it can be used for mutation detection, cloning and sequencing for these 30 monogenic disorders. We obtained around one million primers/oligonucleotides for the given 30 genes. These sequences were then grouped into two categories; one having rating of more than 90 and other less than 90. After extraction of various features these sequences were subjected to feature selection algorithms to obtain the most important and remove the redundant parameters (features). The most important features were then classified through Support Vector Machines (SVM) where the overall efficiency of the data obtained was 70%. The results presented in this paper are preliminary and further investigations are being carried out on the sequences to extract more features and to increase the classification accuracy.

1 Introduction
Primer/oligonucleotide design is one of the most important and critical process in molecular biology techniques which involve polymerase chain reaction (PCR) for mutation detection and cloning to DNA sequencing. There are several important parameters that need to be considered when designing primers/oligonucleotides. Some of the parameters that have shown to affect the primer properties and hence their stability are:

1. Optimal size: It is the desired nucleotide length for primer and is generally taken as 20bp by most of the online tools available. However in most of the tools the length can be varied based on users requirement.

2. Adverse base pair composition:
   - Total number of As and Ts (approx. 10–15) and number of Gs and Cs (approx 6–9), not more than six A or T in a row.
   - GC content should be around 40–60 %.

3. $T_m$ uniformity: It is calculated from the base composition and should be equal for both forward and reverse primers. There are several methods to calculate $T_m$. The fast and less accurate Wallace method [1] is suitable for primers of less than 18bp length and is calculated as in (1):
   $$T_m = 2(A + T) + 4(G + C)$$

Another more accurate method is the Nearest Neighbor method [2]:
   $$T_m = \frac{\Delta H}{\Delta S + R(\ln C_1 - \ln C_2)} - 273.15$$

where, $\Delta H$ and $\Delta S$ is the standard enthalpy and entropy, $C_1$ and $C_2$ is the initial concentration of single and complementary strand, and $R$ is the universal gas constant.

4. GC clamp: It is the number of consecutive Gs and Cs at 3’ end of both the primers. It is important for the formation of complex with the target DNA sequence.

5. Internal stability of primers:
   - Stable 5’ termini and unstable 3’ termini of primers give the best results by reducing false priming on unknown targets.
   - Duplex formation that may initiate DNA synthesis can be prevented by low 3’ stability and 5’ end must also pair in order to form a stable duplex.
   - Optimal terminal $\Delta G$ ∼ 8.5 kcal/mol; variation in this reduces priming efficiency.

6. Product size and $T_m$: Optimal product size of the amplicon is important for designing the PCR protocol. The optimal $T_m$ of the amplicon is very important for proper amplification reaction and is calculated using the formula from Bolton and McCarthy [3].
   $$T_m = 81.5 + 16.6 \log[Na^+] + 0.41(GC\%) - 600/\text{length}$$
7. Salt concentration: It is very important for primer design as alterations in salt concentration can change the \( T_m \), hybridization efficiency and other primer parameters.

Other than these parameters which basically are derived from the sequence composition, there are some structural parameters which determine the efficiency of PCR reaction. These include:

1. Hairpin loop/ primer dimer formation: It is important to exclude the nucleotide which has the tendency to form hairpin loop.

2. Dimer/ cross-dimer formation

3. A palindrome score (a measure of probe self complementarity) of \(< 7\)bp.

There are number of softwares available either as stand alone applications or online versions for primers/ oligonucleotide generation, which are applicable for wide range of PCR including quantitative PCR, nested PCR, multiplexing PCR, methylation specific PCR for SNP detection, site directed mutagenesis, expression analysis etc. and other microarray applications [4] [5] [6]. Some of these tools have advanced options such as validating the primer or providing rating to the generated primers. These ratings depend upon various parameters including the adverse base pair composition, 3’ stability, dimer or duplex formation, palindrome sequence etc. Different tools take into account different parameters and suggest their own best rating, which is usually different from the other tools.

Present work involves the development of a synthetic database (PrimerDB) that can provide efficiency of the primers/ oligonucleotides designed for a particular sequence (our study involved analysis of genes associated with thirty monogenic disorders). This involves the integration of various tools for primer/ oligonucleotide design based on various defined criterion that influences their properties and stability and then obtaining their rating based on secondary structure parameters. In present work the primers/ oligonucleotides are designed through well adapted and popular primer design tool - Primer3 [4] and their ratings are obtained through modified NetPrimer [7]. Both the tools are freely available. In addition to the ratings, some additional parameters have been determined, which might be crucial for primer/ oligonucleotide. After obtaining their ratings and other parameters, these primers were classified and their efficiency was predicted using a classification algorithms explained in Section 2.

2 Methodology

2.1 Disorders and associated genes

Thirty monogenic disorders were taken into consideration, as the main focus of our group is towards mutation detection in these disorders. List of all the genes into consider-

<table>
<thead>
<tr>
<th>Input Data</th>
<th>Feature extraction</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Feature selection</td>
</tr>
<tr>
<td></td>
<td>Classification</td>
</tr>
</tbody>
</table>

Figure 1. Steps in the classification algorithm

ation in this work with description and chromosomal location is given in Table 1. Sequences for these genes were obtained from March 2006 assembly of UCSC Genome Browser. The sequences downloaded includes exons, intron, 5’ Untranslated region (UTR) and 3’ UTR.

2.2 Primer/ Oligonucleotide design

Downloaded sequences were then analyzed through Primer3 v 0.4.0 for primer/ oligonucleotide design. Table 2 lists the input parameters for primer designing tools.

2.3 NetPrimer Rating

The rating of individual primers is calculated as:

\[
\text{Rating} = 100 + (\Delta G(\text{Dimer}) \times 1.8 + \Delta G(\text{Hairpin}) \times 1.4)
\]

Example: Say a primer has worst \(\Delta G(\text{Dimer})\) as -2.4 and worst \(\Delta G(\text{Hairpin})\) as -1.0, that primer will have a rating of 94.

2.4 Classification and Efficiency Prediction

We classify the data into two classes. The classification criterion that we have chosen is that the primers with rating greater than and equal to 90 have been considered in one class and the primers with rating less than 90 have been considered in the other class. The steps in classification method are shown in Fig. 1.

2.4.1 Feature Extraction

A modified version of NetPrimer was used to obtain ratings for the primers obtained. All the properties were kept same except the secondary structure parameters which were made more strict/ robust. In addition to rating, we also generated various other parameters that included:

1. Fraction of A, C, G, or T in a probe (regardless of the size)

2. Fraction of A, C, G, or T around the middle
<table>
<thead>
<tr>
<th>Disease</th>
<th>Gene</th>
<th>Description</th>
<th>Accession Number</th>
<th>Chromosomal Location</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cystic Fibrosis</td>
<td>CFTR</td>
<td>Transmembrane conductance regulator</td>
<td>NM_000492</td>
<td>chr7:116,986, 754-116,992,078</td>
</tr>
<tr>
<td>Microcephaly</td>
<td>ASPM</td>
<td>Human abnormal spindles mRNA</td>
<td>NM_018136</td>
<td>chr1:195,323, 871-195,382,287</td>
</tr>
<tr>
<td>Beta thalassemia</td>
<td>HBB</td>
<td>Beta globin</td>
<td>NM_000518</td>
<td>chr11:5,203, 272-5,204,877</td>
</tr>
<tr>
<td>Sickle cell disease</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Phenylketonuria</td>
<td>PAH</td>
<td>hydroxylase</td>
<td>NM_000277</td>
<td>chr12:101,756, 234-101,835,511</td>
</tr>
<tr>
<td>Homocystinuria</td>
<td>CBS</td>
<td>Cystathionine-beta synthase</td>
<td>NM_000071</td>
<td>chr21:43,346, 370-43,369,493</td>
</tr>
<tr>
<td>Alkaptonuria</td>
<td>HGD</td>
<td>Homogentisate 1,2-dioxygenase</td>
<td>NM_000187</td>
<td>chr3:6,49,506, 956-49,538,811</td>
</tr>
<tr>
<td>Methylmalonic acidemia</td>
<td>MUT</td>
<td>Methylmalonyl Coenzyme A mutase precursor</td>
<td>NM_000255</td>
<td>chr14:87,499, 411-87,529,660</td>
</tr>
<tr>
<td>Krabbe disease</td>
<td>GALC</td>
<td>Galactosylceramidase isoform b precursor</td>
<td>NM_001037525</td>
<td>chr17:75,689, 950-75,708,274</td>
</tr>
<tr>
<td>Galactosemia-classic type</td>
<td>GALT</td>
<td>Galactose-1-phosphate uridylyltransferase</td>
<td>NM_000155</td>
<td>chr3:6,49,506, 956-49,538,811</td>
</tr>
<tr>
<td>Biotinidase deficiency</td>
<td>BTD</td>
<td>Biotinidase precursor</td>
<td>NM_000060</td>
<td>chr3:6,49,506, 956-49,538,811</td>
</tr>
<tr>
<td>Congenital adrenal hyperplasia (21-HD)</td>
<td>CYP21A2</td>
<td>Cytochrome P450, family 21, subfamily A</td>
<td>NM_000500</td>
<td>chr6.qbl.hap2:3, 221, 083-3,224,420</td>
</tr>
<tr>
<td>Niemann-Pick disease</td>
<td>SMPD1</td>
<td>Sphingomyelin phosphodiesterase 1</td>
<td>NM_001007593</td>
<td>chr11:5,203, 272-5,204,877</td>
</tr>
<tr>
<td>Retinoblastoma</td>
<td>RB1</td>
<td>Retinoblastoma 1</td>
<td>NM_000321</td>
<td>chr14:87,499, 411-87,529,660</td>
</tr>
<tr>
<td>Metachromatic leukodystrophy</td>
<td>ARSA</td>
<td>Arylsulfatase A precursor</td>
<td>NM_000487</td>
<td>chr14:87,499, 411-87,529,660</td>
</tr>
<tr>
<td>Pompe’s disease</td>
<td>GAA</td>
<td>alpha-glucosidase preprotein, acid</td>
<td>NM_001079803</td>
<td>chr14:87,499, 411-87,529,660</td>
</tr>
<tr>
<td>Huntington’s disease</td>
<td>HD</td>
<td>Huntingtin</td>
<td>NM_002111</td>
<td>chr14:87,499, 411-87,529,660</td>
</tr>
<tr>
<td>Neurofibromatosis 1</td>
<td>NF1</td>
<td>Neurofibromatosis 1</td>
<td>NM_000267</td>
<td>chr14:87,499, 411-87,529,660</td>
</tr>
<tr>
<td>Marfan syndrome-classic</td>
<td>FBN1</td>
<td>Fibrillin 1 precursor</td>
<td>NM_000138</td>
<td>chr14:87,499, 411-87,529,660</td>
</tr>
<tr>
<td>Holt-Oram syndrome</td>
<td>TBX5</td>
<td>T-box 5 isoform 2</td>
<td>NM_080718</td>
<td>chr14:87,499, 411-87,529,660</td>
</tr>
<tr>
<td>Crouzon and Apert Syndrome</td>
<td>FGFR2</td>
<td>Fibroblast growth factor receptor 2 isoform 2</td>
<td>NM_022970</td>
<td>chr14:87,499, 411-87,529,660</td>
</tr>
<tr>
<td>Achondroplasia</td>
<td>FGFR3</td>
<td>mRNA for fibroblast growth factor receptor 3 isoform 1</td>
<td>NM_000142</td>
<td>chr14:87,499, 411-87,529,660</td>
</tr>
<tr>
<td>Duchenne muscular dystrophy</td>
<td>DMD</td>
<td>Dystrophin</td>
<td>NM_000109</td>
<td>chr14:87,499, 411-87,529,660</td>
</tr>
<tr>
<td>Androgenetic alopecia, Androgen insensitivity syndrome</td>
<td>AR</td>
<td>Androgen receptor isoform 2</td>
<td>NM_001011645</td>
<td>chr14:87,499, 411-87,529,660</td>
</tr>
<tr>
<td>Glucose-6-Phosphodehydrogenase Deficiency</td>
<td>G6PD</td>
<td>Glucose-6-phosphate dehydrogenase</td>
<td>NM_001042351</td>
<td>chr14:87,499, 411-87,529,660</td>
</tr>
<tr>
<td>Fragile X syndrome</td>
<td>FMR1</td>
<td>Fragile X mental retardation 1 variant</td>
<td>NM_002024</td>
<td>chr14:87,499, 411-87,529,660</td>
</tr>
<tr>
<td>Ornithine transcarbamylase</td>
<td>OTC</td>
<td>precursor</td>
<td>NM_000531</td>
<td>chr14:87,499, 411-87,529,660</td>
</tr>
</tbody>
</table>
Table 2. Input parameters for primer designing tools

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Minimum</th>
<th>Maximum</th>
<th>Optimal</th>
</tr>
</thead>
<tbody>
<tr>
<td>$T_m$</td>
<td>45</td>
<td>63</td>
<td>55</td>
</tr>
<tr>
<td>Length</td>
<td>18</td>
<td>27</td>
<td>20</td>
</tr>
<tr>
<td>GC (%)</td>
<td>30</td>
<td>80</td>
<td>50</td>
</tr>
</tbody>
</table>

3. Fraction of dimers at different positions (5’ end, 3’ end and in the middle of the sequence)

4. Fraction of trimers

All these parameters are used as features for the classification. The feature vector obtained by concatenating all these features is of size 159.

2.4.2 Feature Selection

A ideal algorithm should extract the most relevant features and eliminate the irrelevant and redundant ones. It is important because throwing away irrelevant features reduces the risk of overfitting and decreases computational complexity. The selection of an optimal subset of features can be carried out by using an appropriately designed performance measure to evaluate their ability to classify the samples. It could not be done using a brute force method, if number of features are huge. The SVM-RFE algorithm proposed in [8], first takes all the features for training a SVM and access the relative importance of the features in the classifier by the feature ranking criterion. The feature ranking criterion for SVM-RFE is

$$w_r = \sum_{x_i \in S} y_i \lambda_i K(x_i, x)$$  \hspace{1cm} (5)

where, $\lambda_i$ is the lagrange multiplier, $K(\cdot, \cdot)$ is the kernel and $x$ is the input data. The SVM is discussed in detail in the next section. It preserves the higher rank features and eliminates the lower rank features. The process is repeated (using only survived features) until all the features are accessed. At the end, rank of all features is obtained. Now the task is to decide the optimal set of features from this ranked list of features. For that, the different numbers of top-ranked features are selected to form a series of different feature subsets. The optimal set of features is decided by computing the performance of the classifier with these selected features, iteratively.

2.4.3 Classification

There are a lot of classification techniques proposed in the literature and support vector machine is one of them. Support vector machines (SVMs) are a set of supervised learning methods used for classification and regression. SVMs belong to a family of generalized linear classifiers which use structural risk minimization for classification. A significant advantage of SVMs is that it looks for a global and unique solution. Additionally, the computational complexity of SVMs does not depend on the dimensionality of the input space and less prone to overfitting. Therefore, in our current study, support vector machine has been used for classification of primer data. One important factor in implementation of SVM is its selection of a suitable kernel. Here, it is empirically found that on our database, SVM using polynomial kernel with degree two performs superior compared to using kernels such as radial basis function and sigmoidal. A brief description of the SVM is given below:

The SVM [9] [10] is a promising classification and regression technique proposed by Vapnik and co-workers at the AT&T Bell Laboratories. Originally, it was focussed for optical character recognition but soon SVM were used in areas like object recognition, face detection, data mining etc. In theory, SVM classification can be traced back to the classical risk minimization approach, which fixes the classification decision function by minimizing the classification risk. For SVM, the linearly separable problem can be treated as the classical classification decision function as:

$$f_w,b = sign(w.x + b)$$ \hspace{1cm} (6)

where $f$ is the classification decision function, $w$, $x$ and $b$ are weights, input sample and bias respectively. However, the SVM is trying to fix the optimal separating hyperplane by constructing the maximum margin between different classes as:

$$\min \left( \frac{1}{2} w^T W \right),$$ \hspace{1cm} (7)

$$y_i(w.x_i + b) >= 1$$ \hspace{1cm} (8)

For a linearly-non separable case the above formula can be extended by introducing a new set of variables $\{ \psi_i \} = 1, 2, ..., L$ as a measurement of violation of constraints as follows:

$$\min \left( \frac{1}{2} w^T W + C(\sum_{i=1}^{L} s_i)^k \right),$$ \hspace{1cm} (9)

$$y_i(w^T x_i + b) >= 1 - \psi_i$$ \hspace{1cm} (10)

where, parameter $C$ and $k$ are used to penalize variables $\psi_i$ is a non-linear function, which maps the input space.
Minimizing the first term in Eq. 9 amounts to minimizing the VC-dimension of the learning machine, minimizing the second term in Eq. 9 controls the empirical risk. Therefore, in order to solve problem, this method needs to construct a set of functions, and implement classical risk minimization on the function set. Here, a Lagrangian method is used to solve the above problem. The Eq. 9 can be written as

\[
\max_{\lambda} : F(\lambda) = \lambda_1 - \frac{1}{2} \lambda^T D \lambda
\]  

(11)

\[
\lambda_i y_i = 0; \quad \lambda_i < C; \quad \lambda_i > 0
\]  

(12)

where, \( \lambda = (\lambda_1, ..., \lambda_L) \), \( D = y_i y_j x_i x_j \) for binary classification and the decision function can be re-written as

\[
f(x) = \text{sgn} \left( \sum_{i=1}^{L} y_i \lambda_i (x_i x_i + b) \right)
\]  

(13)

For a non-linearly separable case the input data is mapped into higher dimension space using a mapping function \( \phi \). SVM does not explicitly do the mapping, it uses a kernel function \( K(x, y) = \langle \phi(x), \phi(y) \rangle \), which maps the input data to the higher dimension space by simply taking the dot product. Kernel function enables the SVM to operate efficiently in high-dimensional space without being adversely affected by the dimensionality of that space. The decision function for this case is defined as:

\[
f(x) = \text{sgn} \left( \sum_{i=1}^{L} y_i \lambda_i K(x_i x_i + b) \right)
\]  

(14)

3 Results and Discussion

High density array and multiplexing PCR are widely used in molecular research. Most critical step of these methods is to design optimal primer/ oligonucleotide with minimum hybridization free energy for target sequence and maximum hybridization free energy for all other target sequences [11]. Primers/ oligonucleotides obtained for various genes mentioned were in the range of 20-25 base pair in length. Our analysis of the data involves two major approaches. One is the free energy parameters based on the secondary structure which plays a critical role in hybridization and second is the sequence features which again determines the stability of the probe.

One hundred and fifty nine (159) different features (as mentioned in section 2.4.1) have been extracted for the primer/ oligonucleotide obtained from Primer3. Free energy parameters which provide information about the minimum energy hybridization structure between the probe and target sequence were calculated from modified NetPrimer (Table 3). Other structural parameters that provide the information of probe stability based on the sequences were also generated. The efficiency of each probe generated was obtained from NetPrimer.

All these features were then subjected to a feature selection method to eliminate the irrelevant and redundant features. Following this, SVM has been used to classify them and to obtain the overall efficiency of the probe generation. Altogether around one million primers/ oligonucleotides were generated for the given set of genes. Out of these around 71,000 has been used for the classification due to computational complexity. Around, 14,000 primers were used with rating more than 90 and around 57,000 with rating below 90. This set is divided into two parts training and testing. We have used around 8000 primers in each category for training and rest i.e. 6,000 and 49,000 primers with rating more than 90 and with rating below 90, respectively for testing. The accuracy of 70% has been achieved on this database.

Fig. 2 shows the distribution of both the classes (rating greater and less 90) along with variation in various sequence and energy parameters. All the parameters taken (length of the primer, GC%, Tm, \( \Delta H \), \( \Delta S \), 3’ end stability and 5’ end \( \Delta G \)) were made flexible in order to determine their effect on the rating. It is evident from the figure that variation in structural parameters did not affect the distribution of rating as the ratio of distribution of highly rated primers to poorly rated primer were uniformly distributed.

Fig. 3(a) shows the effect of GC% at 3’ end of the primer/ oligonucleotide on the rating generated. Though the rating varied significantly from 40 to 100, the variation in 3’ end GC% remained localized in few regions (0%, 15-20%, 32-40%, 50-60% and very few beyond 60). Similar results were observed with 3’ end stability (Fig. 3(b)) where there was huge variation in 3’ free energy and GC% at 3’ end of the primer/ oligonucleotide. Though the classification accuracy obtained is low (70%), but the distribution of highly rated primers/ oligonucleotide was uniform over the whole range of parameters. These results indicate the flexibility of PCR or array reaction that can be performed with these sets of primers/ oligonucleotides.

Based on the generated rating, the primers were divided into two categories, where primers/ oligonucleotides with rating > 90 were placed in one category and < 90 in other category. The value of 90 was taken as cut-off value based on the observation that the primers/ oligonucleotides below this value formed more secondary structure thereby reducing the efficiency of the primer/ oligonucleotide.

A database of all these primer/ oligonucleotide sequences along with various free energy parameters and other extracted features has been created. The main advantage of this database is that it provides the efficiency of the each sequence based on the secondary structure stability. This database will be made public on the internet. Researchers working with these monogenic disorders will be able to obtain relevant information from this database. This database can be helpful in mutation detection, cloning and sequencing analysis.

4 Conclusion

A synthetic database has been generated for the all the primers obtained for 30 monogenic disorders. Around one
Table 3. Various parameters calculated from NetPrimer for the probes generated from Primer3

<table>
<thead>
<tr>
<th>Sequence</th>
<th>Length (bp)</th>
<th>GC (%)</th>
<th>Tm (°C)</th>
<th>ΔH (kcal/mol)</th>
<th>ΔS (kcal/mol)</th>
<th>3’ Stability (kcal/mol)</th>
<th>5’ end Δ G (kcal/mol)</th>
<th>Rating</th>
</tr>
</thead>
<tbody>
<tr>
<td>ctgtgccccccacaaacgc</td>
<td>20</td>
<td>60</td>
<td>62.97</td>
<td>-150.6</td>
<td>-0.379</td>
<td>-8.84</td>
<td>-6.84</td>
<td>84</td>
</tr>
<tr>
<td>tggccagtcagttgtgctgg</td>
<td>24</td>
<td>54.15</td>
<td>68.95</td>
<td>-178.1</td>
<td>-0.449</td>
<td>-8.63</td>
<td>-11.22</td>
<td>76</td>
</tr>
<tr>
<td>ccttcagactttgtccacattgtg</td>
<td>25</td>
<td>48</td>
<td>66.23</td>
<td>-188.1</td>
<td>-0.479</td>
<td>-7.19</td>
<td>-8.18</td>
<td>100</td>
</tr>
<tr>
<td>gaggaccaaggaagtacgggaag</td>
<td>24</td>
<td>58.33</td>
<td>66.43</td>
<td>-190.1</td>
<td>-0.49</td>
<td>-8.18</td>
<td>-7.809</td>
<td>93</td>
</tr>
</tbody>
</table>

million primers were generated for genes associated with them through Primer3. Various free energy parameters for these primers were calculated using modified NetPrimer which was made more robust in determining the secondary structure. Ratings for these primers/ oligonucleotides were also obtained from NetPrimer. Further several other sequence parameters affecting the stability of the probe were calculated. All these features (around 100) were then subjected to a feature selection method to eliminate the irrelevant and redundant feature. Following this, SVM was used to classify them and to obtain the overall efficiency of the probe generation. The overall efficiency of the classification for the data generated is around 70%. Further work is being carried out to extract more features from these parameters in order to achieve better classification accuracy.

References


