Current Protocols in Bioinformatics

April 7, 2005

Dr. Michael Q. Zhang
Cold Spring Harbor Laboratory
1 Bungtown Road
Cold Spring Harbor, NY 11724

Tel: (516) 367-8393
Fax: (516) 367-8461
mzhang@cshl.org

RE: CPBI Unit 2.9 rough pages

1. Please read the rough pages and mark any changes right in the text.
2. If you have large inserts to add, please supply us with a disk and hard copy of the insert(s) and indicate where they should go.
3. Read and answer all the queries on the Author/Editor Query Sheet and return this sheet with your Rough Pages.
4. If you have not returned your signed contract yet, you must return it at this time.
5. **Dr. Zhang should forward all materials by Tuesday, April 19th via Fed Ex or UPS to:**

   Dr. Gary Stormo
   Dept of Genetics
   Washington University School of Medicine
   4566 Scott Avenue, Box 8232
   St. Louis, MO 63110

   Tel: (314) 747-5534
   Fax: (314) 362-7855
   stormo@genetics.wustl.edu

6. **Dr. Stormo should forward all materials by Friday, April 22nd via Fed Ex or UPS to:**

   Allen Ranz
   Current Protocols Editorial Office
   John Wiley & Sons, Inc.
   111 River St.
   Hoboken, NJ 07030

   Tel: (201) 748-6278
   Fax: (201) 748-6207
   aranz@wiley.com

cc: Liz Miranker, Allen Ranz, Gary Stormo

Author/Editor Queries

1. Generally: Per CPBI style, we have reorganized the manuscript to separate out the Web-based instructions (Basic Protocol) from those pertaining to the stand-alone downloaded version (Alternate Protocol). Text has been rearranged; please check carefully for accuracy.

2. Figures, generally: The legends to these figures were composed primarily by the editors. Please examine carefully for correctness and augment with additional information as you deem fit. Figure 2.9.4, in particular, requires more explanation of exactly what is being depicted.

3. Basic Protocol Necessary Resources, Files, next-to-last sentence: We have edited this to read "… (P2, the major promoter) is at +1162…". Is this corrrect? Original manuscript had +162.

4. Basic Protocol, step 2 annotation: Please explain here how you conclude the core promoter sites are overlapping based on what is shown in figure 2.9.3 as it is not clear from the text. Where does +1162 (the annotated TSS) come from, as it does not appear on the figure?

5. Alternate Protocol Necessary Resources, (a) Software: The FTP address *ftp://cshl.edu/pub/science.mzhanglab/promoter/* leads to an error message. Has the site been reorganized since the unit was written? Please adjust appropriately. Also check to see whether any of the Unix commands in step 1 need to be adjusted accordingly.

   (b) Files: OK to copy over the "files" information from the Basic Protocol.

6. Alternate Protocol, (a) generally: In the courier text representing Unix input/output, we have indicated items that appear to be comments (i.e., intructions not to be typed) in bold. Please check to make sure that all of this is correct.

   (b) step 1: Correct that the password is one's e-mail address? It said "internet address" in the original.

   (c) step 3 and annotation: Instead of simply referencing the README file, please insert some additional steps based on the instructions so that the reader can figure out how to install using this protocol alone.

   (d) step 4 and subsequent steps: Please describe in more detail how to run the CorePromoter command-line version. We have attempted to do this here (steps 4 and 5) by editing your original text—please check to see whether this is correct. Does the

user type in the full command `cpromoter_linux mycEPD2k.seq 1147 2000` What are the parameters `1147` and `2000`, and are these entered by the user or output by the program? Please explain.

7. Guidelines for Understanding Results, 1st paragraph: Please expand this paragraph and describe the output results. Describe what the term "score" means. Is 1.00 the best possible score? What is considered a good score, poor score, etc., and what exactly does the score signify. Keep in mind that our readership typically consists of biology graduate students.

8. Background Information, “Theory,” 1st paragraph, 1st sentence: OK to begin the sentence with the phrase “To create a database for developing CorePromoter”? Please edit as necessary to explain why these sequences were extracted.

Using CorePromoter to Find Human Core Promoters

A core promoter is located at the 5′ end of every protein-coding gene. It is the DNA region designated as (−60,+40), with these numbers representing base pairs relative to the Transcriptional Start Site (TSS) at +1. The core promoter is functionally recognized and bound by the Pre-Initiation Complex (PIC), which consists of RNA polymerase II and basal (general) transcription factors (GTFs). The CorePromoter program (Zhang, 1998a) is based on the Quadratic Discriminant Analysis (QDA) method, and is designed to predict the most probable TSSs within a given 2-kb genomic DNA sequence fragment. The predictors in CorePromoter are selected based on the study of known functional motifs (e.g., TATA box, Inr, DPE) and other statistical sequence features within real core promoter regions (Zhang, 1998b). In its simple mode of operation, the program is run with a single 2-kb input genomic DNA sequence for prediction of top-ranking (default = 20 in the Web version) TSSs, or for outputting the successive TSS profile scores (i.e., at every base pair) in the whole region. More often, it is used in conjunction with other gene-finding programs. Users may run CorePromoter remotely on the Web, or it may be downloaded to be run locally on Unix or Linux computers.

Since the CorePromoter program is trained and optimized using 240 bp, i.e., (−160,+80) with respect to the TSS, it is not designed for scanning large genomic regions. Its input sequence size is limited to 2 kb, maximum.

CorePromoter can be used to analyze sequence data either via the Web interface (Basic Protocol) or using the command-line version running locally on a Unix (or Linux) machine (Alternate Protocol). The latter is the most powerful way of using CorePromoter.

USING THE WEB INTERFACE OF CorePromoter TO LOCATE CORE PROMOTER REGIONS

BASIC PROTOCOL

Necessary Resources

Hardware

Any Internet-connected computer

Software

Web browser

Files

The input sequence file in the standard FASTA format (*APPENDIX 1B*) with a title line starting with > and no more than 80 characters per line of sequence. In this example, the sample data file `mycEPD2k.seq` is used, which is the upstream region of the human proto-oncogene c-Myc, extracted from the Eukaryotic Promoter Database (EPD; *http://www.epd.isb-sib.ch/*). This sequence contains the region (−1000,+1000); when using the EPD extraction tool, one must specify "from position −999 to 1000" because its convention has the position "0" for the immediate upstream base of TSS. The sequence is shown in Figure 2.9.1, and according to the annotation, there are two promoters/TSSs; one (P1) is at +1 and the other (P2, the major promoter) is at +1162.This sequence may also be found on the *Current Protocols in Bioinformatics* Web site (follow link at *http://www.interscience.wiley.com/c_p/index.htm*).

Contributed by Michael Q. Zhang

Current Protocols in Bioinformatics (2005) 2.9.1-2.9.12

```
>EP11146 (+) Hs c-myc P1; range -999 to 1000.
CCCGAGACTGTTGCAAACCGGCGCCACAGGGCGCAAAGGGGATTTGTCTCTTCTGAAACC
TGGCTGAGAAATTGGGAACTCCGTGTGGGAGGCGTGGGGGTGGGACGGTGGGGTACAGAC
TGGCAGAGAGCAGGCAACCTCCCTCTCGCCCTAGCCCAGCTCTGGAACAGGCAGACACAT
CTCAGGGCTAAACAGACGCCTCCCGCACGGGGCCCCACGGAAGCCTGAGCAGGCGGGGCA
GGAGGGGCGGTATCTGCTGCTTTGGCAGCAAATTGGGGGACTCAGTCTGGGTGGAAGGTA
TCCAATCCAGATAGCTGTGCATACATAATGCATAATACATGACTCCCCCCAACAAATGCA
ATGGGAGTTTATTCATAACGCGCTCTCCAAGTATACGTGGCAATGCGTTGCTGGGTTATT
TTAATCATTCTAGGCATCGTTTTCCTCCTTATGCCTCTATCATTCCTCCCTATCTACACT
AACATCCCACGCTCTGAACGCGCGCCCATTAATACCCTTCTTTCCTCCACTCTCCCTGGG
ACTCTTGATCAAAGCGCGGCCCTTTCCCCAGCCTTAGCGAGGCGCCCTGCAGCCTGGTAC
GCGCGTGGCGTGGCGGTGGGCGCGCAGTGCGTTCTCTGTGTGGAGGGCAGCTGTTCCGCC
TGCGATGATTTATACTCACAGGACAAGGATGCGGTTTGTCAAACAGTACTGCTACGGAGG
AGCAGCAGAGAAAGGGAGAGGGTTTGAGAGGGAGCAAAAGAAAATGGTAGGCGCGCGTAG
TTAATTCATGCGGCTCTCTTACTCTGTTTACATCCTAGAGCTAGAGTGCTCGGCTGCCCG
GCTGAGTCTCCTCCCCACCTTCCCCACCCTCCCCACCCTCCCCATAAGCGCCCCTCCCGG
GTTCCCAAAGCAGAGGGCGTGGGGGAAAAGAAAAAAGATCCTCTCTCGCTAATCTCCGCC
CACCGGCCCTTTATAATGCGAGGGTCTGGACGGCTGAGGACCCCCGAGCTGTGCTGCTCG
CGGCCGCCACCGCCGGGCCCCGGCCGTCCCTGGCTCCCCTCCTGCCTCGAGAAGGGCAGG
GCTTCTCAGAGGCTTGGCGGGAAAAAGAACGGAGGGAGGGATCGCGCTGAGTATAAAAGC
CGGTTTTCGGGGCTTTATCTAACTCGCTGTAGTAATTCCAGCGAGAGGCAGAGGGAGCGA
GCGGGCGGCCGGCTAGGGTGGAAGAGCCGGGCGAGCAGAGCTGCGCTGCGGGCGTCCTGG
GAAGGGAGATCCGGAGCGAATAGGGGGCTTCGCCTCTGGCCCAGCCCTCCCGCTGATCCC
CCAGCCAGCGGTCCGCAACCCTTGCCGCATCCACGAAACTTTGCCCATAGCAGCGGGCGG
GCACTTTGCACTGGAACTTACAACACCCGAGCAAGGACGCGACTCTCCCGACGCGGGGAG
GCTATTCTGCCCATTTGGGGACACTTCCCCGCCGCTGCCAGGACCCGCTTCTCTGAAAGG
CTCTCCTTGCAGCTGCTTAGACGCTGGATTTTTTTCGGGTAGTGGAAAACCAGGTAAGCA
CCGAAGTCCACTTGCCTTTTAATTTATTTTTTTATCACTTTAATGCTGAGATGAGTCGAA
TGCCTAAATAGGGTGTCTTTTCTCCCATTCCTGCGCTATTGACACTTTTCTCAGAGTAGT
TATGGTAACTGGGGCTGGGGTGGGGGGTAATCCAGAACTGGATCGGGGTAAAGTGACTTG
TCAAGATGGGAGAGGAGAAGGCAGAGGGAAAACGGGAATGGTTTTTAAGACTACCCTTTC
GAGATTTCTGCCTTATGAATATATTCACGCTGACTCCCGGCCGGTCGGACATTCCTGCTT
TATTGTGTTAATTGCTCTCTGGGTTTTGGGGGGCTGGGGGTTGCTTTGCGGTGGGCAGAA
AGCCCCTTGCATCCTGAGCTCCTTGGAGTAGGGACCGCATATCGCCTGTGTGAGCCAGAT
CGCTCCGCAGCCGCTGACTT
```

Figure 2.9.1 The sequence of the upstream region of the human proto-oncogene c-Myc (extracted from EPD), used here as a sample data file (`mycEPD2k.seq`).

1. Point the browser to *http://rulai.cshl.org/tools/genefinder/CPROMOTER/*. In the CorePromoter main page that appears, click Human to obtain the data-input form. Paste the FASTA sequence (including the header line; see Fig. 2.9.1) into the input panel. Using the sample data file sequence, `mycEPD2k.seq`, the screen will appear as shown in Figure 2.9.2.

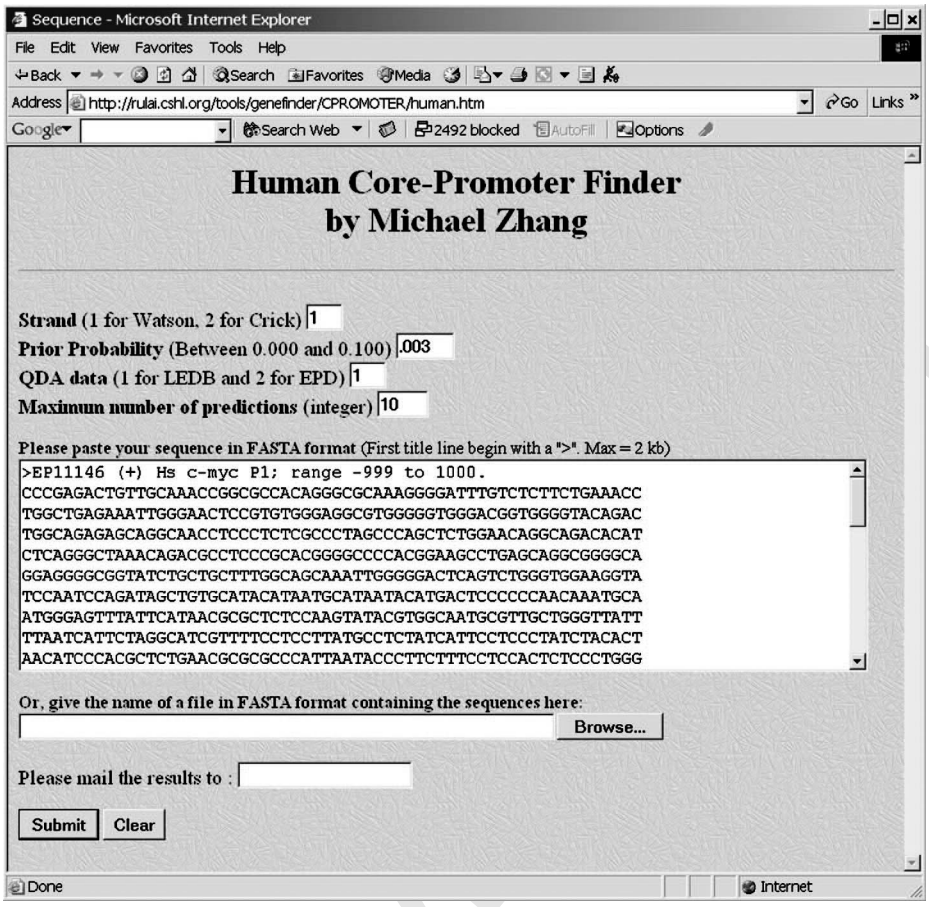

Figure 2.9.2 Screen shot of the Human CorePromoter data-input Web page, with the sequence from the sequence data file `mycEPD2k.seq` pasted into the text field.

Table 2.9.1 Parameters on CorePromoter Data Input Form

Parameter	Description	Explanation given in `readme` file[a]
Strand	`1` for Watson strand (default), `2` for Crick strand	If the Crick strand is chosen, coordinate will be enumerated reversely
Prior Probability	0.003 (default), a positive decimal number less than 1. Prior probability for finding a core promoter.	How likely can a TSS be found by chance?
QDA data	`1` (default) using scores trained from LEDB dataset; `2` using scores trained from EPD dataset[b]	LEDB uses QDA covariant matrix trained by 673 human promoter sequences built from the authors' our nonredundant Lead Exon Database; EPD uses QDA covariant
Maximum number of predictions	`10` (default), an integer less than the length of the sequence − 240	Maximum number of predictions sorted by the profile scores
Output the profile?[c]	`1` for yes; `0` for no (default)	Whether to output the whole profile scores or not

[a] *http://rulai.cshl.org/tools/genefinder/CPROMOTER/readme.htm.*
[b] See Zhang (1998b).
[c] Available only in command-line version (see Alternate Protocol).

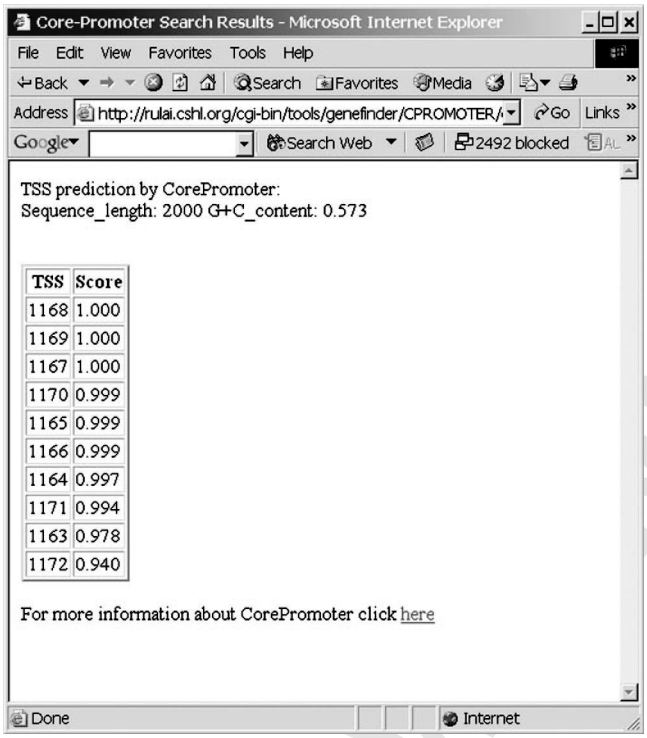

Figure 2.9.3 Screen shot of the resuts from the example run with `mycEPD2k.seq` and default parameters on the Web.

A short description of the CorePromoter program may be viewed in a `readme` *file by clicking the link at the bottom of the CorePromoter main page, which leads to http://rulai.cshl.org/tools/genefinder/CPROMOTER/readme.htm.*

On the data-input form, one can also type in the sequence file name or use the Browse button to upload the sequence file. The program can only take the standard DNA/RNA character symbols (either in upper or lower cases); ambiguous IUPAC symbols will be converted to the standard symbols by a random draw (e.g., `N` *will be converted into* `A`,`C`,`G`, *or* `T` *with equal probability). The five parameters listed on the the data-input form are explained in Table 2.9.1.*

2. Click the Submit button. The result shown in Figure 2.9.3 will appear. One can also ask for the result to be sent back via e-mail by typing in one's e-mail address before submitting.

   It is clear the top 10 predictions are all overlapping, with the major core promoter (P2) where the annotated TSS is at `+1162`.

ALTERNATE PROTOCOL

USING THE COMMAND-LINE VERSION OF CorePromoter TO LOCATE CORE PROMOTER REGIONS

Necessary Resources

Hardware

Any Unix or Linux computer

Software

The CorePromoter software can be downloaded by anonymous FTP at *ftp://cshl.edu/pub/science.mzhanglab/promoter/*. The executable codes for CorePromoter are free for academic users. To obtain source codes (written in Fortran 77) or for

commercial users, one should contact the Cold Spring Harbor Laboratory (CSHL) licensing office (Mr. John Maroney, 1-516-367-8312, *maroney@cshl.edu*).

Files

The input sequence file in the standard FASTA format (*APPENDIX 1B*) with a title line starting with > and no more than 80 characters per line of sequence. In this example, the sample data file `mycEPD2k.seq` is used, which is the upstream region of the human proto-oncogene c-Myc, extracted from the Eukaryotic Promoter Database (EPD; *http://www.epd.isb-sib.ch/*). This sequence contains the region (−1000,+1000); when using the EPD extraction tool, one must specify "from position −999 to 1000" because its convention has the position "0" for the immediate upstream base of TSS. The sequence is shown in Figure 2.9.1, and according to the annotation, there are two promoters/TSSs, one (P1) is at +1 and the other (P2, the major promoter) is at +1162. This sequence may also be found on the Current Protocols in Bioinformatics Web site (follow link at *http://www.interscience.wiley.com/c_p/index.htm*).

Downloading and installing CorePromoter

1. Download the CorePromoter executable file by running an FTP session as follows (items in **bold** are comments).

```
%ftp cshl.edu

Name: anonymous

Password: [your e-mail address]

ftp> cd pub/science/mzhanglab/promoter

ftp> get README (for the updated README file)

ftp> binary

ftp> get cpromoter.tar.Z

ftp> get cpromoter_linux (for linux binary)

ftp> get cppromoter_stat_SUN (for Solaris, statically compiled, does not need the f77 run time library)

ftp> quit
```

2. To `unzip` and `untar` the `tar.Z` file, type:

```
% gunzip cpromoter.tar.Z (or type uncompress cpromoter.tar.Z)

% tar -xvf cpromoter.tar
```

This should create the following files in the `/cp` subdirectory:

```
cp/EPD441_474.DAT

cp/EPD456_504.DAT

cp/EPD471_504.DAT

cp/EPD501_534.DAT
```

```
cp/EPD501_549.DAT
cp/EPD531_564.DAT
cp/EPD546_594.DAT
cp/EPD561_594.DAT
cp/EPD591_624.DAT
cp/EPD591_639.DAT
cp/EPD621_654.DAT
cp/EPD636_684.DAT
cp/EPD651_684.DAT
cp/EPDQDA.DAT
cp/LEDB441_474.DAT
cp/LEDB456_504.DAT
cp/LEDB471_504.DAT
cp/LEDB501_534.DAT
cp/LEDB501_549.DAT
cp/LEDB531_564.DAT
cp/LEDB546_594.DAT
cp/LEDB561_594.DAT
cp/LEDB591_624.DAT
cp/LEDB591_639.DAT
cp/LEDB621_654.DAT
cp/LEDB636_684.DAT
cp/LEDB651_684.DAT
cp/LEDBQDA.DAT
cp/README
cp/cpromoter_OSF
cp/cpromoter_SUN
cp/test.seq
```

The last file (`test.seq`) is just an example of an input DNA sequence for a test run. The format of input sequence file is the standard FASTA format, with a title line starting with > and with no more than 80 characters per line of sequence. A short description of CorePromoter is given in the `README` file found in the abovementioned subdirectory. Before executing the program, one should tell the program where the data files are by defining the environmental variable CPDATA with a correct path, by typing, e.g.:

```
%setenv CPDATA ~/mydatapath/
```

One may have to change the mode and to move the executable to another subdirectory:

```
%chmod u+x cpromoter_linux

%mv cpromoter__linux ~/bin
```

3. Install CorePromoter by doing ??.

   The instruction on how to install CorePromoter is in the `README` *file, which also has a short description about the program and parameters.*

Running CorePromoter

4. Run the command-line version of CorePromoter using the sample data file `mycEPD2k.seq` as follows.

```
% cpromoter_linux mycEPD2k.seq

1147 2000 0
```

5. The results will be printed to the screen as follows:

```
TSS prediction by CorePromoter:

Sequence_length: 2000 G+C_content: 0.573

0.573499978

Top 20 scores are:

TSS Score

1168 1.000

1169 1.000

1167 1.000

1170 0.999

1165 0.999

1166 0.999

1164 0.997

1171 0.994

1163 0.978

1172 0.940

1162 0.937

1161 0.830

1156 0.582

1173 0.550

1157 0.531

1160 0.525
```

```
1158 0.493

1159 0.453

1155 0.441

1153 0.388

Profile statistics: mean = 0.00952324737; Std =
  0.00176317571
```

All these top 20 overlapping predictions are also corresponding to P2, the major core promoter (similar to the Web result of top 10 scores; see Basic Protocol).

6. One can change the default parameters. First, simply type the command name `cpromoter` (i.e., `cpromoter_SUN`, or `cpromoter_OSF`, or `cpromoter_linux`, etc.) alone and hit Enter to see the following output:

```
Usage: cpromoter seqfile strand p0 score top

sequence file in fasta format (required, maximum size =
  2 kb)

strand: 1 (default)- forward; 2 - reverse

p0: prior probability (default 0.003)

score: 1 (default)-LEDB; 2 - EPD

number of output top scores (default = 20)

output the profile? 1 for yes; 0 for no (default = 0)
```

Note that, if one wants to change the default parameters, one must enter all five numbers after the sequence file name (less than five numbers will result in either accepting the default parameters, or the generation of an error message on some systems). The "top" parameter requires two integers, `n1` *and* `n2` *(separated by space); if* `n2` *= 0, it will output top* `n1` *scores; if* `n2` *= 1, it will output all the (profile) scores and ignore* `n1`*. For example:*

```
cpromoter_linux mycEPD2k.seq 1 0.003 1 30 0

1147 2000 0

TSS prediction by CorePromoter:

Sequence_length: 2000 G+C_content: 0.573

0.573499978

Top 30 scores are:

TSS Score

1168 1.000

1169 1.000

1167 1.000

1170 0.999

1165 0.999
```

```
1166 0.999

1164 0.997

1171 0.994

1163 0.978

1172 0.940

1162 0.937

1161 0.830

1156 0.582

1173 0.550

1157 0.531

1160 0.525

1158 0.493

1159 0.453

1155 0.441

1153 0.388

1154 0.376

1152 0.338

1174 0.152

1151 0.144

1150 0.081

1241 0.078

1242 0.072

1002 0.062

987 0.055

1003 0.054

Profile statistics: mean = 0.00952324737; Std =
  0.00176317571
```

Now one can see the minor core promoter P1 at around +1001 when asking for top 30 scores.

GUIDELINES FOR UNDERSTANDING RESULTS

The result output contains sequence length and G + C content followed by the TSS predictions: a coordinate on left and a score on the right (Fig. 2.9.3). The predictions are sorted by the scores unless the whole profile (i.e., scores at all positions and ordered by the positions, when `n2` = 1 in the command-line) is displayed.

Since CorePromoter uses a moving window of 240 bp to scan the input sequence at every position and predicts the likelihood of a TSS starting at the 121-bp position within the window (see Background Information and Zhang, 1998a), overlapping high-scoring predictions often correspond to the same TSS. One could post-process the result by clustering, and this may easily be seen by plotting the profile. Real TSSs may not always have high scores; for instance, the minor core promoter (P1 at the position +1001) in the human c-myc example shown above only has a score of ∼0.6 compared to the score of ∼1.0 for the major core promoter (P2 at the position +1162). Often, the relative signal-to-noise ratio is more indicative than the absolute signal score—i.e., a real TSS may seem to have a low score but be the highest within its surroundings (a relative peak in the profile plot); this is particularly true for TATA-less core promoters.

COMMENTARY

Background Information

Theory

To create a database for developing CorePromoter, 177 human nonredundant promoter sequences were extracted from EPD48 (Bucher and Trifonov, 1986). Each sequence was then extended from the original range (−500,+100) to (−600,+600) by BLASTing GenBank (release no. 100). A few corrections were made after checking against both the original and recent publications. A larger promoter database(673 sequences, called LEDB for Lead Exon Database in CorePromoter options) was extracted (or extended when necessary) from a nonredundant first-exon database (including the flanking regions), which was constructed according to GenBank annotations (Zhang, 1998c). The range for this data was also (−600,+600).

Standard Quadratic Discriminant Analyses (QDA; see, e.g., Zhang, 2000, 2003, and references therein) were used for core-promoter discrimination. All feature variables were 5-tuple scores averaged within a position-specific window. If one defines $f_w(s)$ to be the signal frequency of a 5-tuple s in the window w and $f_b(s)$ to be the background frequency calculated as the average of $f_L(s)$ and $f_R(s)$, where L and R indicate the left and the right nearest-neighbor nonoverlapping windows, then the 5-tuple score is defined by $x(s) = f_w(s)/[f_w(s) + f_b(s)]$. All the f_w values were estimated from the aligned data, and Bayesian priors were used to render all frequencies nonzero (Tanner and Wong, 1987).

In the QDA studies, each sample was a sequence of size 240 bp, which contained two sets of windows of size 30 bp or 45 bp each. As shown in Figure 2.9.4, there were thirteen 5-tuple feature variables altogether. Samples were drawn either from the 177 extended EPD48 sequences or from the 673 data at a 6-bp interval. Again, each sequence was considered to have just one true sample at (−160,+80).

For more information on discriminant analysis and Bayes error, and QDA and its relation to LDA, see *UNIT 4.2*.

Advantages and limitations

Advantages

CorePromoter is simple and fast. It is easily portable and may be incorporated into other programs readily. It can find TSS within a very short genomic DNA sequence, >240-bp, and it can find alternative core promoters.

Limitations

Since it is trained on (−600,+600) promoter sequences, CorePromoter is not designed for genome-wide scans. It only takes a maximum 2-kb input DNA sequence; longer

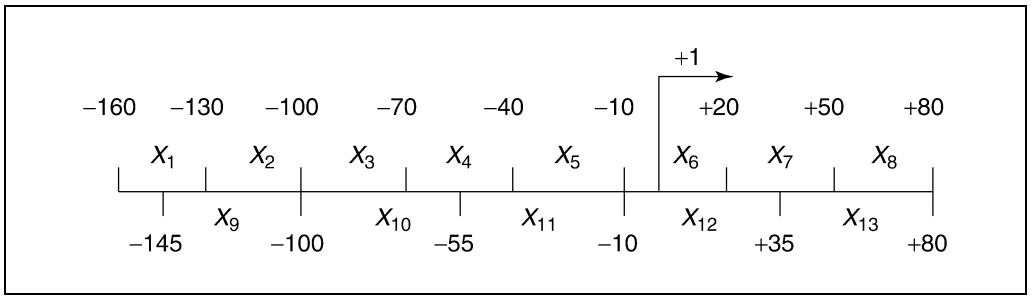

Figure 2.9.4 Graphical depiction of the QDA studies used to create the CorePromoter database.

sequence would have to be truncated and fed in separately.

The user needs to run CorePromoter in conjunction with other gene-finding tools to determine gene structure, and must run the reverse strand separately.

Since the training data (from 1998) may be somewhat limited in comparison to currently available data, the accuracy may be biased toward genes sequenced earlier.

Other options for similar or more comprehensive analysis

There are two programs that are related to CorePromoter—CpGpromoter (Ioshikhes and Zhang, 2000) and FirstEF (Davuluri et al., 2001; *UNIT 4.7*)—that can be used to complement CorePromoter results. CpGpromoter is complementary to CorePromoter in the sense that it can localize a promoter to an ~2-kb region by using CpG islands; CorePromoter may be used for further fine mapping of the TSS at a higher resolution. FirstEF is designed for prediction of the entire first exon of a human protein-coding gene; one would get core promoter and the first intron donor site simultaneously. FirstEF, however, only works for intron-containing genes. For a detailed, worked-out example of how to combine CpGpromoter and other gene-finding tools with CorePromoter, one is referred to Zhang (2000), or *http://rulai.cshl.org/reprints/briefing.pdf*.

Critical Parameters and Troubleshooting

Most often, troubleshooting should start by checking if the input sequence file format is correct (FASTA format). One should always check the sequence length in the output report and see if it is correct. If it is not correct, the discrepancy is most likely caused by extra blank spaces or more than 80 characters per line in the sequence file. One should always test the program with a gene of known structure. If there is no prediction, try to increase "P0" and/or increase the output scores. Plotting the entire profile can also help to spot the score peaks or warn of error (if all scores are zeroes). Normally, if G + C content is low or there is no TATA box, the prediction scores are also be low.

Suggestions for Further Analysis

One should always try to localize a gene region in a large genomic sequence. The best resource would be the full-length cDNAs. Since cDNA library is limited, one must also run several gene- and/or promoter-finding programs (Zhang, 2002). Homology searches against ortholog gene databases and transcription factor binding site analyses are also indispensable.

CorePromoter should also be run in conjunction with other programs that can predict different type of exons and/or different part of the gene structure. Often, the results from these programs can reinforce one another. For example, as described above, one could run CorePromoter (Zhang, 1998a), CpG Promoter (Ioshikhes and Zhang, 2000), FirstEF (a first exon finder; Davuluri et al., 2001), MZEF (an internal coding exon finder; Zhang, 1997), JTEF (A last-exon finder, Tabaska et al. 2001), and Polyadq (a poly(A) site finder, Tabaska and Zhang, 1999). All these programs can be accessed from *http://www.cshl.org/mzhanglab/*. Examples of how one can combine some of these programs for gene-finding may be found in Zhang (2000).

Literature Cited

Bucher, P. and Trifonov, E.N. 1986. Compilation and analysis of eukaryotic POL II promoter sequences. *Nucl. Acids Res.* 14:10009-10026.

Davuluri, R., Grosse, I. and Zhang, M.Q. 2001. Computational identification of promoters and first exons in the human genome. *Nature Genet.* 29:412-417.

Ioshikhes, I. and Zhang, M.Q. 2000. Large-scale human promoter mapping using CpG islands. *Nature Genet.* 26:61-63.

Tanner, M.A. and Wong, W.H. 1987. The calculation of posterior distribution by data augmentation. *J. Am. Stat. Ass*. 82:528-550.

Zhang, M.Q. 1997. Identification of protein coding regions in the human genome based on quadratic discriminant analysis. *Proc. Natl. Acad. Sci. U.S.A.* 94:565-568.

Zhang, M.Q. 1998a. Identification of human gene core-promoters in silico. *Genome Res.* 8:319-326.

Zhang, M.Q. 1998b. A discrimination study of human core-promoters. *In* Proceedings of Pacific Symposium on Biocomputing 1998 (R.B. Altman, A.K. Dunker, L. Hunter, K. Lauderdale, and T.E. Klein, eds.) pp. 240-251. World Scientific, Singapore.

Zhang, M.Q. 1998c. Statistical features of human exons and their flanking regions. *Hum. Mol. Genet.* 7:919-932.

Zhang, M.Q. 2000. Discriminant analysis and its application in DNA sequence motif recognition. *Briefings in Bioinformatics* 1:331-342.

Zhang, M.Q. 2002. Computational prediction of eukaryotic protein coding genes. *Nat. Rev. Genet.* 3:698-709.

Internet Resources

http://rulai.cshl.org/tools/genefinder/CPROMOTER/

CorePromoter Web server.

http://www.cshl.org/mzhanglab

Papers and other related information.

ftp://cshl.org/pub/science/mzhanglab/promoter/

CorePromoter FTP site.

Key References

Zhang, M.Q. 1998a. See above.

This is the original CorePromoter publication.

Zhang, M.Q. 1998b. See above.

This is a statistical analysis of sequence features in human core promoters.

Zhang, M.Q. 2000. See above.

This is a tutorial including more theory and examples on how to combine different gene-structure analysis programs in real applications.

Zhang, M.Q. 2003. See above.

Background on discrimination analysis and QDA.

Contributed by Michael Q. Zhang
Cold Spring Harbor Laboratory
Cold Spring Harbor, New York