BIOINFORMATICS

Bioinformatics, based on National Institutes of Health definition, covers: „Research, development, or application of computational tools and approaches for expanding the use of biological, medical, behavioral or health data, including those to acquire, store, organize, archive, analyze, or visualize such data."

Biological databases can be categorized in several ways, depending mainly on what kind of data they contain (e. g. DNA or protein sequence, 3D structures, gene expression data, metabolic pathways). Since 2004, 'Nucleic Acid Research' has published a yearly issue on databases (http://www3.oup.co.uk/nar/database/c/).

At present, more than 1000 databases exist. Among the main nucleotide databases we find three connected databases developed by the International Nucleotide Sequence Database Collaboration:

1. DDBJ (DNA Data Bank of Japan)/ http://www.ddbj.nig.ac.jp/Welcome-e.html
2. EMBL Nucleotide DB (European Molecular Biology Laboratory) /http://www.ebi.ac.uk/embl/index.html
3. GenBank/NCBI (National Center for Biotechnology Information)/ http://www.ncbi.nlm.nih.gov/ (Fig1)

The NCBI homepage offers access to many important databases (PubMed**,** GenBank, OMIM), as well as some tools. PubMed contains over 17 million biomedical citations and abstracts, while you can get full text journal articles freely in PubMed Central. OMIM (Online Mendelian Inheritance in Man) which deals with genetic disorders is a very useful tool for physicians and genetics researchers.
During the practice sessions we will mainly use NCBI; the aim of the practice is the introduction of practical problems solved with the help of databases.

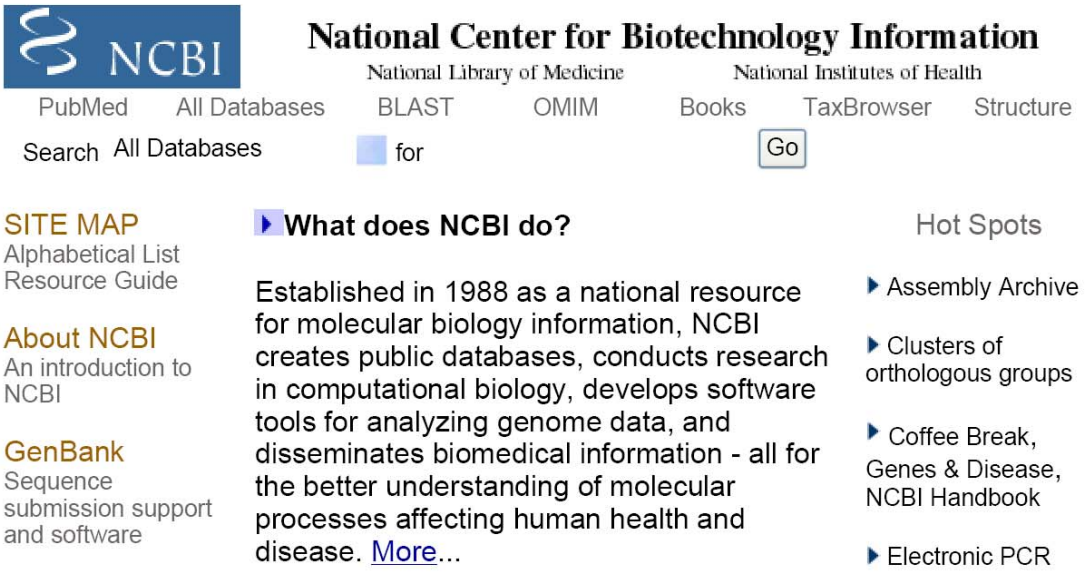


Fig. 1

Applications

A. Identification of *Mycobacterium tuberculosis* by PCR

Tuberculosis is reappearing in most regions of the world, with the estimated rate of new infection being one per second. The diagnosis is usually based on physical, X-ray and laboratory findings. Microbiological diagnosis, namely culturing *Mycobacterium tuberculosis* is the most sensitive and specific method. Unfortunately, since the Mycobacterium multiplies slowly (18 hours/division), culturing gives results only after 2-8 weeks. That is why a PCR based method has been developed. In addition to providing fast results, this method works with very small amounts of specimen sample.
One of the articles based on diagnosis of tuberculosis used the following primers:

Forward primer: 5'-CAC ATG CAA GTC GAA CGG AAA GG-3'
Reverse primer: 5'-GCC CGT ATC GCC CGC ACG CTC ACA-3'

A/I Are these primers specific for tuberculosis?
In order to answer this question we use the NCBI database. The URL address is the following:

http://www.ncbi.nlm.nih.gov/blast/Blast.cgi?PAGE=Nucleotides&PROGRAM=blastn&MEGABLAST=on&BLAST_PROGRAMS=megaBlast&PAGE_TYPE=BlastSearch&SHOW_DEFAULTS=on

This link leads you to *Blastn.* With the help of *Blastn* one can compare the nucleotide query sequence against the nucleotide sequence database.

1. „"*Enter Query Sequence*": copy the sequence of the forward primer (CAC ATG CAA GTC GAA CGG AAA GG)
2. „"*Choose Search Se*t": the others (nr) nucleotide collection has to be chosen (where 'nr' means non-redundant)
3. „"*Program Selection*": Highly similar sequences (megablast)
4. Click on "*Blast*" sign!

After a few seconds searching, the program gives a result page where you can find the sequences most similar to the query sequence ("*Sequences producing significant alignments*"**)**

In detail:

Job Title: Nucleotide sequence (23 letters) (23 bases)
Reference: The original article on blast program
Database: The program looks for matches in the following databases: GenBank+EMBL+DDBJ+PDB
Query= Length=23 base number of the query sequence.

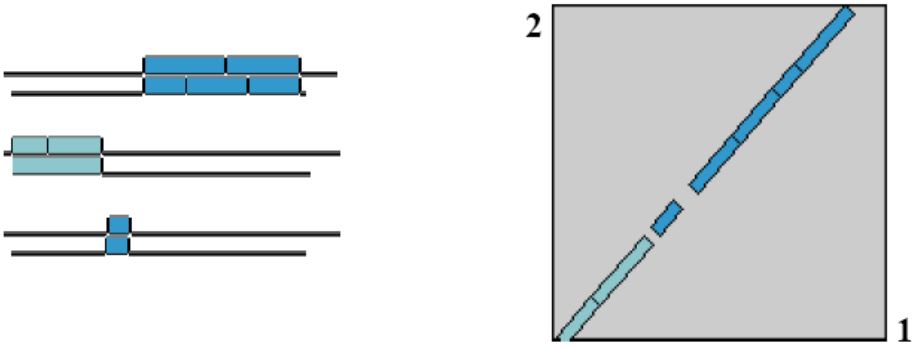
Sequences producing significant alignments:
Accession (accession number): a unique identifier of the sequence
Description: Short description of the sequence (e.g. what kind of gene does it belong to)
Query coverage: 100% means absolute similarity

If you check the descriptions, you find that the 16S rRNA sequence of different bacterial strains show similarity to the query sequence. This means that the primer was designed from the conserved (highly similar, evolutionally important) part of the 16S rRNA sequences, and therefore results in a pan-specific primer.

Check the reverse primer as well with the help of the *Blastn* program!
GCC CGT ATC GCC CGC ACG CTC ACA

A/II Design a Mycobacterium specific primer!
The 16S rRNA gene has a variable part, i.e. it is different in related species. We will find this sequence by comparing two gene sequences, the *Mycobacterium tuberculosis* 16S rRNA gene, (ACCESSION AM283534) and a very similar but non-mycobacterium 16S rRNA gene (ACCESSION EU133135).
On the blast page, you find "*Blast 2 sequences*" program, a BLAST-based tool for aligning two nucleotide sequences.

1. http://www.ncbi.nlm.nih.gov/blast/bl2seq/wblast2.cgi
2. Copy the accession numbers into window "sequence1" and "sequence2".
(sequence1: AM283534 sequence2: EU133135)
3.Click on the "align" sign! The upcoming result page shows which part is similar/identical in the two sequences (Fig. 2)



If we check it in detail, the alignment stops at base 452 of the Mycobacterium sequence, and restarts at base 483. So we can assume that if we design a primer from 452-483, it will be specific for the Mycobacterium. Hence, we shall choose one primer from this segment, and the second one from its flanking sequences to give a product size that is easily amplified in the reaction and conveniently analyzed.

Now that we know the target sequence, we need a primer design program. There are lots of available programs; we will use a freeware program, namely *Primer3.*
Follow the link:
4. *http://frodo.wi.mit.edu/cgi-bin/primer3/primer3_www.cgi*
Open the program and copy the sequence into it (Fig 3). (The sequence starts at base 451, and includes the variable region, and additional nucleotides, so one can amplify a 200-400 bases long DNA fragment.

451 caccatcgac gaaggtccgg gttctctcgg
481 attgacggta ggtggagaag aagcaccggc caactacgtg ccagcagccg cggtaatacg
541 tagggtgcga gcgttgtccg gaattactgg gcgtaaagag ctcgtaggtg gtttgtcgcg
601 ttgttcgtga aatctcacgg cttaactgtg agcgtgcggg cgatacgggc agactagagt
661 actgcagggg agactggaat tcctggtgta gcggtggaat gcgcagatat caggaggaac
721 accggtggcg aaggcgggtc tctgggcagt aactgacgct gaggagcgaa agcgtgggga
781 gcgaacagga ttagataccc tggtagtcca cgccgtaaac ggtgggtact aggtgtgggt
841 ttccttcctt gggatccgtg ccgtagctaa cgcattaagt accccgcctg gggagtacgg
901 ccgcaaggct aaaactcaaa ggaattgacg gggcccgca caagcggcgg agcatgtgga
961 ttaattcgat gcaacgcgaa gaaccttacc tgggtttgac atgcacagga cgcgtctaga

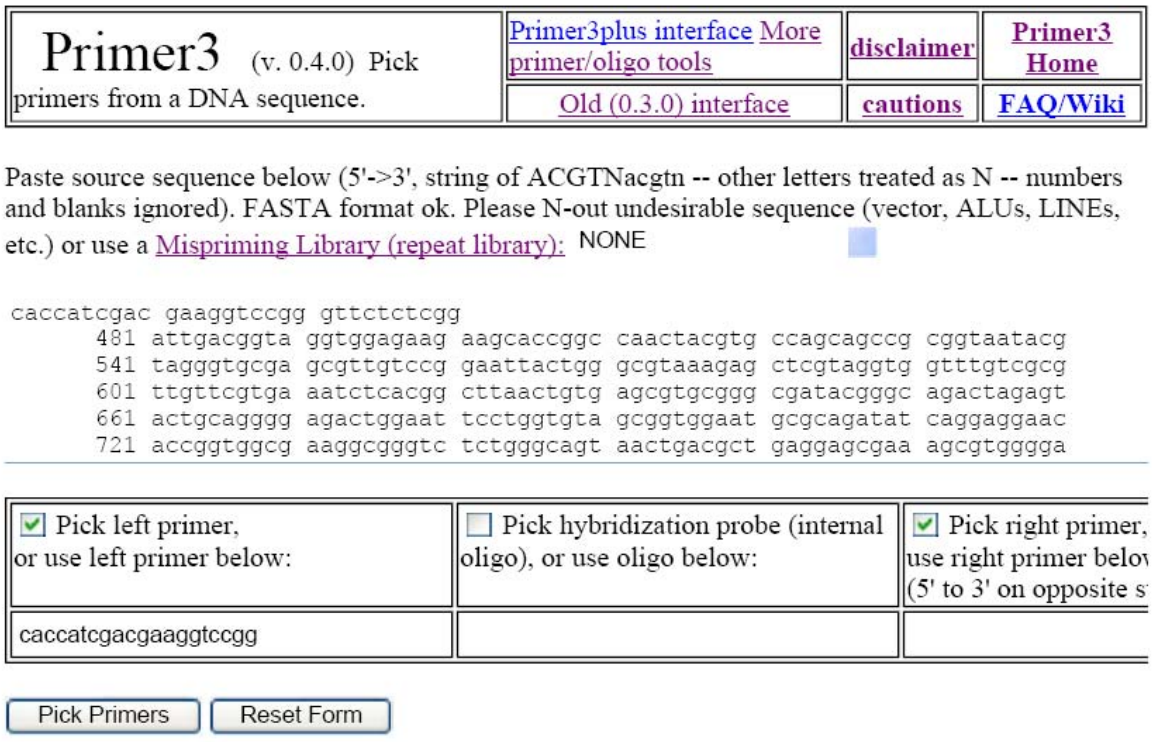


Fig. 3

We will select the forward primer sequence, since it has to be in the variable region. Copy the following sequence into the "*Pick left primer, or use left primer below*" window:
451 caccatcgacgaaggtccgg 470

Click on *pick primers* sign! (Fig. 3)

The program won't accept the chosen left (forward) primer: "*WARNING: Left primer is unacceptable: Tm too high*": which means the difference between the melting temperatures (Tm) of the forward and the possible reverse (right) primers is too big.
The Tm of the primers determines the annealing temperature where primers bind to the single stranded template DNA. Since we use a single annealing temperature during the PCR reaction, the Tm of the primers should be as close as possible.
The Tm also has an important role in the outcome of the reaction: too low a Tm results in non-specific binding, multiplex products, while too high a Tm makes primer binding difficult, resulting in a low yield.
Primer 3 program uses Tm values in the range 57C°-63C°.
The Tm depends on the primer length and GC content.
$Tm = 4(G+C) + 2(A+T)^{o}C$, where G, C, A and T is the number of times a particular nucleotide is present in the sequence.

Let's choose a primer that has a lower Tm!

E.g.:
471 gttctctcggattgacggta 490
The program now accepts the primer, gives the main features of the primer, and shows the DNA fragment that will be amplified during the PCR reaction (Fig. 4).

Primer3 Output

WARNING: Numbers in input sequence were deleted.

```
No mispriming library specified
Using 1-based sequence positions
OLIGO            start  len      tm     gc%   any   3' seq
LEFT PRIMER         21   20   57.74   50.00  3.00  2.00 gttctctcggattga
RIGHT PRIMER       267   20   60.20   50.00  6.00  2.00 cctcctgatatctgc
SEQUENCE SIZE: 570
INCLUDED REGION SIZE: 570

PRODUCT SIZE: 247, PAIR ANY COMPL: 3.00, PAIR 3' COMPL: 1.00

    1 caccatcgacgaaggtccgggttctctcggattgacggtaggtggagaagaagcaccggc
                           >>>>>>>>>>>>>>>>>>>>

   61 caactacgtgccagcagccgcggtaatacgtagggtgcgagcgttgtccggaattactgg


  121 gcgtaaagagctcgtaggtggtttgtcgcgttgttcgtgaaatctcacggcttaactgtg


  181 agcgtgcgggcgatacgggcagactagagtactgcaggggagactggaattcctggtgta


  241 gcggtggaatgcgcagatatcaggaggaacaccggtggcgaaggcgggtctctgggcagt
              <<<<<<<<<<<<<<<<<<<
```

Fig. 4

5. With the help of the *Blastn* program, check if the chosen left/forward primer (gttctctcggattgacggta) is really specific for Mycobacterium. ("*Choose Search Set*": check if the other (nr) nucleotide collection is marked)
6. Is the primer specific for *Mycobacterium tuberculosis*? Do you think it is possible to differentiate between Mycobacterium species or subspecies with the help of PCR?

B. Factor V mutation

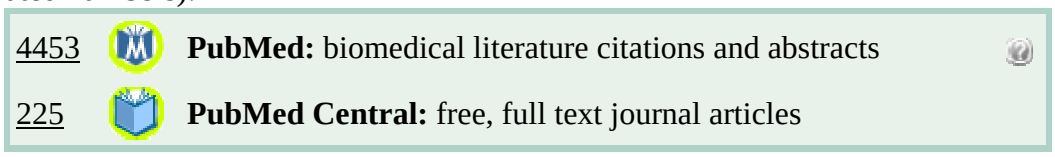
One of the known mutations of factor V is the Leiden mutation, a point-mutation where arginine is replaced by glutamine at position 506. (With one-letter amino acid code: R506Q)

The PCR diagnosis uses two pair of primers: primer pair "H" gives a PCR product if there is no mutation (amino acid 506 is arginine); while primer pair "S" gives a PCR product if residue 506 is replaced by glutamine.

B/I Design the two pairs of primers!

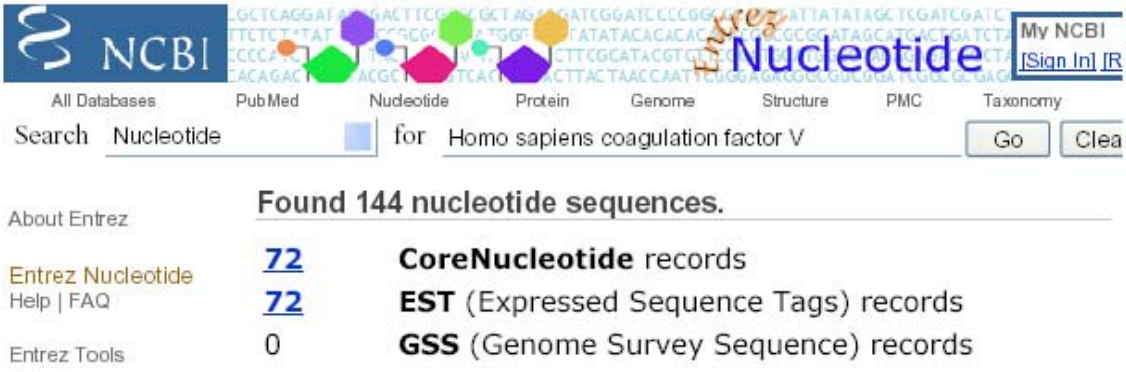
1. Copy the following keywords into NCBI search (All Databases) (Link: *http://www.ncbi.nlm.nih.gov/*): *Homo sapiens coagulation factor V*

At present, there are the following results: (since databases are expanding, you might see elevated numbers):



This means that 4453 articles contain the query words, and among those 225 articles are freely available.

This means that 72 nucleotide sequence names contain the keywords. Among them you can find complete and partial sequences for the cDNA, and splice variants.
We get the same result choosing "nucleotide" in the NCBI search:



(Expressed Sequence Tag or 'EST' is a short stretch of unique sequence derived from the cDNA. It contains only a fragment of the gene.

Click on "*CoreNucleotide records*"!
At the time of writing this practical, the sequence that appeared first was NM_000130:

NM_000130 *Homo sapiens coagulation factor V (proaccelerin, labile factor) (F5), mRNA*
Click on this sequence.

There is some general information:

LOCUS: NM_000130 (accession number) 9179 bp (number of bases) mRNA linear
DEFINITION: Homo sapiens coagulation factor V (proaccelerin, labile factor) (F5), mRNA.
ACCESSION (accession number): NM_000130
SOURCE: Homo sapiens (human)
ORGANISM: Homo sapiens
REFERENCE: there are many references. (Thus, the sequence seems reliable, since many researchers refer to it.)

Features:

Source: 1..9179 (number of bases)
/organism="Homo sapiens"
/mol type="mRNA"
/db_xref="taxon:9606" (taxonomy database: “The NCBI taxonomy database contains the names of all organisms that are represented in the genetic databases with at least one nucleotide or protein sequence”)
/chromosome="1"
/map="1q23"
gene 1..9179
CDS 146..6820 (coding sequence)
Accession number in protein database and other cross-references:
/protein_id="NP_000121.2"
/db_xref="GI:105990535"
/db_xref="CCDS:CCDS1281.1"
/db_xref="GeneID:2153"
/db_xref="HGNC:3542"
/db_xref="HPRD:01964"
/db_xref="MIM:227400"

Translation of the nucleic acid sequence for the protein, based on the one-letter amino acid code (Here we show a shortened version):
/translation="MFPGCPRLWVLVVLGTSWVGWGSQGTEAAQLRQFYVAAQGISWS
YRPEPTNSSLNLSVTSFKKIVYREYEPYFKKEKPQSTISGLLGPTLYAEVGDIIKVH
F
KNKADKPLSIHPQGIRYSKLSEGASYLDHTFPAEKMDDAVAPGREYTYEWSISED
SGP
THDDPPCLTHIYYSHENLIEDFNSGLIGPLLICKKGTLTEGGTQKTFDKQIVLLFAV
F

DESKSWSQSSSLMYTVNGYVNGTMPDITVCAHDHISWHLLGMSSGPELFSIHFN
GQVL
EQNHHKVSAITLVSATSTTANMTVGPEGKWIISSLTPKHLQAGMQAYIDIKNCPK
KTR
NLKKITREQRRHMKRWEYFIAAEEVIWDYAPVIPANMDKKYRSQHLDNFSNQIG
KHYK
KVMYTQYEDESFTKHTVNPNMKEDGILGPIIRAQVRDTLKIVFKNMASRPYSIYP
HGV
TFSPYEDEVNSSFTSGRNNTMIRAVQPGETYTYKWNILEFDEPTENDAQCLTRPY
YSD
VDIMRDIASG
sig_peptide (signal peptide) 146..229
mat_peptide (mature peptide) 230..6817
polyA_signal 6948..6953 ?
polyA_site 6967

Let's find amino acid residue 506 in the nucleotide sequence!
506 amino acid (506x3) = 1518 bases.
We have to add 229 bases since mature factor V starts after the signal peptide.

1518+229=1747

This means amino acid 506 is coded by bases 1745-1747.

Find these 3 nucleotides!

1741 caggcgagga atacagaggg cagcagacat cgaacagcag gctgtgtttg ctgtgtttga

Since CGA codes for arginine, the reference sequence does not contain a mutation.
Conversion of CGA to CAA by point mutation changes arginine to glutamine.

Design primers that amplify the healthy (mutation-free) sequence.
We choose Primer3 again:
http://frodo.wi.mit.edu/cgi-bin/primer3/primer3_www.cgi

Copy sequence NM_000130 into the empty window.

A SNP (single nucleotide polymorphism) can be detected most efficiently by PCR if the location of the possible point mutation is at the 3' end of the primer.
Copy the following sequence into window "*Pick left primer, or use left primer below*":

agcagatccctggacaggcg

Click on "*pick primers*"!

The program won't accept this left (forward) primer:

WARNING: Left primer is unacceptable: Tm too high/High end self complementarity/High 3' stability
It is better to find another primer, since high self complementarity makes secondary structure formation very likely. Secondary structures including hairpins and self-dimers will lower the efficiency of the PCR reaction. Thus, we shall use a reverse primer, that will end with the nucleotide in question at its 3'end.

On the sense strand, the reverse primer will bind to the following sequence (in red/underline):

1741 caggcgagga atacagaggg cagcagacat cgaacagcag gctgtgtttg ctgtgtttga
The primer itself will be complementary (antisense) to this:
3'ctcct tatgtctccc gtcgt 5'

We have to write this sequence (or every other sequence) in the 5' 3' direction in the *Primer3* program. Please copy the following sequence:
5'tgctgccctctgtattcctc 3'into the window *"Pick right primer, or use right primer below"*

Copy into the window *"paste your sequence"* sequence NM_000130.

Click on *pick primers*.
We get a proper primer pair this time.
Check if the primers are really specific (bind only to the sequence we would like to amplify) with the help of *Blastn* program.
Link:
http://www.ncbi.nlm.nih.gov/blast/Blast.cgi?PAGE=Nucleotides&PROGRAM=blastn&MEGABLAST=on&BLAST_PROGRAMS=megaBlast&PAGE_TYPE=BlastSearch&SHOW_DEFAULTS=on
(On the search page you have to choose "Database: Human")

For the "S" pair of primers (which amplifies only the mutated version) we change the 3' C to T (in the coding strand: G to A):
5' tgctgccctctgtattcctt 3'

Check this primer for specificity as well.

B/II Let's examine if this mutation could be detected by PCR-RFLP!
(In this case we amplify the mutation - bearing fragment by PCR, then digest it with a restriction endonuclease (RE). If the mutation changes the RE recognition site, we will get more/fewer fragments after digestion with the appropriate endonuclease compared to the healthy (mutation free) sample.
Is there any RE cleavage site that could be affected by the mutation? We will answer this question with the help of the following program:
http://tools.neb.com/NEBcutter2/index.php (Fig. 5)

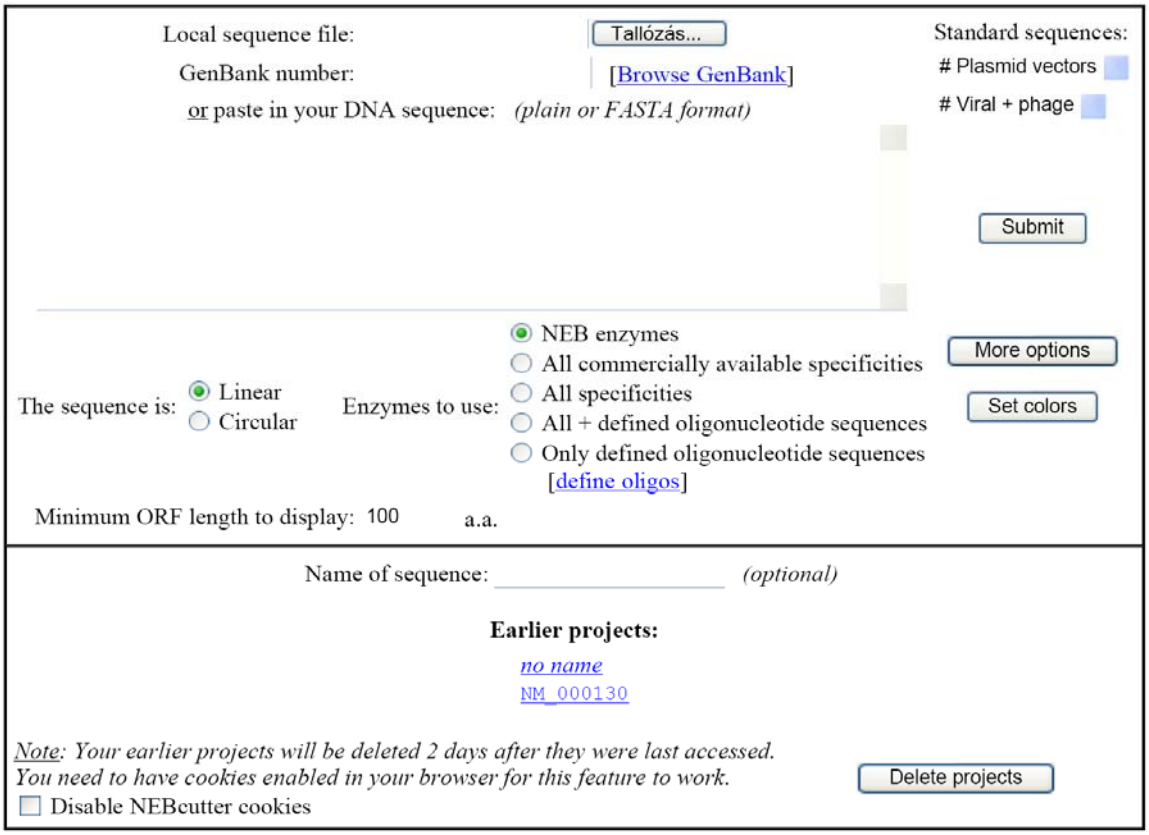


Fig. 5

We could write the accession number into window "*GenBank number*", but we would get a very complicated picture. Let's assume that we amplify the gene from nucleotide 1731 to nucleotide 1750 by PCR (in a real experiment we would work with longer fragments).

Wild type sequence: gatccctgga caggcgagga atacagaggg
Mutant sequence: gatccctgga caggcaagga atacagaggg

Copy the wild type sequence into window "*or paste in your DNA sequence*", then click on "*submit*". On the results page, we have the sequence, the relevant restriction endonucleases and their cleavage sites (Fig. 6). By moving the cursor to the particular RE, its recognition site is appears (red underline).

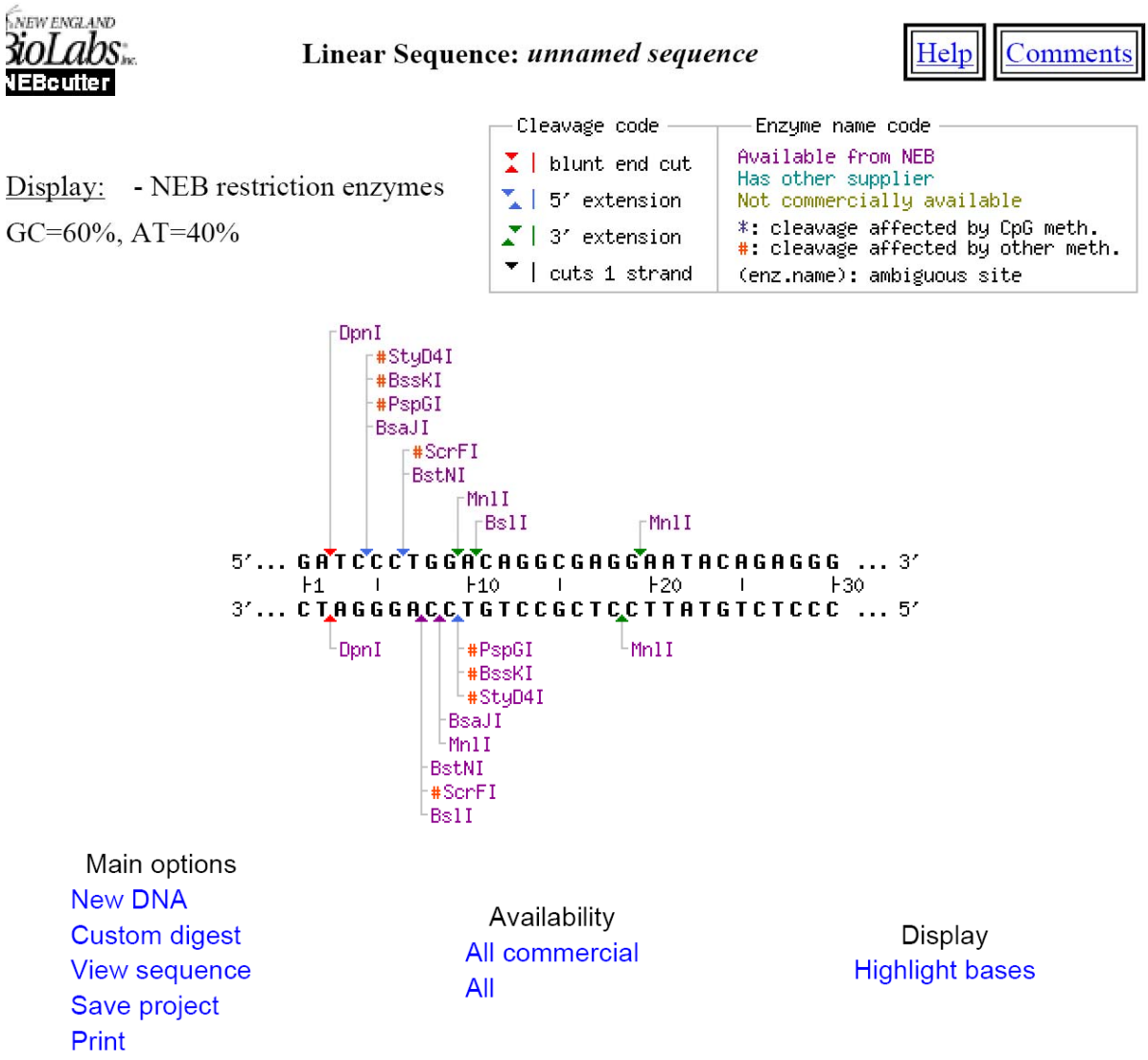


Fig. 6

If the mutation changes the nucleotide in position 16 in the fragments, the *MnlI* RE first recognition site could disappear. (The *MnlI* RE is different from the restriction endonucleases we have met so far; this enzyme recognizes a non-palindromic sequence and its cleavage site is different from the recognition site.)

Select the option "*custom digest*". The program shows which enzyme cuts the fragment, how many times, and what is its favored buffer (1,2,3,4).

Choose *MnlI* RE, than click "*digest*"! The results page shows the fragment with the cleavage site (Fig. 7).

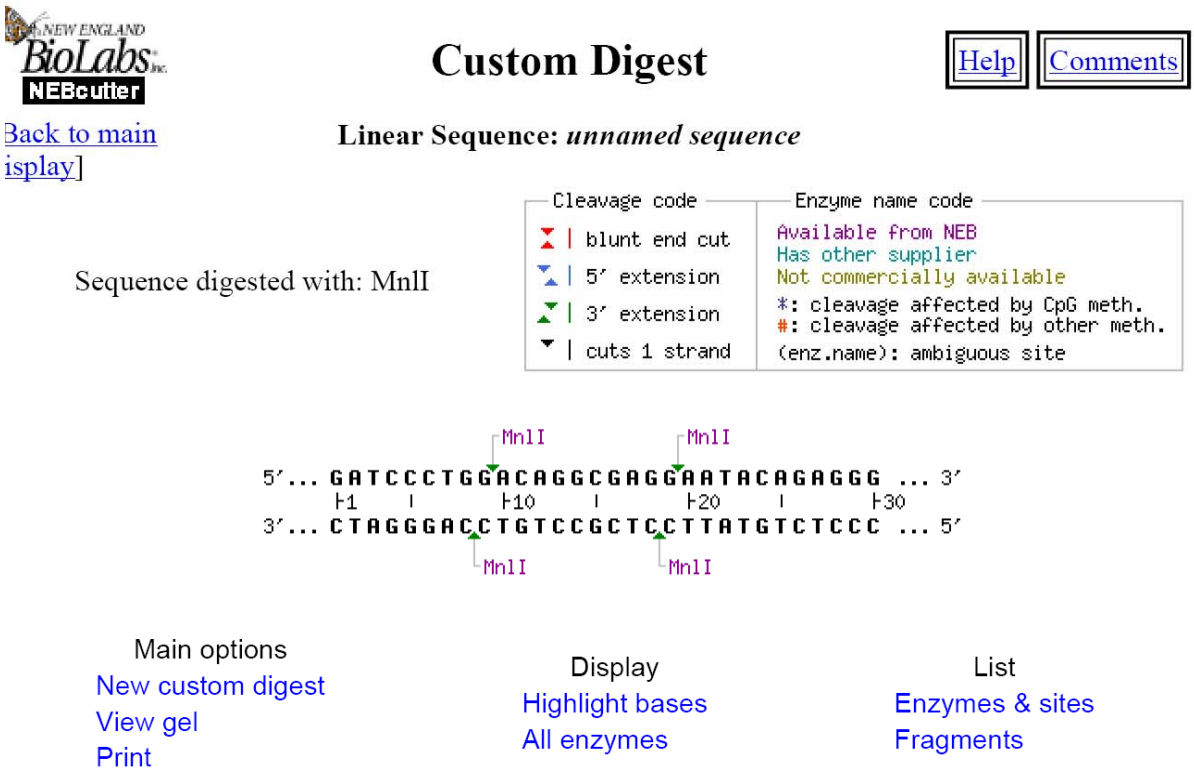
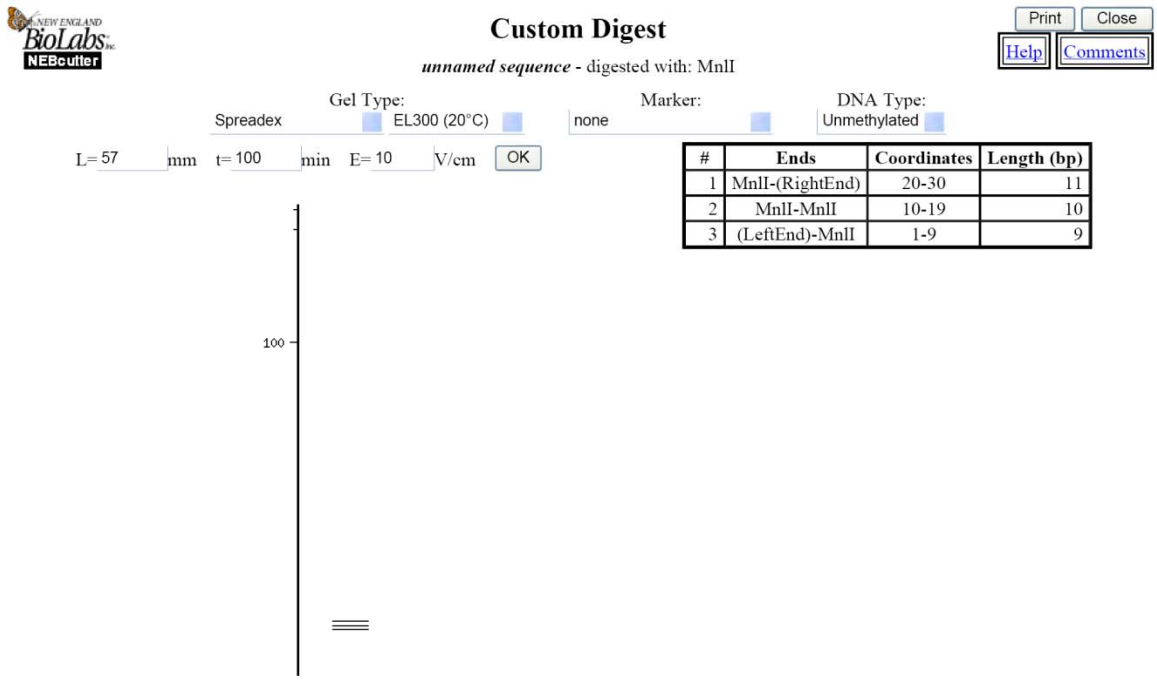


Fig. 7

By clicking on option "*view gel*", the electrophoretic pattern of fragments resulting from digestion is displayed (Fig 8). Since the difference in the length of the fragments is only 1-2 bases, we need to choose the best separating gel to show these differences: ("*gel type*": Spreadex). Now we can visualize the three fragments formed by the cleavages at the two MnII sites.



#	Ends	Coordinates	Length (bp)
1	MnlI-(RightEnd)	20-30	11
2	MnlI-MnlI	10-19	10
3	(LeftEnd)-MnlI	1-9	9

Fig 8

Copy the mutant sequence into window "*or paste in your DNA sequence:*", then click on "*submit*". The first recognition site of *MnlI* disappeared, so we get two fragments instead of three. Choosing the option "*custom digest"* shows a cut at only one *MnlI* site. Choose the RE *MnlI*, then click on "*digest*". The resulting page shows the sequence with one cleavage site, confirmed by the "view gel" / Spreadex option as well.