BIOINFORMATICS

Bioinformatics, based on National Institutes of Health definition, the following: „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 differently; mainly depending on what kind of data they are containing (e. g. DNA or protein sequence, 3D structures, gene expression data, metabolic pathways). Nucleic Acid Research has published a yearly issue on databases since 2004 (http://www3.oup.co.uk/nar/database/c/).

To date more than 1000 databases exist. Among the main nucleotide databases we find three connected database 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)

NCBI homepage offers many important databases (PubMed**,** GenBank, OMIM) and some tools as well. 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) containing genetic disorders is a very useful tool concerning physicians and genetics researchers.
During the practice we will mainly use NCBI; the aim of the practice is the introduction of practical problems solved by the help of databases.

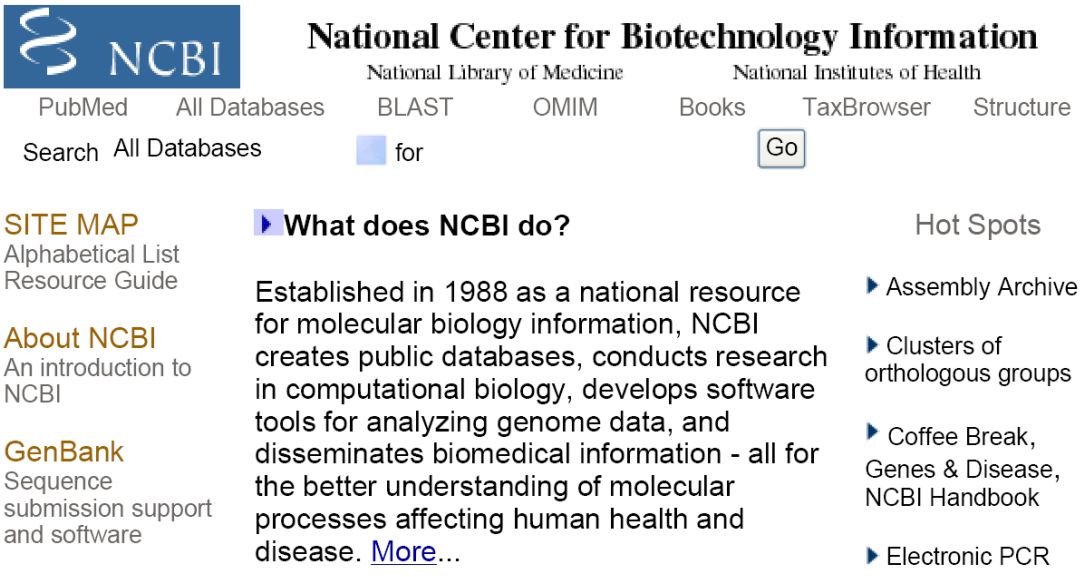


Fig 1

Applications

A. Diagnosis of *Mycobacterium tuberculosis* by PCR

Tuberculosis is a resurgent disease in most regions of the world, leading to new infections one per second. The diagnosis based on physical, X-ray and laboratory findings. Microbiological diagnosis namely culturing Mycobacterium tuberculosis the most sensitive and specific method. Unfortunately, since Mycobacterium multiplies slowly (18 hours/division), culturing gives result after 2-8 weeks. That was the reason to work out a quick PCR based method that requires small amount of specimen as well.
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 link 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

The link lead 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 Set*": the others (nr) nucleotide collection has to be chosen (where 'nr' means nonredundant)
3. „*Program Selection*": Highly similar sequences (megablast)
4. Click on „*Blast*" sign!

The program after a quick search gives a result page where you can find the most similar sequences to the query sequence. („*Sequences producing significant alignments*"**)**

In details:

Job Title: Nucleotide sequence (23 letters) (23 bases)
Reference: The original article on blast program
Database: The program looking 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): 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 to the conserved (highly similar, evolutionally important) part of the 16S rRNA sequences, results 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!
16S rRNA gene has a variable part that is different among the species. We will find this sequence in a simplified way. Let's compare two gene sequence, Mycobacterium tuberculosis 16S rRNA gene, (ACCESSION AM283534) and a very similar but not 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 „blast" sign! The upcoming result page shows which part is similar/identical between the two sequences (Fig2)

```
Query  403   CAGCGACGCCGCGTGGGGGATGACGGCCTTCGGGTTGTAAACCTCTTTCACCATCGACGA  462
             ||||||||||||||| |||| || ||||||||||||||||||||||||  |  |||||
Sbjct  361   CAGCGACGCCGCGTGAGGGAAGAAGGCCTTCGGGTTGTAAACCTCTTTCAGGAGGGACGA  420

Query  463   AGGTCC--GGGT---TCTCTCGGAT--TGACGGTAGGTGGAGAAGAAGCACCGGCCAACT  515
             ||   |  ||||   |  | | ||   ||||||||  |  ||||||||| ||||| ||||
Sbjct  421   AGCCACTCGGGTGAATAGCCCAGAGGGTGACGGTACCTCCAGAAGAAGCCCCGGCTAACT  480
```

Fig2

If we check it in details, the alignment stops at the 452. bases of the Mycobacterium sequence, and restarts at the 483. bases. So we can assume the primer that will bind somewhere between 452-483 will be specific for Mycobacterium.

Now that we know the target sequence we need a primer designer program. From the lots available software we will use a freeware, namely *Primer3*.

Follow link:

4. *http://frodo.wi.mit.edu/primer3/*

Open the program and copy the sequence into it (Fig3) (the sequence starts at the 451. bases, includes the variable part, 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

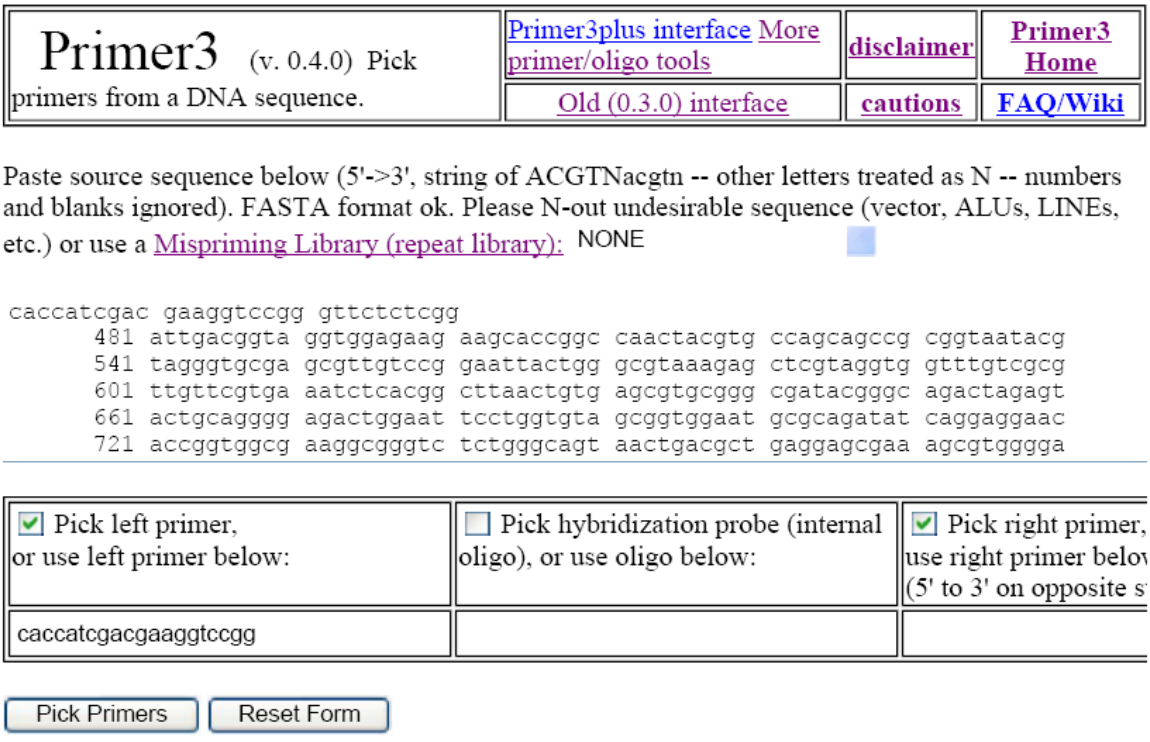


Fig3

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

Click on *pick primers* sign! (Fig3)

The program won't accept the chosen left (forward) primer:
„*WARNING: Left primer is unacceptable: Tm too high*": that means the difference is too big between the melting temperatures (Tm) of the forward and the possible reverse (right) primers.
The Tm of the primers determines the annealing temperature where primers bind to the single stranded template DNA. Since we use one 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: low Tm results in non-specific binding, multiplex products, while high Tm makes primer binding difficult, resulting low yield.
Primer 3 program uses Tm between 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 the regarding nucleotides

Let's choose a primer that has lower Tm!

Eg:
471 gttctctcggattgacggta 490
The program accept the primer, gives the main features of the primer and shows the DNA fragment that will be amplified during the PCR reaction. (Fig4)

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 *Blastn* program check if the chosen left/forward primer (gttctctcggattgacggta) is really specific for Mycobacterium. („*Choose Search Set*": check if the others (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 the PCR?

B. Factor V mutation

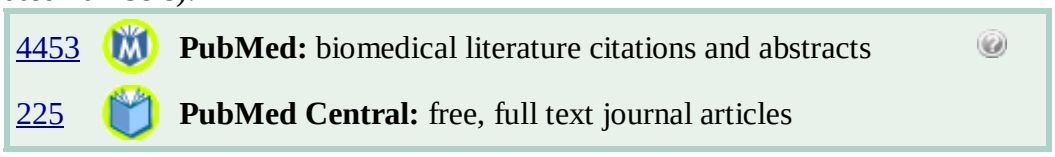
One of the known mutations of factor V is the Leiden mutation; a point-mutation where the arginine will be replaced by glutamine at the 506. amino acid. (With one-letter amino acid code: R506Q)

The PCR diagnosis uses two pair of primers; primer pair "H" gives PCR product if there are no mutation (amino acide 506. is arginine), while primer pair "S" gives PCR product if the 506. amino acide replaced by glutamine.

B/I Design the two pair of primers!

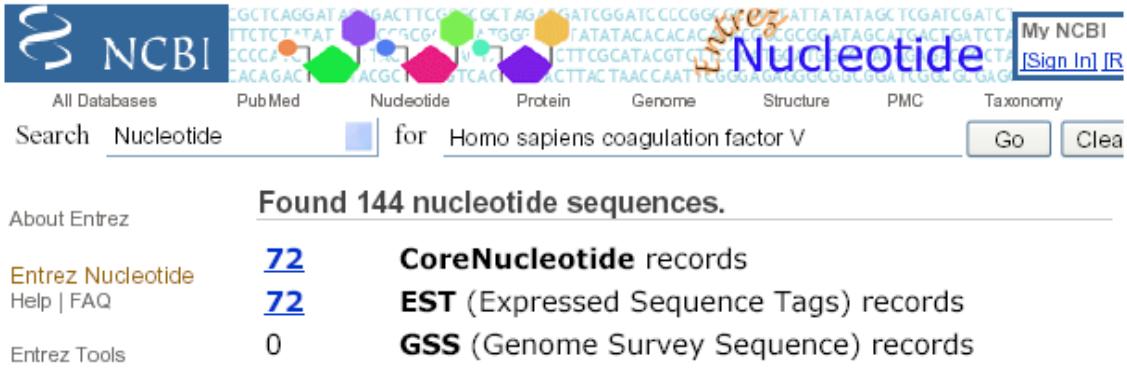
1. Copy the following keywords into NCBI search (All Databases) (Link: *http://www.ncbi.nlm.nih.gov/*): *Homo sapiens coagulation factor V*

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



That means 4453 articles contain the query words and among those 225 articles can be reached by anyone.

That means 72 nucleotide sequence name contains the keywords. Among them you find complete and partial cDNA, splice variant, so on.
We get the same result choosing „nucleotide" in NCBI search:



(Expressed Sequence Tag or EST is a short cDNA derived not necessarily protein coding sequence.)
Click on „*CoreNucleotide records*"!

Search for the following sequence: Accession: NM_000130.4
NM_000130 *Homo sapiens coagulation factor V (proaccelerin, labile factor) (F5), mRNA*

Click on this sequence.

There are general informations:

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: The sequence seems reliable since there is high number of reference (it was published earlier by many).

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= Translation of nucleic acid sequence for protein, based on one-letter amino acid code
sig_peptide (signal peptide) 146..229
mat_peptide (mature peptide) 230..6817
polyA_signal 6948..6953
polyA_site 6967

Let's find amino acid 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 mutation.
Conversion 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/primer3/

Copy sequence NM_000130 into the empty window.

(SNP, single nucleotid polimorfism) can be detected succesfully by PCR if the possible place of 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 comlementarity makes secondary structure formation very possible. Secondary structures including hairpins, self-dimers will lower PCR reaction specificity, and came-out.
So be the possibly mutated nucleotide at the 3' end of the reverse/right primer.

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

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

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

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

Click on *pick primers*.

We get a proper primer pair this time.
Check with the help of *Blastn* program if the primers are really specific (bind only to the sequence we would like to amplify). 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 primer (that amlifies the mutated version only) 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 amlify the mutation bearing fragment by PCR, then digest it with restriction endonuclease (RE). If the mutation changes the RE recognition site (appeared/disappeared), we will get more/less fragment after the digestion compared to the healthy (mutation free) sample.
Is there any RE 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)

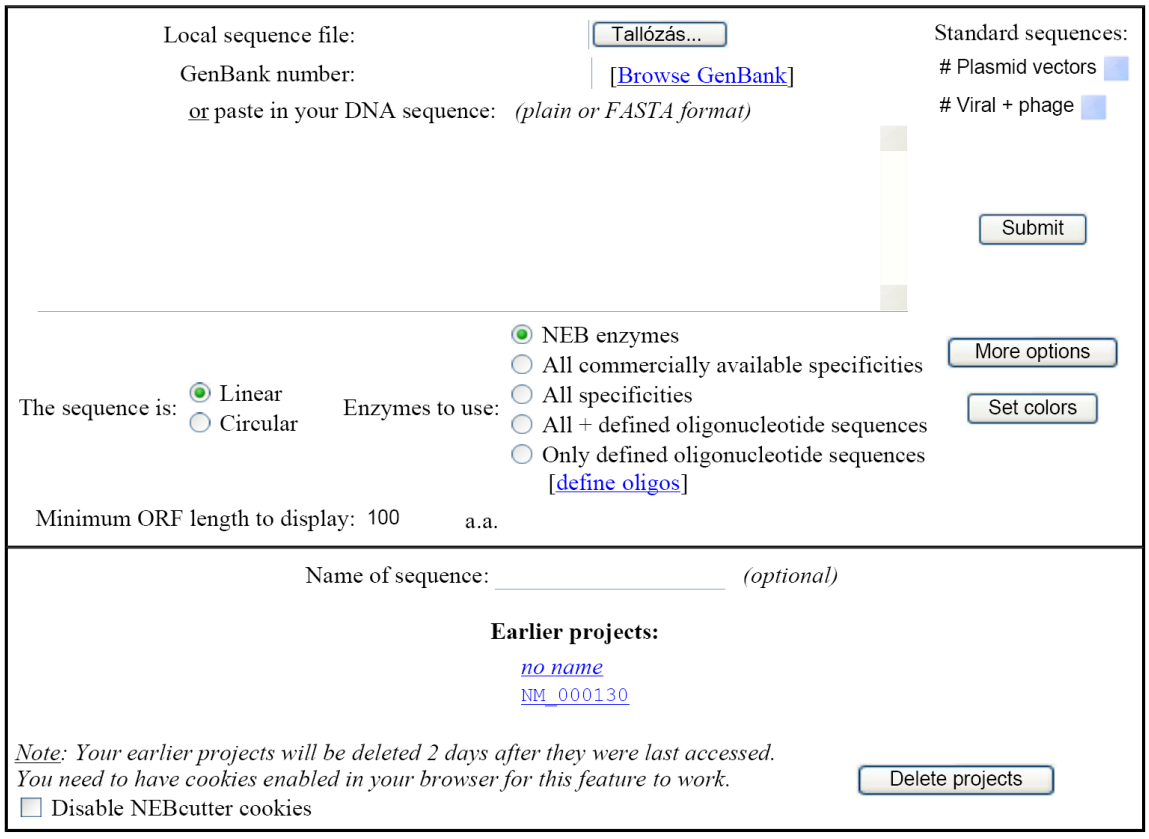


Fig 5

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 1731 to 1750 by PCR (in a real experiment we are working 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 result page we have the sequence, the related restriction endonucleases and their cleavage site (Fig 6). By moving the cursor to the certain RE, its recognition site is appearing (red underline).

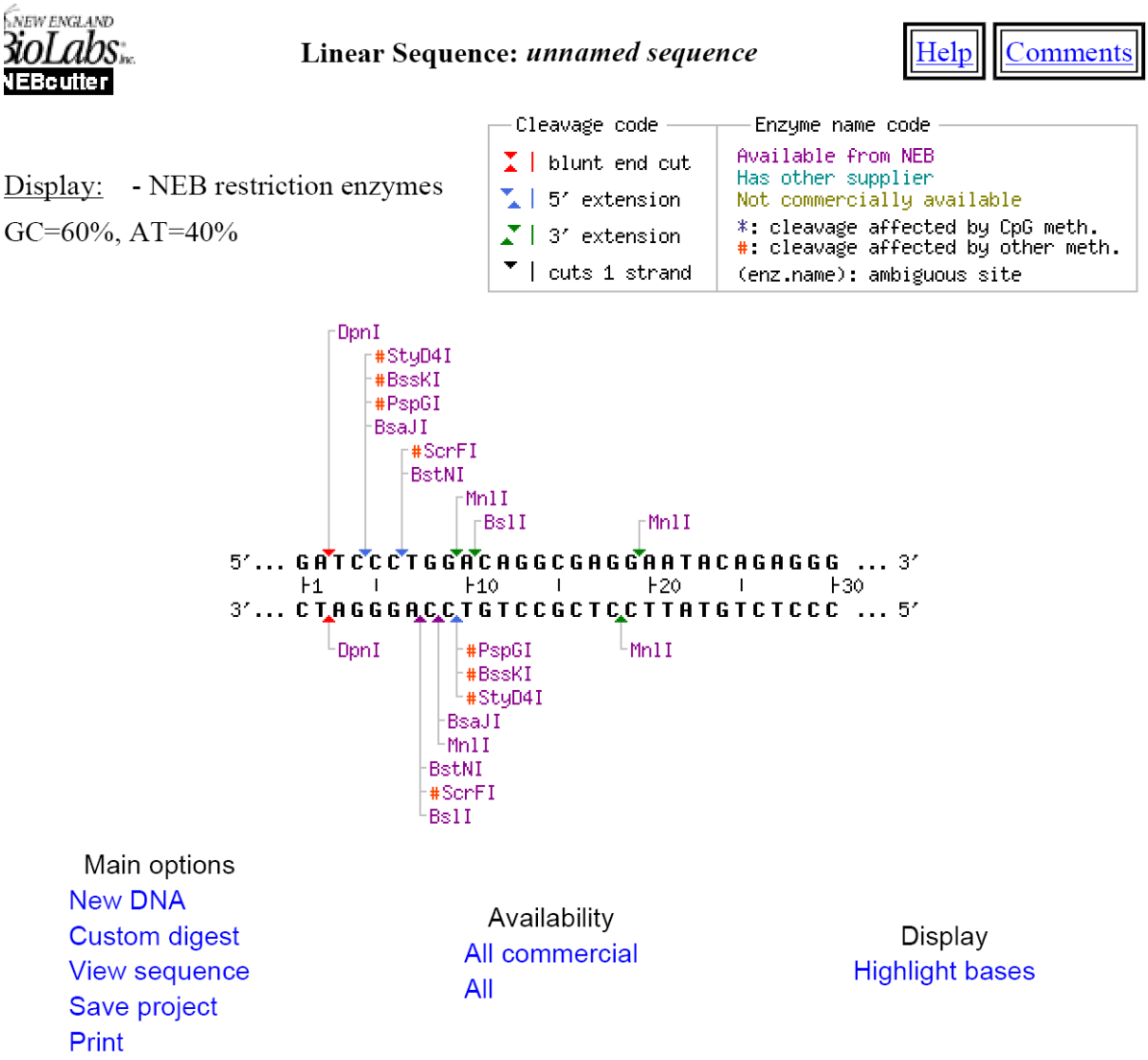


Fig 6

If the mutation change the nucleotide 16 in the fragments, the *MnlI* RE first recognition site could disappear. (The *MnlI* RE different from the restriction endonucleases we learnt

so far; the enzyme recognize a non-palindrome sequence and its cleavage site is different from the recognition site.)
Go for option "*custom digest*". The program shows which enzyme cuts the fragment, how many times, and what is the favoured buffer (1,2,3,4).
Choose *MnlI* RE, than click *„digest"*! The result page shows the fragment with the cleavage site. (Fig 7)

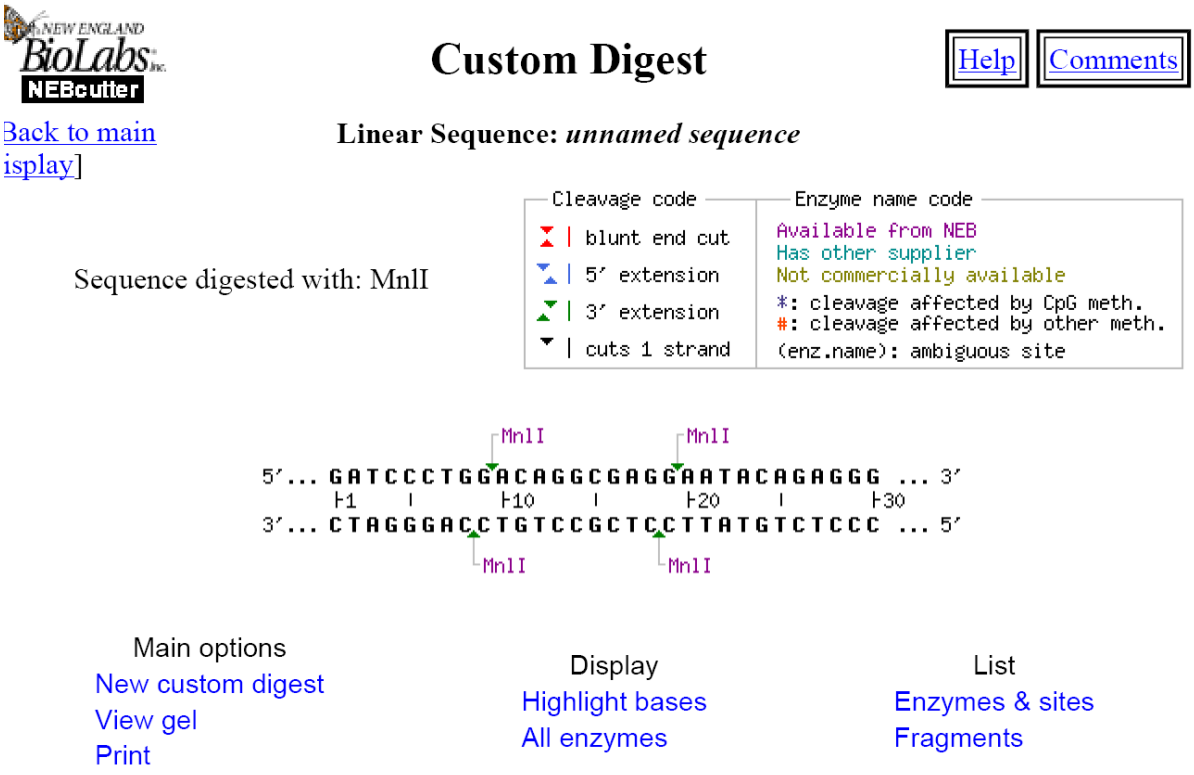
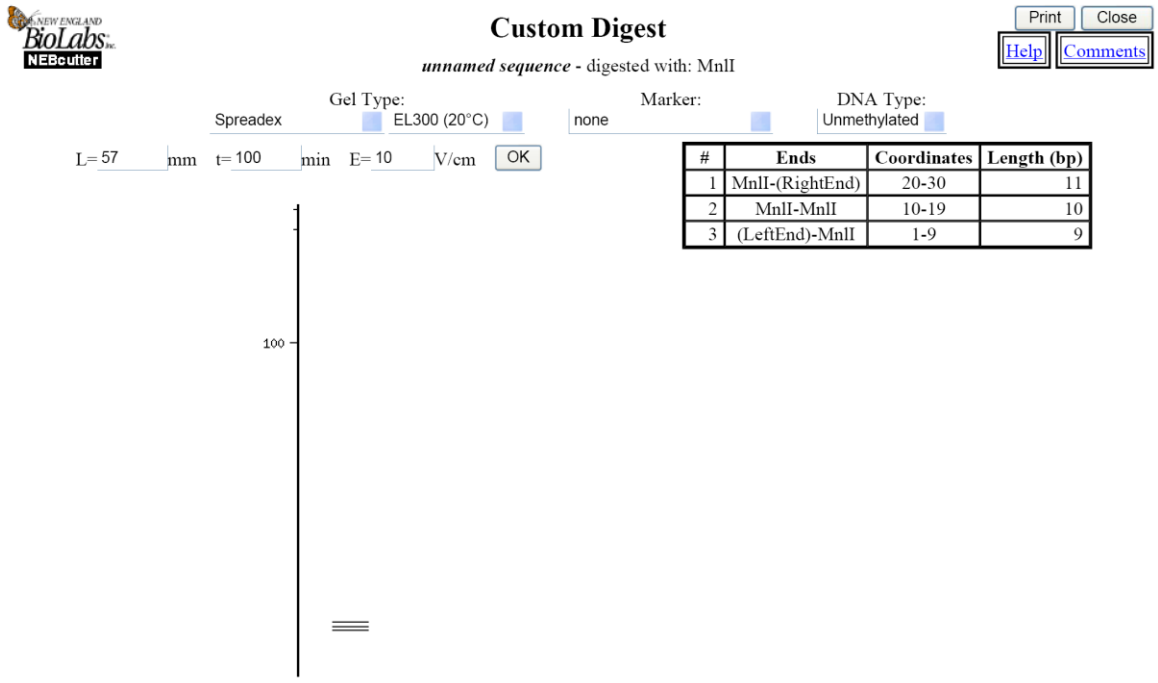


Fig 7
By click on option „*view gel*", the resulted fragments after the digestion and there electroforetic image become apparent (Fig 8). Since there is only 1-2 bases difference in the length of the fragments, we need to choose the best separating gel to show these differences: („*gel type*": Spreadex). Now we can see the three fragments resulting by the two cleavage 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, we will get two fragment instead of three. Choosing option *„custom digest"* shows one *MnlI* cut only. Choose RE *MnlI* then click on *"digest*". The result page shows the sequence with one cleavage site, confirmed by „view gel" / Spreadex option as well.