Introduction to Bio-Informatics
Lecture 8

Lecture by: Ahmad R. Naghsh-Nilchi, PhD
Department of Artificial Intelligence
Faculty of Computer Engineering
University of Isfahan

So far, we discussed:

- The goal of bioinformatics, types of bioinformatics data; protein, DNA, RNA structures and sequences,
- DNA to protein conversion, genetic coding, its challenges,
- Going from protein to DNA, its challenges, with hemoglobin example.
- Comparing protein sequences, using BLAST search,
- Protein information resources (PIR), BioEdit, JalView capabilities,
- Protein multiple alignments: procedure and analysis,
- Nucleotide sequence databases: Primary Type (GenBank, EMBL, DDBJ), Secondary Type (UniProt, SRA, NCBI RefSeq) and their applications, challenges, and considerations.
- Classes of living: Prokaryotes, Achaea, Eukaryotes, their structure, similarities, differences, and their evolutionary implications,
- GenBank information entries on prokaryotic, eukaryotic (mRNA and genomic types), and viral genomics are introduced.
- Their similarities and differences and some examples; X01714, U90223, AH005568, NC045512 (SARS-CoV-2).
  - GenBank information entries on prokaryotes, eukaryotic, and viral organisms
- Gene-Centric database advantages,

Working With the DNA Sequence

- Proteins perform most biological functions.
- Direct work in amino-acid sequences of a protein is hell !
- DNA contains information that the cell knows how to put to effective use.
- Sequencing its gene (DNA), and deducing the protein sequence from the genetic code is much easier!!
- Think of orthogonal mappings (i.e. Fourier Transform) in mathematics and their use to solve difficult problems.
- DNA sequence analysis could act as a transitional step!
- DNA sequences are the mother of all sequences !!!

The integrity Concerns

- The integrity of genetic information is paramount to the proper functioning of biological systems.
- DNA is subject to errors during replication (duplication), transcription, and environmental influences.
- Identifying and correcting these errors is critical for genomic stability and preventing diseases, such as cancer.
- Errors in a DNA sequence can be in various forms, including:
  - Point mutations,
  - Insertions,
  - Deletions,
  - Larger structural anomalies.
- Point mutations, which involve a change in a single nucleotide, may result from replication mistakes or exposure to mutagens.

Mutagen

- A mutagen is a chemical or physical agent that can cause irreversible changes in a cell's DNA.
- DNA changes caused by mutagens may harm cells and cause certain diseases, such as cancer.
- Examples of mutagens: radioactive substances, x-rays, ultraviolet radiation, and certain chemicals.
- Thus, the reliability of DNA replication is essential,
- Numerous mechanisms have evolved to catch and correct these errors.

Detection and Repairing

- One of the primary systems for detecting and repairing errors is the DNA repair mechanism which includes several pathways: base editing repair, nucleotide editing repair, and mismatch repair.
- These sophisticated systems involve the recognition of erroneous bases, removal of damaged sections, and precise re-synthesis of the correct sequence.
- Additionally, cellular enzymes such as DNA polymerases possess proofreading abilities that facilitate identifying and correcting mismatched nucleotides during DNA synthesis.

Solution(s)?

- Bioinformatics uses computational algorithms and statistical models to analyze biological data, enhancing our ability to identify inconsistencies within DNA sequences.
- One widely used method is the implementation of sequence alignment algorithms, such as Smith-Waterman and Needleman-Wunsch.
- These algorithms allow comparing a query sequence against a reference sequence, providing insights into possible insertions, deletions, or substitutions that may indicate errors.
- Moreover, machine learning techniques are powerful tools for detecting anomalies within large genomic datasets.
- By training models on known datasets, researchers can create predictive algorithms that identify likely error patterns.
- These approaches reduce the rate of false positives and enhance the overall efficiency of genomic analysis.

First thing, first!

- Experimental sequences (from labs) may be contaminated with problems and errors.
- The first task is to make sure the obtained sequence is okay so you do not waste a month or so trying to make sense of something completely wrong and end up nowhere!
- How to check for the most common problems encountered in sequences fresh from the lab is the focus here.

Steps in Sequencing a DNA Fragment

- Sequencing a DNA fragment involves:
  - Purifying it,
  - Cloning it into a vector such as a plasmid,
  - Amplifying it into a biological host, most often a bacterium like E. coli,
  - Submitting it to various sequencing protocols, such as primer extension or dye-termination.
- Many unintended events can occur during this process, eroding the sequence!
- Errors in DNA sequences can arise from various sources, including sequencing technology limitations, sample contamination, or human error during data entry.

Vector Contamination!

- The most common problem in DNA sequencing is that the sequence has been contaminated with the sequences of the vectors previously used in the lab.
- Usual Lab Procedure: The DNA (or cDNA) you send out to the laboratory for sequencing is inserted into a cloning vector – plasmid, phage, cosmid, BAC, PAC, or YAC – so that it can be worked out.
- The sequence you get back from the lab may include segments derived from these vectors.
- Knowing how to deal with these unwelcome vectors is vital.

Contaminated Sequence Example:

- ATGCGTACGTAGCTAGCTAGCTAGCTAGCTGACTAGC TAGCTAGCTAGCTAGC
- ATGCGTACGTAGCXXXXXXGCTAGCTAGCTAGCTAGCT AGCTAGCTAGCTAGC
- In this example, the second sequence contains 'XXXXXX', indicating potential contamination in that sequence.
- Note that you should verify contamination levels through appropriate bioinformatics tools.

Contaminated Vectors Removal

- Furthermore, your sequence might be cross-contaminated by somebody else's vector – maybe from the bench next to yours.
- So, your solution to remove these vectors, not only should deal with expected vectors but also should include other possible vector contaminations!
- A search and similarity finder software is required to detect these vectors from your sequence.
- This is done against metadata maintained by NCBI, the UniVect Database.

Software and Tools for Vector Removal

- For large sequences (usually much larger than 100K pairs), it is necessary to use sophisticated techniques and advanced machine learning algorithms, such as deep neural networks, to do this task.
- For small-size sequences (less than 100K), it is possible to run a search for similarity against the sequence of the vector using NCBI interactive tool and facility named VecScreen available at: https://www.ncbi.nlm.nih.gov/tools/vecscreen/
- The first webpage of VecScreen is shown next.

An official website of the United States government Here's how you know

NIH National Library of Medicine
National Center for Biotechnology Information

Log in

VecScreen All Databases Search

VecScreen UniVec Contamination

VecScreen: Screen a Sequence for Vector Contamination

VecScreen is a system that quickly finds segments of a nucleic acid sequence that may be of vector origin. It helps researchers identify and remove any segments of vector origin before they analyze or submit sequences. more...

Enter your query sequence below as an Accession or FASTA.

Run VecScreen Clear Input

Links

- About VecScreen
- Interpretation of Results
- Contamination
- The UniVec Database
- Current UniVec Statistics
- Current UniVec Content

VecScreen for Vector Contamination.

- Failure to remove foreign segments in a sequence can:
  - lead to erroneous conclusions about the biological significance of the sequence
  - waste time and effort in the analysis of contaminated sequence
  - delay the release of the sequence in a public database
  - pollute public databases with contaminated sequence
- Note that GenBank “Annotation Staff” use VecScreen to verify that sequences submitted for inclusion in the database are free of vector contamination.

How/What VecScreen Does?

- VecScreen searches a query sequence for segments that match any sequence in the "UniVec Database".
- UniVec Database is HUGE, and it is a specialized non-redundant vector database.
  - The search uses **BLAST** with parameters preset for optimal detection of vector contamination.
  - Those segments of the query that match vector sequences are categorized according to the strength of the match, and their locations are displayed
  - A VecScreen search will not identify the vector that is the most likely source of the contamination, but this can usually be deduced from the cloning history of the sequenced DNA.

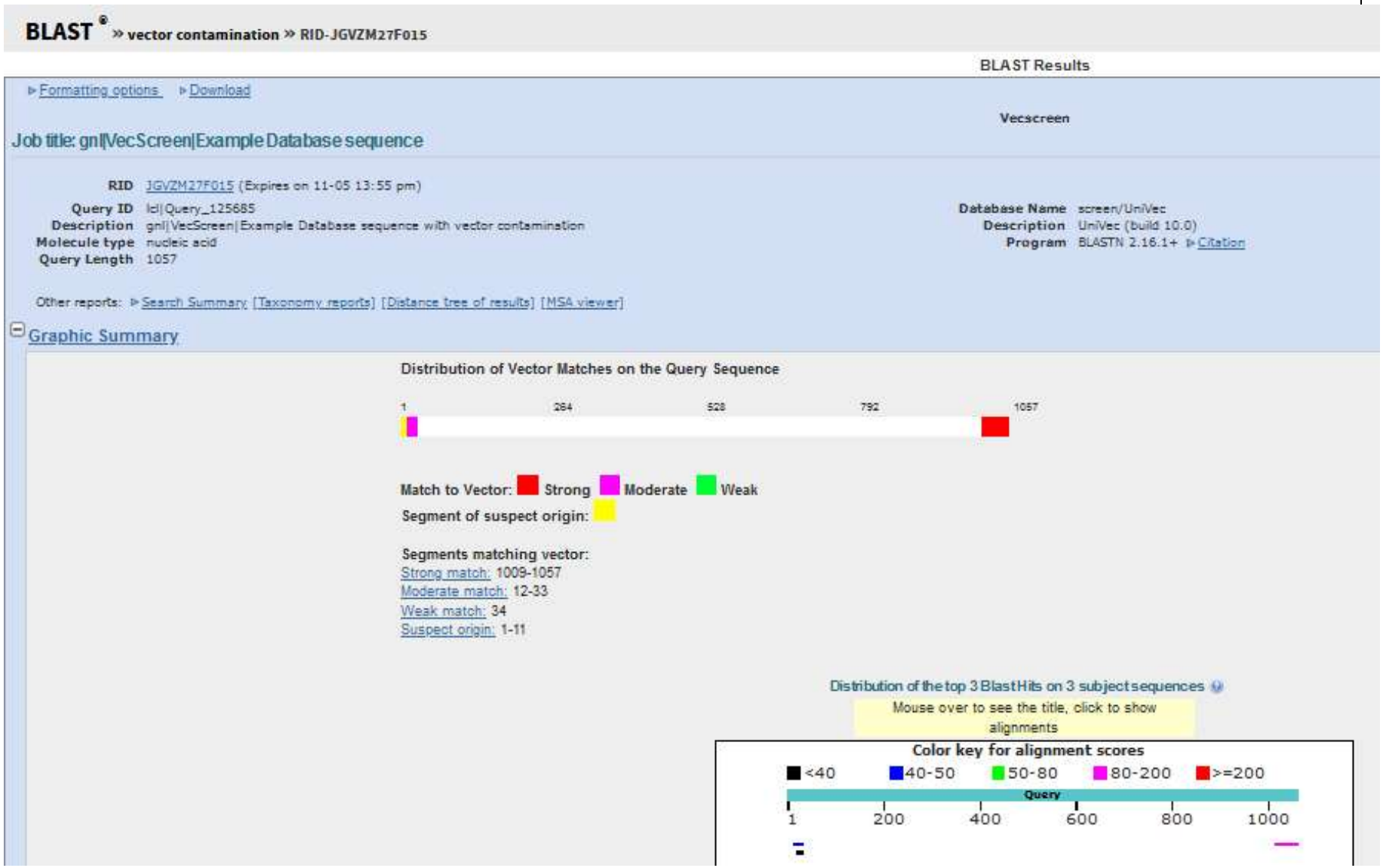
Two Possible Outcomes:

- Non-Significant Similarity found
  - This is usually good news!
  - This means that the submitted sequence does not resemble any known vector.
  - You can proceed to the next stage of your analysis.
- An output listing matches of some kind
  - Usually a bad news.
  - Your sequence may be contaminated.

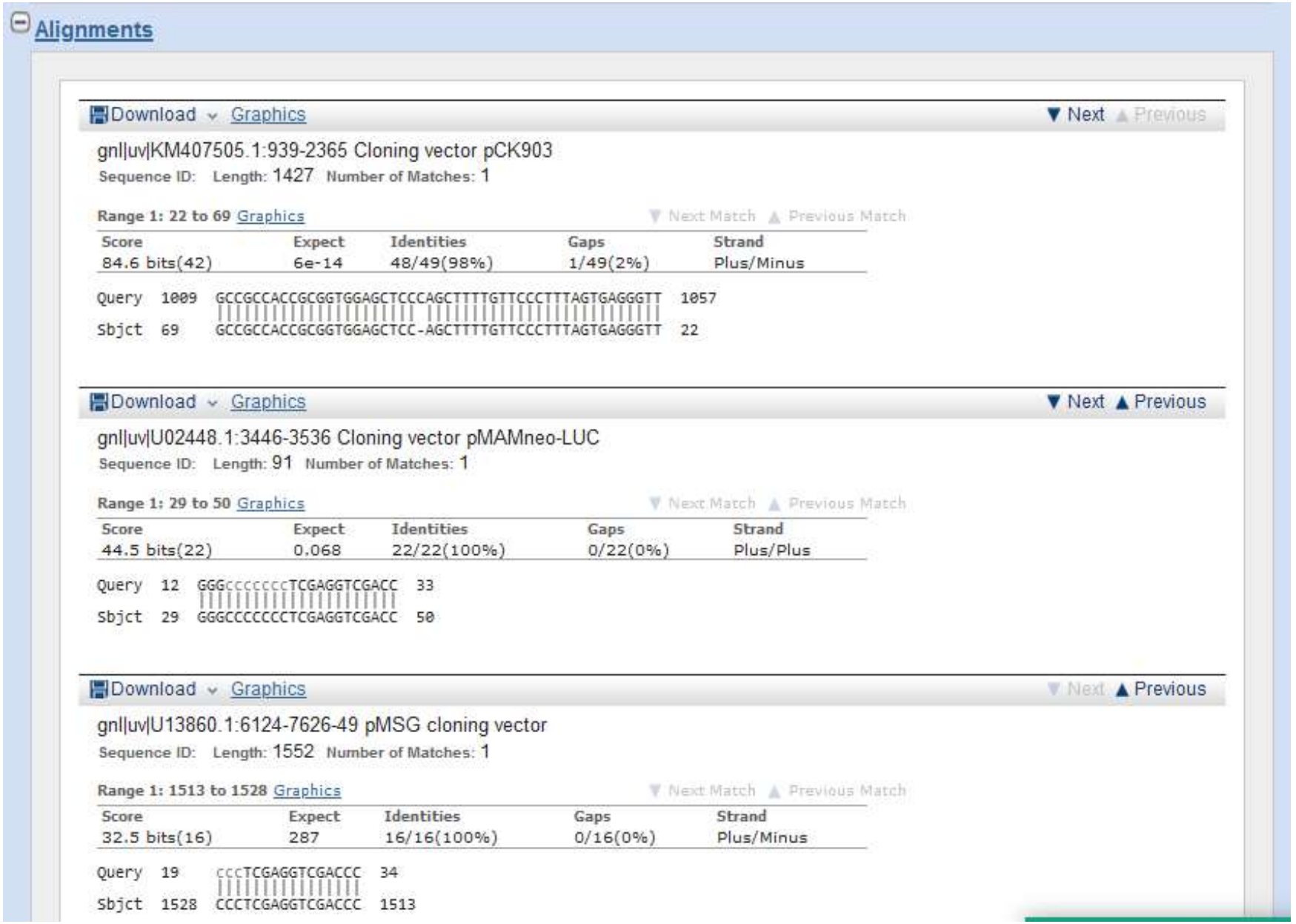
Interpretation of VecScreen Results

- When VecScreen alarms for vector contamination, results are displayed in 4 categories:
  - 1. Strong match, 2. moderate match, 3. Weak match, and 4. "Suspect Origin".
- VecScreen summarizes the BLAST output with a graphical representation of the sequence.
- The query sequence is color-coded to show the location of segments that match vector sequences.
- The matches are color-coded at four significance levels: strong (red), moderate (purple), weak (green), and suspicious (yellow).
  - For details click on "Interpretation of Results" bottom.

VecScreen Usage Example



Continued …. BlastN Alignments:



Alignments

Download Graphics ▼ Next ▲ Previous

gnl|uv|KM407505.1:939-2365 Cloning vector pCK903

Sequence ID: Length: 1427 Number of Matches: 1

Range 1: 22 to 69 Graphics ▼ Next Match ▲ Previous Match

Score	Expect	Identities	Gaps	Strand
84.6 bits(42)	6e-14	48/49(98%)	1/49(2%)	Plus/Minus

```
Query  1009  GCCGCCACCGCGGTGGAGCTCCCAGCTTTTGTTCCCTTTAGTGAGGGTT  1057
             ||||||||||||||||||||| |||||||||||||||||||||||||||
Sbjct  69    GCCGCCACCGCGGTGGAGCTCC-AGCTTTTGTTCCCTTTAGTGAGGGTT  22
```

Download Graphics ▼ Next ▲ Previous

gnl|uv|U02448.1:3446-3536 Cloning vector pMAMneo-LUC

Sequence ID: Length: 91 Number of Matches: 1

Range 1: 29 to 50 Graphics ▼ Next Match ▲ Previous Match

Score	Expect	Identities	Gaps	Strand
44.5 bits(22)	0.068	22/22(100%)	0/22(0%)	Plus/Plus

```
Query  12  GGGcccccccTCGAGGTCGACC  33
           ||||||||||||||||||||||
Sbjct  29  GGGCCCCCCCCTCGAGGTCGACC  50
```

Download Graphics ▼ Next ▲ Previous

gnl|uv|U13860.1:6124-7626-49 pMSG cloning vector

Sequence ID: Length: 1552 Number of Matches: 1

Range 1: 1513 to 1528 Graphics ▼ Next Match ▲ Previous Match

Score	Expect	Identities	Gaps	Strand
32.5 bits(16)	287	16/16(100%)	0/16(0%)	Plus/Minus

```
Query  19    cccTCGAGGTCGACCC  34
             ||||||||||||||||
Sbjct  1528  CCCTCGAGGTCGACCC  1513
```

Pitfalls

- VecScreen may not detect entire vector contamination. This can happen if:
  - Query sequences are not covered with UniVec database
  - VecScreen underestimates vector contamination region
  - VecScreen predicts the results as lower categories

PCR Primers

- The Polymerase Chain Reaction (PCR) is a common experimental laboratory technique used to amplify a DNA segment.
- PCR involves mixing a DNA template (containing the sequence for being amplified), the primers, a cocktail of nucleotide and other biochemical compounds, and a heat resistance enzyme called DNA polymerase
- All in a single little plastic tube.

Why PCR?

- The tube is put in a benchtop machine called THERMAL CYCLER.
  - This machine makes the tube go up and down in temperature.
- The amount of copies of the DNA segment to amplify is doubled for each temperature cycle.
- For example, after 30 cycles, 2^{30} (over 1 billion) more of it is generated.
- This is why PCR is great for forensic science:
  - Lick a stamp, and scientists will be able to get your DNA sequence!

How to Make a PCR?

- Identify the DNA sequence to be amplified.
- Order Primers from a DNA synthesis company,
  - Primers are small pieces of DNA (20-30 bases) that match the boundaries of the complete sequence of interest.
  - Designing good primers is the most delicate step in a PCR
- Run PCR experiment.

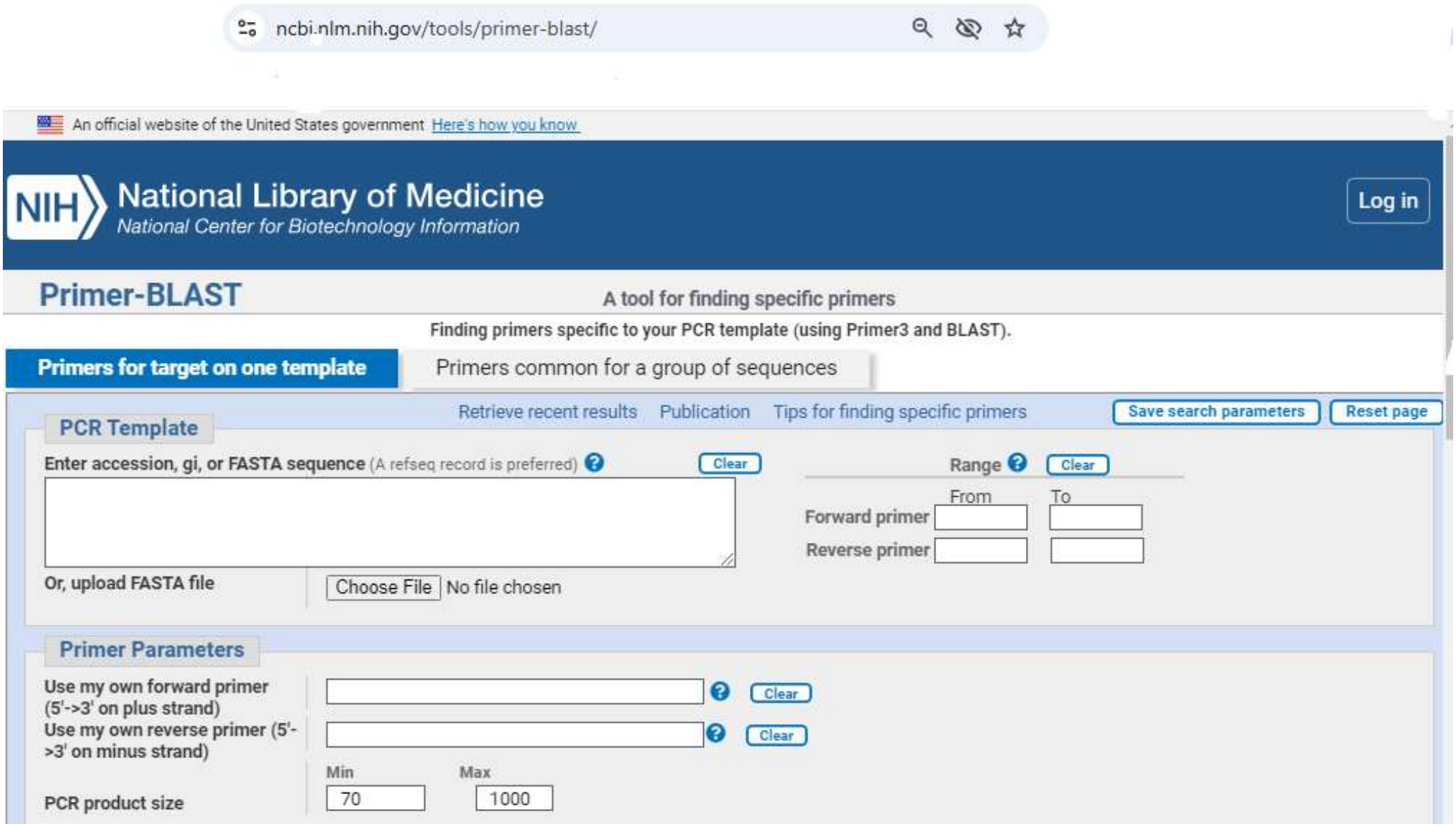
Designing the Primers

- The trickiest step in PCR procedure is the design of the primers.
  - Primers - the two small DNA fragments that are capable of firmly hybridizing on each side of the gene in a highly specific manner.
- Primer design software and programs are available to help decide which portion of your large sequence makes the best primers.

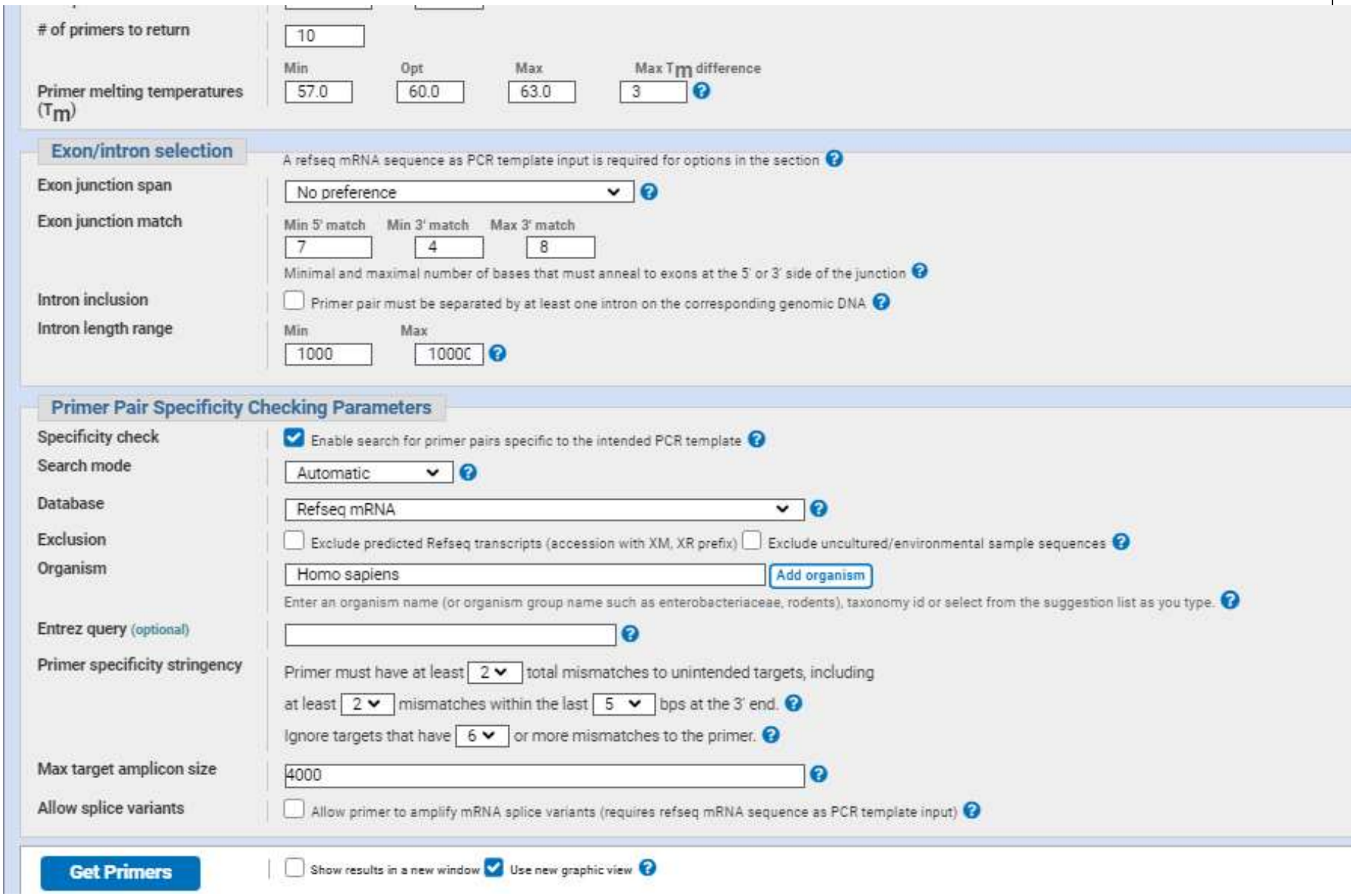
PCR Design Procedure

- NCBI also provides tools for PCR Design
- The NCBI address:
  - https://www.ncbi.nlm.nih.gov/tools/primer-blast/
- The page looks rather intimidating!
- Do not panic!
- Paste your sequence in the sequence window.
- Use defaults (for start, and for most cases) for every parameter.
- Click the “Get Primers” button.

NCBI Source for Primers Design



Continued ....


# of primers to return
10
Primer melting temperatures (Tm)
Min
Opt
Max
Max Tm difference
57.0
60.0
63.0
3
Exon/intron selection
A refseq mRNA sequence as PCR template input is required for options in the section
Exon junction span
No preference
Exon junction match
Min 5' match
Min 3' match
Max 3' match
7
4
8
Minimal and maximal number of bases that must anneal to exons at the 5' or 3' side of the junction
Intron inclusion
Primer pair must be separated by at least one intron on the corresponding genomic DNA
Intron length range
Min
Max
1000
10000
Primer Pair Specificity Checking Parameters
Specificity check
Enable search for primer pairs specific to the intended PCR template
Search mode
Automatic
Database
Refseq mRNA
Exclusion
Exclude predicted Refseq transcripts (accession with XM, XR prefix)
Exclude uncultured/environmental sample sequences
Organism
Homo sapiens
Add organism
Enter an organism name (or organism group name such as enterobacteriaceae, rodents), taxonomy id or select from the suggestion list as you type.
Entrez query (optional)
Primer specificity stringency
Primer must have at least 2 total mismatches to unintended targets, including
at least 2 mismatches within the last 5 bps at the 3' end.
Ignore targets that have 6 or more mismatches to the primer.
Max target amplicon size
4000
Allow splice variants
Allow primer to amplify mRNA splice variants (requires refseq mRNA sequence as PCR template input)
Get Primers
Show results in a new window
Use new graphic view

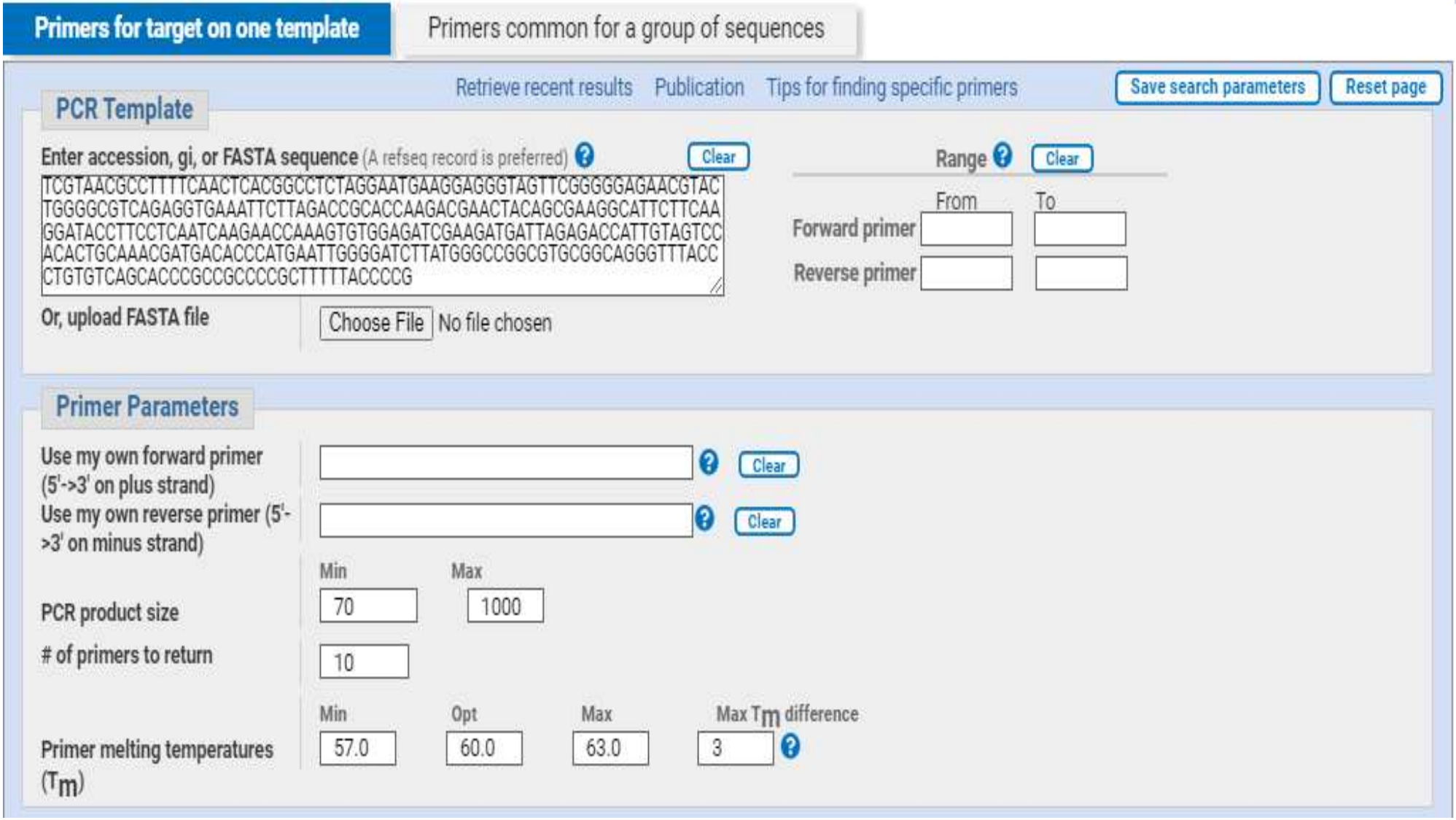
The Results include:

- Include a map of the best oligonucleotide pair.
  - Oligonucleotide pair is left and right primers.
- In addition, 10 alternative solutions are proposed for each oligonucleotide pair
- The position,
- Length,
- Predicted melting temperature,
- G-C percentage (the fraction on guanine and cytosine nucleotides)

Notes on G-C pairs

- G-C pairs in nuclotides are significant because they are more stable and have a higher thermal stability than A-T:
- **Hydrogen bonding**:
  - G-C pairs have three hydrogen bonds, A-T pairs have two. This stronger interaction makes G-C pairs more stable.
- **Thermal stability**:
  - G-C pairs have a stronger stacking interaction than A-T. This makes RNA with high G-C content more resistant to high temperatures.
- **Denaturation (a process of breaking up a protein):**
  - Strands with more G-C content are more stable and have a greater resistance to denaturation.

An Example


Primers for target on one template
Primers common for a group of sequences
Retrieve recent results
Publication
Tips for finding specific primers
Save search parameters
Reset page
PCR Template
Enter accession, gi, or FASTA sequence (A refseq record is preferred)
Clear
TCGTAACGCCTTTTCAACTCACGGCCTCTAGGAATGAAGGAGGGTAGTTCGGGGGAGAACGTAC
TGGGGCGTCAGAGGTGAAATTCTTAGACCGCACCAAGACGAACTACAGCGAAGGCATTCTTCAA
GGATACCTTCCTCAATCAAGAACCAAAGTGTGGAGATCGAAGATGATTAGAGACCATTGTAGTCC
ACACTGCAAACGATGACACCCATGAATTGGGGATCTTATGGGCCGGCGTGCGGCAGGGTTTACC
CTGTGTCAGCACCCGCCGCCCCGCTTTTTACCCCG
Range
Clear
From
To
Forward primer
Reverse primer
Or, upload FASTA file
Choose File
No file chosen
Primer Parameters
Use my own forward primer (5'->3' on plus strand)
Use my own reverse primer (5'->3' on minus strand)
Clear
Clear
Min
Max
PCR product size
70
1000
# of primers to return
10
Min
Opt
Max
Max Tm difference
Primer melting temperatures (Tm)
57.0
60.0
63.0
3

Sample Results:

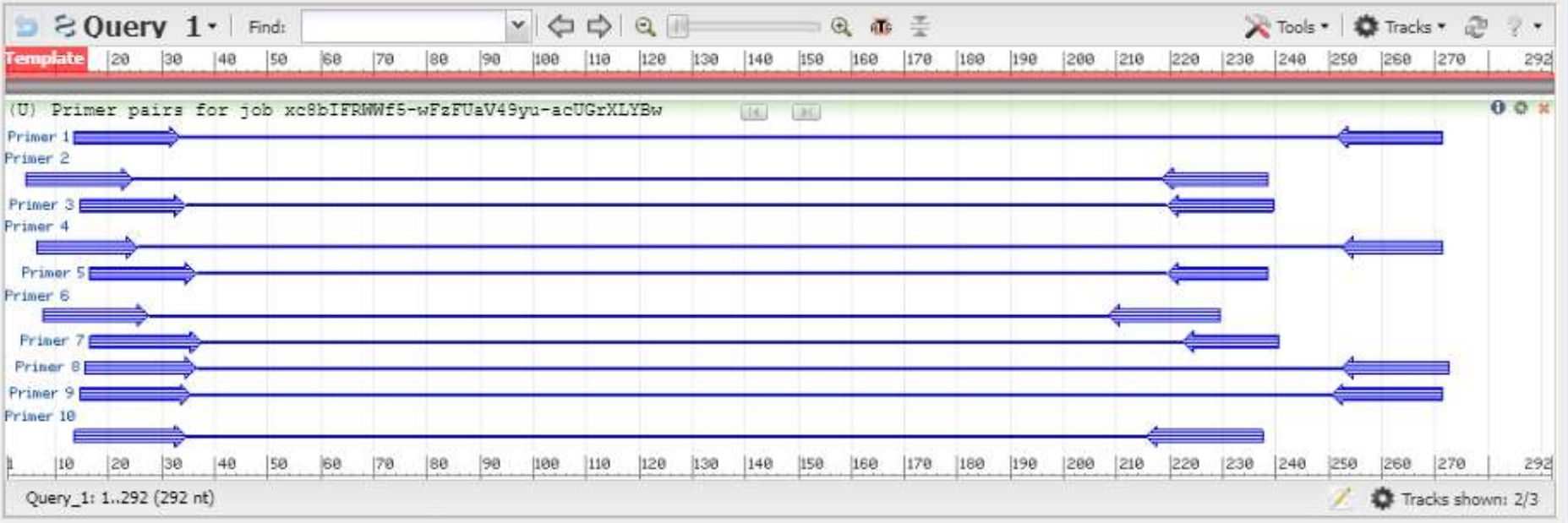
Primer-BLAST » JOB ID:xc8bIFRWWf5-wFzFUaV49yu-acUGrXLYBw

Primer-BLAST Results

Edit search NEW

Input PCR template	lcl|Query_1
Range	1 - 292
Specificity of primers	Primer pairs are specific to input template as no other targets were found in selected database: Refseq mRNA (Organism limited to Homo sapiens)
Other reports	▹Search Summary

— Graphical view of primer pairs



— Detailed primer reports

Continued …

Detailed primer reports

Download primer pairs

Primer pair 1

	Sequence (5'->3')	Template strand	Length	Start	Stop	Tm	GC%	Self complementarity	Self 3' complementarity
Forward primer	TCAACTCACGGCCTCTAGGA	Plus	20	14	33	59.96	55.00	4.00	2.00
Reverse primer	GGGTGCTGACACAGGGTAAA	Minus	20	271	252	59.89	55.00	3.00	1.00
Product length	258								

Primer pair 2

	Sequence (5'->3')	Template strand	Length	Start	Stop	Tm	GC%	Self complementarity	Self 3' complementarity
Forward primer	AACGCCTTTTCAACTCACGG	Plus	20	5	24	59.34	50.00	4.00	1.00
Reverse primer	CGGCCCATAAGATCCCCAAT	Minus	20	238	219	59.59	55.00	4.00	2.00
Product length	234								

Primer pair 3

	Sequence (5'->3')	Template strand	Length	Start	Stop	Tm	GC%	Self complementarity	Self 3' complementarity
Forward primer	CAACTCACGGCCTCTAGGAA	Plus	20	15	34	59.10	55.00	4.00	0.00
Reverse primer	CCGGCCCATAAGATCCCCAA	Minus	20	239	220	61.72	60.00	4.00	0.00
Product length	225								

Primer pair 4

	Sequence (5'->3')	Template strand	Length	Start	Stop	Tm	GC%	Self complementarity	Self 3' complementarity
Forward primer	CGCCTTTTCAACTCACGGC	Plus	19	7	25	60.08	57.89	3.00	2.00
Reverse primer	GGGTGCTGACACAGGGTAA	Minus	19	271	253	59.24	57.89	3.00	3.00
Product length	265								

Primer pair 5

	Sequence (5'->3')	Template strand	Length	Start	Stop	Tm	GC%	Self complementarity	Self 3' complementarity
Forward primer	ACTCACGGCCTCTAGGAATG	Plus	20	17	36	58.88	55.00	4.00	0.00
Reverse primer	CGGCCCATAAGATCCCCAA	Minus	19	238	220	59.16	57.89	4.00	0.00
Product length	222								

Primer pair 6

Continued ....

Product length 222

Primer pair 6

	Sequence (5'->3')	Template strand	Length	Start	Stop	Tm	GC%	Self complementarity	Self 3' complementarity
Forward primer	GCCTTTTCAACTCACGGCCT	Plus	20	8	27	61.17	55.00	4.00	2.00
Reverse primer	AGATCCCCAATTCATGGGTGT	Minus	21	229	209	59.07	47.62	4.00	2.00
Product length	222								

Primer pair 7

	Sequence (5'->3')	Template strand	Length	Start	Stop	Tm	GC%	Self complementarity	Self 3' complementarity
Forward primer	ACTCACGGCCTCTAGGAATGA	Plus	21	17	37	60.34	52.38	4.00	2.00
Reverse primer	GCCGGCCCATAAGATCCC	Minus	18	240	223	59.97	66.67	6.00	0.00
Product length	224								

Primer pair 8

	Sequence (5'->3')	Template strand	Length	Start	Stop	Tm	GC%	Self complementarity	Self 3' complementarity
Forward primer	AACTCACGGCCTCTAGGAATG	Plus	21	16	36	59.52	52.38	4.00	0.00
Reverse primer	CGGGTGCTGACACAGGGTAA	Minus	20	272	253	62.11	60.00	3.00	3.00
Product length	257								

Primer pair 9

	Sequence (5'->3')	Template strand	Length	Start	Stop	Tm	GC%	Self complementarity	Self 3' complementarity
Forward primer	CAACTCACGGCCTCTAGGAAT	Plus	21	15	35	59.52	52.38	4.00	2.00
Reverse primer	GGGTGCTGACACAGGGTAAAC	Minus	21	271	251	61.15	57.14	3.00	3.00
Product length	257								

Primer pair 10

	Sequence (5'->3')	Template strand	Length	Start	Stop	Tm	GC%	Self complementarity	Self 3' complementarity
Forward primer	TCAACTCACGGCCTCTAGGAA	Plus	21	14	34	60.55	52.38	4.00	0.00
Reverse primer	GGCCCATAAGATCCCCAATTCA	Minus	22	237	216	60.16	50.00	4.00	1.00
Product length	224								

Other PCR Design Options:

- You may modify default settings using the form.
- The form allows all reasonable experimental situations and constraints to be imposed upon the primer selection process, including:
  - Searching for only a left or right primer,
  - Searching for a single hybridization probe,
  - Proposing your own left or right primer,
  - Selecting sequence positions to be excluded,
  - Selecting a range of product sizes,
  - Imposing a range of oligonucleotide sizes, G-C percentages, or melting inputs,
  - And many more options

Other Considerations

- The final responsibility is to verify that primers will not hybridize anywhere – except where you intend them to hybridize.
- For instance, you want to make sure that the selected oligonucleotide sequences are not found outside the gene you are interested in.
- Or you may want to resemble a frequent repeat in the DNA you are going to amplify.
- Avoiding these problems involves BLAST searches against the vector sequences, the relevant genomes, and their most common repeats.

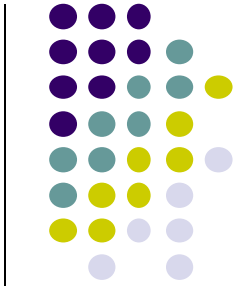
تکلیف۳

- با مراجعه به منابع اینترنتی، یک دنباله DNA آلوده را پیدا کنید و یا سعی کنید یک دنباله سالم در GenBank انتخاب و آن را بطور دستی آلوده کنید
- میزان آلودگی دنباله را با استفاده از VecScreen مشخص و نتایج گرافیکی و alignment آنرا گزارش کنید.
- برای دنباله سالم انتخابی با استفاده از ابزارهای موجود در NCBI یک PCR طراحی کنید. بهترین map و چهار map جانشین آن را گزارش کنید.

Analyzing the DNA Composition

- After all the going back and forth from computer to bench, you are now made sure that the sequence in your test tube is indeed the right stuff.
- Now you are ready to analyze the purified and amplified sequence.
- Having a new piece of a new genome, the first question is: What is the percentage of nucleotides pairing between adenosine (A), thymidine (T), guanosine (G), and cytosine (C) in the DNA ladder?
- This statistical information determines how the DNA will behave in your experiments.
- The first level of analysis, thus, is to count the number of A, G, C, T in your sequence.

Counting words in DNA sequence



- A DNA sequence is made up of overlapping "words". For example, consider:
- ATCAGGCTAGATG …
- You can read it as (simple nucleotides):
  - A, T, C, A, G, G, C, T, A, G, A, T, G, …

  and counting the A, C, G, and T components.

  Or read and count it as di-nucleotides (2-letter words):
  - AT, TC, CA, AG, GG, GC, CT, TA, AG, GA, AT, TG, …

  Or read and count it as tri-nucleotides (triplet words):
  - ATC, TCA, CAG, AGG, GGC, GCT, CTA, TAG, AGA, GAT, ….
- Or larger sizes of nucleotide words.

پروژه پایانی

- با مراجعه به GenBank یک دنباله دارای طول حداقل ۲۰۰۰ نوکلئوتیدی DNA یک ژن را انتخاب کنید.
- برنامه کامپیوتری تحت Python یا R بنویسید بطوریکه تعداد و درصد نوکلئوتیدهای ساده، کلمات دو حرفی، و کلمات سه حرفی را به فرم جداول آمده در اسلاید های بعدی گزارش کند.
- با توجه به نتایج بدست آمده مدرج در جداول، پایداری ژن و همچنین میزان استقامت در مقابل گرمای DNA مورد مطالعه را تحلیل و گزارش کنید.

فرمت جداول خروجی برنامه شما

Mono-Nucleotides:

	In bps	In %
A		
C		
G		
T		

Di-Nucleotides (in bps):

first nucl. \ second nucleotide	A	C	G	T
A				
C				
G				
T				

Di-Nucleotides (in %):

first nucl. \ second nucleotide	A	C	G	T
A				
C				
G				
T				

فرمت جداول ادامه ...

Tri-Nucleotide (in bps):

first nucl.	second/third nucleotide															
	AA	AC	AG	AT	CA	CC	CG	CT	GA	GC	GG	GT	TA	TC	TG	TT
A																
C																
G																
T																

Tri-Nucleotide (in %):

first nucl.	second/third nucleotide															
	AA	AC	AG	AT	CA	CC	CG	CT	GA	GC	GG	GT	TA	TC	TG	TT
A																
C																
G																
T																

Thank you

for your attention