

BIOINFORMATICS

Biological information is encoded in the nucleotide sequence of DNA. Bioinformatics is the field that identifies the information in DNA using computer-based tools. Some bioinformatics algorithms aid the identification of genes, promoters, and other functional elements of DNA. Other algorithms help determine the relationships between DNA sequences.

Because of the large number of tools and DNA sequences available on the Internet, experiments done *in silico* ("in silicon," or computers) complement experiments done *in vitro* (in glass, or test tubes). This integration of biochemistry and computation is a key feature of modern biological research.

In Part I you will use the *Basic Local Alignment Search Tool (BLAST)* to predict the size of products amplified by the four CAPS primer sets. In Part II you will extract the predicted sequences of the CAPS markers and store them for further manipulations. In Part III you will use an electronic tool to predict the products of restriction digests of the PCR products. In Part IV you will identify the *AGO1* gene and learn about the function of the *argonaute* protein. In Part V you will electronically map *ago1* and the four CAPS markers on *Arabidopsis* Chromosome 1.

I. Use **BLAST** to Predict the Size of the CAPS Amplicons

The following primer sets were used in the experiment:

g4026	5'-GGGGTCAGTTACATTACTAGC-3'	(Forward Primer)
	5'-GTACGGTCTCTTCCCTTA-3'	(Reverse Primer)
H77224	5'-GGATTGGGGAAGAGGAAGTAA-3'	(Forward Primer)
	5'-TCCTTAGCCTTGCTTGATAGT-3'	(Reverse Primer)
m235	5'-GAATCTGTTTCGCCTAACGC-3'	(Forward Primer)
	5'-AGTCCACAACAATTGCAGCC-3'	(Reverse Primer)
UFO	5'-GTGGCGGTTTCAGACGGAGAGG-3'	(Forward Primer)
	5'-AAGGCATCATGACTGTGGTTTTTC-3'	(Reverse Primer)

1. Initiate a *BLAST* search.
 - a. Open the Internet site of the National Center for Biotechnology Information (NCBI), www.ncbi.nlm.nih.gov/.
 - b. Click on *BLAST* in the top speed bar.
 - c. Click on Nucleotide-nucleotide *BLAST* (blastn).
 - d. Enter the g4026 primer set into the Search window.
 - e. Delete any non-nucleotide characters from the window.

- f. Click on *BLAST!*.
 - g. Click on *Format!* and wait for your results.
2. The results of the *BLAST* search are displayed in three ways as you scroll down the page:
 - a. A graphical overview illustrating how significant hits align with the query sequence,
 - b. A list of *significant alignments* with *Geneinfo Identifier (gi)* – links, and
 - c. A detailed view of the primer sequences (*Query*) aligned to the nucleotide sequence of the search hit (*sbjct*).
 3. What is the predicted length of the product that the primer set would amplify in a PCR reaction (*in vitro*)?
 - a. In the list of significant alignments, notice the scores in the *E-Value* column on the right. The Expectation or *E-Value* is the number of alignments with the query sequence that would be expected to occur by chance in the database. The lower the *E-Value* the higher the probability that the hit is related to the query. For example, $3e-4$ denotes equals 3×10^{-4} or 0.0003. Shorter queries, such as primers, produce higher *E-values*.
 - b. Note any significant alignment that has an *E-Value* less than 0.1.
 - c. Scroll down to the *Alignments* section, and examine the corresponding alignments with the two primers.
 - d. The lowest and highest nucleotide positions in the subject sequence indicate the borders of the amplified sequence. Record these start and stop positions in the table below.

Marker	Start Position	Stop Position	Amplicon (bp)	gi-number
g4026				gi-2252639
H77224				gi-18410209
m235				gi-2098816
UFO				gi-9690193

- e. Calculate the predicted length of the amplicon and record in the table. Don't forget to add one nucleotide to your result!
- f. Repeat Steps a-e above for the remaining three CAPS primer sets.

II. Extract and Store Predicted CAPS Marker Sequences

1. Follow the *gi* – link for the g4026 marker to open a Genbank record, which contains basic information about each DNA sequence submitted to the Genbank database. Accession, number source, references, and other general information is at the top of the page, while the actual DNA sequence is at the bottom. In between are annotations of identifiable features – including gene, mRNA, coding sequence (cds), and repeats. (Follow this link for a clickable record, with descriptions
<http://www.ncbi.nlm.nih.gov/Sitemap/samplerecord.html>.)
2. Scroll down to the sequence record. Extract the predicted CAPS marker by highlighting and copying the region between the start and stop positions listed on your chart.
3. Paste the predicted CAPS marker into a text file, such as MS Word, Wordpad, or Notepad.
4. Repeat steps 1–3 with the remaining three primer sets, and save the file. This sequence repository will be used in the following sections.

III. Digest CAPS Markers Electronically (*In Silico*)

1. Open
http://www.dnalc.org/bioinformatics/nucleotide_analyzer.htm#clipper.
2. Copy the g4026 marker from your sequence repository, and paste it into the sequence window.
3. Enter the name of the restriction enzyme used to produce the CAPS marker in the experiment.
4. Record the length of the predicted restriction fragments in the table below.

Marker	Restriction Enzyme	Amplicon (bp)	Restriction Fragments (bp)
g4026	<i>RsaI</i>		
H77224	<i>TaqI</i>		
m235	<i>HindIII</i>		
UFO	<i>TaqI</i>		

5. Repeat steps 1–4 with the remaining three markers.
6. Compare the results of your electronic (*in silico*) and electrophoretic (*in vitro*) analyses. Which fragments are missing and why?
7. Which *Arabidopsis* ecotype was used for the *Arabidopsis* Genome Sequencing Project?

IV. Use **BLAST** to Identify the **AGO1** Gene

1. Open the NCBI Internet site, www.ncbi.nlm.nih.gov/.
2. Click on *Map Viewer from the Hot Spots* list on the right.
3. In *Search* window at the top of the page, select *Arabidopsis thaliana* (*thale cress*), enter *ago1*, and click on *Go!*
4. Click on the link to the chromosome that shows a hit.
5. Click on the *genes seq* link or use the *zoom* feature on the left to enlarge the structure of the *AGO1* gene. How long is the gene, and how many exons and introns does it have?
6. Follow links in the pink speed bar to learn about the function of the argonaute protein and its functional domains.
7. Click on At2g23380 (*Arabidopsis thaliana* gene ID).
8. Scroll down the page to Related Sequences, and click on *mRNA U91995*.
9. Extract (highlight and copy) the coding sequence at the bottom of the page. Then paste this *AGO1* gene sequence into the sequence repository you started in Part II.

V. Map **AGO1** and **CAPS** Markers Electronically

1. Open the NCBI Map Viewer, <http://www.ncbi.nlm.nih.gov/mapview/>.
2. Click on the *BLAST* link to the left of *Arabidopsis thaliana* (*thale cress*) in the *Plants list*.
3. Copy all of the CAPS markers and the *AGO1* gene from your sequence repository, and paste them into the sequence window.
4. Click on *Begin Search*.
5. Click on *Format!* and wait for your results.
6. Click *Genome View*.
7. Click on the link to the chromosome that shows the hits.
8. Click on the pink *Blast hit* link to identify the locations of the *AGO1*.

9. Use the chart below to record the answers to the following questions:

- a. What is the order of the five search sequences on the chromosome?
- b. Use the ruler on the left to determine the map positions for the CAPS markers and the *AGO1* gene.

M = Mb = megabasepairs = 1 millions basepairs

Locus	Chromosome Order	Map Position (Mb)
AGO1		
g4026		
H77224		
m235		
UFO		

10. Which CAPS marker is located most closely to *ago1*? How many Mb separate them?