

The Institute for Genomic Research

Gene Hunting

Introduction to bioinformatics

Copyright ©2004

The Institute for Genomic Research

9712 Medical Center Drive, Rockville, MD 20850

Phone 301-795-7000 ♦ Fax 301-838-0229 ♦ www.tigr.org/edutrain

Questions or comments:

Education & Training ♦ training@tigr.org

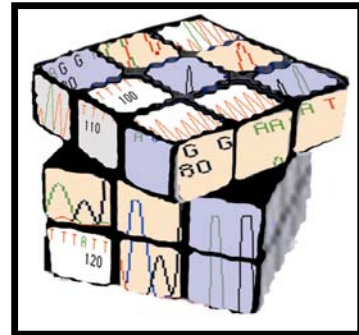
This activity was written and developed by Jennifer Jenkins, Jennifer Colvin and Lisa McDonald at The Institute for Genomic Research.

Introduction

Gene Hunting

Although knowing the sequence of an organism's DNA seems useful, without further analysis and interpretation the actual base sequence provides little information. Knowing the sequence of DNA can be likened to having a book in a language you do not understand.

Large-scale sequencing has become quite routine - every day more and more sequencing data is available for researchers to use for everything from a clearer understanding of human genes to previously unknown genes in mice, potato and tomato. In this Post-Sequencing Era, the challenge for genomics is deciphering the sequence data.



Bioinformatics

The first issue to address in instituting a high-throughput analysis of any genome is data management. In the 1990's, bioinformatics sought to manage and manipulate the unprecedented amount of sequence data by designing computer programs that can perform quadrillions of calculations to better understand biology. In what some are calling The Human Genome – Part 2, bioinformatics is playing a much larger role. Increasingly more powerful computers and computer programs are allowing researchers to compare the genetic sequences of different organisms to help determine the functions and relationships of their genes.

Genomic researchers manipulating large amounts of data are using computers, software tools, and databases – this science is called bioinformatics. It is increasingly being accepted that computer are essential tool for the lab scientist and laboratory and computational hypothesis will be how all biomedical research will be done in the future.

Research is no longer only performed in a laboratory; biomedical research now depends on computer tools and databases

With growing database resources, many bioinformatics scientists are doing comparative genomics. Using computers, they analyze DNA sequence patterns of different organisms side by

side to identify genes and determine functions. In addition, they are looking for similar genes in different species to determine possible relationships and genomic variations. These comparisons are giving scientists a better understanding of traits that have influenced evolution.

Genomic-scale technologies will be needed to fully analyze and interpret all of the genomic and functional genomic data. Organizations are now setting out to study and compare entire genomes, sets of expressed RNAs or proteins, families of genes from a large number of species, variation among individuals, and the classes of elements that regulate how often genes are turned on and off.

For the following activity we will be analyzing the complete genome sequence of *Deinococcus radiodurans* R1. The genome of this organism consists of two circular chromosomes as well as a megaplasmid and a small plasmid. *Deinococcus* was originally found in irradiated meats but is not a pathogen of humans. Understanding the biology of these bacteria will help us understand the unique genes it has that allow it to survive high levels of radiation that can kill other organisms such as humans, mice, cockroaches, and *E.coli*.

The activity

This activity demonstrates several tools available for DNA sequence data manipulation on Comprehensive Microbial Resource website developed at The Institute for Genomic Research (TIGR). TIGR is committed to the continued expansion of sequence information and to the application of this data in medicine, agriculture, and basic biological research.

This computer-based activity allows participants to analyze a section of genomic sequence data and identify potential open reading frames. An open reading frame, typically called an ORF, is a region of DNA that contains a series of bases coding for amino acids without any termination codons. The sequence is potentially translatable into protein.

Proteins are fairly large molecules made up of strings of amino acids linked together like a chain precisely arranged in a 3-D structure that is unique for each protein; as the protein is made, the amino acids "fold" into a specific shape. The specific sequence of nucleotides in the gene directs the cell which amino acid to link together and form the chain.

The genetic code

Nucleotides are read in triplet; a group of three nucleotides is called a codon and each codon indicates to the cell which of the 20 amino acids to use in building the protein. Any one of the four possible nucleotides can occupy any of the three positions of the codon; there are $4^3 = 64$ possible combinations of codons.

The striking feature of the genetic code is that it is deliberate – almost every amino acid is represented by several codons. Codons representing the same or related amino acids tend to be similar in sequence. Often the base in the third position of a codon is not significant, because the similar codons differ only in the third base represent the same amino acid. This reduced specificity at the last position is known as third-base degeneracy.

The benefit of having a tendency for similar amino acids to be represented by related codons minimizes the effect mutations have on the function of the gene– it increases the probability that a single random base change will result in no amino acid substitution or in one involving amino acids of similar character.

		SECOND POSITION							
		T	C	A	G				
T	TTT	Phe	TCT	Ser	TAT	Tyr	TGT	Cys	T
	TTC	Phe	TCC	Ser	TAC	Tyr	TGC	Cys	C
	TTA	Leu	TCA	Ser	TAA	Stop	TGA	Stop	A
	TTG	Leu	TCG	Ser	TAG	Stop	TGG	Trp	G
C	CTT	Leu	CCT	Pro	CAT	His	CGT	Arg	T
	CTC	Leu	CCC	Pro	CAC	His	CGC	Arg	C
	CTA	Leu	CCA	Pro	CAA	Gln	CGA	Arg	A
	CTG	Leu	CCG	Pro	CAG	Gln	CGG	Arg	G
A	ATT	Ile	ACT	Thr	AAT	Asn	AGT	Ser	T
	ATC	Ile	ACC	Thr	AAC	Asn	AGC	Ser	C
	ATA	Ile	ACA	Thr	AAA	Lys	AGA	Arg	A
	ATG	Start	ACG	Thr	AAG	Lys	AGG	Arg	G
G	GTT	Val	GCT	Ala	GAT	Asp	GGT	Gly	T
	GTC	Val	GCC	Ala	GAC	Asp	GGC	Gly	C
	GTA	Val	GCA	Ala	GAA	Glu	GGA	Gly	A
	GTG	Val	GCG	Ala	GAG	Glu	GGG	Gly	G

The first step of the laboratory activity is to create a restriction digest of the segment of genomic data. A restriction enzyme cuts the DNA at a specific and unique location based on the sequence of nucleotides. Restriction enzymes are important tools in genetic engineering, and enable researchers to manipulate DNA. The computational restriction digest in this activity creates many fragments of various lengths. The output is in FASTA format. One of these fragments is selected.

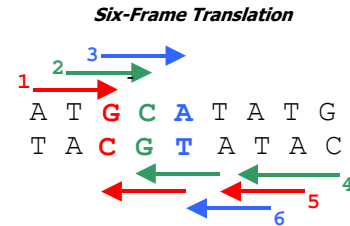
FASTA format description

A sequence in FASTA format begins with a single-line description, followed by lines of sequence data. The description line is distinguished from the sequence data by a greater-than ("**>**") symbol in the first column. It is recommended that all lines of text be 60 characters in length. An example sequence in FASTA format is:

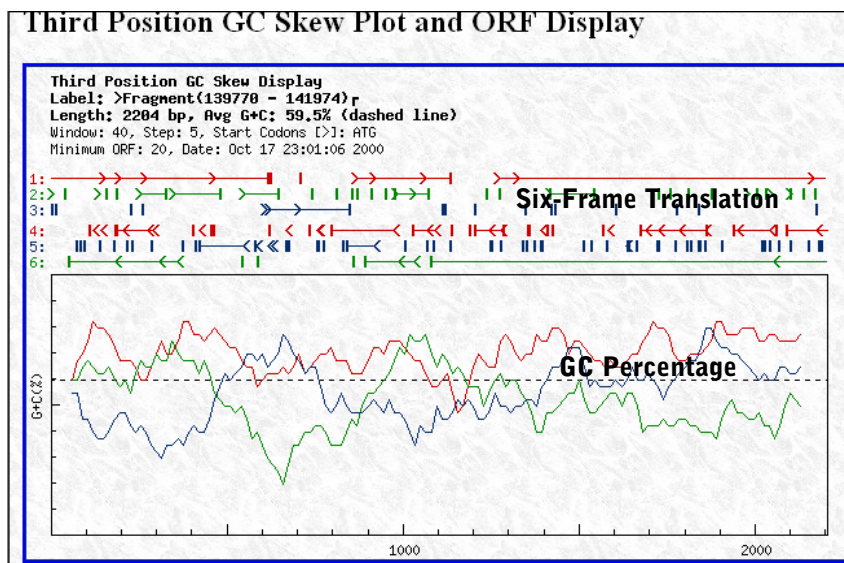
```
>Fragment 33(11617 - 12728)
GAATTCGCCGTTTCAGTCCGCTGGATTTCCGACTTAAAGCCGCCTAAAACGACGAACCA
TTGTTTATCGCGTCCCGCGCAAATTTGCCCGGACATGTCGGTACTCAATTTGAAACCG
TGTGAAGGGTGGCGATCAGTTCGGCGGCGCCTTCCCGACTGTGCAGGAATTAAGCGGAAT
CCGTATGAGACCAGCGGCCGTCTTCCGAGAGGTTCTTCTGCATCACCTCGAGCTCA
AGCCTGCTGCCCGGCGAGTTGGGAGGCGACAGCTTCTTTCAGGCTGGTGCTGGGCCTTC
AATCAGTTCGCGCAGGTCGCTGTTGGGAGCTCCAGTTCGCCCGCGCCACCCCTTCGTC
CGAGGAAACGAGCATCTGCGCGACAGGCGCAGGCAACCCGAAGCCTTCCAGGGTGCCTG
ATAGTCTTCGGCGCGAAATTTGCGGTATGACAGTCCCTGCCGCTCTGCTTCGACAGTTC
GGCGGCCAGTTCGGACAGGCTGAACGCTTCGTCGCCCCCAGTTCGTACACTTTTCCGGC
CTGTCGCTCAGCGGTTCAGTTCGCGCGGACAGCGGCTCGGCCAGGTCCTTGCAGGAGGCGG
GCTGATGCGGCCGTTCCGGCGGCTCCCAGGACACTTCCGGTCTGCAATGCCGCGCCAG
ATCGTAGTTTCCAGATACCAGCCGTTCCGACAGGAGGACGAACGGCACGCGCCGTTTCGGT
CAGCAATTTCTCGGTGGCTGATGGTTCGGCGGCCAGGACATCCGGGCGCTGTCGGCCTT
GAGCAGGCTGGGTAGCGGAGCAGTTCCACGCCCGCTGCGCTGGCCGCTCGATCACCGT
GCGGTGCTGGTCCGCGCGGTCATTTCAGGTCGCTCGACGAAATCAGCAACAGCCGCTGCAC
CCCCGCCAGTGCCTGTCCAGCCTTCCGCGCAGTTGTAGTCGGCCTGACGAATCTGCAC
GCCGGGCGCAACAGACCCCTGCGCCTTGGCAGGATTCGGACGATGGCGACGATGTGGTC
GGCAGGCACACCGCGTTCCAGCAACGACGAAACGACGAGCTGACCGAGCTTCCGGTGGC
GGCGGTGACGGCGATGGTTCGGGATGAGGTG
```

Third position GC skew display

We will then take one of the fragments of DNA and read it to see if there are potential genes there. There are no markers in the DNA sequence to indicate where one codon ends and the next one begins. Consequently, unless the location of the start codon is known ahead of time, a double-stranded DNA sequence can be interpreted in any of six ways. To account for this uncertainty, when predicting a protein the DNA sequence is translated into all six possible amino acid sequences. This exhaustive translation is called a "six-frame translation."





The DNA sequence of the selected fragment is analyzed to determine the percent of G's and C's occurring in the third position of each reading frame. The output is a Third Position GC Skew Display.



This display is a tool designed for predicting genes by comparing possible open reading frames to a third position GC plot. The tool is most effective in organisms with GC rich genomic sequence but it also works on all microbial sequences.

This plot shows possible open reading frames (ORFs) and graphs the GC content for those areas along the DNA molecule. On top of the plot possible open reading frames

(ORFs) in the six possible frameshifts. The arrowhead  indicates the start codon and the vertical lines  indicate the stop codon. All six of the possible reading frames (frameshifts) are represented in the first six lines. Below these lines is a graph that denotes areas that have higher GC content than the average. The average GC content is displayed by the dotted horizontal line in the center of the plot.

Once a potential reading frame is identified, the user can select that area and a new window will appear with information about the potential open reading frame. The new window also has several search options to do sequence queries against two different biological databases: TIGR and NCBI (National Center for Biotechnology Information). To submit the sequence of nucleotides or amino acids simply click the SEARCH button below the database search you are interested in.

Start: 3184 End: 3859 Frame: 1

3rd-Letter GC: 88.9%

Base Composition: A: 99 C: 247 G: 236 T: 96

>#1 (3184..3861) 678bp

```

ATGTGGCTCTGACCTTCCCGGAGGCGACCGAGGTGGGGTGGGGT
GGACGTGGGGCTTGGCCATGCACGGGGGAGTTGGGGTGTCTGG
TGCAGCGGGGAGTTGCCACGCCAGACTGGGCACTGGCGGGCGC
TTCCGTGACCGGGGAGGAACCTGATGAGGGGGCTGGCGAACTGAG
GACCGAAACAGGCTCAGCCTGGAGCCCGGACCTGGAGCAGTTTTTA
CCTTGGCGAGGTGGGGCGACCCGGCGGGCGGTCTTAAGCGTGGC
CACCTGGCGCTTGGCGACGGCAGGTGGAGCGGGCGCGAGGGGGCA
CACGTTGGGGGAGTGGCTGGCGGACACGCTGGGGCTGGCT
TCGACCACAGGGGATTCGGACCGCCATCAAGGGCTGCAACTGGC
CTGACTAGCCACCTGGCGCTGAAATTCGGCGACACCTTCAACCT
GCCGAGCTGCAAGGGGTACGAGGCCATTGGCCACGGCACTGCACA
AGCGTAATTCGGTAAGCGCATTCTGGCGGGGATCTGACCCCTGC
GGCGAGGGCGCAGCGGCTGGGACGGCCCGCAGCTCTACCGCGCGC
CAAGGCACCGCAGCGGGCGCTGTA

```

Blast nucleotide sequence against TIGR databases:

Search with blastx against

ignore Hypotheticals filtered for low complexity regions.

Searching the databases - BLAST

A common application of sequence alignment is searching a database for sequences that are similar to a query sequence. In these searches, an alignment of a sequence is matched against a database. By far the most popular tool for searching sequence databases is a program called BLAST, or Basic Local Alignment Search Tool. This is the most common algorithm used for online search tools.

Sequences are expected to be represented in the standard IUB/IUPAC amino acid and nucleic acid codes, with these exceptions: lower-case letters are accepted and are mapped into upper-case; a single hyphen or dash can be used to represent a gap of indeterminate length; and in amino acid sequences, U and * are acceptable letters (see below).

Before submitting a request, any numerical digits in the query sequence should either be removed or replaced by appropriate letter codes (e.g., N for unknown nucleic acid residue or X for unknown amino acid residue). The nucleic acid codes supported are:

A --> adenosine	M --> A C (amino)	C --> cytidine
S --> G C (strong)	G --> guanine	W --> A T (weak)
T --> thymidine	B --> G T C	U --> uridine
D --> G A T	R --> G A (purine)	H --> A C T
Y --> T C (pyrimidine)	V --> G C A	K --> G T (keto)
N --> A G C T (any)		

BLAST

For those programs that use amino acid query sequences (BLASTP and TBLASTN), the accepted amino acid codes are:

A alanine	P praline	B aspartate or asparagines
Q glutamine	C cystine	R arginine
D aspartate	S serine	E glutamate
T threonine	F phenylalanine	U selenocysteine
G glycine	V valine	H histidine
W tryptophan	I isoleucine	Y tyrosine
K lysine	Z glutamate or glutamine	L leucine
X any	M methionine	N asparagine

Programs available for the BLAST search

BLAST® (Basic Local Alignment Tool) is a set of similarity search programs designed to explore all of the available sequence databases regardless of whether the query is protein or DNA. The BLAST programs have been designed for speed, with a minimal sacrifice of sensitivity to distant sequence relationships. BLAST uses a heuristic algorithm that seeks local as opposed to global alignments and is therefore able to detect relationships among sequence that share only isolated regions of similarity.

blastp	compares an amino acid query sequence against a protein sequence database
blastn	compares a nucleotide query sequence against a nucleotide sequence database
blastx	compares a nucleotide query sequence translated in all reading frames against a protein sequence database
tblastn	compares a protein query sequence against a nucleotide sequence database dynamically translated in all reading frames
tblastx	compares the six-frame translations of a nucleotide query sequence against the six-frame translations of a nucleotide sequence database.

Nucleotide Sequence Databases

nr	All GenBank+EMBL+DDBJ+PDB sequences (but no EST, STS, GSS, or phase 0, 1 or 2 HTGS sequences). No longer "non-redundant".
Month	All new or revised GenBank+EMBL+DDBJ+PDB sequences released in the last 30 days.
Drosophila	Drosophila genome provided by Celera and Berkeley Drosophila Genome Project (BDGP) .
Dbest	Database of GenBank+EMBL+DDBJ sequences from EST Divisions
Dbsts	Database of GenBank+EMBL+DDBJ sequences from STS Divisions
Htg	Unfinished High Throughput Genomic Sequences: phases 0, 1 and 2 (finished, phase 3 HTG sequences are in nr)
Gss	Genome Survey Sequence , includes single-pass genomic data, exon-trapped sequences, and Alu PCR sequences.
Yeast	Yeast (<i>Saccharomyces cerevisiae</i>) genomic nucleotide sequences
E. coli	Escherichia coli genomic nucleotide sequences
Pdb	Sequences derived from the 3-dimensional structure from Brookhaven Protein Data Bank
Kabat	[kabatnuc] Kabat's database of sequences of immunological interest
Vector	Vector subset of GenBank(R), NCBI, in ftp://ncbi.nlm.nih.gov/blast/db/
Mito	Database of mitochondrial sequences
Alu	Select Alu repeats from REPBASE, suitable for masking Alu repeats from query sequences. It is available by anonymous FTP from ncbi.nlm.nih.gov (under the /pub/jmc/alu directory). See "Alu alert" by Claverie and Makalowski, Nature vol. 371, page 752 (1994).
Epd	Eukaryotic Promotor Database found on the web at http://www.genome.ad.jp/dbget-bin/www_bfind?epd

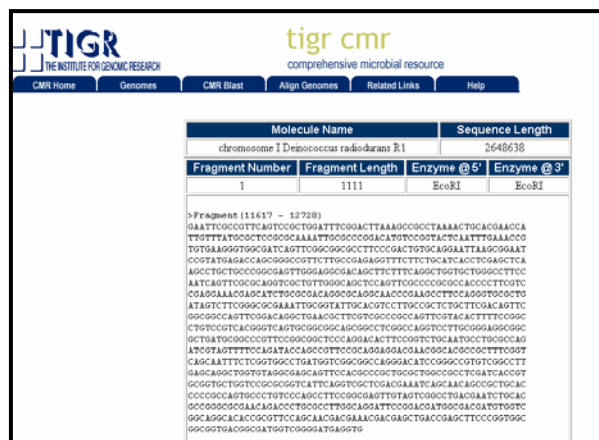
Identifying Potential Open Reading Frames

Gene Hunting

1. The time required to search an entire genome for potential open reading frames may be well over one hour – for this activity, it is suggested to use only a section of the genome. To select a section to analyze, perform a pseudo-restriction digest of a DNA molecule. Enter TIGR's Web Site at <http://www.tigr.org>.
2. At TIGR's home page, select Comprehensive Microbial Resource from the left hand tool bar under Genome Databases.
3. Scroll to the bottom of the page. Under Multi-Genome Applications, select Restriction Digest.
4. A new window will open: Restriction Digest Tool. Many restriction enzymes are available in the scroll window. Select EcoRI. Click on Add>>.
5. Select Genome from the scroll down menu. Select "Deinococcus radiodurans R1"
6. Select DNA Molecule from the scroll down menu. Select "chromosome I Deinococcus radiodurans R1."
7. Under Output, select "pseudo-restriction gel". No other optional criteria are selected. Click on submit.
8. A new window appears, showing the Psuedo-Restriction Digest. To improve resolution, Change the upper bound to 80,000 and select run again. Note the increase in fragment resolution.

Which fragments are larger on the pseudo-restriction digest: The fragments at the top or the fragments at the bottom of the image?



9. Click the Back button twice to return to the previous page and refresh the page. In the Restriction Digest Tool window set up the same restriction enzyme, genome and DNA molecule and under output, select "Fasta file genomic digest" and click on Submit.
10. A new window appears with the fragments in fasta format.
11. Scroll down the page to the fragment that is 4479 nucleotides in length (you may enter the find command to search for the correct fragment).
12. Using the mouse, carefully select only the nucleotide sequence. Enter the copy command and close the window.
13. Go back to the Compressive Microbial Resource window by selecting the CMR Home tab at the top of the page. Under Multi-Genome Application, select Third Position GC Skew Display.
14. Scroll down to "Enter Nucleotide Sequence in FASTA or Raw Format." Click mouse in the field and select the paste command.



15. Under additional options, keep as default. The options are explained below:

- Start codon: Choose the start codon commonly used in the organism searching.
- Window Size: Choose size of sliding window.
- Step Size: Window is moved along the sequence by the chosen step size. Larger step size makes the program run much faster. It will produce a low-resolution plot but using the default value gives a satisfactory result.
- Minimum ORF Size: The program recognizes ORFs, which are larger than the minimum ORF size as ORFs.
- Image width: Choose the width of the output image. By default, Image width is adjusted automatically.
- Incomplete ORF: The program recognizes incomplete ORFs that don't have apparent start or stop codons.

16. Select Generate Plot. A new window will open with the Third Position GC Skew Plot and ORF Display. This plot shows possible open reading frames (ORFs) and graphs the GC content for those areas along the DNA molecule. All six of the possible reading frames (frameshifts) are represented:

17. In this window, scroll down and note areas that have higher GC content than the average. The average GC content is displayed by the dotted horizontal line in the center of the plot. On top of the plot, the possible open reading frames (ORFs) in the six possible frameshifts are displayed. The arrowhead  indicates the start codon and the vertical lines  indicate the stop codon.

18. Along the frame with the highest GC content, move the mouse over the line and a small window will appear with coordinates. Note them (3184:3859 Frame 1).

Which other open reading frames may be potential genes based on the GC percentage (note the coordinates)?

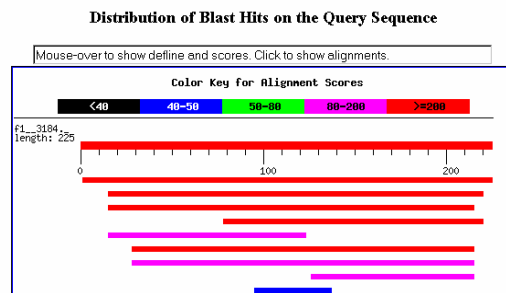
19. Scroll down the page and select the ORF coordinates that were noted in step 18. A new window will appear describing the ORF and reading the nucleotide sequence in that frame.

20. Under the peptide sequence section, "Blast Peptide sequence against NCBI database." Select the BLAST program to be used (select blastp for this exercise) with the arrow scroll down menu.

21. Using the scroll menu, select the database to search against, "nr".

22. Select search. A new window will appear. Scroll to the bottom of the page to the Format section. Select, under show, "Graphical Overview." Then, select "BLAST!". A new window appears. select "Format" and then wait for your results.

The scores are calculated based on the statistical probability of an incorrect match and the length of the match compared to the overall query length. Move the mouse over the different bars in the window and it will note what gene it is in the window. Click on the first bar and view the alignment.



The ORF matches several genes from Deinococcus best and several other organisms (results may vary):

Sequences producing High-scoring Segment Pairs:	High Score	Probability P(N)	N
EGAD 173798 DR0192 MutT/nudix family protein (Deinococcus...	1076	5.6e-108	1
GP 5738483 emb CAB52831.1 AL109848 putative DNA hydrolas...	374	1.4e-33	1
EGAD 49922 slr1690 hypothetical protein (Synechocystis PC...	285	3.7e-24	1
GP 3036882 emb CAA18515.1 AL022374 putative DNA hydrolas...	261	1.3e-21	1

What organism(s) did your open reading frame match?

Were any of the matches previously identified genes (name at least one if there were any)?

23. Click on the scores numbers (the blue links on the right) to view the alignment data. Repeat for other ORFs.
24. To find more information on the protein found from the BLAST search, go to the main CMR page and under Multi-Genome search select, Name. When a new window opens enter the name of the protein found in the previous step in the box.

Which organisms also have this type of gene?

25. Click on the locus link for any of the genes listed under *Deinococcus radiodurans* to find more information on the gene's function.

What are the cellular roles of this gene?

What are some potential uses for this type of information and analysis?