



Database Searching and Finding Genes - Part 1

Introduction to Bioinformatics
September 2002

CRC-CGF Bioinformatics Course, 2002



Talk Outline

- Part 1
 - Blast Similarity Searching Basics:
 - Algorithm, Programs, Options
 - Statistics and Interpretation
 - General Search Strategy
 - Information Retrieval with SRS
- Part 2
 - ORF translation
 - Database searching
 - Finding genomic and mapping information
 - NCBI, EBI, Celera, Golden Path

CRC-CGF Bioinformatics Course, 2002



Basic Search Process

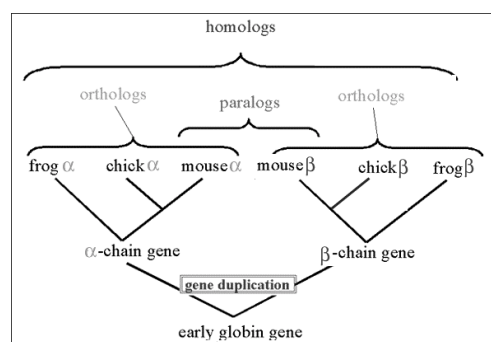
- Align query sequence against all database entries
- Score each alignment
- Rank the database entries by score
- Report the most interesting hits

- As you decide what question you are asking, then you will need to decide which program and database to use.

CRC-CGF Bioinformatics Course, 2002



Glossary




Similarity

Relatedness by some calculated measure

Statistically significant

Significant according to some statistical model


CRC-CGF Bioinformatics Course, 2002



Blast

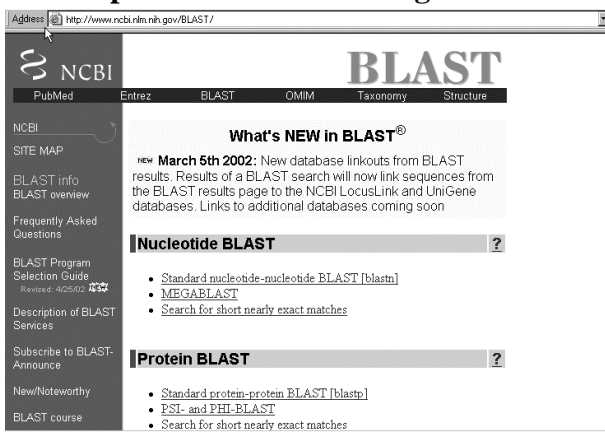
- Blast algorithm
- List of Blast programs
- Program options
- Interpreting the results
- Extreme-value statistics
- Outline of a general search strategy

CRC-CGF Bioinformatics Course, 2002



Blast at NCBI

<http://www.ncbi.nlm.nih.gov/BLAST/>



The screenshot shows the NCBI BLAST website interface. At the top, there is a navigation bar with links for PubMed, Entrez, BLAST, OMIM, Taxonomy, and Structure. The main content area features a 'What's NEW in BLAST' section with a date of March 5th 2002, followed by a 'Nucleotide BLAST' section with links to 'Standard nucleotide-nucleotide BLAST (blastn)', 'MEGABLAST', and 'Search for short nearly exact matches'. Below that is a 'Protein BLAST' section with links to 'Standard protein-protein BLAST (blastp)', 'PSI- and PHI-BLAST', and 'Search for short nearly exact matches'. A left sidebar contains various utility links like 'SITE MAP', 'BLAST info', and 'Frequently Asked Questions'.

CRC-CGF Bioinformatics Course, 2002



Blast Algorithm

CRC-CGF Bioinformatics Course, 2002



Blast Algorithm

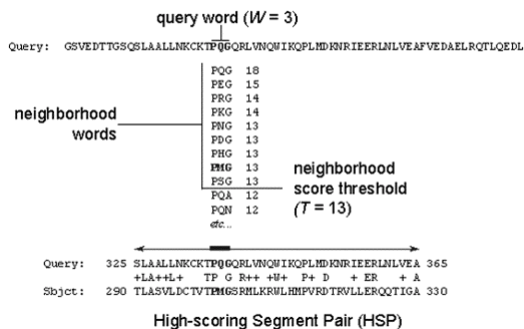
- Blast uses pre-indexed databases
 - The location of every 'word' of each database entry is remembered
- To perform a query
 - Break the query sequence into 'words'
 - Scan the database for these 'words'
 - (For each 'word' don't just use the 'word', use all similar 'words')
 - Where two non-overlapping 'words', within a certain distance of each other in the query, are matched against a database entry and with no gap required then this region of the two sequences is called a segment pair
 - The segment pair is extended until the score drops by **X** below its maximum value
 - The score for each database entry is calculated in this way
 - Database entries with statistically significant scores are reported

CRC-CGF Bioinformatics Course, 2002



Neighbourhood Word Threshold

The BLAST Search Algorithm



CRC-CGF Bioinformatics Course, 2002



Scoring the Alignment

Query: 119 EGVDIIMGSHGKTNLKEILLGSVTENVIKKSNKPVLVVK 158
E V +II+ S GK +L LGS V++K+ KPVL++K

Sbjct: 116 ENVSLIILPSRGKLSLSHEFLGSTVMRVLRLKTKKPVLIK 155

	A	C	D	E	F	G	H
A	4	0	-2	-1	-2	0	-2
C	0	9	-3	-4	-2	-3	-3
D	-2	-3	6	2	-3	-1	-1
E	-1	-4	2	5	-3	-2	0
F	-2	-2	-3	-3	6	-3	-1
G	0	-3	-1	-2	-3	7	-1
H	-2	-3	-1	0	-3	-1	7

BLOSUM 62

Score =

$$\begin{aligned}
 &5+0+4+0+2+4+4+2-2 \\
 &+4+0+6+5-2+1+4+0 \\
 &+0-3+0+4+6+4-2-2 \\
 &-2+0+4+2+2+5+1+0 \\
 &+5+7+4+4+1+1+5 \\
 &= 81
 \end{aligned}$$

CRC-CGF Bioinformatics Course, 2002



Blast Algorithm

- Lowering the Neighbourhood Word Threshold (T) allows the search to find more distantly related sequences
 - Raising the value of the Segment Extension Cutoff (X) allows the search to extend the region that may be considered as an HSP
 - Changing the Expectation parameter (E) does not change any scores or HSPs, just the score threshold for reporting
-
-

CRC-CGF Bioinformatics Course, 2002



Blast Programs

CRC-CGF Bioinformatics Course, 2002



Blast at NCBI

- **Nucleotide BLAST**
 - Standard nucleotide-nucleotide Blast [blastn] MegaBlast Search for short nearly exact matches
- **Protein BLAST**
 - Standard protein-protein Blast [blastp] PSI and PHI Blast Search for short nearly exact matches
- **Translated BLAST Searches**
 - Nucleotide query - Protein db [blastx] Protein query - Translated db [tblastn]
 - Translated nucleotide query - Translated db [tblastx]
- **Search for conserved domains**
 - Search the Conserved Domain Database using RPS-Blast
 - Search by domain architecture [CDART]
- **Pairwise BLAST**
 - Blast 2 sequences
- **Genomic BLAST pages**
 - Human, Mouse, Rat, Fugu, Zebrafish, *Anopheles gambiae*, *Arabidopsis thaliana*, *Oryza sativa*, Microbes
- **Specialized BLAST pages**
 - VecScreen, IgBlast, Trace Blast
- **Retrieve results for an existing Request ID**
 - Retrieve results for an existing Request ID
- **JavaScript free BLAST pages**
 - Get the Blast home page with JavaScript free links

CRC-CGF Bioinformatics Course, 2002



Blast Programs

	Sequence	Database	Comparison
blastn	nucleotide	nucleotide	nucleotide
blastp	protein	protein	protein
blastx	nucleotide	protein	protein
tblastn	protein	nucleotide	protein
tblastx	nucleotide	nucleotide	protein
PSI-Blast	protein	protein	protein
PHI-Blast	protein	protein	protein

CRC-CGF Bioinformatics Course, 2002



PSI-Blast and PHI-Blast


CRC-CGF Bioinformatics Course, 2002



PSI Blast Profile Searching

- Position specific scoring matrix
- Represents the weighting for each amino acid or nucleic acid to contribute to the score for an alignment
- Requires a previously constructed multiple alignment
- More sensitive than searching with a single sequence because real residue distributions at each position are taken into account

CRC-CGF Bioinformatics Course, 2002




PSSM Search

a	15	226	(45)	182	(167)	(18)				
c	(30)	18	32	39	60	21				
g	24	18	42	42	24	15				
t	229	(36)	179	(35)	47	244				

A	C	T	A	T	A	A	T	C	G		
	30	36	45	35	167	18			331		
		229	226	179	182	167	244			1227	
			15	36	45	182	47	21			346

CRC-CGF Bioinformatics Course, 2002



PSI-Blast

- Position Specific Iterative Blast
- Generate a multiple sequence alignment from first search pass
- Build a position-specific score matrix
- Use this construct to search database
- Iterate until all significant matches against the PSSM have been found

- Profile style search
- Only available for protein-protein search

CRC-CGF Bioinformatics Course, 2002



PSI-Blast output

```

Sequences producing significant alignments:                                (bits) Value

sp|P49767|VEGC_HUMAN  VASCULAR ENDOTHELIAL GROWTH FACTOR C PRECU...  396  e-109
gnl|PID|e1215492  (Y15837) vascular endothelial growth factor C ...  384  e-106
sp|P97953|VEGC_MOUSE  VASCULAR ENDOTHELIAL GROWTH FACTOR C PRECU...  381  e-105
bbs|177252  VEGF-C=32 kda vascular endothelial growth factor/Flt...  290  8e-78
gi|2323339  (AF014827) vascular endothelial growth factor D [Rat...  255  2e-67
gnl|PID|e283242  (X99572) member of PDGF /VEGT family of growth ...  255  2e-67
gnl|PID|d1025175  (D89630) VEGF-D [Homo sapiens] >gi|2879834|gnl...  255  3e-67

gnl|PID|e4974  (X00560) B chain fragment (aa 3-71) (7 is 2nd bas...  88  8e-17
bbs|85194  (S85224) vascular endothelial growth factor; VEGF 206...  58  9e-08
gi|3139081  (AF062645) vascular endothelial growth factor 183 [H...  56  3e-07
gnl|PID|e1309800  (Y15921) COL1A1 and PDGFB fusion transcript [H...  52  4e-06
gnl|PID|e293780  (Y08643) COL1A1 and PDGFB fusion transcript [Ho...  52  5e-06
gi|3133495  (AF062644) vascular endothelial growth factor [Rattu...  48  7e-05

Sequences with E-value WORSE than threshold
pir||A60706  vascular endothelial growth factor - guinea pig (fr...  40  0.016
sp|P22762|GLHA_MACMU  GLYCOPROTEIN HORMONES ALPHA CHAIN PRECURSO...  38  0.048
gi|332625  (J02396) v-sis transforming protein p28 [Simian sarco...  37  0.14
    
```

CRC-CGF Bioinformatics Course, 2002



PHI Blast Motifs and Patterns

- What does a motif look like?
 - WSXWS
 - G-x(3)-[LIVMF]-x(2)-[GSA]-[LIVMF](2)-G-C-x-[GA]-[STA]-x(2)-[EG]-x(2)-[CWN]-[LIVM](2)
 - C-x-[GNQ]-x(1,3)-G-x-C-x-C-x(2)-C-x-C
 - CX[GNQ] GXCXCXXCXC (C^CC)

CRC-CGF Bioinformatics Course, 2002



PHI-Blast

- Pattern Hit Initiated Blast
- Define an amino acid pattern (WSXWS)
- Only Blast search against database entries containing that pattern
- Tests for similarity including and around the defined motif, i.e. segment pairs must include the defined motif
- No motif - No hit

CRC-CGF Bioinformatics Course, 2002



Blast Program Options

CRC-CGF Bioinformatics Course, 2002

Victorian Bioinformatics Consortium

Program Options

Search in one organism or division

Options for advanced blasting

Limit by entrez query or select from: (none)

Choose filter Low complexity Human repeats Mask for lookup table only Mask lower case

Expect: 10

Word Size: 11

Other advanced: -G 8 -y 25 -W 8 -q -4

Filter out non-specific sequences

Change the word size used to generate matching segments

Change the score threshold at which hits are reported

CRC-CGF Bioinformatics Course, 2002

Victorian Bioinformatics Consortium

Program Options

Format

Show Graphical Overview Linkout NCBI Alignment HTML Format

Number of: Descriptions 100 Alignments 50

Alignment view: Pairwise

Limit results by: query-anchored with identities (none)

Expect value range: flat query-anchored without identities

Layout: Two Windows Formatting options on page with results: None

Autofomat: Semi-auto

Show no more than this many hits

Control the display of alignments

Display the PSSM rather than the alignment

Include links to other databases

CRC-CGF Bioinformatics Course, 2002



Alignment Format

```

1      1      atgctgggtcatggcgccccaaccgtcctcctgctgctctcggcgccctggccctgacc 60
402303 1      .....a.....c..... 60
8571981 1      .....a.....c..... 60
22871 1      .....g.....c.....c..... 60
493158 1      .....g.....c.....c.....g..g..ag.. 60
8886026 1      .....a.....c.....tc..... 60
3643696 1      .....g.....c..a.....c.....g..g..ag.. 60
7958605 1      .....g.....c.....g..g..ag.. 60
187820 1      .....g.....c.....a..... 60
8571993 5      .....a..c..t..... 60
184122 1      .....g.....c.....a.....c.....g..g.. 60
454779 2      .....g.....c.....g..g..ag.. 36
2384743 1      .....g.....c.....ga..... 60
32184 1      .....g.....c.....a..... 60
1827489 1      .....g.....c.....a.....c.....g..g.. 60
1399310 5      .....t.....g..... 59
1261809 1      .....g.....c.....g..g..ag.. 60
187693 11      ..... 67
187680 3      .....c.....g..g..ag.. 51
22884 1      .....g.....c..a.....ga..... 60
22880 1      .....g.....c..a.....c.....ga..... 60
6746370 1      .....g.....c..a.....g..g..ag.. 60
    
```

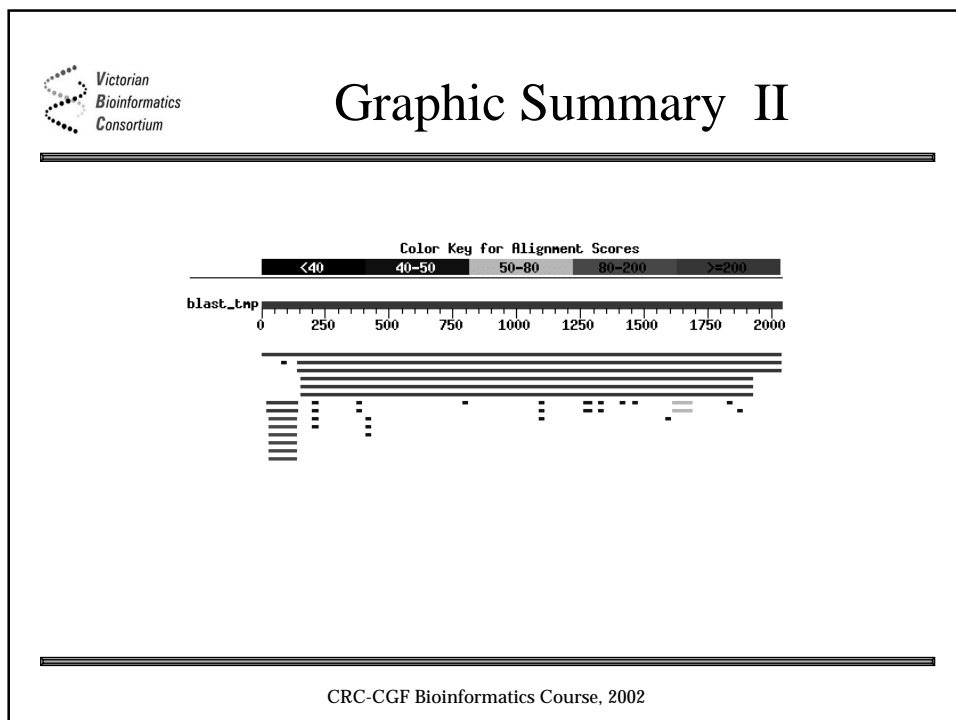
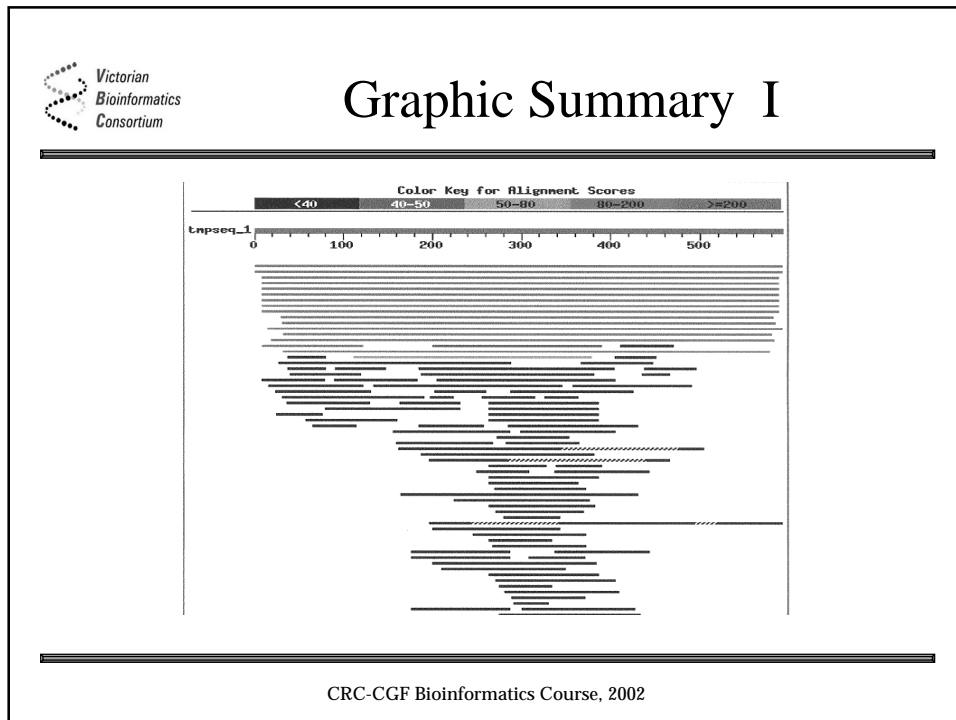
Flat query-anchored with identities


CRC-CGF Bioinformatics Course, 2002



Interpreting Blast Output

CRC-CGF Bioinformatics Course, 2002






Blast Output

TTHY_CHICK TRANSTHYRETIN PRECURSOR (PREALBUM	280	8e-75	137	150
TTHY_CROPO TRANSTHYRETIN PRECURSOR (PREALBUM	263	1e-69	125	150
CAA61120 S.scrofa mRNA for transthyretin	207	7e-53	100	144
BAB31352 Mus musculus adult male kidney cDNA1-l...	206	2e-52	102	149
TTHY_HUMAN TRANSTHYRETIN PRECURSOR (PREALBUM(TT...	203	8e-52	100	150
BAA77579 Xenopus laevis mRNA for transthyrete cds	183	8e-46	91	151
AAC26108 Sparus aurata transthyretin precursNA,...	156	1e-37	77	143
AAA40709 Rat transthyretin gene, exon 4	69	3e-11	32	34
AAD14937 Homo sapiens transthyretin (transthne, ...	52	3e-06	22	44
AAC33718 Salmonella dublin regulatory protein (copR)	49	3e-05	34	107
AAF34358 Hystrix africaeaustralis SP7702 transthyret	39	0.032	20	31
AAA36784 Human thyroxine binding prealbumin exon 4	32	3.2	15	17
AAB64289 Zea mays translation initiation factor ..	31	9.4	15	56

CRC-CGF Bioinformatics Course, 2002



cDNA vs Genomic

>emb|AL008726.2|HS337018 Human DNA sequence from ... on chromosome 20q12-13.1,
 Length = 86080

Query: 1364 agggcgccggccacatggttcccaccgacaagcccctcgctgccttcaccatgttctccc 1423
 |||

Sbjct: 70761 agggcgccggccacatggttcccaccgacaagcccctcgctgccttcaccatgttctccc 70820

Query: 449 aggtcgcccagagcaattttgaggcccttcaagatttcttccgcctctttccggagtaca 508
 |||

Sbjct: 65119 aggtcgcccagagcaattttgaggcccttcaagatttcttccgcctctttccggagtaca 65178

Query: 63 agtgtcctggcgctcccaggcgaggcagcccccgaccaggacgagatccagcgcctccc 122
 |||

Sbjct: 64021 agtgtcctggcgctcccaggcgaggcagcccccgaccaggacgagatccagcgcctccc 64080

Query: 952 caggcaactgctgcgctcaggggataaagtgcgcatggannnnnntgcaccaacacaaca 1011
 |||

Sbjct: 67387 caggcaactgctgcgctcaggggataaagtgcgcatggacccccctgcaccaacacaaca 67446

CRC-CGF Bioinformatics Course, 2002



Statistics and Blast

CRC-CGF Bioinformatics Course, 2002



Statistics

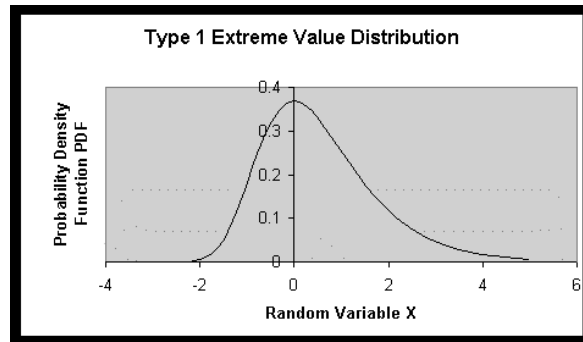
- The Blast program looks to find the highest-scoring segments of local similarity between the query sequence and database entries.
- If the query sequence and database entries are strings of random letters then the distribution of these **maximum** scores has an Extreme Value Distribution
- This is different to the distribution of the scores from **all** of the segments of local similarity between a query and a database
- Biological databases consist of very many sequences unrelated to the query and maybe a few that are, so the Extreme Value Distribution is a good approximation of the scores we find when database searching

CRC-CGF Bioinformatics Course, 2002



Extreme Value Distribution

$$E = Kmne^{-\lambda S}$$



CRC-CGF Bioinformatics Course, 2002



Statistics

- Blast uses the extreme value distribution to approximate the statistical significance of scores found in the real search of a biological database
- Given a score, one can calculate the number of entries that would be expected to attain that score or better, aligning a random query against a random database
- The E value is the number of high-scoring segment pairs that would be expected to achieve the given local similarity score given the random sequence - random database model
- NOT - the probability that the query and database entry are not biologically related

CRC-CGF Bioinformatics Course, 2002



Statistics


TTHY_CHICK TRANSTHYRETIN PRECURSOR (PREALBUM	280	8e-75	137	150
TTHY_CROPO TRANSTHYRETIN PRECURSOR (PREALBUM	263	1e-69	125	150
CAA61120 S.scrofa mRNA for transthyretin	207	7e-53	100	144
BAB31352 Mus musculus adult male kidney cDNA1-1...	206	2e-52	102	149
TTHY_HUMAN TRANSTHYRETIN PRECURSOR (PREALBUM(TT...	203	8e-52	100	150
BAA77579 Xenopus laevis mRNA for transthyrete cds	183	8e-46	91	151
AAC26108 Sparus aurata transthyretin precursNA,...	156	1e-37	77	143
AAA40709 Rat transthyretin gene, exon 4	69	3e-11	32	34
AAD14937 Homo sapiens transthyretin (transthne, ...	52	3e-06	22	44
AAC33718 Salmonella dublin regulatory protein (copR)	49	3e-05	34	107
AAF34358 Hystrix africaeustralis SP7702 transthyret	39	0.032	20	31
AAA36784 Human thyroxine binding prealbumin exon 4	32	3.2	15	17
AAB64289 Zea mays translation initiation factor ..	31	9.4	15	56

CRC-CGF Bioinformatics Course, 2002



General Search Strategy


CRC-CGF Bioinformatics Course, 2002



Blast Searching

- Start by deciding which program and database will give the best information
- Pick a search program
 - What type of sequence is the query, nucleic/protein - short/long - cDNA/genomic?
 - PSI-Blast gives very sensitive protein searches
- Pick a database
 - Are you looking for a protein match or genomic/mapping information?

CRC-CGF Bioinformatics Course, 2002



Blast Searching

- Pick the right search parameters
 - Use the default parameters to start with
 - Be prepared to try different scoring matrices, word lengths, gap penalties and extension cut offs
 - Changing the E value will affect how many hits are reported but not the score of any database hit

★ **Interpret the output according to your biological knowledge** ★

✦ ✦ ✦ ✦ ✦

- You can run Blast on your own PC if you have a local database of interest

CRC-CGF Bioinformatics Course, 2002

Victorian
Bioinformatics
Consortium

Things you never see

CRC-CGF Bioinformatics Course, 2002