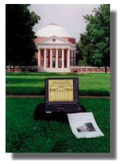
A Web-based Tool for Visual Identification of Novel Genes

Yanlin Huang, William Pearson and Gabriel Robins

www.cs.virginia.edu



Department of **Computer Science**

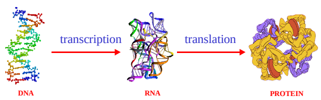
School of Engineering and Applied Science

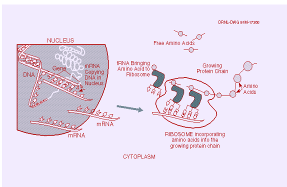
University of Virginia

(434) 982-2207 {yh4h,wrp, robins}@cs.virginia.edu

Molecular Genetics and Bioinformatics

CENTRAL DOGMA OF LIFE





GENETIC CODE

```
DNA      ACGCCAACCAGCACCAT GCCCATGATACTGGGGTACTGG
RNA      ACGCCAACCAGCACCAU GCCCAUGAUACUGGGGUACUGG
PROTEIN  T  P  T  S  T  M  P  M  I  L  G  Y  W
```

Second position

First position (5'-end)	U	C	A	G	Third position (3'-end)
U	UUU phe UUC phe UUA leu UUG leu	UCU ser UCC ser UCA ser UCG ser	UAU tyr UAC tyr UAA Stop UAG Stop	UGU cys UGC cys UGA Stop UGG trp	U C A G
C	CUU leu CUC leu CUA leu CUG leu	CCU pro CCC pro CCA pro CCG pro	CAU his CAC his CAA gln CAG gln	CGU arg CGC arg CGA arg CGG arg	U C A G
A	AUU ile AUC ile AUA ile AUG met	ACU thr ACC thr ACA thr ACG thr	AAU asn AAC asn AAA lys AAG lys	AGU ser AGC ser AGA arg AGG arg	U C A G
G	GUU val GUC val GUA val GUG val	GCU ala GCC ala GCA ala GCG ala	GAU asp GAC asp GAA glu GAG glu	GGU gly GGC gly GGA gly GGG gly	U C A G

Initiation Termination

Mutation in DNA

GAC : Asp (D)

↓

GAG : Glu (E)

Change in Protein

GOAL: FIND NEW GENES

EVOLUTION & SEQUENCE ALIGNMENT

```
                TASILNLCAIALDRYW
     TASILNLCAIALDRYW      TASILNLCAISLDRYW
#1                 TASILNLCAISLDRYW   #2 #3 TASILNLCAISLDRYT
TSSILNLCAIALDRYW
          #4                     #5
   TASILNLCVISLDRYW      TASILNLCIISLDRYW
```

EVOLUTION

```
#1 TSSILNLCAIALDRYW
#2 TASILNLCAISIDRYT
#3 TASILNLCAISIDRYT
#4 TASILNLCVISVDRYW
#5 TASILNLCIISVDRYW
```

These five proteins belong to the same protein family

SIMILARITY SCORING MATRIX (SM)

```
G S H K I L A R          G S H K I L A R
: : : : . : . :          :   : : . : . :
G S H K V L G R          G W H K V L G R
+5+4+7+5+3+6+1+7 = 38    +5-3+7+5+3+6+1+7 = 31
```

	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V	B	Z	X	*
A	4	-3	-1	-1	-3	-2	0	1	-3	-2	-3	-3	-2	-5	1	1	1	-7	-4	0	-1	-1	-1	-9
R		7	-2	-4	-5	1	-3	-5	1	-3	-5	2	-1	-6	-1	-1	-3	1	-6	-4	-3	-1	-2	-9
N			5	3	-5	-1	1	-1	2	-3	-4	1	-4	-5	-2	1	0	-5	-2	-3	4	0	-1	-9
D				5	-7	0	4	-1	-1	-4	-6	-1	-5	-8	-3	-1	-2	-9	-6	-4	4	3	-2	-9
C					9	-8	-8	-5	-4	-3	-8	-8	-7	-7	-4	-1	-4	-9	-1	-3	-6	-8	-5	-9
Q						6	2	-3	3	-4	-2	0	-2	-7	-1	-2	-2	-7	-6	-3	0	5	-2	-9
E							5	-1	-1	-3	-5	-1	-4	-8	-2	-1	-2	-9	-5	-3	3	4	-2	-9
G								5	-4	-5	-6	-3	-4	-6	-2	0	-2	-9	-7	-3	-1	-2	-2	-9
H									7	-4	-3	-2	-4	-3	-1	-2	-3	-4	-1	-3	1	1	-2	-9
I										6	1	-3	1	0	-4	-3	0	-7	-3	3	-3	-3	-2	-9
L											6	-4	3	0	-4	-4	-3	-3	-3	0	-5	-4	-3	-9
K												5	0	-7	-3	-1	-1	-6	-6	-4	0	-1	-2	-9
M													9	-1	-4	-3	-1	-6	-5	1	-4	-2	-2	-9
F														8	-6	-4	-5	-1	4	-3	-6	-7	-4	-9
P															7	0	-1	-7	-7	-3	-3	-1	-2	-9
S																4	2	-3	-4	-2	0	-2	-1	-9
T																	5	-7	-4	0	-1	-2	-1	-9
W																		12	-2	-9	-6	-8	-6	-9
Y																			9	-4	-4	-6	-4	-9
V																				5	-4	-3	-2	-9
B																					4	2	-2	-9
Z																						5	-2	-9
X																							-2	-9
*																								1

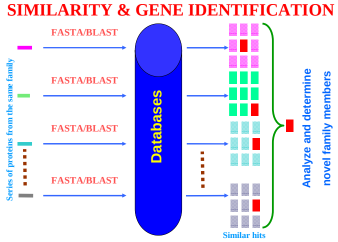
WHY: BETTER DRUGS, DISEASE CURES, LONGER LIFE, ETC.

Panning for Genes

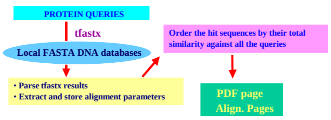
BEST SM-BASED ALIGNMENT PROGRAMS

Name	Alignment	Type
blastp/BLAST	Protein vs. Protein Database	pairwise
tblastn/BLAST	Protein vs. DNA Database	pairwise
fasta/FASTA	Protein vs. Protein Database	pairwise
tfastx/FASTA	Protein vs. DNA Database	pairwise
genewise	Protein vs. DNA sequence	pairwise
CLUSTALW	Protein or DNA	multiple

SIMILARITY & GENE IDENTIFICATION



FAST_PAN AUTOMATES THE PROCESS




```
>>gi|10873260|gb|BF079430.1|BF079430 MARC 2PIG Sus scrofa cDNA 5', mR (557 aa)
Frame: f initn: 778 init1: 778 opt: 786 Z-score: 1837.8 bits: 348.5 E(): 2.2e-98
Smith-Waterman score: 786; 75.824% identity in 182 aa overlap (1-182:10-555)

>gi|108 1- 182:-----------------------------------------------------------------:
         10        20        30        40        50        60        70        80
MPMILGYWNVRGLTHPIRMLLEYTDSSYDEKRYTMGDAPDFDRSQWLNEKFKLGLDFPNLPYLIDGSHKITQSNAILRYL
: .::::..:::.: ::.:::::::::.::.:::::::::::::::::::::::::.::::::::::::::::::.::::
MTLILGYWDIRGLAHAIRLLLEYTDSSYEEKKYTMGDAPDYDRSQWLSDKFKLGLDFPNLPYLIDGAHKLTQSNAILRYI
         10        40        70       100       130       160       190       220

         90       100       110       120       130       140       150       160
ARKHHLDGETEEERIRADIVENQVMDTRMQLIMLCYNPDFEKQKPEFLKTIPEKMKLYSEFLGKRPWFAGDKVTYVDFLA
::::.. ::::::.::.:..:::. ::   :  :::.:::: :: .:: ::::: .::::::::::::::::::::::
ARKHNMCGETEEEKIRVDVLENQANDTSEALASLCYSPDFEKLKPGYLKEIPEKMKPFSEFLGKRPWFAGDKLTYVDFLA
        250       280       310       340       370       400       430       460
```

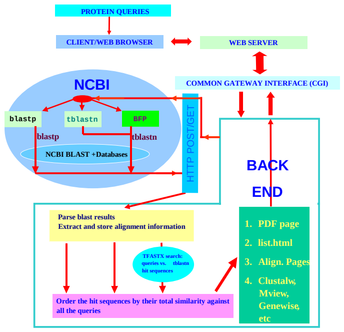
FAST_PAN DRAWBACKS

- Command-line on UNIX/LINUX not user-friendly
- Can only query local DNA databases
- Can not query against a specific organism's sequences
- Limited computational power
- Small user base
- Alternative sequence alignment capabilities needed
  (e.g., CLUSTALW, GENEWISE, MVIEW, etc.)

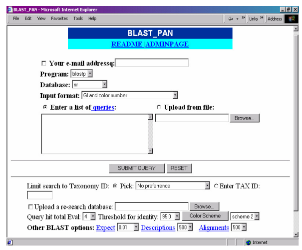
MOTIVATION

- BLAST_PAN: send the queries to NCBI BLAST server
- Search against the NCBI databases (DNA or protein)
- Parse and plot out the high-scoring database sequences
- WWW access
- Front end ➔ CGI ➔ back end ➔ output
- Integrate other sequence alignment capabilities
  e.g. CLUSTALW, MVIEW, GENEWISE, TFASTX, etc

BLAST_PAN STRATEGY



EXAMPLE WEB INTERFACE



INPUT:

```
gtm1_human 2
gtm1_mouse 2
gtm3_human 2
gt27_fashe 3
gtp_human 7
gtp_caeel 7
gts1_caeel 6
gts_ommsl 6
gta1_human 1
gta1_mouse 1
gta2_mouse 1
gta2_human 1
gtt1_human 5
gtt2_human 5
dcma_metsp 5
gtt1_drome 4
gtt1_anoga 4
gta_plepl 0
gth3_arath 0
gth1_arath 3
gth3_maize 3
gth4_maize 3
gtxa_tobac 2
gtxa_arath 2
gtx2_maize 2
sspa_ecoli 1
gtx1_soltu 6
lige_psepa 6
gt_haein 7
```

INPUT FORMATS

scheme	0	1	2	3	4	5	6	7
0	0	1	2	3	4	5	6	7
1	0	1	2	3	4	5	6	7
2	0	1	2	3	4	5	6	7
3	0	1	2	3	4	5	6	7
4	0	1	2	3	4	5	6	7
5	0	1	2	3	4	5	6	7

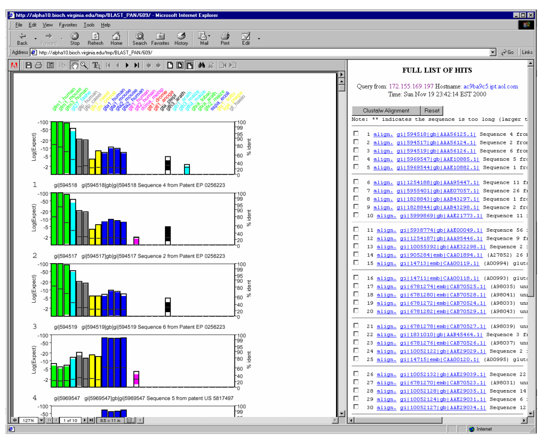
gene_identifier color_no

```
gtm1_mouse 2
gtm2_mouse 2
```

>fasta_format_description_line <color: color_no>

```
>GTM1_HUMAN GLUTATHIONE S-TRANSFERASE MU 1 (GSTM1-1) <color:1>
PMILGYWDIRGLAHAIRLLLEYTDSSYEEKKYTMGDAPDYDRSQWLNEKFKLGLDFPNLPYLIDGAHKI
TQSNAILCYIARKHNLCGETEEEKIRVDILENQTMDNHMQLGMICYNPEFEKLKPKYLEELPEKLKLYS
EFLGKRPWFAGNKITFVDFLVYDVLDLHRIFEPKCLDAFPNLKDFISRFEGLEKISAYMKSSRFLPRPV
FSKMAVWGNK
>GTT1_DROME GLUTATHIONE S-TRANSFERASE 1-1 (CLASS-THETA) <color:4>
MVDFYYLPGSSPCRSVIMTAKAVGVELNKKLLNLQAGEHLKPEFLKINPQHTIPTLVDNGFALWESRAI
QVYLVEKYGKTDSLYPKCPKKRAVINQRLYFDMGTLYQSFANYYYPQVFAKAPADPEAFKKIEAAFEFL
NTFLEGQDYAAGDSLTVADIALVATVSTFEVAKFEISKYANVNRWYENAKKVTPGWEENWAGCLEFKKY
FE
```

SAMPLE OUTPUT



Post-processing Capabilities

CLUSTALW CAPABILITY

- **Identify same clone with different annotations**

```
gi|4504176|ref|NM_000849.1|      CTCGGAAGCCCGTCACCATGTCGTGCGAGTCGTCTATGGTTCTCGGGTAC
gi|183680|gb|J05459.1|HUMGSTM3   CTCGGAAGCCCGTCACCATGTCGTGCGAGTCGTCTATGGTTCTCGGGTAC
                                 **************************************************
```

- **Compare similarity among different database sequences**

```
gi|399829|sp|Q00285|GTMU_CRILO   MPMILGYWNVRGLTNPIRLLLEYTDSSYEEKKYTMGDAPDSDRSQWLNEK
gi|121720|sp|P19639|GTM3_MOUSE   MPMTLGYWNTRGLTHSIRLLLEYTDSSYEEKRYVMGDAPNFDRSQWLSEK
gi|121719|sp|P08010|GTM2_RAT     MPMTLGYWDIRGLAHAIRLFLEYTDTSYEDKKYSMGDAPDYDRSQWLSEK
gi|232206|sp|P30116|GTMU_MESAU   MPVTLGYWDIRGLAHAIRLLLEYTDTSYEEKKYTMGDAPNFDRSQWLNEK
gi|232204|sp|P28161|GTM2_HUMAN   MPMTLGYWNIRGLAHSIRLLLEYTDSSYEEKKYTMGDAPDYDRSQWLNEK
                                 **: ****: ***::.***:*****:***:*:* *****: ******.**
```

MVIEW CAPABILITY

Colored by: property

Identities computed with respect to: (1)gi|399829|sp|Q00285|GTMU_CRILO

```
1 gi|399829|sp|Q00285|GTMU_CRILO 100.0%   MPMILGYWNVRGLTNPIRLLLEY
2 gi|232204|sp|P28161|GTM2_HUMAN  78.0%   MPMTLGYWNIRGLAHSIRLLLEY
3 gi|232206|sp|P30116|GTMU_MESAU  79.8%   MPVTLGYWDIRGLAHAIRLLLEY
4 gi|121720|sp|P19639|GTM3_MOUSE  82.1%   MPMTLGYWNTRGLTHSIRLLLEY
5 gi|121717|sp|P04905|GTM1_RAT    89.0%   MPMILGYWNVRGLTHPIRLLLEY
```

MVIEW assigns different colors to biochemically different types of amino acids so the user will be able to tell whether there is a type change at a certain position according to the colors

SAMPLE ALIGNMENT PAGE

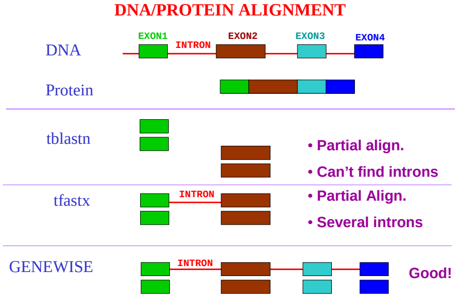
Query sequences matching: gi|594518

```
Match to gtm1_human (218 aa)
gi|594518|gb|AAA56125.1| Sequence 4 from Patent EP 0256223
Length = 218
Score = 394 bits (1001), Expect = e-110 Identities = 178/218 (81%), Positives =
201/218 (91%)

Query: 1 MPMILGYWDIRGLAHAIRLLLEYTDSSYEEKKYTMGDAPDYDRSQWLNEKFKLGLDFPNL 60
          MPM LGYWDIRGLAHAIRL LEYTD+SYE+KKY+MGDAPDYDRSQWL+EKFKLGLDFPNL
Sbjct : 1 MPMTLGYWDIRGLAHAIRLFLEYTDTSYEDKKYSMGDAPDYDRSQWLSEKFKLGLDFPNL 60

Query: 61 PYLIDGAHKITQSNAILCYIARKHNLCGETEEEKIRVDILENQTMDNHMQLGMICYNPEF 120
           PYLIDG+HKITQSNAIL Y+ RKHNLCGETEEE+IRVD+LENQ MD +QL M+CY+P+F
Sbjct : 61 PYLIDGSHKITQSNAILRYLGRKHNLCGETEEERIRVDVLENQAMDTRLQLAMVCYSPDF 120

Match to gtm2_human (218 aa)
```

⋮

Alignment page contains the alignments between the hit sequence and each of the queries

DNA/PROTEIN ALIGNMENT



TFASTX ALIGNMENT CAPABILITY

```
Match to gtm2_mouse (218 aa)
>>gi|467622|emb|X78316.1|ASPGST Artificial sequence plasmid GST-fusion vecto  (4905 aa)
```

GENEWISE ANALYSIS

```
Frame: f   BLAST SCORES:  Score =  185 bits (464), Expect = 3e-45
Smith-Waterman score: 457;      43.137% identity in 204 aa overlap (5-208:1067-1663)

>gi|467     5- 208: --------------------------------------------------------------------  :

            10        20        30        40        50        60        70        80
gi|121 LGYWDIRGLAHAIRLLLEYTDTSYEDKKYTMGDAPDYDRSQWLSEKFKLGLDFPNLPYLIDGSHKITQSNAILRYLARKH
       :::: :.::..  :::::: . .::.  :   ..    ..:  ..: ..::.:::.:::::: :::. :.::: ::.::.: ::
gi|467 LGYWKIKGLVQPTRLLLEYLEEKYEEHLYERDEG-----DKWRNKKFELGLEFPNLPYYIDGDVKLTQSMAIIRYIADKH
       1070      1100      1130      1160      1190      1220      1250      1280

            90       100       110       120       130       140       150       160
gi|121 NLCGETEEERIRVDILENQAMDTRIQLAMVCYSPDFEKKKPEYLEGLPEKMKLYSEFLGKQPWFAGNKVTYVDFLVYDVL
       :. :   .:: ....::. ..: :  .. ..:: :::  : .:.. :::  : .:::.. :  .  .. :. ::. ::..::.:
gi|467 NMLGGCPKERAEISMLEGAVLDIRYGVSRIAYSKDFETLKVDFLSKLPEMLKMFEDRLCHKTYLNGDHVTHPDFMLYDAL
        1310      1340      1370      1400      1430      1460      1490      1520

           170       180       190       200
gi|121 DQHRIFEPKCLDAFPNLKDFMGRFEGLKKISDYMKSSRFLSKPI
       :    ..: ::::::.:  :  :.:.. .:. :.:::.... :.
gi|467 DVVLYMDPMCLDAFPKLVCFKKRIEAIPQIDKYLKSSKYIAWPL
        1550      1580      1610      1640
```

GENEWISE ALIGNMENT CAPABILITY

```
gtm2_mouse          1 MPMTLGYWDIRG                LAHAIRLLLEYTDT
                      M MTLGYWDIRG                LAHAIRLLLEYTD+
                      MSMTLGYWDIRG                LAHAIRLLLEYTDS
gi|11321913|em-88256 ataacgttgacgGTGAGTG  Intron 1  CAGcgcgaccccgtagt
                      tctctgagatgg                tcactgtttaacac
                      gcgaggcgcccg                gcccccgcgacaca

gtm2_mouse         27 SYEDKKYTMGD                 PDYDRSQWLSEK
                      SYE+KKYTMGD                 PDYDRSQWL+EK
                      SYEEKKYTMGD                 PDYDRSQWLNEK
gi|11321913|em-87900 atggaataaggGGTAATGA  Intron 2  CAGCTcgtgaactcaga
                      gaaaaaactga                 caaaggagtaaa
                      ccgaggtgggc                 tctcacgggtaa

gtm2_mouse         51 FKLGLDFPN                   LPYLIDGSHKITQSNAI
                      FKLGLDFPN                   LPYLIDG+HKITQSNAI
                      FKLGLDFPN                   LPYLIDGAHKITQSNAI
gi|11321913|em-87401 tacgcgtcaGTAGGTG  Intron 3  CAGccttagggcaaacaaga
                      tatgtatca                   tcattagcaatcagact
                      cggcgctct                   gccgttgtcgccgcccc
```

Genewise does a good job of finding introns and exons!

SUMMARY: BLAST_PAN

BLAST PAN

- Web-based, platform-independent, user-friendly
- Utilizes NCBI BLAST Server/database
- tfastx alignment capabilities like `tfastx` (BFP)
- Integrates CLUSTALW, GENEWISE and MVIEW
- Other useful features including organism-specific search, etc

Conclusion: BLAST_PAN offers a rapid, visual, powerful and comprehensive approach for identifying novel genes.

ACKNOWLEDGEMENTS: NIH, NSF, Packard Foundation