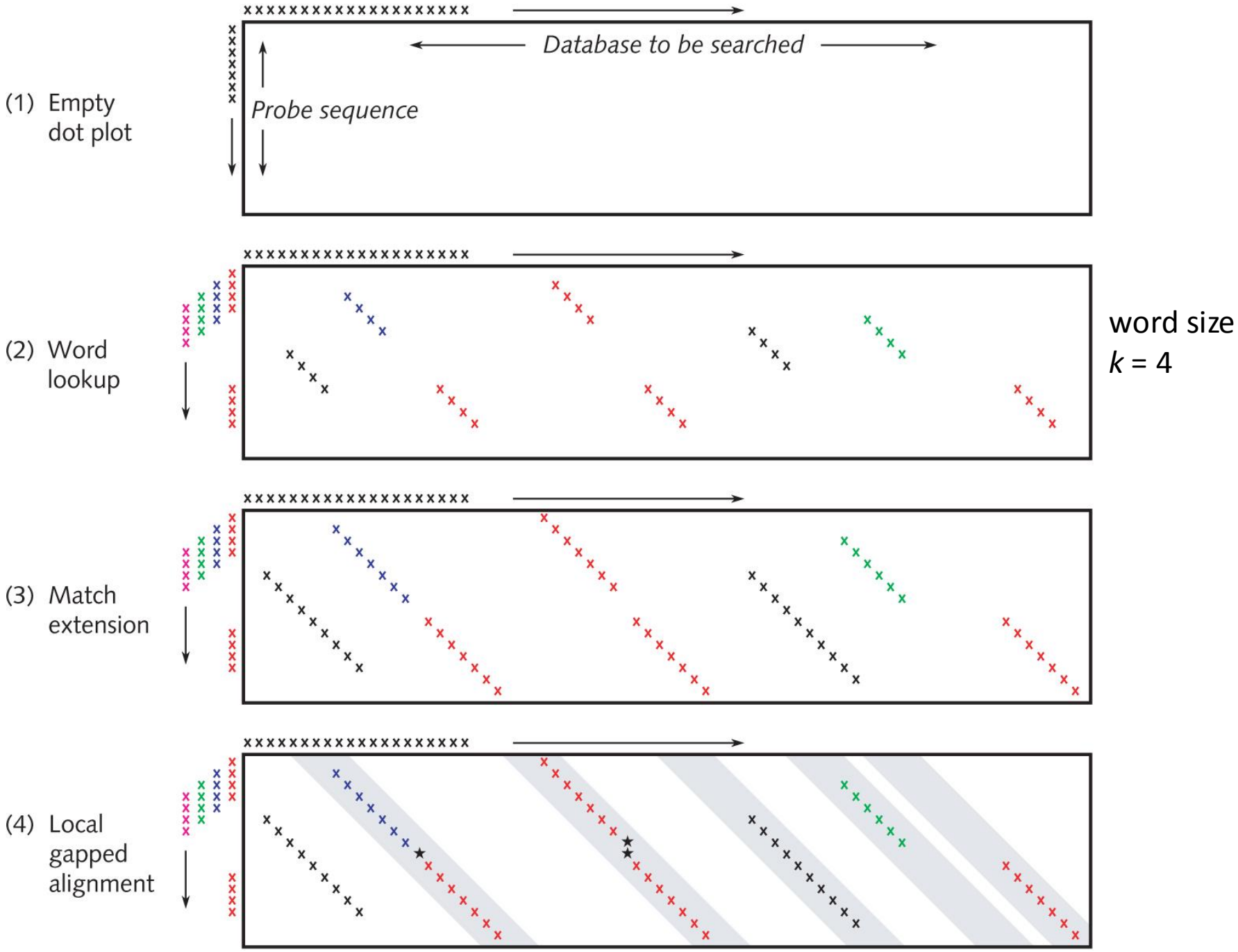
LECTURE 7
Blast

Using BLAST to search sequence databases

- Aims
  - Learn how to use BLAST (blast.ncbi.nlm.nih.gov)

    BLASTP, BLASTN, TBLASTN, BLASTX
  - Learn what's in the NCBI sequence databases
    - Refseq
    - Accession numbers
    - Genome, WGS, single-gene, EST
  - Concept of annotation


(1) Empty dot plot
Database to be searched
Probe sequence
(2) Word lookup
word size
k = 4
(3) Match extension
(4) Local gapped alignment

What BLAST does

(BLAST was developed by Stephen Altschul *et al*, 1990. It is the most-cited scientific paper ever.)

BLAST looks for HSPs:

HSP: "High-Scoring Pair" = a grey region in the previous slide, i.e. a region of matching between your **Query** and a database entry (the **Subject**). HSPs usually don't have gaps in the alignment between Query and Subject, or have only small gaps.

A Query can have several HSPs to the same Subject.

For each Subject in the database (millions of them), BLAST asks:

Does the Subject match the Query with at least *k* identical letters?

(by default, "word size" *k* = 8 for DNA; *k* = 3 for protein)

If yes, BLAST then extends each *k*-matching region out as far as it can, to make an HSP. The HSP is given a **score**, which is:

for DNA, the score is just 2x the number of matching letters, minus gap penalties.

for proteins, the score is calculated from a BLOSUM62 matrix.

What BLAST does

When a search is run, BLAST keeps a list of the database Subjects whose HSPs had the highest scores to your Query. (Typically 1000 are kept).

The **score** of each HSP in the list is then converted into an **E-value** ("expect" value).

An E-value is the number of HSPs expected to have this score or higher, purely by chance, taking into account:

- the size of the database
- the composition of the Query (e.g. a query that is AAAAAAAAAAA will have a lot of spurious hits).

Low E-values mean strong hits.

In theory, any HSP with E < 1 is significant.

In practice, a hit is only "convincing" if E is 1×10^{-6} or lower. This is written as 1.0e-6.

The output from BLAST is a sorted list of the Subjects with the lowest E-values in the database.

Note that

-- An E-value is not a probability.

-- In any search, <u>something</u> has to be the best hit. The trick is figuring out if the hit is a coincidence or due to shared ancestry (homology) of the sequences.

Exercise

- Find the sequences of EPO genes in as many different species as we can.

- By sequence similarity searching.

- Starting with human EPO:
  - Nucleotide database accession number X02157
  - Protein database accession number CAA26094

blast.ncbi.nlm.nih.gov

BLAST® Basic Local Alignment Search Tool

Home | Recent Results | Saved Strategies | Help

NCBI/ BLAST Home

BLAST finds regions of similarity between biological sequences. more...

New DELTA-BLAST, a more sensitive protein-protein search Go

BLAST Assembled RefSeq Genomes

Choose a species genome to search, or list all genomic BLAST databases.

- Human
- Mouse
- Rat
- *Arabidopsis thaliana*
- *Oryza sativa*
- *Bos taurus*
- *Danio rerio*
- *Drosophila melanogaster*
- *Gallus gallus*
- *Pan troglodytes*
- *Microbes*
- *Apis mellifera*

Basic BLAST

Choose a BLAST program to run.

nucleotide blast	Search a **nucleotide** database using a **nucleotide** query *Algorithms:* blastn, megablast, discontiguous megablast
protein blast	Search **protein** database using a **protein** query *Algorithms:* blastp, psi-blast, phi-blast, delta-blast
blastx	Search **protein** database using a **translated nucleotide** query
tblastn	Search **translated nucleotide** database using a **protein** query
tblastx	Search **translated nucleotide** database using a **translated nucleotide** query

Specialized BLAST

Choose a type of specialized search (or database name in parentheses.)

- Make specific primers with Primer-BLAST
- Search trace archives
- Find conserved domains in your sequence (cds)
- Find sequences with similar conserved domain architecture (cdart)
- Search sequences that have gene expression profiles (GEO)
- Search immunoglobulins (IgBLAST)
- Search using SNP flanks
- Screen sequence for vector contamination (vecscreen)
- Align two (or more) sequences using BLAST (bl2seq)
- Search protein or nucleotide targets in PubChem BioAssay
- Search SRA transcript and genomic libraries
- Constraint Based Protein Multiple Alignment Tool
- Needleman-Wunsch Global Sequence Alignment Tool
- Search RefSeqGene

BLAST®

Basic Local Alignment Search Tool

Home | Recent Results | Saved Strategies | Help

NCBI/ BLAST/ blastn suite

Standard Nucleotide BLAST

blastn | blastp | blastx | tblastn | tblastx

BLASTN programs search nucleotide databases using a

Enter Query Sequence

Enter accession number(s), gi(s), or FASTA sequence(s) Clear

X02157

Query subrange

From

To

Or, upload file Browse...

Job Title

Enter a descriptive title for your BLAST search

Align two or more sequences

Choose Search Set

Database: Human genomic + transcript / Mouse genomic + transcript / Others (nr etc.):

Human genomic plus transcript (Human G+T)

Exclude Optional: Models (XM/XP) Uncultured/environmental sample sequences

Entrez Query Optional

Enter an Entrez query to limit search

Program Selection

Optimize for:

- Highly similar sequences (megablast)
- More dissimilar sequences (discontiguous megablast)
- Somewhat similar sequences (blastn)

Choose a BLAST algorithm

BLAST

Search database Human G+T using Blastn (Optimize for somewhat similar sequences)

Show results in a new window

Algorithm parameters

4 types of BLAST search: #1, BLASTN (≈megablast)

		Query	
		DNA	Protein
Database	DNA	**BLASTN** megablast	TBLASTN
	Protein	BLASTX	BLASTP

BLASTN: Searches a DNA Query vs. a DNA database.

Typical use: to find highly-similar DNA sequences.

Advantages: It's the only option for sequences that are not protein-coding.

Disadvantages:

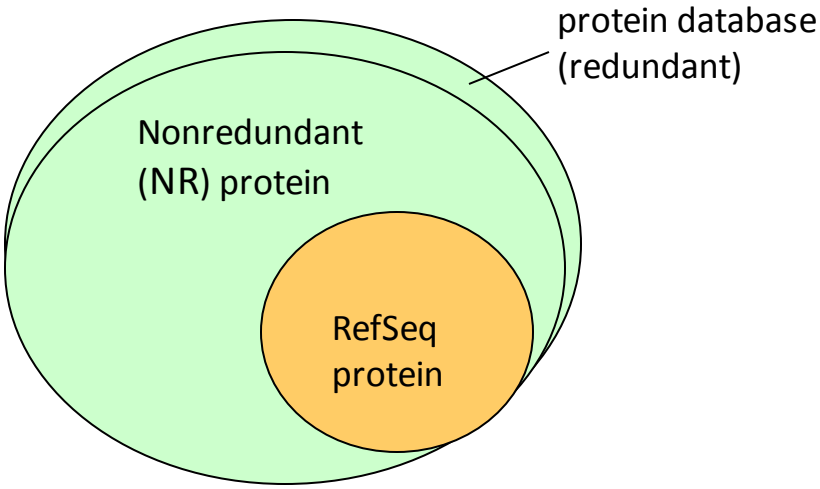
- It will miss genes whose sequences have diverged a lot.
- Repetitive DNA sequences cause problems (e.g. human Alu repeats).

Nucleotide databases for BLAST (BLASTN, TBLASTN)

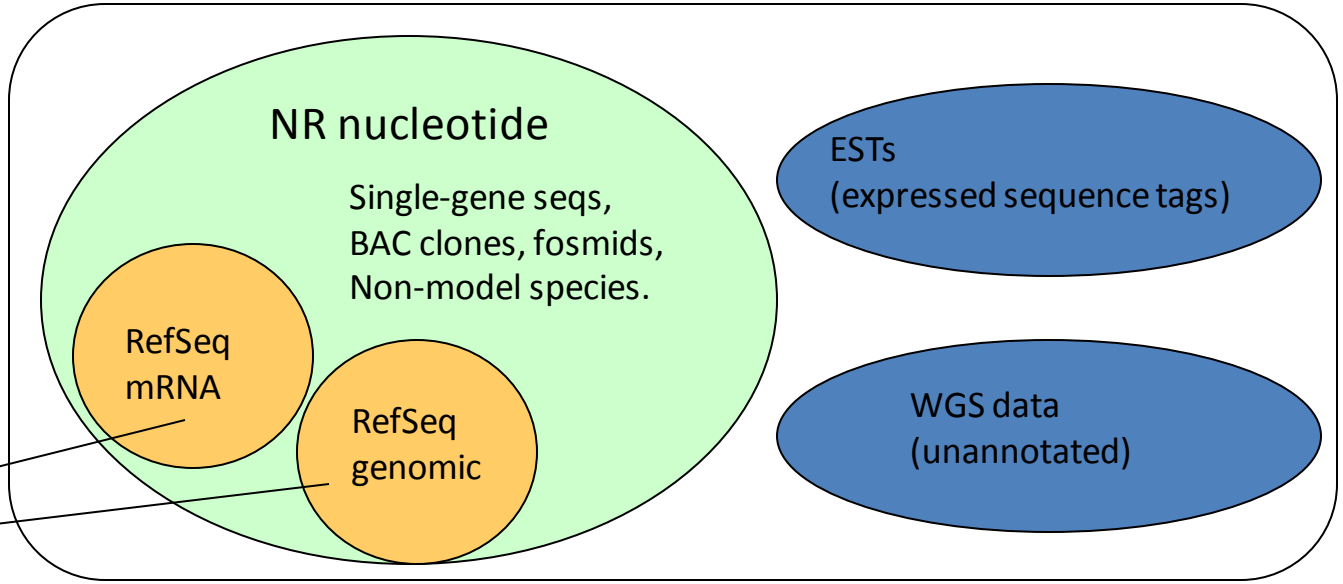
- Human Genomic + Transcript
- Mouse Genomic + Transcript
- Nucleotide collection (nr/nt) ("nonredundant nucleotide" db)
- Reference RNA sequences (refseq_RNA)
- Reference genomic sequences (refseq_genomic)
- Expressed sequence tags (EST)
- Whole genome shotgun contigs (WGS)
- and others…

NCBI sequence databases

protein



nucleotide



For genome-project species

Example: 1A: BLASTN: Query is human EPO cDNA. Database is Human Genomic + Transcript.

Edit and Resubmit | Save Search Strategies | ▷Formatting options | ▷Download

emb|X02157| (1342 letters)

Query ID gi|31229|emb|X02157.1|
Description Human mRNA for fetal erythropoietin
Molecule type nucleic acid
Query Length 1342

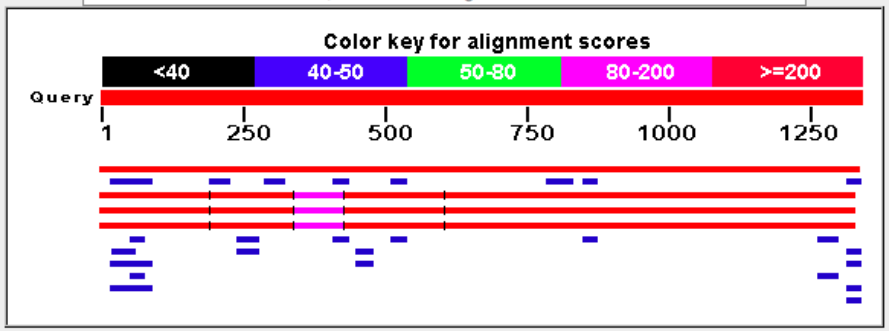
Database Name Human G+T (2 databases)
Description ▷See details
Program BLASTN 2.2.27+ ▷Citation

Other reports: ▷Search Summary [Taxonomy reports] [Distance tree of results] [genome view]

Graphic Summary

Distribution of 42 Blast Hits on the Query Sequence

Mouse over to see the defline, click to show alignments



Descriptions

Legend for links to other resources: U UniGene E GEO G Gene S Structure M Map Viewer PubChem BioAssay

Score of best individual HSP

Total score of all HSPs

E-value of best individual HSP. Sorted: lowest first, for each database.

Sequences producing significant alignments:

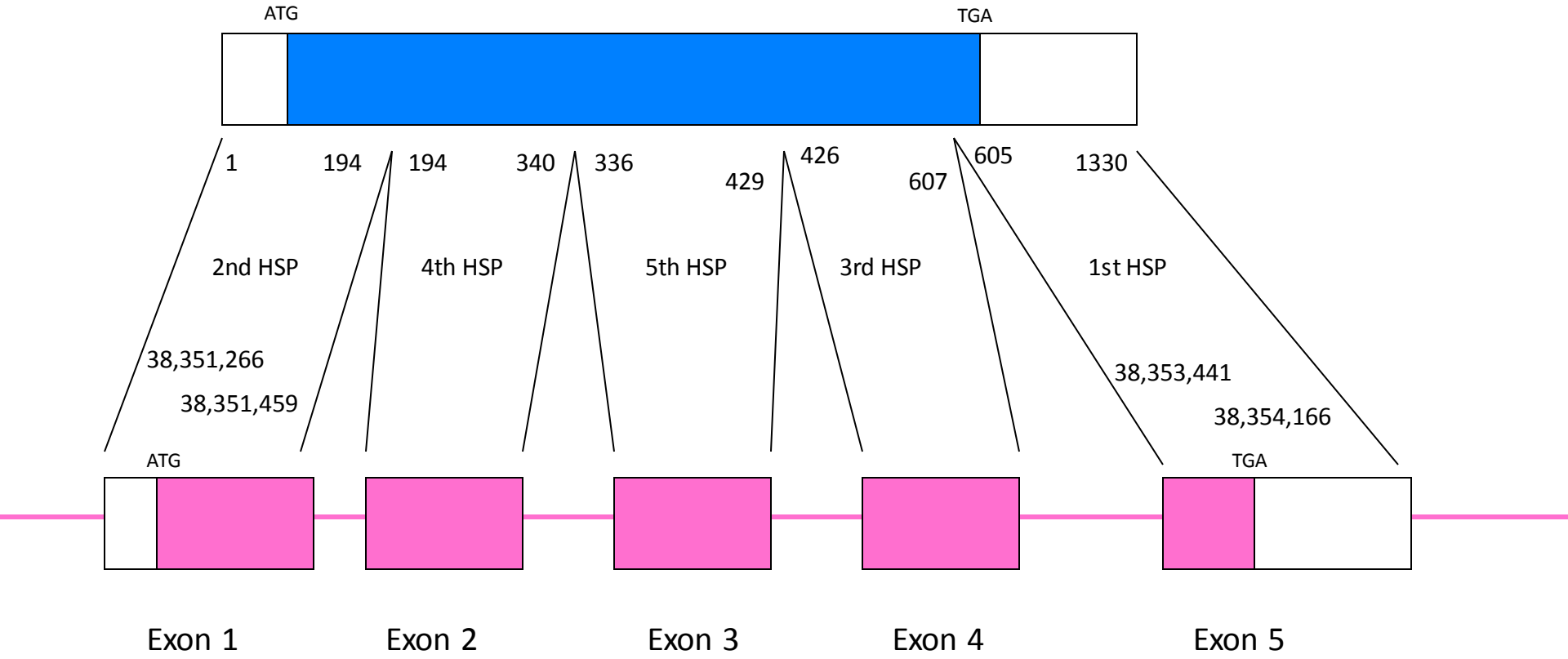
Accession	Description	Max score	Total score	Query coverage	E value	Max ident	Links
Transcripts							
NM_000799.2	Homo sapiens erythropoietin (EPO), mRNA	2412	2412	99%	0.0	99%	U E G M
NM_173587.3	Homo sapiens REST corepressor 2 (RCOR2), mRNA	42.8	42.8	5%	1.1	75%	U E G M
NM_003703.1	Homo sapiens NOP14 nucleolar protein homolog (yeast) (NOP14), mR	41.0	41.0	2%	3.7	83%	U E G M
Genomic sequences[show first]							
NW_001839065.2	Homo sapiens chromosome 7 genomic contig, alterna	1310	2417	99%	0.0	100%	
NT_007933.15	Homo sapiens chromosome 7 genomic contig, GRCh3	1305	2412	99%	0.0	100%	
NT_079595.2	Homo sapiens chromosome 7 genomic contig, alterna	1305	2412	99%	0.0	100%	
NT_011526.7	Homo sapiens chromosome 22 genomic contig, GRCh	42.8	42.8	2%	1.1	90%	
NT_011515.12	Homo sapiens chromosome 21 genomic contig, GRCh	42.8	42.8	1%	1.1	96%	
NT_010783.15	Homo sapiens chromosome 17 genomic contig, GRCh37.p5 Primary A	42.8	42.8	3%	1.1	82%	
NT_010194.17	Homo sapiens chromosome 15 genomic contig, GRCh37.p5 Primary A	42.8	42.8	3%	1.1	79%	
NT_167190.1	Homo sapiens chromosome 11 genomic contig, GRCh37.p5 Primary A	42.8	42.8	5%	1.1	75%	

Hyperlinks down page to each alignment

Example: 1A: One of the genomic hits from this search, marked by green arrow on previous slide

Query:

Human EPO cDNA sequence (GenBank X02157)



Subject:

Human genomic sequence

(human chromosome 7 from Refseq: NT_007933.15)

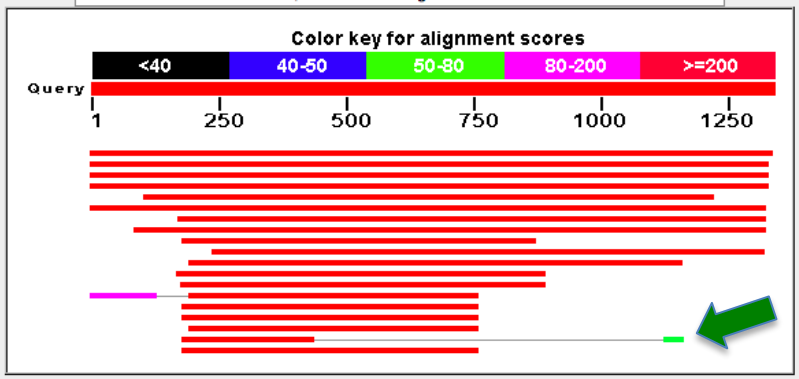
(version 37 of the reference human genome seq.)

Query ID	gi|31229|emb|X02157.1|	**Database Name**	refseq_rna
Description	Human mRNA for fetal erythropoietin	**Description**	NCBI Transcript Reference Sequences
Molecule type	nucleic acid	**Program**	BLASTN 2.2.27+ ▷Citation
Query Length	1342		

Example: 1B: BLASTN: Query is human EPO cDNA. Database is Refseq_RNA (=more species).

Distribution of 21 Blast Hits on the Query Sequence

Mouse over to see the defline, click to show alignments



Descriptions

Legend for links to other resources: U UniGene E GEO G Gene S Structure M Map Viewer PubChem BioAssay

Sequences producing significant alignments:

Accession	Description	Max score	Total score	Query coverage	E value	Max ident	Links
NM_000799.2	Homo sapiens erythropoietin (EPO), mRNA	2470	2470	99%	0.0	99%	U E G M
XM_003812904.1	PREDICTED: Pan paniscus erythropoietin (EPO), mRNA	2407	2407	99%	0.0	99%	G
XM_519268.2	PREDICTED: Pan troglodytes erythropoietin (EPO), mRNA	2401	2401	99%	0.0	99%	G M
XM_003278104.1	PREDICTED: Nomascus leucogenys erythropoietin-like (LOC10060743	2141	2141	99%	0.0	96%	G M
XM_003895802.1	PREDICTED: Papio anubis erythropoietin (EPO), mRNA	1688	1688	83%	0.0	94%	G
XM_003934171.1	PREDICTED: Saimiri boliviensis boliviensis erythropoietin (EPO), mRN	1642	1642	98%	0.0	89%	G
NM_001081825.1	Equus caballus erythropoietin (EPO), mRNA >dbj|AB100030.1| Equus	1099	1099	86%	0.0	84%	U G M
NM_214134.1	Sus scrofa erythropoietin (EPO), mRNA >emb|AJ249745.1| Sus scrofa	1051	1051	92%	0.0	82%	U G M
NM_001042736.1	Macaca mulatta erythropoietin (EPO), mRNA >gb|L10609.1|MACERYT	1007	1007	51%	0.0	93%	U G M
NM_173909.2	Bos taurus erythropoietin (EPO), mRNA >gb|U44762.1|BTU44762 Bos	878	878	80%	0.0	82%	U G M
NM_001006646.1	Canis lupus familiaris erythropoietin (EPO), mRNA >gb|AY572971.1| (	782	782	72%	0.0	82%	U E G M
NM_001009269.1	Felis catus erythropoietin (EPO), mRNA >gb|U00685.1|FDU00685 Feli	776	776	54%	0.0	86%	G
XM_002927297.1	PREDICTED: Ailuropoda melanoleuca erythropoietin-like (LOC1004837	769	769	53%	0.0	86%	G M
XM_002743991.2	PREDICTED: Callithrix jacchus erythropoietin (EPO), mRNA	745	940	52%	0.0	93%	G
XM_003422486.1	PREDICTED: Loxodonta africana erythropoietin-like (LOC100676284),	627	627	43%	1e-176	86%	G M
NM_001024737.1	Ovis aries erythropoietin (EPO), mRNA >emb|Z24681.1| O.aries eryth	579	579	43%	3e-162	85%	U G
XM_003798904.1	PREDICTED: Otolemur garnettii erythropoietin (EPO), mRNA	562	562	42%	3e-157	[illegible]	G
XM_002817776.2	PREDICTED: Pongo abelii erythropoietin (EPO), mRNA	451	530	22%	7e-124	100%	G
XM_003470146.1	PREDICTED: Cavia porcellus erythropoietin-like (LOC100712648), mR	440	440	43%	2e-120	81%	G M

>ref|XM_002817776.2| PREDICTED: Pongo abelii erythropoietin (EPO), mRNA
Length=297

GENE ID: 100459890 EPO | erythropoietin [Pongo abelii]

Sort alignments for this subject sequence b
E value Score Percent identity
Query start position Subject start posit

```
 Score =   451 bits (244),  Expect = 7e-124
 Identities = 254/259 (98%), Gaps = 0/259 (0%)
 Strand=Plus/Plus

Query  182  ATGGGGGTGCACGAATGTCCTGCCTGGCTGTGGCTTCTCCTGTCCCTGCTGTCGCTCCCT  241
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  1    ATGGGGGTGCACGAATGTCCTGCCTGGCTGTGGCTTCTCCTGTCCCTGCTGTCGCTCCCT  60

Query  242  CTGGGCCTCCCAGTCCTGGGCGCCCCACCACGCCTCATCTGTGACAGCCGAGTCCTGGAG  301
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  61   CTGGGCCTCCCAGTCCTGGGCGCCCCACCACGCCTCATCTGTGACAGCCGAGTCCTGGAG  120

Query  302  AGGTACCTCTTGGAGGCCAAGGAGGCCGAGAATATCACGACGGGCTGTGCTGAACACTGC  361
            |||||||||||||||||||||||||||||||| ||||||||||||||||| |||||||||
Sbjct  121  AGGTACCTCTTGGAGGCCAAGGAGGCCGAGAATGTCACGACGGGCTGTGCCGAACACTGC  180

Query  362  AGCTTGAATGAGAATATCACTGTCCCAGACACCAAAGTTAATTTCTATGCCTGGAAGAGG  421
            ||||||| ||||||||||||| |||||||||||||||||||| |||||||||||||||||||
Sbjct  181  AGCTTGAGTGAGAATATCACCGTCCCAGACACCAAAGTTAACTTCTATGCCTGGAAGAGG  240

Query  422  ATGGAGGTCGGGCAGCAGG  440
            |||||||||||||||||||
Sbjct  241  ATGGAGGTCGGGCAGCAGG  259

 Score = 78.7 bits (42),  Expect = 1e-11
 Identities = 42/42 (100%), Gaps = 0/42 (0%)
 Strand=Plus/Plus

Query  1125  CAGGGACAGGATGACCTGGAGAACTTAGGTGGCAAGCTGTGA  1166
             ||||||||||||||||||||||||||||||||||||||||||
Sbjct  256   CAGGGACAGGATGACCTGGAGAACTTAGGTGGCAAGCTGTGA  297
```

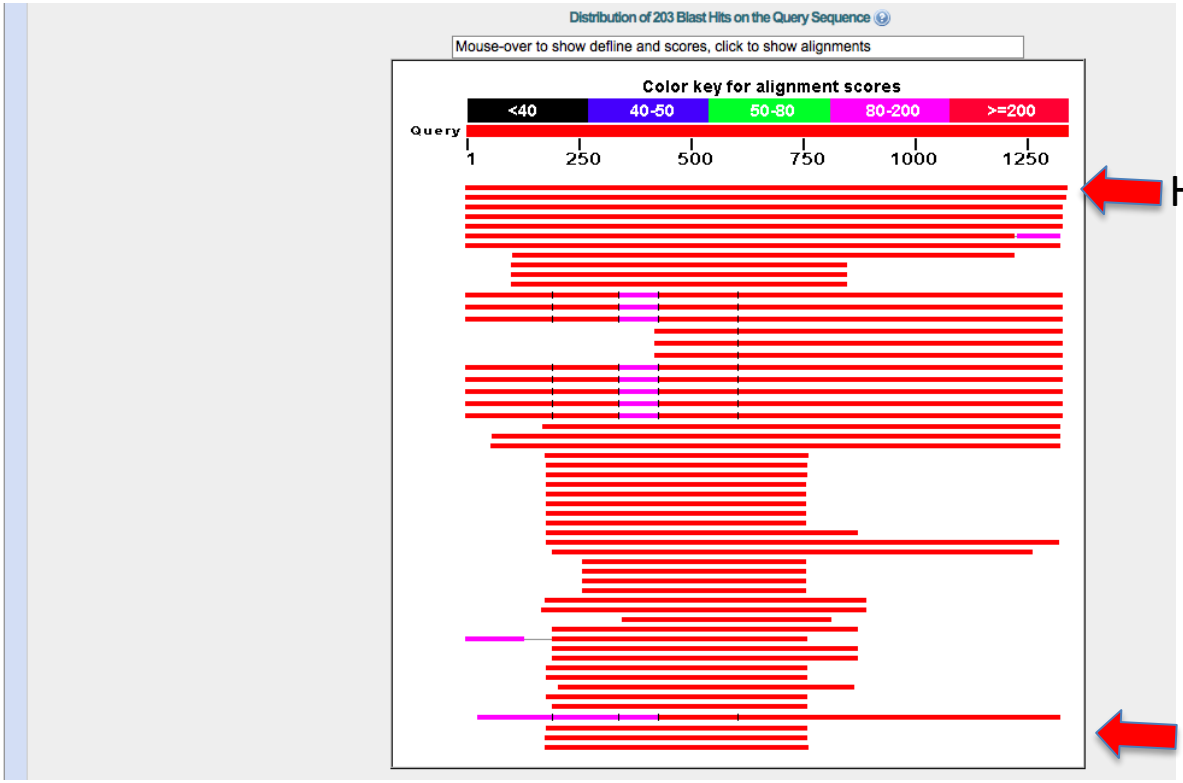
NCBI/ BLAST/ blastn suite/ Formatting Results - 920RK1X8016

Edit and Resubmit | Save Search Strategies | Formatting options | Download

emb|X02157| (1342 letters)

Query ID	gi|31229|emb|X02157.1|	**Database Name**	nr
Description	Human mRNA for fetal erythropoietin	**Description**	Nucleotide collection (nt)
Molecule type	nucleic acid	**Program**	BLASTN 2.2.27+ Citation
Query Length	1342		

Example: 1C: BLASTN: Query is human EPO cDNA. Database is NR (=lots of species).



Human, top hit, E = 0

Eospalax, 50th hit, E = 2.1e-152

M12930.1	Mouse erythropoietin gene, complete cds	237	452	51%	2e-58	83%
AY092019.1	Saguinus oedipus erythropoietin gene, partial cds	233	555	29%	2e-57	

100th

BLAST Search database Nucleotide collection (nr/nt) using Blastn (Optimize for somewhat similar sequences)

☑ Show results in a new window

Algorithm parameters

Note: Parameter values that differ from the default are highlighted in yellow and marked with ♦ sign

General Parameters

Max target sequences	♦ 1000 — Select the maximum number of aligned sequences to display
Short queries	☑ Automatically adjust parameters for short input sequences
Expect threshold	10
Word size	11
Max matches in a query range	0

Scoring Parameters

Match/Mismatch Scores	2,-3
Gap Costs	Existence: 5 Extension: 2

Filters and Masking

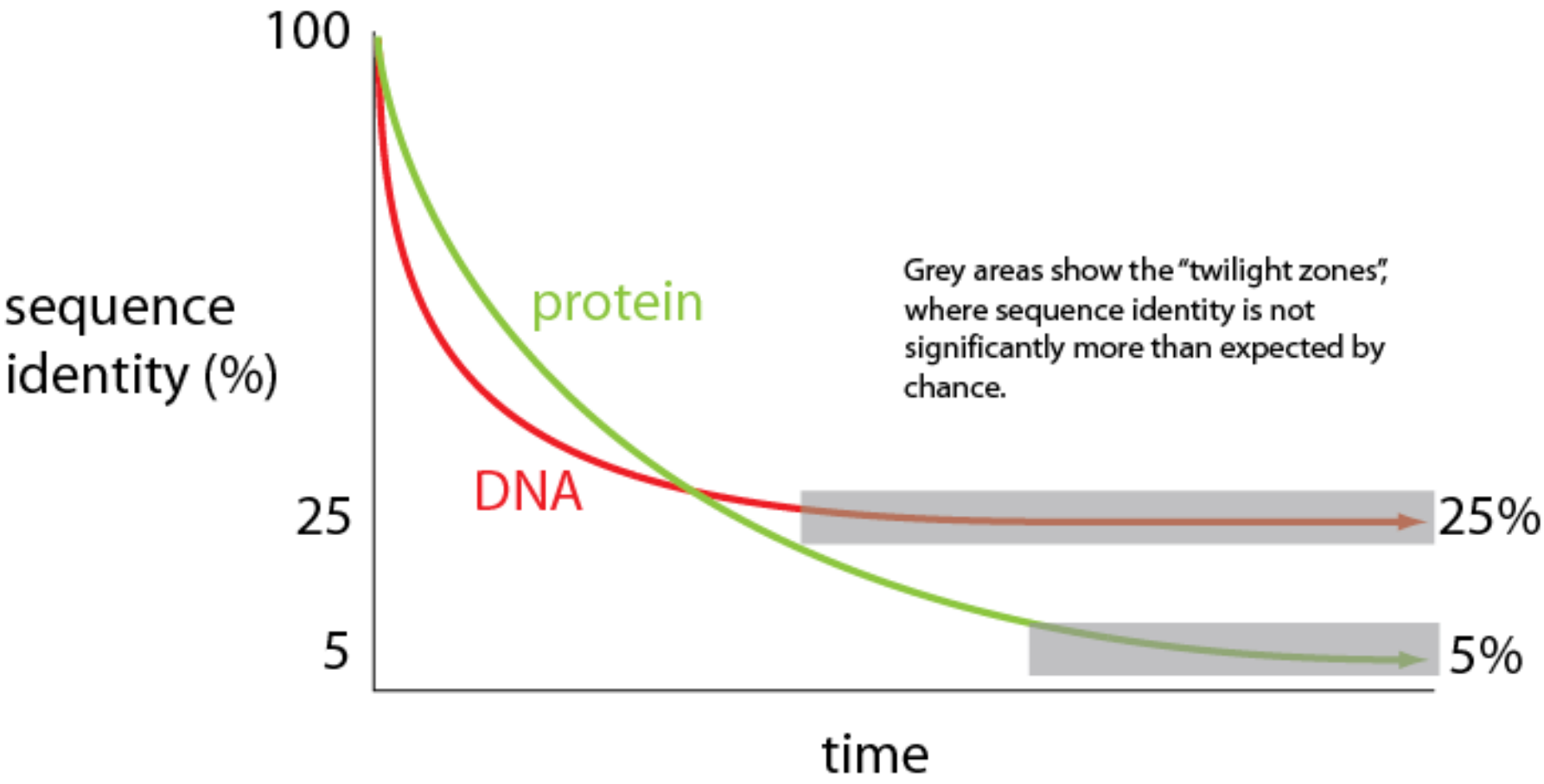
Filter	☑ Low complexity re ☐ Species-specific r
Mask	☑ Mask for lookup ta ☐ Mask lower case l

M12482.1	Mouse erythropoietin gene, complete	241	463	51%	1e-59
M12930.1	Mouse erythropoietin gene, complete cds	237	452	51%	2e-58
AY092019.1	Saguinus oedipus erythropoietin gene, partial cds	233	555	29%	2e-57
Y11971.1	M.musculus mRNA Epo (abnormal Epo allele)	187	187	23%	3e-43
XM_001371448.2	PREDICTED: Monodelphis domestica erythropoietin-like (LOC1000181	167	167	22%	3e-37
EF087949.1	Physeter catodon clone EPO erythropoietin-like gene, partial sequence	158	280	14%	1e-34
DQ465472.1	Pantholops hodgsonii erythropoietin mRNA, partial cds	156	156	10%	5e-34
FJ176349.1	Neophocaena phocaenoides erythropoietin-like (EPO) gene, partial se	149	244	14%	7e-32
AJ278715.1	Cloning vector pAEC-SPE3, partial	147	147	6%	2e-31
JQ002761.1	Tursiops truncatus erythropoietin (EPO) gene, partial	86.0	132	7%	7e-13
AF202312.1	Homo sapiens erythropoietin (EPO) gene, exon 1	55.4	55.4	2%	0.001
AF202306.1	Homo sapiens erythropoietin (EPO) gene, exon 1	55.4	55.4	2%	0.001
XM_002926807.1	PREDICTED: Ailuropoda melanoleuca tubulin beta-3 c	48.2	48.2	3%	0.17
X73471.1	M.musculus 3'flanking region of gene for erythropoietin	48.2	48.2	3%	0.17
L13456.1	Mouse erythropoietin gene	48.2	48.2	3%	0.17
XM_001519767.2	PREDICTED: Ornithorhynchus anatinus frizzled-2-like (LOC100090775	46.4	46.4	3%	0.61
FR845719.1	Streptomyces venezuelae ATCC 10712 complete genome	46.4	46.4	3%	0.61
CR790366.19	Zebrafish DNA sequence from clone DKEY-245M3 in linkage group 5 C	46.4	46.4	1%	0.61
AK123083.1	Homo sapiens cDNA FLJ41088 fis, clone ASTRO2002459	46.4	46.4	1%	0.61
BC110175.1	Bos taurus cDNA clone IMAGE:8068452	46.4	46.4	2%	0.61
BX908798.1	Parachlamydia-related symbiont UWE25, complete genome	46.4	46.4	2%	0.61
CP002399.1	Micromonospora sp. L5, complete genome	44.6	44.6	3%	2.1
CP002162.1	Micromonospora aurantiaca ATCC 27029, complete genome	44.6	44.6	3%	2.1
AC134912.5	Mus musculus BAC clone RP23-162E15 from chromosome 14, complet	44.6	44.6	2%	2.1

100th

107th

109th


100
sequence identity (%)
protein
DNA
Grey areas show the "twilight zones", where sequence identity is not significantly more than expected by chance.
25
25%
5
5%
time

Protein databases for BLAST (BLASTP, BLASTX)

- Nonredundant protein sequences (nr)
- Reference proteins (refseq_protein)
- UniProtKB (Swiss-prot)
- Protein Databank proteins (pdb) ← with known 3D structures
- and others…

4 types of BLAST search: #2, BLASTP

		Query	
		DNA	Protein
Database	DNA	BLASTN megablast	TBLASTN
	Protein	BLASTX	**BLASTP**

BLASTP: protein query vs. protein database.

Typical use: to find hits in <u>annotated</u> protein databases.

Advantages : Much more sensitive than BLASTN.
Disadvantages : It will miss unannotated genes (they're not in protein database).

Example: 2: BLASTP: Query is human EPO protein. Database is NR proteins.

E-values.
Sorted: lowest first.

XP_001371485.2	PREDICTED: erythropoietin-like [Monodelphis domestica]	171	171	89%	3e-50	55%
XP_002817822.2	PREDICTED: erythropoietin [Pongo abelii]	140	140	49%	2e-39	89%
NP_001184210.1	erythropoietin precursor [Xenopus laevis] >[illegible]AI82351.1| erythrop	120	120	85%	7e-31	39%
NP_001233194.1	erythropoietin precursor [Xenopus (Silurana) tropicalis] >gb|ADJ6800	110	110	93%	6e-27	37%
NP_001108599.1	erythropoietin isoform S [Danio rerio] >gb|ABQ41210.1| erythropoiet	106	106	87%	1e-25	37%
NP_001108600.1	erythropoietin isoform L1 precursor [Danio rerio] >gb|ABQ41209.1| e	106	106	92%	1e-25	37%
CAH39855.1	erythropoietin-I [Cyprinus carpio]	106	106	92%	2e-25	37%
NP_001033098.1	erythropoietin isoform L2 precursor [Danio rerio] >sp|Q2XNF5.1|EPO_	106	106	92%	2e-25	37%
ABB83930.1	erythropoietin [Cyprinus carpio]	105	105	87%	7e-25	36%
ADD13992.1	erythropoietin, partial [Cyprinodon variegatus]	101	101	85%	8e-24	38%
Q5IGQ0.1	RecName: Full=Erythropoietin; Flags: Precursor >gb|AAW29029.1| er	99.4	99.4	86%	8e-23	37%
XP_003457688.1	PREDICTED: erythropoietin-like [Oreochromis niloticus]	98.2	98.2	86%	3e-22	36%
Q4T554.1	RecName: Full=Erythropoietin; Flags: Precursor >emb|CAF91978.1| u	97.1	97.1	90%	6e-22	37%
AAR25698.1	erythropoietin [Tetraodon nigroviridis]	97.1	97.1	86%	8e-22	38%
AAQ72466.1	erythropoietin brain specific isoform [Takifugu rubripes]	96.3	96.3	86%	1e-21	37%
Q6JV22.1	RecName: Full=Erythropoietin; Flags: Precursor >gb|AAQ72467.1| er	95.9	95.9	86%	2e-21	37%
XP_001342254.1	PREDICTED: erythropoietin-like [Danio rerio]	94.0	94.0	83%	9e-21	36%
CAH39856.1	erythropoietin-II [Cyprinus carpio]	85.1	85.1	65%	7e-18	38%
ABB89952.1	erythropoietin [Oncorhynchus mykiss]	84.3	84.3	74%	1e-17	34%
AAB29659.1	erythropoietin, Epo {N-terminal} [rats, Wistar, blood, Peptide Partial,	77.0	77.0	25%	1e-15	78%
ABF01021.1	erythropoietin [Pantholops hodgsonii]	73.6	73.6	23%	2e-14	82%
ABD73008.1	erythropoietin, partial [Oryzias melastigma]	75.1	75.1	75%	3e-14	34%
AFH89746.1	erythropoietin, partial [Tursiops truncatus]	61.2	61.2	17%	4e-10	79%
CAA72707.1	erythropoietin [Mus musculus]	53.9	53.9	13%	2e-07	92%
NP_001001784.1	thrombopoietin p[illegible]or [Gallus gallus] >gb|AAT45554.1| thrombop	43.5	43.5	78%	0.011	24%
P42705.1	RecName: Full=Thrombopoietin; AltName: Full=C-MPL ligand; Short=	42.4	42.4	42%	0.045	31%
XP_003209225.1	PREDICTED: thrombopoietin-like [Meleagris gallopavo]	39.7	39.7	78%	0.23	23%

4 types of BLAST search: #3, BLASTX

		Query	
		DNA	Protein
Database	DNA	BLASTN megablast	TBLASTN
	Protein	**BLASTX**	BLASTP

BLASTX: DNA query vs. protein database.

Typical use: What does this piece of DNA code for? e.g. an EST.

Advantages : Like BLASTP, but the Query doesn't need to be annotated.
Disadvantages : It will miss unannotated genes (they're not in protein database).

6 reading frames:

6 ways that the same DNA sequence could potentially encode a protein

... S H L V E A L Y L V C G E R G F F... frame +1
...L T P G G S S L P S V R G T R L L ... frame +2
... H T W W K L S T * C A G N E A S ... frame +3
1 tcacacctggtggaagctctctacctagtgtgcggggaacgaggcttcttc 51

51 gaagaagcctcgttccccgcacactaggtagagagcttccaccaggtgtga 1
... E E A S F P A H * V E S F H Q V * ... frame -1
... K K P R S P H T R * R A S T R C ... frame -2
... R S L V P R T L G R E L P P G V ... frame -3

Bothrops alternatus (common pit viper)

What does the EST with accession number GW576306 code for?

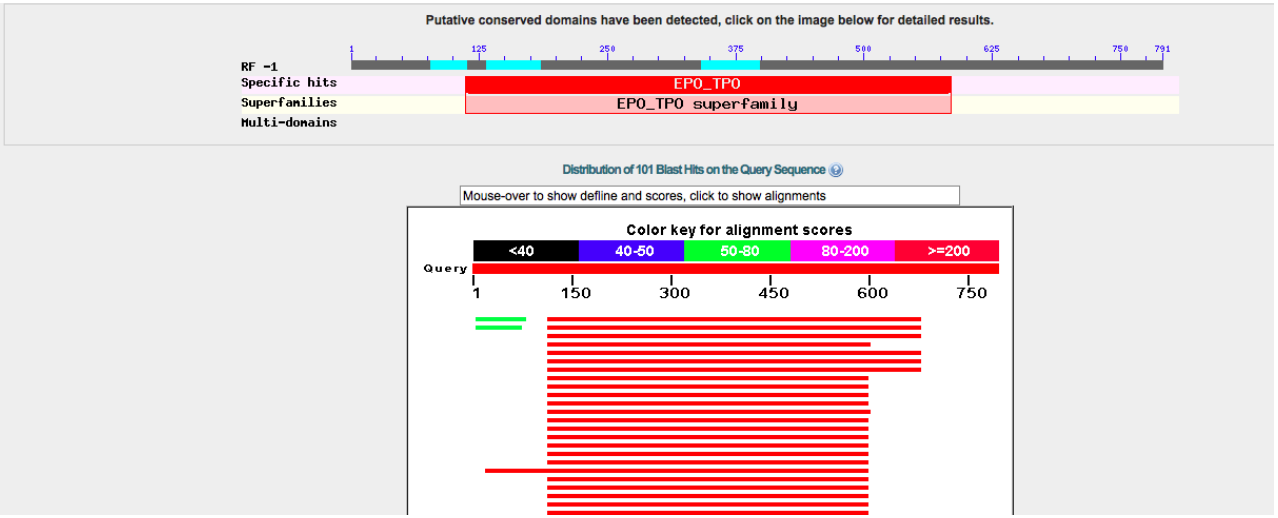
Or GW576313 ?

Or GW576315 ?

An EST (expressed sequence tag) is a single sequencing read from a random clone in a cDNA library = a randomly sampled mRNA.

Query ID	gi|289588961|gb|GW576306.1|	**Database Name**	nr
Description	BACCGV3035B12.b Bothrops alternatus venom gland Bothrops alternatus cDNA clone BACCGV3035B12.	**Description**	All non-redundant GenBank CDS translations+PDB+SwissProt+P environmental samples from WGS projects
Molecule type	rna	**Program**	BLASTX 2.2.27+ ▷Citation
Query Length	791		

Example: 3: BLASTX: Query is snake EST EPO GW576306. Database is NR proteins.



Sequences producing significant alignments:

Accession	Description	Max score	Total score	Query coverage	E value
NP_031968.1	erythropoietin precursor [Mus musculus] >sp|P07321.1|EPO_MOUSE RecName:	337	337	70%	5e-114
AAI44888.1	Epo protein [Mus musculus]	330	330	70%	2e-111
NP_058697.1	erythropoietin precursor [Rattus norvegicus] >sp|P29676.1|EPO_RAT RecName:	320	320	70%	9e-108
AAA41126.1	erythropoietin, partial [Rattus norvegicus]	315	315	61%	1e-105
XP_003510685.1	PREDICTED: erythropoietin-like [Cricetulus griseus] >gb|EGW06331.1| Erythrop	307	307	70%	2e-102
ABY56032.1	erythropoietin [Eospalax baileyi]	306	306	70%	3e-102
Q0Z956.1	RecName: Full=Erythropoietin; Flags: Precursor >gb|ABG47336.1| erythropoieti	303	303	70%	8e-101
Q6H8T2.1	RecName: Full=Erythropoietin; Flags: Precursor >emb|CAG29397.1| erythropoi	290	290	61%	1e-95
Q6H8S9.1	RecName: Full=Erythropoietin; Flags: Precursor >sp|Q6H8T0.1|EPO_SPAJD Rec	288	288	61%	8e-95
ABF01021.1	erythropoietin [Pantholops hodgsonii]	71.2	71.2	17%	3e-13
CAA72707.1	erythropoietin [Mus musculus]	62.8	62.8	19%	3e-10
AFH89746.1	erythropoietin, partial [Tursiops truncatus]	62.0	62.0	12%	6e-10
AAL59385.1	IntI [Citrobacter freundii]	60.1	60.1	9%	3e-08
ACE77052.1	LacZ alpha peptide [Cloning vector pSMARTGC Blue]	57.0	95.9	27%	2e-07
CCM44347.1	CcdB toxin fusion [Cloning vector pSAWloxP-K]	57.4	57.4	8%	4e-07

4 types of BLAST search: #4, TBLASTN

		Query	
		DNA	Protein
Database	DNA	BLASTN megablast	**TBLASTN**
	Protein	BLASTX	BLASTP

TBLASTN: Searches a protein query vs. DNA database.

Typical use: Can I find any new homologs of my gene?

Advantages：Like BLASTP, but the database entry doesn't need to be annotated.
Disadvantages：Your query needs to be a protein.

4 types of BLAST search: #5, TBLASTX

		Query	
		DNA	Protein
Database	DNA	BLASTN **TBLASTX**	TBLASTN
	Protein	BLASTX	BLASTP

TBLASTX: DNA query vs. DNA database, 6-frame translations.

(Comparing all proteins that could possibly be encoded by the Query,
to all proteins that could possibly be encoded by each sequence in the database.)

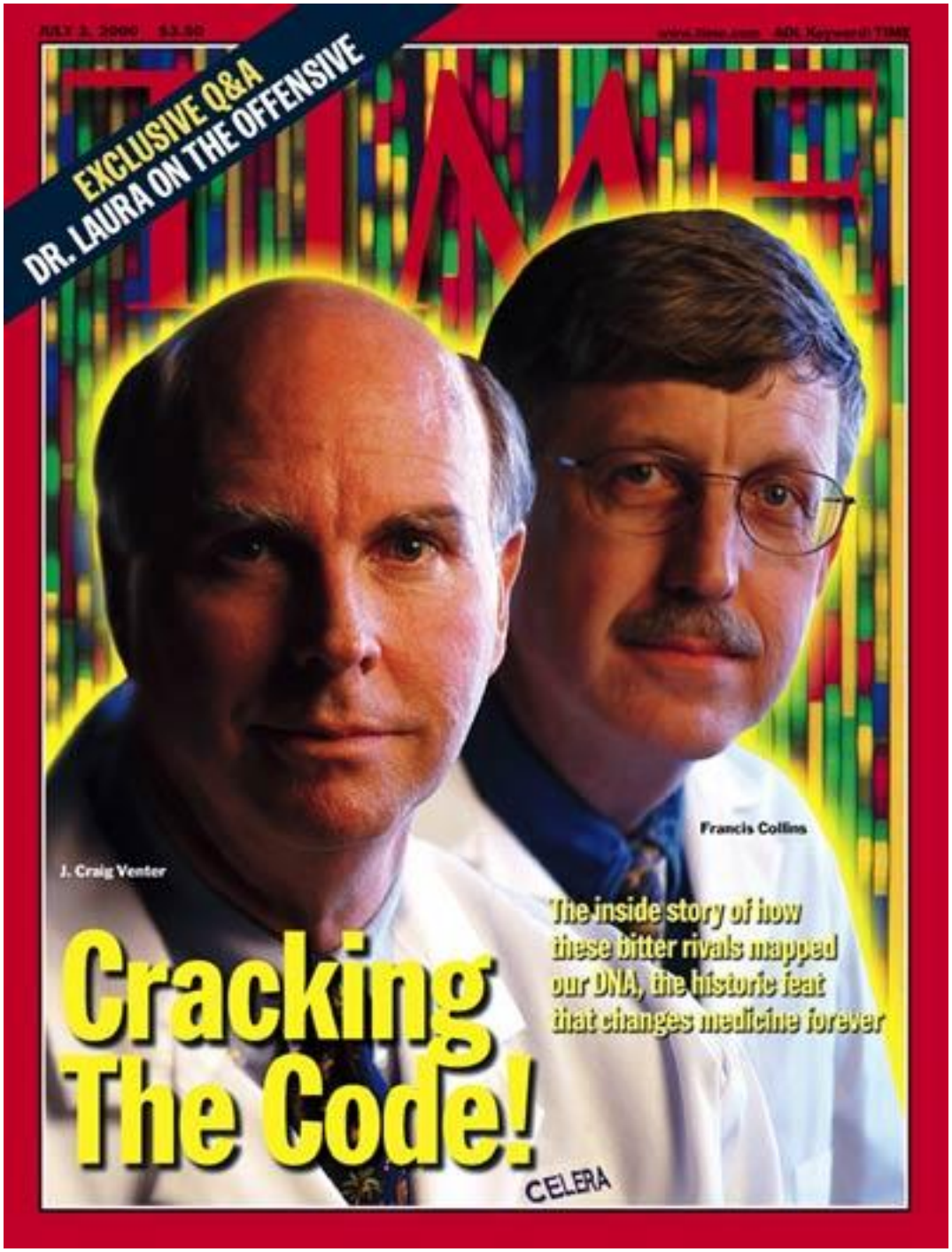
Typical use: I'm desperate!

Advantages: Query and database can both be unannotated.
Disadvantages: Dreadfully slow. TBLASTX searches against most databases are banned on the NCBI server. Results can be hard to interpret.

J. Craig Venter

Celera Genomics

2001



Francis Collins

Intl. Human Genome Sequencing Consortium

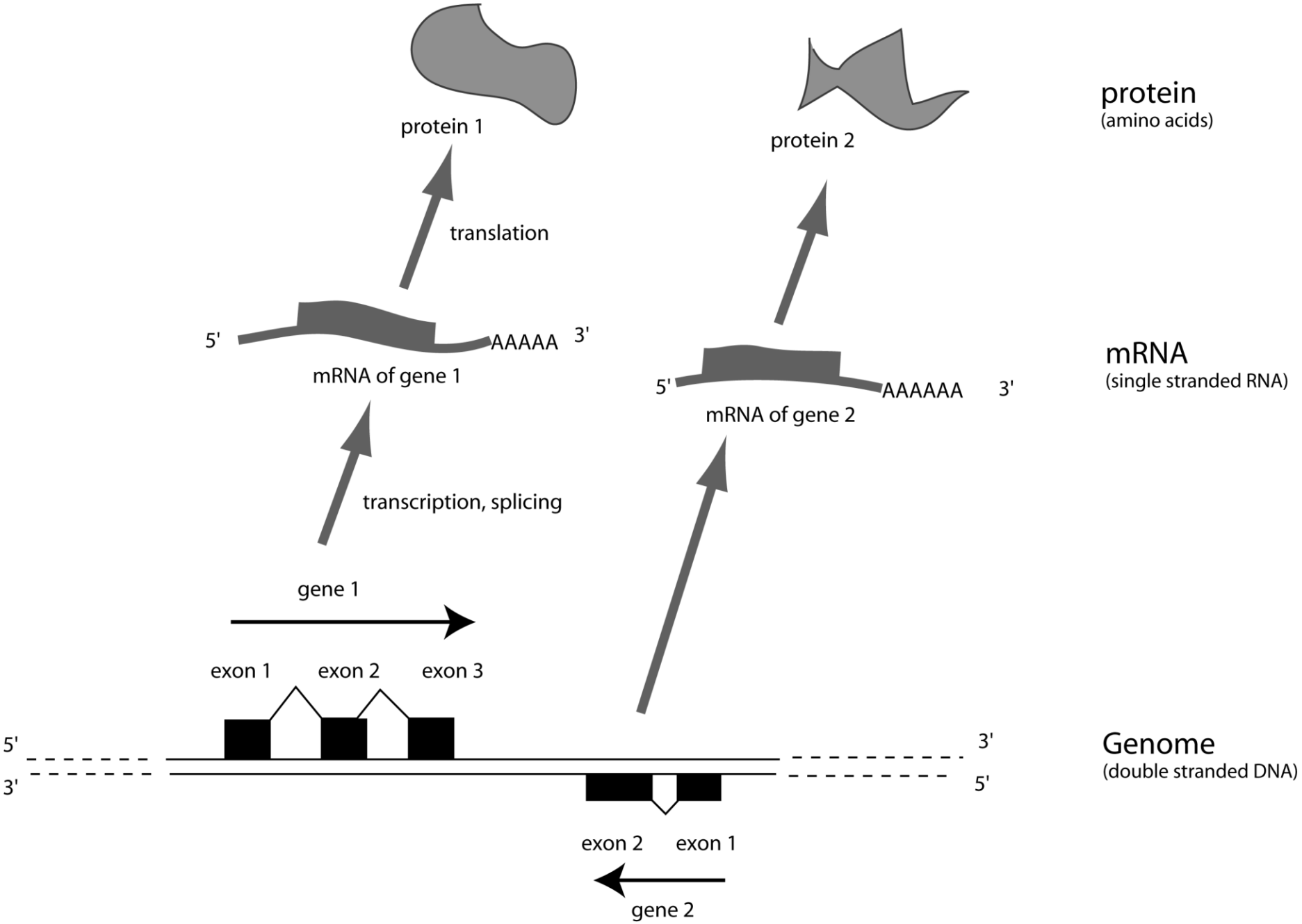
Complementary DNA Sequencing: Expressed Sequence Tags and Human Genome Project

MARK D. ADAMS, JENNY M. KELLEY, JEANNINE D. GOCAYNE, MARK DUBNICK, MIHAEL H. POLYMEROPOULOS, HONG XIAO, CARL R. MERRIL, ANDREW WU, BJORN OLDE, RUBEN F. MORENO, ANTHONY R. KERLAVAGE, W. RICHARD MCCOMBIE, J. CRAIG VENTER*

Automated partial DNA sequencing was conducted on more than 600 randomly selected human brain complementary DNA (cDNA) clones to generate expressed sequence tags (ESTs). ESTs have applications in the discovery of new human genes, mapping of the human genome, and identification of coding regions in genomic sequences. Of the sequences generated, 337 represent new genes, including 48 with significant similarity to genes from other organisms, such as a yeast RNA polymerase II subunit; *Drosophila* kinesin, *Notch*, and *Enhancer of split*; and a murine tyrosine kinase receptor. Forty-six ESTs were mapped to chromosomes after amplification by the polymerase chain reaction. This fast approach to cDNA characterization will facilitate the tagging of most human genes in a few years at a fraction of the cost of complete genomic sequencing, provide new genetic markers, and serve as a resource in diverse biological research fields.

EST category	Hippocampus	
Database match—human		
Mitochondrial genes	48	(12.8)
Repeated sequences	39	(10.4)
Ribosomal RNA	10	(2.7)
Other nuclear genes	32	(8.6)
Database match—other	32	(8.6)
No database match	160	(42.8)
Polyadenylate insert	53	(14.1)
No insert	1	(0.3)

1991


protein 1
protein 2
protein
(amino acids)
translation
5'
AAAAA
3'
mRNA of gene 1
5'
AAAAAA
3'
mRNA of gene 2
mRNA
(single stranded RNA)
transcription, splicing
gene 1
exon 1
exon 2
exon 3
5'
3'
3'
5'
Genome
(double stranded DNA)
exon 2
exon 1
gene 2

RESEARCH

Generation and Analysis of 280,000 Human Expressed Sequence Tags

LaDeana Hillier,[1,4] Greg Lennon,[2] Michael Becker,[1] M. Fatima Bonaldo,[3] Brandi Chiapelli,[1] Stephanie Chissoe,[1] Nicole Dietrich,[1] Treasa DuBuque,[1] Anthony Favello,[1] Warren Gish,[1] Maria Hawkins,[1] Monica Hultman,[1] Tamara Kucaba,[1] Michelle Lacy,[1] Maithao Le,[1] Nha Le,[1] Elaine Mardis,[1] Bradley Moore,[1] Matthew Morris,[1] Jeremy Parsons,[1] Christa Prange,[3] Lisa Rifkin,[1] Theresa Rohlfing,[1] Kurt Schellenberg,[1] M. Bento Soares,[2] Fang Tan,[1] Jean Thierry-Meg,[1] Evanne Trevaskis,[1] Karen Underwood,[1] Patricia Wohldman,[1] Robert Waterston,[1] Richard Wilson,[1] and Marco Marra[1]

[1]Genome Sequencing Center, Washington University School of Medicine, St. Louis, Missouri 63108; [2]Human Genome Center, Lawrence Livermore National Laboratories, Livermore, California 94550; [3]Department of Psychiatry, College of Physicians and Surgeons of Columbia University, and the New York State Psychiatric Institute, New York, New York 10032

We report the generation of 319,311 single-pass sequencing reactions (known as expressed sequence tags, or ESTs) obtained from the 5′ and 3′ ends of 194,031 human cDNA clones. Our goal has been to obtain tag sequences from many different genes and to deposit these in the publicly accessible Data Base for Expressed Sequence Tags. Highly efficient automatic screening of the data allows deposition of the annotated sequences

1996