Searching info from public (genetic) databases

Päivi Onkamo 9.11.2006 BfMS, Genetics module

Main "sequence" databases

• DNA:

- EMBL / Genbank / DDBJ – RefSeq
- Protein:
- UniProt / SWISS-PROT – RefSeq
- ReiSe
- Genomes:
 Ensembl
 - UCSC Genome Browser

Nucleotide sequence databases

- EMBL / Genbank / DDBJ – Primary sequence databases
- RefSeq
- NRDB
- UniGene







	niProt e	ntry, sequence
File Edit View Parvarites 1	oola Help	
🔇 Beck + 🔘 · 💌 🕻	🖞 🏠 🔎 Search 👷 Foverites 🐼 👔	3• 🖕 🖂 🛄 🖏
Address 🚵 http://www.abi.unipro	t.org/uniprot-ans/uniProtView.do/protein3d=0R.CC1_P	LMANipager.offset=0 👻 🛃 Ge Laiss *
	STROND	620 620 1
		FeatureTableViewer Feature aligner
	Sequence information	
	Length	633.44
	morecular weight	2010234512451245128522 [This is a checksum on the converse]
	HPEIRLEBVY SCHEDDETHC ABNILK	ADTY REWRAAKAGE RTIDVVLOLE 50
	KEEQIRGVDI GEDGGAFVEV 19686A	GGAG RODVEVILVT SEPHEPEESE 100
	SCSRPRRVRB FCPD0LVRAA AEKRVD	RVKI VCSQFYSKDS FFGLSFVBFH 150
	SPFDEDEAEA PSORVTVTEL GOFEVE	EEDE GANGLEDGAL FFGRINETSP 200
	VTASDPAGPS YAAATLOASS AASSAS	EVER AIGSTSKEDE SERGRERIDI 250
	NQEEKKTPSK PPAQLSPSVP ERPELP	AFTE TFATAFVFAR AQGAVTGEFE 300
	GEGTEPRRPR AGPEELGRIL OGVVVV	ISGF GNPFRSELRD KALELGARVR 350
	FDWTRDSTHL ICAFANTPRY SQWIGL	GGRI VEKEVVLDIE SMERELPIER 400
	VLNAGPOSSS EEDEASHSOG SODEAP	KIPO KOPOTKTKPT QAAGPSSPOK 450
	PPTPEETKAA SPVLORDIDI EGVOSE	SQDN GAEDSGDTED ELERVARORE 500
	HRLPPGQEEN GEDPVAGSTD ENTDEE	ENGE PPDLPVPELP DFFOGRNFFI 550
	VGEFFGDERR RLIRYVTAFN GELEDY	MEDE VORVITAGEN DESERAIND 600
	NPSLAFVRPR WIVSCNEROK LLPHQL	VGVV PQA 633
	Pop-tjp Fasta Wex	(IniParc UniPartseo UniPartse) UniPartse











oache web-server: bester	Ensembl release 41: Homo sapiens 0			
ensembl Hur	nan MapView	Search efflument AvyBac;		
nbl release 41 - Oct 2006		Help		
omosome 6	Known Genes SNPs Chromosome 6 Genes	Chromosome 6		
New Chromosomo 6 New Chr 6 Systemy Rey your data ords onto this hromosome Ensembl to an a BLAST search ach Ensembl database ata mining (BioMart) siphay your own data sport data ownboad data		Length: 170.896,992 bos Novem Protein-coding Geness: 1,13 Novem Protein-coding Geness: 45 Paradaka Geness: 16 netWake comixic 16 NoVed Resem: 16 NoVed Resem: 26 Mit Ma Comess: 26 Strict: 26 Strict: 72,014		
elp & Documentation	10.3 421 423.1	For more information on gene statistics, see the <u>MapView helo</u> page		
ome Remap Tew previous release of age in Archive! Table Archive! link for this	422.01 422.01 422.02	Chromosome 6 Y		
age		Jump to ContigView		
sanger 👔	Click on the image above to zoom into that point	Choose two features from this chromosome as anchor points and display the region between them. Both features must be mapped to the surrent Ensembla assembly. If you select 'Nove' for the second feature, the display will be based around the first feature.		
ow in osembl Drol		Please note that if you select widely spaced features there may be an significant delay while Ensembl builds the DNA display.		

pache web-server: bester		Ensembl rele	ase 41: Homo sapiens F			
Ensembl Hui	man	ContigV	îew	Search #!	Human Asystang .	778, AL355340.17.1.11244
nbi release 41 - Oct 2006						Help
omosome 6 80,265 - 42,680,266	🗆 Ch	romosome	6			
iew of Chromosome 6 raphical view cavitical overview		Chr. 6	E228		q21	g27
iew alignment with iew alongside	Ξov	erview			1	
iew Syntenic regions iew region at UCSC		Chr. 6 band	42.20 Mb 42.40	p21.1 Mb 42.60 Mb	42.80 Mb	43.00 Mb
rowser lew region in UCSC rowser		Markers	D65400 D651463		D651227E D65199 D651872	0 D65322E D65322E D65322E D651374
ort data sport information about egion sport sequence as FASTA sport EMBL file		Ensembl Genes	VTEN VS8080 VPecudopine VSTC25_HUMAN VGUCA1A VGUCA18 VMRPS 10 VTRENT 1	NOVEL	⁴ RDS ¹ KIA40240 ⁴ 1BCC	INP.001008739.1 INP.001008739.1 INPLCIA-POPERS INPLCIA-DOTEND.00 I
xport SNP info in region xport Vega info in region		ncRNA Genes		LENS G00000210840	⁴ ENS-G00000206648 0210848	HEN HEN
Ensembl to un a BLAST search earch Ensembl database eta mining (BioMart) isplay your own data		Vega Genes	¹ 4RP11-7K24 1 ⁴ GUCA1A ¹ 4RP1-7K24 2 ⁴ RP1-13508.5 ¹ 4RP1-13508.6 ¹ GUCA1B ¹ GUCA1B ¹ MRPS 10	98P11286G - UB1	8.1 ⁴ RDS ⁴ TBCC ⁴ KU ⁴ RP3.382M17.2 ⁴ RP11-3	440240 4893-4758162 4893-4758163 4405A 01182/4893-4758163 4 4893-4758163 4 4893-475816 4893-475816 4993-475816 499285 499285
xport data			Novel Protein coding	F	utative Processed transcript	HP

che web-server: bester	C Ensembl release 41: Homo sapiens F	
emap 📕	Features Comparative DAS Sources Repeats I	Decorations ▼ Export ▼ Image size ▼ Help •
w previous release of le in Archivel ble Archive! link for this	Jump to region 6 42530265 42680266 R	afresh Band: Rafresh
8	<< 5MB < 2MB < 1MB < Window + Z00m	Window> 1MB> 2MB> 6MB>>
sanger 👔 🖻	Chr. 6 42.59 Mb 42.61 M Length Porward shared	іь 42/63 Мь 42/65 Мь 42/67 Мь 100/00 Кь
mpanzee	🖶 EMBL mRNAs	
PanTro 2.1	E Unigene	
CO S	UniProtK8	
	Genscan GENSCAN000004448 > Abinitio Genscan trans	GENS CANGOCOCO444 50 > Ab initio Genscan trans
	EST trans.	ENSESTT00000040631 >
Ensembl		Q8IWV8-4 > Ensembl Known Protein Coding
		UBR2_HUMAN > Ensembl Known Protein Coding
_	Ensemblitrares	080WV8.2 > Havana Known Protein coding
		UBR2_HUMAN > Havana Known Protein coding
		OSTWV84 > Havana Known Protein coding
		Q8WW8-2 > Ensembli Known Protein Coding
		UBR2.001 > Known Ptotein coding
	🖻 Vega trans.	UBR2-003 > Known Protein coding
		UBR2-002 >





pache web-server: beste	10 00	me pag	o . a. a. a. a	• • • •		
xport data ownload data er Links	Peptide st	ats Charge: Isoelectria Molecular Number o	-10.5 point: 5.4004 weight: 63,626.65 fresidues: 551			
elp & Documentation	🖃 Protei	n DAS Report				
itemap itemap age in Archivel table Archivel link for this age	⊟ DAS So	ArrayG	gress (Gene Expression Data and (CBS Protein function and and (CBS Protein function and and (CBS Protein softing protein col (CBS Protein softing protein col (CBS Protein softing protein and) (Stati Colora of prot and) (Stati Colora d'omain assig and) (Stati Colora d'omain assig)	base) structure predi idification site tions) committee) n sequences t sins into hiera niments to prof	ctions) predictions) o protein structures. Bas rohical clusters) lein sequences)	ed on Compare, MSD.)
Animated tutorials	Doma No domains Other No other feat Variat	Update Maxage ins on ENSEST on this peptide features on EN ures on this peptide ions on ENSES	500745 IP00000040631 NSESTP0000004063	1		
			5120000040031			
	Residue	SNP ID	SNP type	Alleles	Ambiguity code	Alternate residues
	Residue 15	SNP ID rs3749897	SNP type Synonymous	Alleles C/T	Ambiguity code	Alternate residues
	Residue 15 154	SNP ID rs3749897 rs9381199	SNP type Synonymous Non-synonymous	Alleles C/T C/T	Ambiguity code Y Y	Alternate residues T, I
	Residue 15 154 172	SNP ID re3749897 re9381199 re6905054	SNP type Synonymous Non-synonymous Non-synonymous	Alleles C/T C/T A/T	Ambiguity code Y Y W	Alternate residues T, I D, E
	Residue 16 154 172 184	SNP ID <u>re3749897</u> <u>re9381199</u> <u>re5905054</u> re16895863	SNP type Synonymous Non-synonymous Non-synonymous Synonymous	Alleles C/T C/T A/T A/T	Ambiguity code Y Y W W	Alternate residues , T, I D, E
	Residue 15 154 172 184 239	SNP ID rs3749897 rs9381199 rs5905054 rs18895863 rs17855250	SNP type Synonymous Non-synonymous Non-synonymous Synonymous Synonymous	Alleles C/T C/T A/T A/T T/C	Ambiguity code Y Y W W Y	Alternate residues - T, I D, E -



Gene Ontology (GO)

- A controlled vocabulary of gene product roles in cells and the role associations
- The roles can be applied to all organisms
- Three main hierarchies: biological process, cellular component and molecular function include currently about 19,000 classes (=roles)

-usually only a small portion of these classes is in use with one organism









Sequence databases

- Many different types of databases, consisting of tens of millions of sequences already exist:
 - Nucleotide data banks EMBL (in Europe, web address www.embl.org), GenBank (USA), DDBJ (Japan)
 - genome builds
 - ENSEMBL (www.ensembl.org)
 - UCSC (www.genome.ucsc.edu/)
 - Efficient means of finding homologous sequences from these???

Searching databases

• Different types of queries:

- Find DNA sequences that are *homologous* to your sequence (the same evolutionary origin)
- Find gene families inside a species
- Align an mRNA sequence to a genome assembly (what gene is my mRNA coding?)
- Design primers for PCR
- Search a database for a specific motif

What is BLAST?

- BLAST
 - Basic Local Alignment Search Tool
 - A set of several computer programs (blastn, blastx...)
 - Optimized for finding local alignments between two sequence

Ref: Altschul, S.F., Gish, W., Miller, W., Myers, E.W. & Lipman, D.J. (1990) "Basic local alignment search tool." J. Mol. Biol. 215:403-410.

User site: http://www.ncbi.nlm.nih.gov/BLAST/

How BLAST works?

- All BLAST programs work more or less similarly, and computationally a BLAST search consists of three phases.
 - Seeding
 - Extension
 - Evaluation

- How BLAST works
- The query sequence is divided into subsequences of a given length.
 – word size 3 for proteins, 11 for nucleotides.
- These are used to look for exact or nearly exact matches in the sequence database. – Fast to do = computationally inexpensive.
- When a match is found, it is extended further.



Threshold in seeding

• Word hit

- Hit is two matching, identical words, one in database, another in the query sequence (used in blastn)
- Hit is a neighborhood (used in protein-related searches)
 The neighborhood of a word contains the word itself and all other words whose score is at least as big as T (threshold) when compared via the scoring matrix.
 - For example, if T=13, word=PQG, matrix=Blosum62, only words getting a score over 13 will be scored as hits: PQG-PEG (15) is accepted, but PQG-PQA (12) is not.
 - Setting T higher will remove more word hits, making BLAST run faster, but increases the chance of missing an interesting alignment.
 - Setting W (wordsize) higher will decrease sensitivity (chance of finding the alignment), but increase speed of the search.

Extension

- Word hits found during seeding are extented from their ends.
- Extension is stopped when the alignment score drops, or in newer implementations, when the alignment score has dropped enough (drop-off score) compared to its previous maximum.

Alignment Word hit Extension

Extension, example

KRISTIAN	gap=0, X=1
-RISTISANA	BLOSUM62
0544541 <mark>2</mark> 00	<- score
00000002	<- drop off score

• Extension terminates when drop off score fall below X.

Evaluation

- When the extension stage has produced the alignments, they will be evaluated to determine whether they are statistically significant.
- Statistical significance is determined using Karlin-Altschul statistics (the E-score)
 - Some simplifying assumptions are made (such as sequences inifinitely long, no gaps), but in practice, K-A statistics is nicely generalizable.

E-score

- The lower the score, the more significant the alignment
- The E-score is dependent on both the database size and the scoring system (substitution matrix, gap penalties).
 - If these are changed, the E-score for a specific alignment will also change.

Karlin-Altschul statistics

- E = Kmn*e*^{-λS}
 - E, the number of alignments expected by chance
 - K, minor constant
 - m, the lenght of query sequence
 - n, the lenght of the database
 - *e*, about 2,71
 - $\lambda S,$ normalized alignment score (S is the score, lambda is the normalization factor)

Karlin-Altschul, example

- What is the chance that when two equally long (250) amino acid sequences are aligned using PAM250 matrix, the alignment score is 75?
- E = Kmn $e^{\lambda S}$ = 0,1*250*250*2,71-(0,229*75) = 0,000217

BLAST programs

Query	Database	Program	Typical uses
DNA	DNA	blastn	Annotation, mapping oligo- nucleotides to genome
protein	protein	blastp	Identifying common regions between proteins
translated DNA	protein	blastx	Finding protein-coding genes in genomic DNA
protein	translated DNA	tblastn	Identifying transcripts, possibl from multiple organisms
translated DNA	translated DNA	tblastx	Cross-species gene prediction searching for genes not yet in protein databases

- A homologous gene or a sequence is derived from a common ancestor to all (or most of) the daughter species, and thus can be aligned accordingly:
 Species 1: A--GTACCTGA
 Species 2: AC-GTACCTTA
 Species 3: ACGGTA-CTTA
 Species 4: -CGGTA-CTTA
 Each homologous position can be thought of as an ethicity to (a surjeticle) observed mean an loss mean and any sector.
 - attribute (i.e., variable) showing more or less random changes which have taken place *since speciation*
- Thus, each variable position gives, in principle, independent information of speciation order – used for example in *molecular systematics*
- Or, vice versa, high-enough sequence similarity between an unknown and previously known genes implicates a possible evolutionary relationship – which can be utilised in deriving the identity and function of the newly sequenced, unknown protein (or DNA)



