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
- Genomes:
  - Ensembl
  - UCSC Genome Browser

Nucleotide sequence databases

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

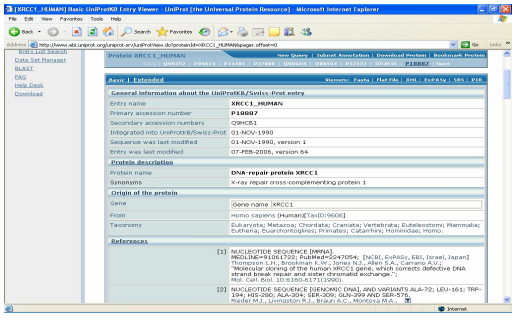
Modified from a Finnish slide by Eija Korpelainen

UniProt

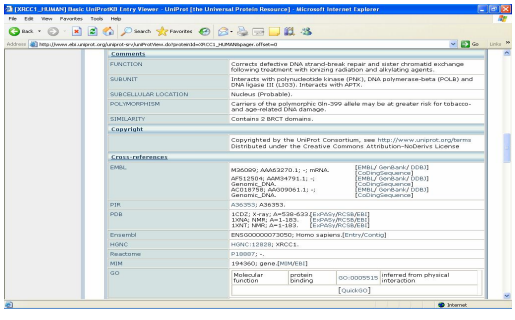
- Good quality annotations
  - Curated, manually edited
- Minimal redundancy
- Extensive cross-links to other databases

Modified from a Finnish slide by Eija Korpelainen

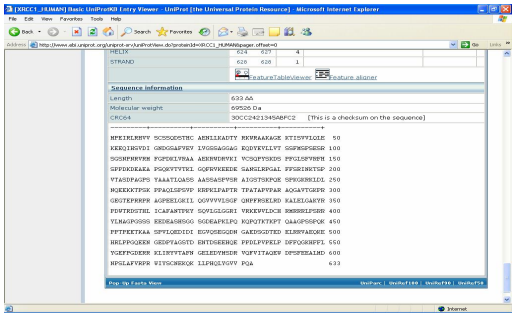
UniProt entry, description



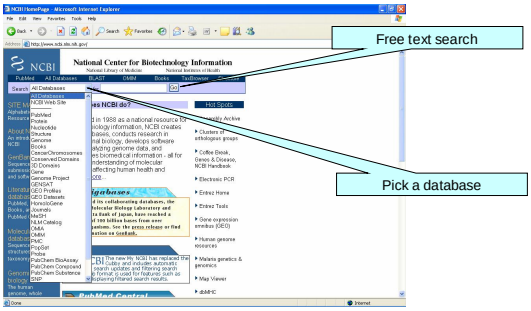
UniProt entry, cross-refs



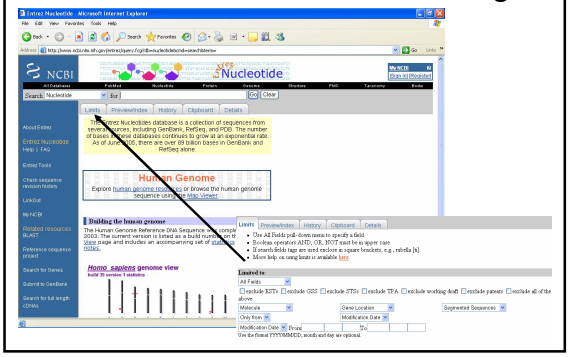
UniProt entry, sequence



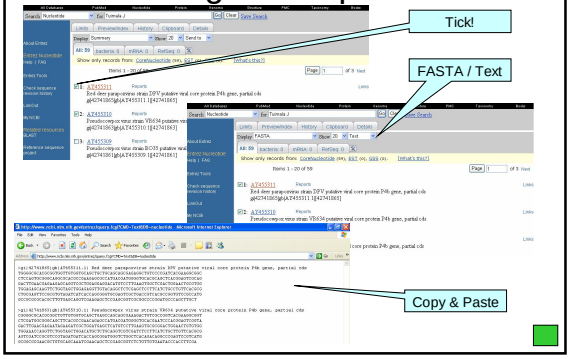
NCBI Entrez – main page



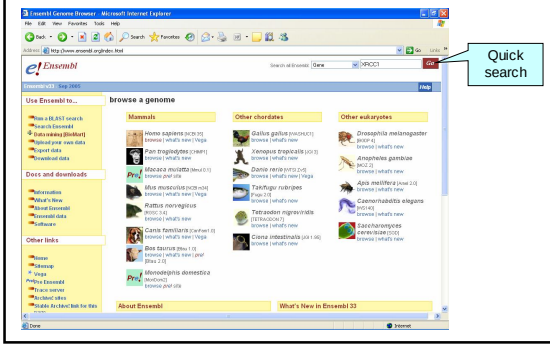
Entrez – use limits for filtering

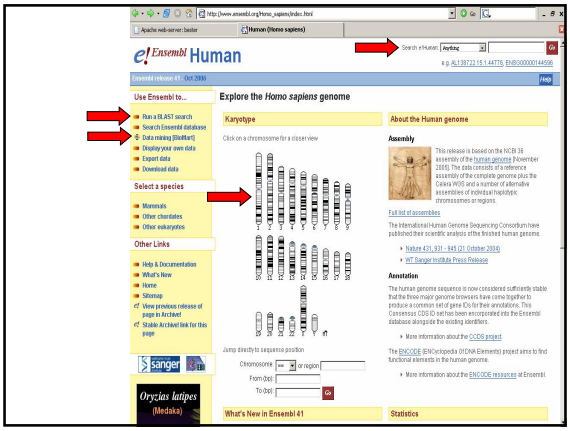


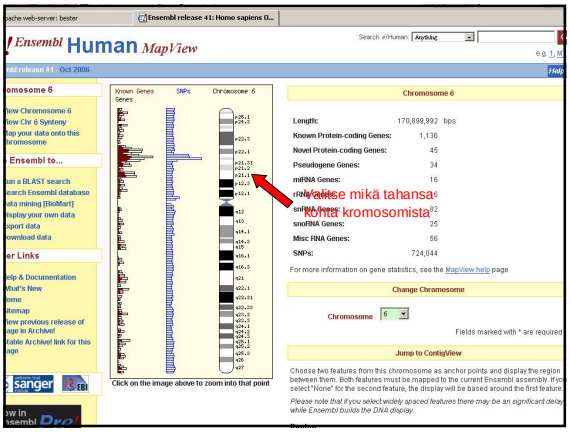
Getting the sequences

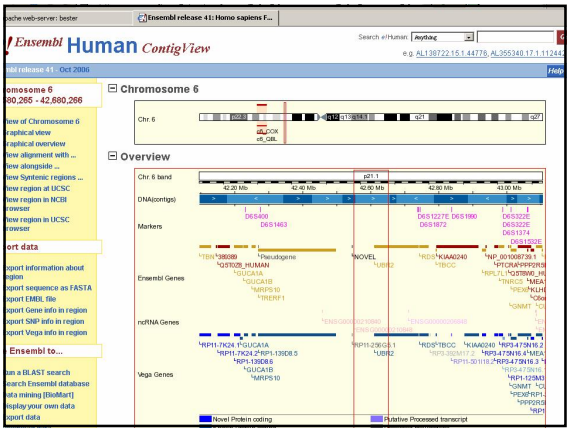


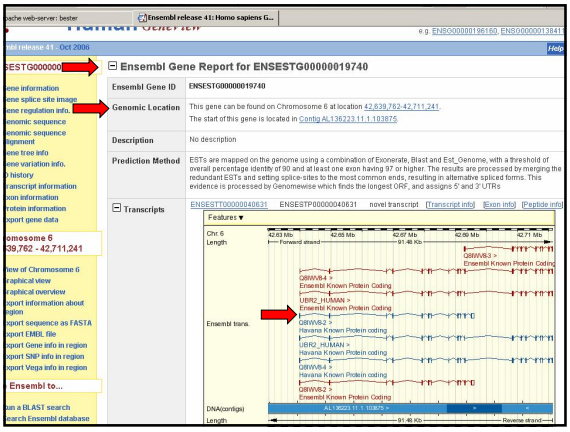
Ensembl front page

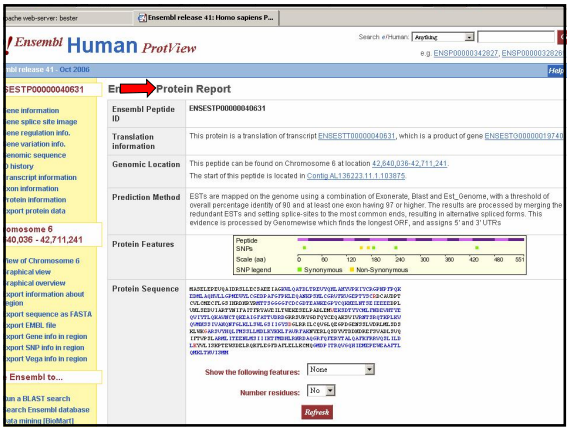


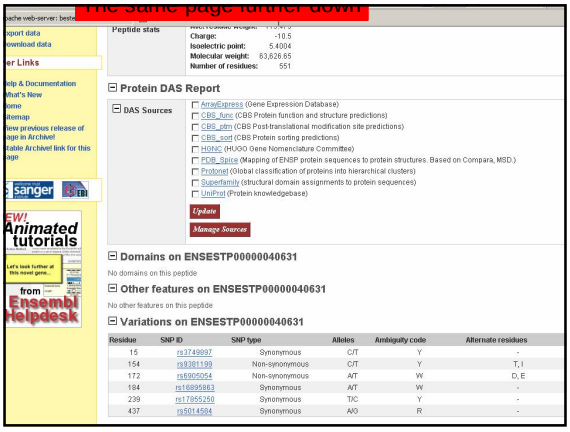


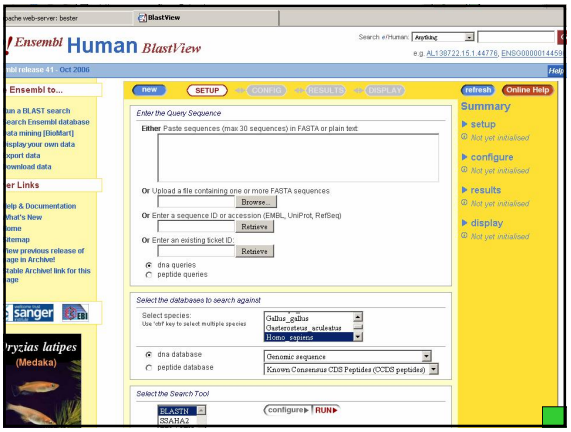

Valitse mikä tahansa
kohta kromosomista








The same page further down



Gene Ontology (GO)

www.geneontology.org

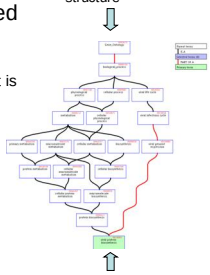
- A controlled vocabulary of gene product roles in cells and the role associations
- The roles can be applied to all organisms
- Three main hierarchies: biological process, cellular component and molecular function include currently about 19,000 classes (=roles)
  - -usually only a small portion of these classes is in use with one organism

Structure of GO

GO graph:

- Hierarchical structure of linked nodes
  - -each node presents one class that is part of its parental class
- Direct Acylic Graph (DAG)
  - -this means a tree-structure where branches do not just split but also merge when going from parental nodes to child nodes.

root of hierarchical structure



Starting node

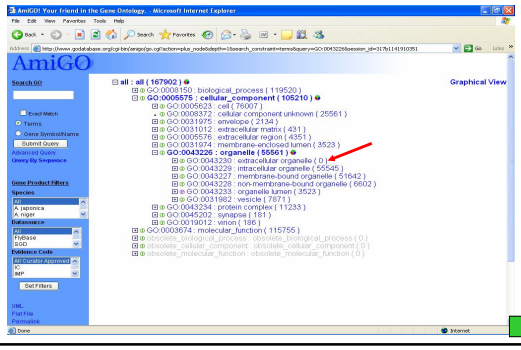
How GO helps?

- GO presents a terminology for presentation of known information of the gene
  - -fixed terminology when presenting information obtained from genomic databases. Easy automatic analysis
  - -same GO structure can be used although the functional information would come from various sources
- GO classifies genes according to their known/predicted functions
  - -these can be analyzed to see if the classes are over-represented in the results

How GO helps

- Classes present various details
  - -for example: cytoplasm, protein synthesis, ribosomal protein, large ribosomal sub-unit…
  - -analysis can find information with various details

Gene Ontology (GO)



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/

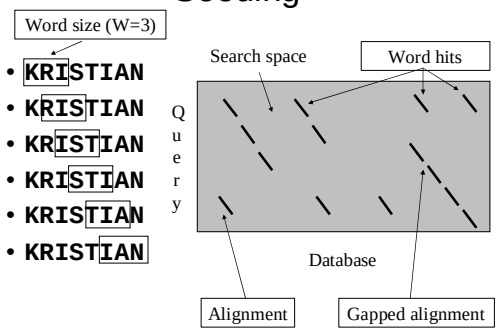
How BLAST works?

- 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.

Seeding

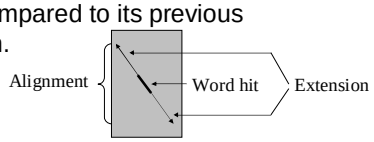


Threshold in seeding

- 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.



Extension, example

KRISTIAN gap=0, X=1
-RISTISANA BLOSUM62
0544541200 **<**- score
0000000**2** **<**- 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

- $E = Kmne^{-\lambda S}$
  - 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
  - λS, 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?
- $E = Kmne^{-\lambda S} = 0{,}1*250*250*2{,}71^{-(0{,}229*75)}$
  $= 0{,}000217$

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, possibly from multiple organisms
translated DNA	translated DNA	tblastx	Cross-species gene prediction, searching for genes not yet in protein databases

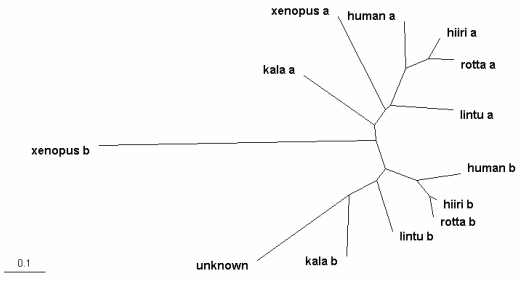
- A *homologous* gene or a sequence is derived from a common ancestor to all (or most of) the daughter species, and thus can be *aligned* accordingly:

  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 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)

Alfa- ja beta-globins



Some HI-virus epidemiology