Bioinformatics databases and applications

Eitan Rubin, December 2002
Outline

• Introduction
• A day in the life of a biologist
• Major databases
• Major tools
Outline

• Introduction
• A day in the life of a biologist
• Major databases
• Major tools
Life as a simple CS problem
A more realistic view

- Input1
- Input2
- Algorithm1
- Output
- Decision
- Algorithm2
- Algorithm3
A typical real-life view
The life cycle of a bioinformatics project

• Clearly define the goals
• Define a strategy
• Run the process
• QA & optimize
  – Controls
  – External knowledge
  – Re-sampling
  – Correlation
Outline

• Introduction

• A day in the life of a biologist

• Major databases

• Major tools
Positional cloning of disease X

XM-417-L16  XM-417-L15
Genome browser @ UCSC
Looking at the region of interest

chrX:98100000-98500000

Gene prediction program suggest there are 6-8 genes in the region
Get mRNA @ NCBI

LOCUS    BC029575
DEFINITION Homo sapiens, Similar to LOC203562, clone NGC:39655 IMAGE:5267530, mRNA, complete cds.
ACCESSION BC029575
VERSION    BC029575.1  GI:20009511
KEYWORDS  NGC.
SOURCE     Homo sapiens (human)
ORGANISM Homo sapiens
Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi; Mammalia; Eutheria; Primates; Catarrhini; Hominidae; Homo.
REFERENCE  1  (bases 1 to 738)
AUTHORS    Strausberg,R.
TITLE  Direct Submission

>unkown_protein
MRLTEKSEGEQQLKPNNSNAPNEDQEEEIQQSEQHTPARQRTQRADTQPSRCRLPSR
RTPTTSSDRTINLLEVLPWPTEWIFNPYRLPALFELYPEFLVFKKEAFHDISHCLKA
LELLVEGPLPWPTEWIFNPYRLPALFELYPEFLVFKKEAFHDISHCLKALLELLVEGPLPWPTEWIFNPYRLPALFELYPEFLVFKKEAFHDISHCLKALELLVEGPLPWPTEWIFNPYRLPALFELYPEFLVFKKEAFHDISHCLKALLELLVEGPLPWPTEWIFNPYRLPALFELYPEFLVFKKEAFHDISHCLKALELLVEGPLPWPTEWIFNPYRLPALFELYPEFLVFKKEAFHDISHCLKALELLVEGPLPWPTEWIFNPYRLPALFELYPEFLVFKKEAFHDISHCLKA
BLAST @ NCBI
Distribution of 5 Blast Hits on the Query Sequence

Mouse-over to show defline and scores. Click to show alignments

Color Key for Alignment Scores

<table>
<thead>
<tr>
<th>Score Range</th>
<th>Color</th>
</tr>
</thead>
<tbody>
<tr>
<td>&lt;40</td>
<td>Black</td>
</tr>
<tr>
<td>40-50</td>
<td>Blue</td>
</tr>
<tr>
<td>50-80</td>
<td>Green</td>
</tr>
<tr>
<td>80-200</td>
<td>Red</td>
</tr>
<tr>
<td>&gt;200</td>
<td>Orange</td>
</tr>
</tbody>
</table>

1.5733

Sequences producing significant alignments:

<table>
<thead>
<tr>
<th>GI</th>
<th>Ref</th>
<th>Description</th>
<th>Score</th>
<th>E (bits)</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>gi</td>
<td>20553864</td>
<td>ref</td>
<td>XP 117553.1</td>
<td>hypothetical protein XP_117553...</td>
<td>196</td>
</tr>
<tr>
<td>gi</td>
<td>20809912</td>
<td>gb</td>
<td>AAH29575.1</td>
<td>Similar to LOC203562 [Homo sapiens]</td>
<td>196</td>
</tr>
<tr>
<td>gi</td>
<td>126580</td>
<td>sp</td>
<td>P19839</td>
<td>LXA1_PHOLU</td>
<td>Alkanal monooxygenase alpha ...</td>
</tr>
<tr>
<td>gi</td>
<td>15430755</td>
<td>gb</td>
<td>AAK98554.1</td>
<td>AF403784_3</td>
<td>luciferase alpha subunit ...</td>
</tr>
<tr>
<td>gi</td>
<td>282045</td>
<td>pir</td>
<td>C42951</td>
<td>alkanal monooxygenase (FMN-linked)</td>
<td>30</td>
</tr>
</tbody>
</table>

Alignments
Search for domains @Interpro

This form allows you to query your protein sequence against InterPro. For more detailed information see the documentation (Readme file or FAQs), or the InterPro user manual or help pages. If you wish to use this facility during a conference then please contact support@ebi.ac.uk.

Please either enter (or cut and paste) your protein sequence into the text box below, or, if you have the sequence on your computer, click the 'Browse' button to upload it directly (you will be given a file selection window if you option). If you need help on sequence formats, this page details various common formats. Enter or cut and paste sequence here.

>unknown_protein
MRLTEKSEGEQQLKPNNSAPNEDQEEEIQSEQHTPARQRTQRAIRTINLLEVLPPWTEWIFNPRYLPAFLLEYPEFLLVFKEAFHDISHCSTLYFYKFFLPITILSLSFILLVLVLFLFIIVFILLF
Search for domains @Interpro
Get predicted protein @ UCSC

Assembly Gene Predictions With Alt-splicing (naharu.b)

Chromosome: X
Band: Xq22.2
Begin in Chromosome: 98232207
End in Chromosome: 98243370
Genomic Size: 11164
Strand: -

Links to sequence:
- Predicted Protein
- Predicted mRNA may be different from the genomic sequence.

> naharu.b
MSSRKQGQSOPRGRQPSAEENFKKPTSRNMQRNKMRGASSGKKTAGPQQKN
LEPALPGRWGGRSAPPSGSVRKTRKणKQKTPGNDGGSTSEAPQPPRK
KRAALQPTVESEAFKNRMKVKIPEELKPWLVEDWDLVTRQKQLFQLP
AKKNVDAILEEYANCKKSQGNVMSKEYAVNVVAGIKEYFNVMLGTQLLY
KFERPQYAEILLAHDPAMPSQVYGAHLRLFVRIGAMLAYTPLDEKSLA
LLLGYLHDFLKYLAKNSASLFTASDYKVASAEYHRKAL
<table>
<thead>
<tr>
<th>Accession</th>
<th>Description</th>
<th>Score</th>
<th>E-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>gi</td>
<td>6912448</td>
<td>ref</td>
<td>NP_036418.1</td>
</tr>
<tr>
<td>gi</td>
<td>9790223</td>
<td>ref</td>
<td>NP_062742.1</td>
</tr>
<tr>
<td>gi</td>
<td>26354781</td>
<td>dbj</td>
<td>BAC41017.1</td>
</tr>
<tr>
<td>gi</td>
<td>26344852</td>
<td>dbj</td>
<td>BAC36075.1</td>
</tr>
<tr>
<td>gi</td>
<td>5803102</td>
<td>ref</td>
<td>NP_006782.1</td>
</tr>
<tr>
<td>gi</td>
<td>25069552</td>
<td>gb</td>
<td>AAN65331.1</td>
</tr>
<tr>
<td>gi</td>
<td>12746239</td>
<td>gb</td>
<td>AAK07407.1</td>
</tr>
<tr>
<td>gi</td>
<td>4406689</td>
<td>gb</td>
<td>AAD20058.1</td>
</tr>
<tr>
<td>gi</td>
<td>11360195</td>
<td>pir</td>
<td>T46285</td>
</tr>
<tr>
<td>gi</td>
<td>26347795</td>
<td>dbj</td>
<td>BAC37546.1</td>
</tr>
<tr>
<td>gi</td>
<td>10441952</td>
<td>gb</td>
<td>AAG17251.1</td>
</tr>
<tr>
<td>gi</td>
<td>6996928</td>
<td>ref</td>
<td>NP_006783.2</td>
</tr>
<tr>
<td>gi</td>
<td>20890584</td>
<td>ref</td>
<td>XP_123435.1</td>
</tr>
<tr>
<td>gi</td>
<td>20892155</td>
<td>ref</td>
<td>XP_147201.1</td>
</tr>
<tr>
<td>gi</td>
<td>25046274</td>
<td>ref</td>
<td>XP_195418.1</td>
</tr>
<tr>
<td>gi</td>
<td>20846898</td>
<td>ref</td>
<td>XP_124668.1</td>
</tr>
<tr>
<td>gi</td>
<td>25030960</td>
<td>ref</td>
<td>XP_205276.1</td>
</tr>
<tr>
<td>gi</td>
<td>20983284</td>
<td>ref</td>
<td>XP_141646.1</td>
</tr>
<tr>
<td>gi</td>
<td>21288856</td>
<td>gb</td>
<td>EAA01149.1</td>
</tr>
<tr>
<td>gi</td>
<td>20878112</td>
<td>ref</td>
<td>XP_123230.1</td>
</tr>
<tr>
<td>gi</td>
<td>20876894</td>
<td>ref</td>
<td>XP_123229.1</td>
</tr>
<tr>
<td>gi</td>
<td>21357515</td>
<td>ref</td>
<td>NP_650442.1</td>
</tr>
<tr>
<td>gi</td>
<td>20875666</td>
<td>ref</td>
<td>XP_143489.1</td>
</tr>
<tr>
<td>Accession</td>
<td>Description</td>
<td>Length</td>
<td>E-value</td>
</tr>
<tr>
<td>-----------</td>
<td>--------------------------------------------------</td>
<td>--------</td>
<td>---------</td>
</tr>
<tr>
<td>gi</td>
<td>179757755</td>
<td>male-specific lethal 3-like 1</td>
<td>38</td>
</tr>
<tr>
<td>gi</td>
<td>13276691</td>
<td>hypothetical protein [Homo sapiens]</td>
<td>36</td>
</tr>
<tr>
<td>gi</td>
<td>765735</td>
<td>MYB binding protein la; p53-actin</td>
<td>36</td>
</tr>
<tr>
<td>gi</td>
<td>25045871</td>
<td>similar to nuclear protein [Ensembl]</td>
<td>36</td>
</tr>
<tr>
<td>gi</td>
<td>7949086</td>
<td>MYB binding protein (P160) la</td>
<td>36</td>
</tr>
<tr>
<td>gi</td>
<td>19070182</td>
<td>hypothetical protein [AF345640_1]</td>
<td>36</td>
</tr>
<tr>
<td>gi</td>
<td>18570703</td>
<td>similar to IGF-II mRNA-binding</td>
<td>35</td>
</tr>
<tr>
<td>gi</td>
<td>23490720</td>
<td>Papain family cysteine protease</td>
<td>35</td>
</tr>
<tr>
<td>gi</td>
<td>22968103</td>
<td>hypothetical protein [Rhodospirillum]</td>
<td>35</td>
</tr>
<tr>
<td>gi</td>
<td>7487134</td>
<td>hypothetical protein T18B16.160 - Ar...</td>
<td>35</td>
</tr>
<tr>
<td>gi</td>
<td>13278576</td>
<td>Similar to MYB binding protein</td>
<td>35</td>
</tr>
<tr>
<td>gi</td>
<td>11358805</td>
<td>RNA helicase RH26 [imported] - Arabidopsis</td>
<td>35</td>
</tr>
<tr>
<td>gi</td>
<td>23508021</td>
<td>citrate synthase, mitochondria</td>
<td>35</td>
</tr>
<tr>
<td>gi</td>
<td>25025497</td>
<td>hypothetical protein XP_206172</td>
<td>35</td>
</tr>
<tr>
<td>gi</td>
<td>12653713</td>
<td>Similar to MYB binding protein</td>
<td>34</td>
</tr>
<tr>
<td>gi</td>
<td>11282566</td>
<td>hypothetical protein DDKFZp434B1021</td>
<td>34</td>
</tr>
<tr>
<td>gi</td>
<td>15242321</td>
<td>DEAD/DEAH box RNA helicase, pu...</td>
<td>34</td>
</tr>
<tr>
<td>gi</td>
<td>15426118</td>
<td>possible hypothetical 186.0 Kd</td>
<td>34</td>
</tr>
<tr>
<td>gi</td>
<td>19115307</td>
<td>hypothetical serine-rich protein</td>
<td>33</td>
</tr>
<tr>
<td>gi</td>
<td>22967609</td>
<td>hypothetical protein [Rhodospirillum]</td>
<td>33</td>
</tr>
<tr>
<td>gi</td>
<td>15615015</td>
<td>BH2452~unknown conserved protein</td>
<td>33</td>
</tr>
<tr>
<td>gi</td>
<td>20892917</td>
<td>similar to DDKFZp586L2024 protein</td>
<td>33</td>
</tr>
<tr>
<td>gi</td>
<td>15924208</td>
<td>conserved hypothetical protein</td>
<td>32</td>
</tr>
<tr>
<td>gi</td>
<td>21282830</td>
<td>conserved hypothetical protein</td>
<td>32</td>
</tr>
</tbody>
</table>

**Alignments**
1: NP_006782. MORF-related gene...[gi:5803102]

**LOCUS** MRG15

**DEFINITION** MORF-related gene 15 [Homo sapiens].

**ACCESSION** NP_006782

**VERSION** NP_006782.1 GI:5803102

**DBSOURCE** REFSEQ: accession NM_006791.1

**KEYWORDS** .

**SOURCE** Homo sapiens (human)

**ORGANISM** Homo sapiens

Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
Mammalia; Eutheria; Primates; Catarrhini; Hominidae; Homo.

**REFERENCE** 1 (residues 1 to 323)

**AUTHORS** Bertram,M.J., Berube,N.G., Hang-Swanson,X., Ran,Q., Leung,J.K.,
Bryce,S., Spurgers,K., Bick,R.J., Baldini,A., Ning,Y., Clark,L.J.,
Parkinson,E.K., Barrett,J.C., Smith,J.R. and Pereira-Smith,O.M.

**TITLE** Identification of a gene that reverses the immortal phenotypic of a
Query: gi|5803102 MORF-related gene 15 [Homo sapiens]
Matching gi: 4454704, 4808625, 6841360, 8895208, 17160904, 18605583, 12643874, 12746237, 12804159, 12855127, 13277348, 13278084, 14275916

81 BLAST hits to 11 unique species Sort by taxonomy proximity

0 Archaea 0 Bacteria 73 Metazoa 7 Fungi 7 Plants 7 Viruses 7 Other Eukaryotae

Keep only Cut-Off 100 Select Reset

323 aa

<table>
<thead>
<tr>
<th>SCORE</th>
<th>P</th>
<th>ACCESSION</th>
<th>GI</th>
<th>PROTEIN DESCRIPTION</th>
</tr>
</thead>
<tbody>
<tr>
<td>1683</td>
<td>27</td>
<td>CAB70879</td>
<td>6808057</td>
<td>hypothetical protein [Homo sapiens]</td>
</tr>
<tr>
<td>1656</td>
<td>21</td>
<td>AAK07407</td>
<td>12746239</td>
<td>histone acetylase complex subunit MRG15-2</td>
</tr>
<tr>
<td>1656</td>
<td>27</td>
<td>AAN65338</td>
<td>25069552</td>
<td>putative transcription factor MRG15-2 [Homo sapiens]</td>
</tr>
<tr>
<td>1629</td>
<td>27</td>
<td>AAG17253</td>
<td>10441952</td>
<td>unknown [Homo sapiens]</td>
</tr>
<tr>
<td>1363</td>
<td>21</td>
<td>XP_147201</td>
<td>20892155</td>
<td>similar to testis expressed gene 189 [Mus musculus]</td>
</tr>
<tr>
<td>1268</td>
<td>27</td>
<td>AAD20058</td>
<td>4406680</td>
<td>Unknown [Homo sapiens]</td>
</tr>
</tbody>
</table>
MG15_HUMAN
- **Size:** 323 amino acids; 37230 Da
- **Function:** PLAYS A ROLE IN GROWTH REGULATION AND REPLICATIVE SENESCENCE. U PROGRESSION.
- **Subcellular location:** Nuclear.
- **Developmental stage:** CELL CYCLE REGULATED.

**MIPS Pedant Viewer:** 62441

**REFSEQ proteins:** NP_006782.1

**InterPro Domains and Families:**

IPR000953; Chromo

[Graphical View of Domain Structure for SP Entry Q9UBU8]

**Gene Ontology (GO) terms (tree view):**

GO:0005634
UPTREGULATION AND REPlicative SeneScence. UPGREGULATION OF \textbf{MRG15} MAY BE NEEDED FOR CELL CYCLE TRANSITION.

Entry Q9UBU8
Outline

• Introduction
• A day in the life of a biologist
• Major databases
• Major tools
BIND; MINT; BRITE ... 

PDB

HSSP

Swissprot; interpro; LAMA; GO

StackDB; Gencarta; Ensembl

???

INSD (genbank, EMBL, DDJB)

Specialized databases: Flybase, YPD, UCSC,
INSD

- Genbank, EMBL, DDJB
- CleanBank
- Divisions (EST, HTG)
- Specialized databases
Major tools

• Transcript modelling from ESTs
  – Sequencher, Staden, StackPACK

• Database searching
  – Blast
  – BLAT
  – Fasta

• Multiple Sequence Alignment
  – ClustalX
  – MAGAW
Major tools

- Gene prediction
- (EST) assembly
- Promoter Finding
- ORF identification
- Similarity searching
- MSA
- Phylogenetic analysis
- Structure prediction
- Docking
ClustalX

• Stepwise tree-guided alignment
• “Bag full of tricks”
• Demo
The effect of parameters

Default parameters

Modified parameters

Eitan Rubin

Bioinformatics & Biological Computing Unit
Department of Biological Services
The effect of parameters
Major tools

- Gene prediction
- (EST) assembly
- Promoter Finding
- ORF identification
- Similarity searching
- MSA
- Phylogenetic analysis
- Structure prediction
- Docking
Similarity searching

- **SW (accelerated)**
- **BLAST**
  - The NCBI environment, Fast, wide dynamic range, availability
  - DNA very bad stats, poor for proteins
  - Highly local FASTA
- **BLAT**
  - Lightening fast, focused
  - Limited dynamic range
MSA

- *ClustalX*
  + Fast; familiar
  - Global; One, not very accurate algorithm
- *Macaw*
  + Very interactive; outstanding GUI; multiple algorithms
  - Immature; runs on PCs; incompatible
- *BLOCKS maker*
  + Fully automated; fast
  - Poor control; many mistakes