Bioinformatics for biomedicine

More annotation, Gene Ontology and pathways

Lecture 6, 2006-10-24

Per Kraulis

http://biomedicum.ut.ee/~kraulis

Course design

1. What is bioinformatics? Basic databases and tools
2. Sequence searches: BLAST, FASTA
3. Multiple alignments, phylogenetic trees
4. Protein domains and 3D structure
5. Seminar: Sequence analysis of a favourite gene
6. **More annotation, Gene Ontology and pathways**
7. Gene expression data, methods of analysis
8. Seminar: Further analysis of a favourite gene

Note: 6 and 7 swapped!

Task from previous lecture

- Unknown sequences
- Which analysis?
- How interpret results?
- Medical relevance?

Unk4

- Danio rerio (zebrafish) sequence
- BLAST
  - Ensembl (species specific): wee1
  - UniProt: Q1LYE1, Q6DRI0 (WEE1 homolog)
- Pfam: Protein kinase
- Cell cycle regulation
  - Negative regulator of G2 to M transition
  - Orthologues in human, and many other species
  - S. pombe: “wee”=small; mutant goes fast into M
  - Against cancer: activation wanted (difficult)

Unk5

- Human, fragment
- BLAST
  - Ensembl (species specific)
  - UniProt: SRBP1_HUMAN
- Pfam: basic helix-loop-helix
- Transcription factor
  - Regulation of lipid synthesis, colesterol
  - Orthologues in many species
  - Very interesting pathway
  - Not itself a good drug target; maybe indirectly through protease regulation

Annotation, so far

- Similarity, homology, orthology,…
- Domains, families
  - Structure, function

- Individual annotation (in a certain sense)
  - Each protein considered on its own
  - Functional relationships: only limited (indirect) information

Functional relationships

- Interactions
  - Genetic: indirect or direct
  - Physical: direct
- Regulation
  - Structural: Cell or tissue type
  - Temporal: Developmental stage
  - Co-regulation patterns
- Processes
  - Involvement, role

How to obain functional data?

- Traditional single-gene approaches
  - Works: good data, cross-checked
  - But: does not scale easily

- Genome-wide screens are now possible
  - Allows novel types of investigations
  - Approximate, rarely cross-checked

Genome-wide screening 1

- Genomes specify all genes
  - List of all genes is available, in principle
- Technologies for screening
  - Design a specific setup to identify all genes that are active
- This allows genome-wide screening

Genome-wide screening 2

- Different from one-gene investigations
  - All genes investigated at once; overview
  - “Fishing expedition”

- Produces novel annotation
  - Genes X,Y,Z were “active” in condition W
  - Co-annotates proteins of unrelated sequence
  - Previously un-annotated genes highlighted

Screens give functional data

- Annotation from screens is functional
- Sequence only indirectly involved
- Experimental setup is crucial
  - Materials
  - Conditions
  - Detection scheme, measurement
- Requires new types of data analysis

Function (or, biology) of a gene

- = Activity of the gene product

- Functional properties
  - Chemistry
  - Interaction partners
- Biological process(es)
  - Location: tissues, cells
  - Temporal: development, conditions

Gene activity depends on… 1

- Activity requires gene product
  - Protein (occasionally RNA)
- Expression (mRNA synthesis) is a
  - necessary,
  - but not sufficient condition
- mRNA abundance: Sum of
  - Synthesis (expression)
  - Degradation

Gene activity depends on… 2

- Activity requires a functional state
  - mRNA synthesized, spliced, located,…
  - Protein synthesized, modified, located, folded,…
- Activity requires context
  - Interaction partners
  - Conditions
  - Signals

Gene expression 1

- Expression = mRNA level
  - Definition, or approximation?
- First genome-wide screen approach
- mRNA is technically easy to
  - Harvest
  - Identify
  - Quantify

Gene expression 2

- Note: mRNA levels not really of intrinsic biological relevance!
  - mRNA is just a messenger, after all

- Focus would be on proteins, if and when those as easy to handle

Gene expression is a proxy

- Reasonable first approximation
  - Expression required for activity
  - Protein vs. mRNA levels: correlated
  - Evolutionary optimization: why synthesize mRNA if not needed?
- Several confounding factors
  - Protein synthesis rate
  - mRNA degradation rate
  - Protein modification, folding, transport,…

Gene expression technologies 1

- EST
  - Expressed Sequence Tag
  - Short (200-500 bp) sequence of DNA
  - DNA derived from mRNA samples
  - Sequence analysis required for identification
  - Coarse quantification of mRNA levels
  - Large databases exist
- SAGE
  - Serial Analysis of Gene Expression
  - Very short (10-14 bp) sequences
  - Some data, not as popular as ESTs

Gene expression technologies 2

- Microarrays
  - Predefined sets of sequences
  - Many spots on a slide/chip, each of which contains a specific sequence
    - cDNA: spotted glass slides (Brown, Stanford)
    - Oligonucleotide: microchips (Affymetrix)
  - Usually relative quantification (up, down)
  - Many different applications
- Other technologies exist
  - New inventions quite possible

Gene expression data

- Analysis is easy, but difficult…
  - Relative differences between samples
    - Statistically significant?
    - Comparison depends on technology
    - Compare between genes, or just gene to itself?
  - Experimental design
    - Crucial for proper analysis; statistics
- Databases combine data sets for analysis
  - Many issues; watch out

Gene Ontology 1

- Keywords in annotation
  - Labels for properties, function, etc
- Problem: different sets of keywords
  - Comparison difficult, or impossible
- 1998 GO Consortium
  - Project to specify a consistent set of keywords for gene annotation
- http://www.geneontology.org/

Gene Ontology 2

- Ontology: Existence and relationships
- Controlled vocabulary
  - Artificial Intelligence
  - Knowledge Representation
- Example: MeSH (Medical Subjects Headings) at NLM

Gene Ontology 3

- Hierarchy of keywords (=terms)
  - Specific terms are instances of general terms
  - Something may be part of something else
- Why?
  - Natural description in biology
  - To reflect level of knowledge
- Example: DNA metabolism
  - DNA integration
  - DNA ligation

Gene Ontology 4

- Three separate hierarchies

- Molecular function
  - Chemical activity
- Cellular component
  - Location: nucleus, cytoplasm,…
- Biological process
  - Role in cellular events

Gene Ontology: applied

- For each gene, give the GO terms for which there is evidence: mapping
- As specific as evidence allows
- Evidence codes: source of statement
- Mapping produced by genome projects, not GO consortium
- Available in Ensembl, UniProt, etc

GO limitations

- Does not describe process, function or cellular component
  - Little spatial info
  - No temporal info
- Lacks many specific details
- Strictly limited: Is about terms (keywords)
- Extensions have been discussed
  - Should be done in independent way...

Proteomics 1

- Protein instead of mRNA
- Intrinsically biologically more interesting
- Technically much more difficult
  - Chemical properties
    - Different: no single protocol applicable
    - Same: hard to separate
  - How to identify? Sequencing not as simple
  - Quantification difficult

Proteomics: 2D gels

- Gel to separate proteins
- 2D: isoelectric point vs. mass
- Each spot detected, identified
  - Sequencing
  - Mass spectrometry
- Issues: many
  - Sensitivity
  - Comparison
- SwissProt http://www.expasy.org/ch2d/

Proteomics: in situ

- Detect protein in tissue sample
  - Issue: Tissue sample treatment?
- Antibodies against peptides
  - Issue: Which peptide?
- Immunostaining to visualize
- Protein Atlas http://www.hpr.se/

Protein interactions

- Use bait to fish out all interaction partners
  - Two-hybrid method
  - Coimmunoprecipitation
  - Colocalization, GFP
  - Other techniques…
- IntAct at EBI
  http://www.ebi.ac.uk/intact/site/

Pathways

- Metabolic
  - Conversion of “small” molecules
  - Energy, synthesis, degradation
  - KEGG http://www.genome.ad.jp/kegg/
  - BioCyc http://www.biocyc.org/
- Signalling, regulation
  - Reactome http://www.reactome.org/
- Work in progress…

Task

- Locate gene/protein in GO, Reactome, etc
- Wee1
- SREBP1
- Your own