Prokaryota: comparative genomics

Martin John Bishop

Cambridge, UK and Milan, Italy

Molecular biology needs bioinformatics

- Biological data - molecules
  - Sequences
  - Structures
  - Gene expression
  - Proteomes
  - Pathways
  - Evolution
- Computer analysis – methods
  - Comparison
  - Modelling
  - Co-regulation
  - Mass spectrometry
  - Knowledge bases
  - Phylogenetics

Molecular biology is about information

- Central dogma
  - DNA
    <-> RNA
    -> protein
    -> phenotype
    <- DNA
- Molecules
- Processes

- Central paradigm
  - Genome repository
    <-> RNA world
    -> Protein sequence
    -> Protein structure
    -> Protein function
    -> Phenotype
  <- Fed back to genome
- Information processing

Bioinformatics books

- http://www.bioplanet.com/bioinformatics_books.php
- http://www.geocities.com/bioinformaticsweb/bioinformaticsbooks.html

Bioinformatics journals

- Briefings in Bioinformatics. Oxford University Press. ISSN 1467-5483. http://bib.oxfordjournals.org/
- Bioinformatics. Oxford University Press. ISSN 1367-4803. http://bioinformatics.oxfordjournals.org/
- http://www.brc.dcs.gla.ac.uk/~actan/bioinformatics/journals.html

Bioethics Resources on the Web

- NIH site has many links
  http://bioethics.od.nih.gov/
- Ethnicity and Genetics
- Gene Patenting
- Gene Testing/Counselling
- Gene Therapy/Transfer

Nuffield Council on Bioethics

- Web site at http://nuffieldfoundation.org/bioethics/
- Genetic Screening
- Human Tissue
- Animal to Human Transplants
- Mental Disorders and Genetics
- Stem cell therapy

Biological literature

- One of the largest biological datasets
- Medline/PubMed catalogues over 15 million entries
- http://dan.corlan.net/medline-trend.html
  Number Year
  - 676175 2005
  - 621965 2004
  - 582277 2003
  - 551900 2002
  - 532420 2001
  - 520717 2000

Bioinformatics data acquisition

- Genomics
- Comparative Genomics
- Functional genomics
  - Transcriptomics
  - Proteomics
  - Metabolomics
- Systems biology

Nucleotide and amino acid sequences

- Probably exceeds the literature in volume
- 394 fully sequenced genome are public
- Many more genomes are being determined
- EMBL Nucleotide Sequence Database contains 134,602,904,495 nucleotides including ESTs, STSs and GSSs at May 2006.

Completely Sequenced Genomes

- http://www.nslij-genetics.org/seq/links.html
- http://www.ebi.ac.uk/2can/genomes/genomes.html
- **http://www.genomesonline.org/**
  Complete Published Genome Projects: 394

  Archaeal:27 Bacterial:326 Eukaryal:41 (Viral:184)

Genome sequence data - TIGR

- **The Comprehensive Microbial Resource**
- **Jeremy D. Peterson, Lowell A. Umayam, Tanja Dickinson, Erin K. Hickey and Owen White**
- The Institute for Genomic Research, 9712 Medical Center Drive, Rockville, MD 20850, USA
- Nucleic Acids Research, 2001, Vol. 29, No. 1 **123-125**

Microbial genome data - TIGR

- Complete
- http://www.tigr.org/tdb/mdb/mdbcomplete.html
- In progress
- http://www.tigr.org/tdb/mdb/mdbinprogress.html

Microbial genome data - Sanger

- http://www.sanger.ac.uk/Projects/Pathogens/
- In 1995 the Wellcome Trust took the decision to set up a Pathogen Sequencing Unit (PSU) at what was then the Sanger Centre, to sequence the genomes of organisms relevant to human and animal health

GENOMICS APPLICATIONS

- Linkage Analysis
- Radiation Hybrid Mapping
- Sequence Ready Clone Maps
- Genome Databases
- Polymorphisms
- Sequence Analysis
- Gene Prediction
- Expression Profiling
- Phylogenetic Analysis

Bacterial gene identification by signal

- Find open reading frames
- Find start and stop
- Find promoters
- Reasonably successful (unlike eukaryotes where it is difficult to impossible - splicing)

Genomic sequences by content

- Training set of genes for which promoter sites are known
- Markov chain models to discriminate the sets of sequence types
- Apply to newly sequenced DNA

Fairly successful methodology

Genomic sequences by similarity

- Sequence comparison
  - Smith-Waterman algorithm
  - FASTA heuristic
  - BLAST heuristic
  - HMMs a full probabilistic methodology
  - EST comparison

Gene finding by similarity is successful

Comparative method

- Life is more uniform at the molecular level than we might have imagined
- Genes from bacteria can be informative in human biology
- Model organisms are extremely informative
  - many bacteria compared
  - yeast, nematode, fruit fly
  - rice, thale cress (a brassica)
  - puffer fish, chicken, mouse

Comparative genomics

- Most powerful method for gene identification
- Protein families by HMMs (PFAM)
- Attempt to compile lists of orthologs and paralogs
- Clusters of orthologous groups (COGs – NCBI)

COGs

- "COG" stands for Cluster of Orthologous Groups of proteins. The proteins that comprise each COG are assumed to have evolved from an ancestral protein, and are therefore either orthologs or paralogs. Orthologs are proteins from different species that evolved by vertical descent (speciation), and typically retain the same function as the original. Paralogs are proteins from within a given species that are derived from gene duplication, and may evolve new functions that are related to the original.

COG species **Archaea (13):**

- Afu *Archaeoglobus fulgidus*
- Hbs *Halobacterium sp. NRC-1*
- Mac *Methanosarcina acetivorans*
- Mth *Methanothermobacter thermautotrophicus*
- Mja *Methanococcus jannaschii*
- Mka *Methanopyrus kandleri AV19*
- Tac *Thermoplasma acidophilum*
- Tvo *Thermoplasma volcanium*
- Pho *Pyrococcus horikoshii*
- Pab *Pyrococcus abyssi*
- Pya *Pyrobaculum aerophilum*
- Sso *Sulfolobus solfataricus*
- Ape *Aeropyrum pernix*

COG species ***Eukaryota (3):***

- Sce *Saccharomyces cerevisiae*
- Spo *Schizosaccharomyces pombe*
- Ecu *Encephalitozoon cuniculi*

COG species ***Bacteria (10):***

- Aae *Aquifex aeolicus*
- Tma *Thermotoga maritima*
- Ctr *Chlamydia trachomatis*
- Cpn *Chlamydophila pneumoniae CWL029*
- Tpa *Treponema pallidum*
- Bbu *Borrelia burgdorferi*
- Syn *Synechocystis*
- Nos *Nostoc sp. PCC 7120*
- Fnu *Fusobacterium nucleatum*
- Dra *Deinococcus radiodurans*

COG species
Actinobacteria (4):

- Cgl *Corynebacterium glutamicum*
- Mtu *Mycobacterium tuberculosis H37Rv*
- MtC *Mycobacterium tuberculosis CDC1551*
- Mle *Mycobacterium leprae*

COG species gramplus (12)

- Cac *Clostridium acetobutylicum*
- Lla *Lactococcus lactis*
- Spy *Streptococcus pyogenes M1 GAS*
- Spn *Streptococcus pneumoniae TIGR4*
- Sau *Staphylococcus aureus N315*
- Lin *Listeria innocua*
- Bsu *Bacillus subtilis*
- Bha *Bacillus halodurans*
- Uur *Ureaplasma urealyticum*
- Mpu *Mycoplasma pulmonis*
- Mpn *Mycoplasma pneumoniae*
- Mge *Mycoplasma genitalium*

COG species **gamma (11):**

- Eco *Escherichia coli K12*
- EcZ *Escherichia coli O157:H7 EDL933*
- Ecs *Escherichia coli O157:H7*
- Ype *Yersinia pestis*
- Sty *Salmonella typhimurium LT2*
- Buc *Buchnera sp. APS*
- Vch *Vibrio cholerae*
- Pae *Pseudomonas aeruginosa*
- Hin *Haemophilus influenzae*
- Pmu *Pasteurella multocida*
- Xfa *Xylella fastidiosa 9a5c*

COG species
Proteobacteria (6):

- Nme *Neisseria meningitidis MC58*
- NmA *Neisseria meningitidis Z2491*
- Rso *Ralstonia solanacearum*
- Hpy *Helicobacter pylori 26695*
- jHp *Helicobacter pylori J99*
- Cje *Campylobacter jejuni*

COG species ***alpha (7):***

- Atu *Agrobacterium tumefaciens strain C58 (Cereon)*
- Sme *Sinorhizobium meliloti*
- Bme *Brucella melitensis*
- Mlo *Mesorhizobium loti*
- Ccr *Caulobacter crescentus CB15*
- Rpr *Rickettsia prowazekii*
- Rco *Rickettsia conorii*

Gene order

- Gene order conserved in operons
- Other genes may occur in more or less random orders
- Conservation therefore hints at function
- Gene fusion can also be used
  - Genes that occur separately in one organism may be joined into one in another

KEGG PATHWAY Database

Current knowledge on molecular interaction networks, including metabolic pathways, regulatory pathways, and molecular complexes

http://www.genome.ad.jp/kegg/pathway.html

Genomic associations predict pathways

- Huynen MA & Snel B (2003) In Galperin MY & Koonin EV (Eds) Frontiers in computational genomics. Caister Academic Press, Wymondham, UK
- Using COG and KEGG show that genomic association is correlated with metabolic distance

Evolution of metabolism

- The two most common models for the evolution of metabolism are the patchwork evolution model, where enzymes are thought to diverge from broad to narrow substrate specificity, and the retrograde evolution model, according to which enzymes evolve in response to substrate depletion. Analysis of the distribution of homologous enzyme pairs in the metabolic network can shed light on the respective importance of the two models.

Patchwork model dominates

- The evolution of the metabolism in E. coli viewed as a single network using data from EcoCyc indicates that, while the retrograde evolution model may have played a small part, the patchwork evolution model is the predominant process of metabolic enzyme evolution.

Functional genomics

- Challenge of quantitative time course measurement of cellular components
  - Transcripts
  - Proteins
  - Metabolites

Gene expression data

- EST libraries
- Serial Analysis of Gene Expression (SAGE) libraries
- DNA Microarray Data
  - Oligonucleotides
  - PCR products

Transcripts

- EST clustering
  - Microarrays
  - Alternative splicing
  - SNPs
- Analysis of expression data
  - Co-regulated genes
  - Hints about function

Microarrays

- cDNA or oligonucleotide based
- Permit high throughput studies of multiple conditions
- Require statistical treatment of data to ensure reproducibility
- Require sophisticated analysis to interpret

Protein identification

- 2D-PAGE
- Mass spectometry

PROTEOMICS APPLICATIONS

- Protein Sequence Analysis
- Protein Structure Analysis
- Protein Structural Modeling
- Proteome Databases
- Tools for Peptide Sequence Determination
- Protein Cellular Localization
- Protein Functional Studies
- Pathways and Protein Interactions

Protein expression and interaction

- Essential for understanding of biology
- Appears to correlate poorly with mRNA
- 4000 protein interactions observed for yeast proteins
- Database of Interacting Proteins (DIP) http://dip.doe-mbi.ucla.edu/ lists 19,053 proteins and 55,733 interactions

Metabolite identification

- Mass Spectrometry
- Nuclear Magnetic Resonance Spectrometry
- Fourier Transform Infrared Spectrometry

Metabolic profiling

- Functional dissection of metabolic pathways
  - NMR
  - GC/MS
- Assign metabolic phenotypes

Pathways

- **Metabolic Pathways**
  - EcoCyc: Encyclopedia of E. coli Genes and Metabolism
  - HinCyc: Encyclopedia of H. influenzae Genes and Metabolism
  - PUMA
  - Biocatalysis/Biodegradation Database
  - Enzyme Database
  - Ligand Database
  - Klotho: Biochemical Compounds Declarative Database
  - Kyoto Encyclopedia of Genes and Genomes page, and Pathways page
  - NetBiochem Welcome Page

Pathways databases

- TRANSFAC regulation of gene expression
- Signalling maps
  - Woodgett Lab
  - SPAD
- **Genetic Regulatory Circuits**
- BRITE

Systems biology

- The synergistic application of experiment, theory and modeling towards understanding biological processes as whole systems instead of isolated parts.
- **SBW – Systems Biology Workbench**
- **SBML – Systems Biology Markup Language**

SBW and SBML

- **SBW and SBML are being developed in collaboration with the groups developing the simulation packages BioSpice, Cellerator, DBsolve, E-CELL, Gepasi, Jarnac, ProMoT/DIVA, StochSim and Virtual Cell (amongst others)**
- http://sbml.org/index.psp