Genome-level understanding of functional microbial system (I)

Spring, 2007

Guangyi Wang, Ph.D.
POST103B
guangyi@hawaii.edu

http://www.soest.hawaii.edu/marinefungi/OCN403webpage.htm
Definition and Classification

- **Genome**
  - The entire complement of genetic material of an organism, virus or organelle.
  - The haploid set of chromosomes (DNA) of an eukaryotic organism.

- **Genomics**
  - First proposed by Thomas H. Roderick in 1986.
  - Described a new scientific discipline concerned with mapping, sequencing, and analyzing genomes.
  - The structure and functions of biological systems.

- **Genetic Map (gene linkage)**
  - Tell you how much DNA separates two genes and is measured in basepairs.
Definition and Classification (cont.)

• Sequencing
  – Determine sequences of DNA or cDNA fragment.
  – Give you detailed genetic codes
  – CCGGAACGTATTCACCGT

• Analyzing genome
  – Computational genome annotation.
  – Include, not limited to, prediction of protein coding region, promoter identification, operon identification, functional categories of genes, and other features in a genomes.
-Annotation of *Vibrio parahaemolyticus*

Annotation of *Vibrio parahaemolyticus* (cont.)

<table>
<thead>
<tr>
<th></th>
<th>Chromosome 1</th>
<th>Chromosome 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Length of sequence (bp)</td>
<td>3 288 558</td>
<td>1 877 212</td>
</tr>
<tr>
<td>G+C ratio</td>
<td>45.4%</td>
<td>45.4%</td>
</tr>
<tr>
<td>Number of coding sequences</td>
<td>3080</td>
<td>1 752</td>
</tr>
<tr>
<td>Protein coding region</td>
<td>86.9%</td>
<td>86.9%</td>
</tr>
<tr>
<td>Average length of coding sequence (bp)</td>
<td>926-9</td>
<td>931-3</td>
</tr>
<tr>
<td>Hypothetical proteins</td>
<td>1 090 (35%)</td>
<td>756 (43%)</td>
</tr>
<tr>
<td>rRNA operons</td>
<td>10</td>
<td>1</td>
</tr>
<tr>
<td>tRNA</td>
<td>112</td>
<td>14</td>
</tr>
</tbody>
</table>

**Genome features of *V parahaemolyticus***

**Figure 2**: Distribution of functional classes of predicted genes on *vibrio* chromosomes 1 and 2

Percentages of genes assigned according to the clusters of orthologous groups of proteins were compared between chromosomes 1 and 2. The percentage of genes in each class on chromosome 2 was divided by the percentage of genes in the same class on chromosome 1. Functional classes: J=translation, ribosomal structure, and biogenesis; K=transcription; L=DNA replication, recombination, and repair; D=cell division and chromosome partitioning; Q=post-translational modification, protein turnover, chaperones; M=cell-envelope biogenesis, outer membrane; N=cell motility and secretion; P=inorganic ion transport and metabolism; T=signal transduction mechanisms; C=energy production and conversion; G=carbohydrate transport and metabolism; E=amino acid transport and metabolism; F=nucleotide transport and metabolism; H=coenzyme metabolism; I=lipid metabolism; O=biosynthesis, transport, and catabolism of secondary metabolites; R=general function prediction only; S=function unknown.

**Figure 3**: Comparison of relative positions of conserved genes between *V parahaemolyticus* and *V cholerae*

Gene pairs were generated by Blastp analysis of predicted genes from each genome. Pairs with an expectation value of $1 \times 10^{-5}$ or less are shown. Red lines connect both conserved genes between the two organisms. The chromosomes are depicted with the putative replication terminus (predicted by the G2 draw) at the center. VP1=V cholerae chromosome 1; VP2=V cholerae chromosome 2; VP3=V parahaemolyticus chromosome 1; VP4=V parahaemolyticus chromosome 2; VP5=V parahaemolyticus chromosome 2.
Definition and Classification (cont.)

• Classification of genomic studies
  – System attributes

1) Structural genomics
   A. The genome wide structural study of genes, proteins, and other biomolecules, including genome mapping, sequencing, and organization as well as protein structure characterization.
   B. The large-scale determination of macromolecular structures, principally those of proteins (structural proteomics) (Brenner, 2001; Burley, 2000)
C. Features of structural genomics

Table 1
List of structural genomics consortia from TargetDB (http://targetdb.pdb.org)

<table>
<thead>
<tr>
<th>Centre/consortium</th>
<th>Location</th>
<th>Organism/focus</th>
</tr>
</thead>
<tbody>
<tr>
<td>Berkeley Structural Genomics Center (BSGC)</td>
<td>USA</td>
<td>Structural representation of the genomes of two pathogens, <em>Mycobacterium genitalium</em> and <em>Mycobacterium pneumoniae</em></td>
</tr>
<tr>
<td>Midwest Center for Structural Genomics (MCSG)</td>
<td>USA</td>
<td>Proteins from all three kingdoms of life</td>
</tr>
<tr>
<td>Northeast Structural Genomics Consortium (NEGS)</td>
<td>USA</td>
<td>Small proteins from eukaryotic model organisms such as <em>Saccharomyces cerevisiae</em>, <em>Caenorhabditis elegans</em> and <em>Drosophila melanogaster</em></td>
</tr>
<tr>
<td>NewYork Structural Genomics Research Consortium (NYSGXRC)</td>
<td>USA</td>
<td>Biologically interesting proteins from model organisms and humans</td>
</tr>
<tr>
<td>Southeast Collaboratory for Structural Genomics (SECSG)</td>
<td>USA</td>
<td><em>C. elegans</em> and <em>Pyrococcus furiosus</em> and selected human proteins</td>
</tr>
<tr>
<td>TB Structural Genomics Consortium (TB)</td>
<td>USA</td>
<td><em>Mycobacterium tuberculosis</em> proteins, particularly potential drug targets and novel folds</td>
</tr>
<tr>
<td>Joint Center for Structural Genomics (JCSG)</td>
<td>USA</td>
<td><em>Thermotoga maritima</em>, novel folds from <em>C. elegans</em>, and human proteins</td>
</tr>
<tr>
<td>Center for Eukaryotic Structural Genomics (CESG)</td>
<td>USA</td>
<td><em>Arabidopsis thaliana</em> proteins; technology development for eukaryotic proteins</td>
</tr>
<tr>
<td>Structure 2 Function Project (S2F) University of California Berkeley</td>
<td>USA</td>
<td>Functional characterization of hypothetical proteins from <em>Haemophilus influenzae</em></td>
</tr>
<tr>
<td>Structural Genomics of Pathogenic Protozoa consortium (SGPP)</td>
<td>USA</td>
<td>Proteins from major global pathogenic protozoa, <em>Leishmania major</em>, <em>Trypanosoma brucei</em> and <em>Plasmodium falciparum</em></td>
</tr>
<tr>
<td>Montreal-Kingston Bacterial Structural Genomics Initiative (BSGI)</td>
<td>Canada</td>
<td>Structures of potential virulence factors from pathogenic bacteria</td>
</tr>
<tr>
<td>Bacterial Targets at IGS-CNRS (BIGS)</td>
<td>France</td>
<td>Discovery of new antibacterial gene targets among evolutionary conserves genes of uncharacterized function</td>
</tr>
<tr>
<td>Oxford protein production facility (OPPF)</td>
<td>UK</td>
<td>Biomedical relevance of human pathogens, in particular Herpes viruses</td>
</tr>
<tr>
<td>Structural Proteomics in Europe (SPINE)</td>
<td>UK</td>
<td>Targets include human proteins implicated in cancer and neurodegenerative diseases</td>
</tr>
<tr>
<td>The Israel Structural Proteomics Center (ISPC)</td>
<td>Israel</td>
<td>Increase the efficiency of protein structure determination</td>
</tr>
<tr>
<td>Marseilles Structural Genomics Programe (MSGP)</td>
<td>France</td>
<td>Focus on bacterial, viral, and human ORFs</td>
</tr>
<tr>
<td>Mycobacterium Tuberculosis Structural Proteomics Project (XMTB)</td>
<td>Germany</td>
<td>Identify lead compounds against XMTB, using a structure based approach</td>
</tr>
<tr>
<td>Protein Structure Factory (PSF)</td>
<td>Germany</td>
<td>Human proteins to understand health and disease</td>
</tr>
<tr>
<td>Paris-Sud Yeast Structural Genomics (YSG)</td>
<td>France</td>
<td>Non-membrane proteins of unknown structure</td>
</tr>
<tr>
<td>RIKEN Structural Genomics Initiative (RSGI)</td>
<td>Japan</td>
<td>Proteins of biological and medical interest from mouse, <em>A. thaliana</em> and <em>Thermus thermophilus</em></td>
</tr>
<tr>
<td>Integrated Center for Structure and Function Innovation (ISFI)</td>
<td>USA</td>
<td>Developing and applying technologies to overcome bottlenecks of production of soluble protein and protein crystallization</td>
</tr>
</tbody>
</table>
Shematic diagram showing the strategy of protein structural genomics (crystal structure of latexin, the only known mammalian carboxypeptidase inhibitor, and its structure related to systatins, cysteine protease inhibitors, Aagaard et al., 2005)

The zinc finger structure (a) and the leucine zipper structure (b). Structure and their functions.
2) Functional Genomics

A. A system-level understanding of the functional aspects of biological systems, this is, gene functions and regulatory networks, using genome-wide approaches (Fields et al, 1999; Hieter & Boguski, 1997).

B. Gene function in different levels of biological functions
   • Biochemical function (e.g. phosphorylation by protein kinase)
   • Cellular function (e.g. role in cell division, DNA replication)
   • Developmental function (e.g. a role in cell-type differentiation)
   • Adaptive function (e.g. gene and its product contributing to the fitness of an organism)

C. Characterized by large-scale experimental approaches combined with statistical analysis, mathematical modeling, and computational analysis of the experimental results (Hieter & Boguski, 1997)

D. Aimed to link genome sequences to biological function, and consequently to provide new insights into the behavior, dynamics of living organisms, and exploration of biotech benefits (via DNA microarray and proteomics).
E. Functional genomics of *Listeria* spp.

a) The genus *Listeria* comprises a group of non-sporulating, Gram-positive, soil bacteria belonging to the low G + C group of microorganisms.

b) The genus consists of only six species, *L. monocytogenes*, *L. ivanovii*, *L. seeligeri*, *L. innocua*, *L. welshimeri*, and *L. grayi*.

c) *L. monocytogenes* and *L. ivanovii* are the only known pathogens of this group.

---

*Phylogenetic tree of the genus Listeria based on 16S and 23S rRNA (Schmid et al., 2005). Non-pathogenic and pathogenic strains are marked in blue and red, respectively.*
d) Comparative whole-genome sequencing of representative strains comprising the entire genus is currently being performed and nearing completion

- Features of their genomes
  - circular chromosomes with sizes that vary between 2.7 and 3.0 Mb in length.
  - have a G + C− content of between 36.4% and 41.5%.
  - The genome sequence encodes approximately 2800 putative protein coding genes, of which 65% genes have an assigned function.
  - all Listeria genomes revealed a highly conserved synteny in gene organization and content.
  - The lack of inversions or shifting of large genome segments could be due to the low occurrence of transposons and insertion sequence (IS) elements in all sequenced Listeria genomes.
  - DNA rearrangements of single gene loci are present in Listeria genomes even though it is a rather rare event.
e) pathogenetic *Listeria*

Intracellular life cycle of *L. monocytogenes*. (A) Schematic representation of the virulence factors involved in the cellular infection process adapted from Vazquez-Boland et al. (2001b). (B) Immunofluorescence microscopic image of infected P388D1 macrophages with wild type EGD-e 8 h post-infection (magnification, ×63) as described by Chatterjee et al. (2006).
f) In the genus *Listeria*, genome reduction has led to the generation of non-pathogenic species from pathogenic progenitor strains. Indeed, many of the regions absent in the non-pathogenic species represent commonly deleted genes.

Virulence gene cluster of *L. monocytogenes* EGD-e and its orthologs among other *Listeriae* (Hain et. al. 2006, J. Biotech.)
Definition and Classification (cont.)

• Classification of genomic studies
  – Relationships to other scientific disciplines

1) Basic genomics
  – Uses whole-genome sequence data and genomics technologies to understand basic cellular processes from a genome-wide perspectives.

2) Applied genomics
  – Applies genomic sequence information and associated high-throughput technologies to solving practical problems in various fields.
Definition and Classification (cont.)

• Classification of genomic studies
  – Types of organisms studies

Comparative genomics
  – The comparison of genome information (e.g., genomic sequences, mRNA, and protein expression profiles) from a variety of organisms for the purpose of obtaining a genome-wide understanding of biological processes and phenomena using both computational high-throughput approaches.
Summary

- Terminology (should know them)
- Classification based on system attributes, relationship to other scientific disciplines, and objects (can you define marine functional genomics??)