**BINF 630: Introduction to Bioinformatics**

Iosif Vaisman

Email: ivaisman@gmu.edu

---

**What is Internet**

Global Internet  
DNS Internet  
IP Internet

---

**Domain Name System**

- **Top-level domain**
- **Domains**
- **Hosts**

Data from ISC  
(January 2000)

---

**Client - Server Model**

CLIENT ➔ PROTOCOL ➔ SERVER

---

**Uniform Resource Locator (URL)**

protocol://host.domain[:port]/path/filename

http - a World Wide Web server (http://mmlin4.pha.unc.edu/~cmb96)  
telnet - a telnet session (telnet://nun.oit.unc.edu)
Network applications in science

- Virtual Laboratory
- Virtual Library
- Virtual Conference
- Virtual Classroom

Network collaboration

Real-time data sharing -- exchange of information between remote participants in the project

Resources sharing -- remote access to the instruments and computers

Resources integration -- simultaneous use of remote instruments and computers

Bioinformatics servers

Remote data access -- database search, cross-links between the databases

Remote computing -- use of server’s processing capabilities (sequence alignment, structure prediction, homology modeling)

Infospace navigation -- pointers to the available resources

Real-time

Asynchronous

Digital information cycle

Creation and capture
Storage and management
Rights management
Search and access
Distribution

Electronic publishing

Quality (peer review, retrospective evaluation)
Reliability (stability of servers, control over alterations, proper archiving and mirroring)
Hypertext Functionality in Scientific Literature

Active references
Forwarding references
Dynamic publishing

Ethical, Legal, and Economical Issues of Electronic Publishing

Intellectual property rights
Ownership of information
Information as a commodity

Molecular Databases

Nucleic acid sequences: GenBank, DNA Data Bank of Japan, EMBL Nucleotide Sequence Database
Nucleic acid structures: NDB - Nucleic Acid Database

Protein sequences: PIR - Protein Information Resource, SWISS-PROT
Protein structures: PDB - Protein Data Bank, NRL_3D

Physical properties: Biological Macromolecule Crystallization Database, BioMagResBank

Molecular images: SWISS-3DIMAGE, Molecules RUS

NCBI integrated search and retrieval system

Entrez

Database

example of a GenBank entry
Example of a Genbank entry

FEATURES             Location/Qualifiers

Gene             707..1774
/gene="luxA"
CDS             707..1774
/gene="luxA"
/codon_start=1
/product="luciferase alpha subunit"
/db_xref="PID:g155175"
/transl_table=11
/translation="MKFGNFLLTYQPPELSQTEVMKRLVNLGKASEGCGFDTVWLLE
HTFEFGLLGNPYVAAAHLLGATETLNVGTAAIVLPTAHPVRQAEDVNLLDQMSKGRFR
FGICRGLYDKDFRVFGTDMDNSRALMDCWYDLMKEGFNEGYIAADNEHIKFPKIQLNP
SAYTQGGAPVYVVAESASTTEWAAERGLPMILSWIINTHEKKAQLDLYNEVATEHGYD
VTIDHCLSYITSVDHDSNRAKDICRNFLGHWYDSYVNATKIFDDSDQTKGYDFNKGQ
WRDFVLKGHKDTNRRIDYSYEINPVGTPEECIAIIQQDIDATGIDNICCGFEANGSEE
EIIASMKLFQSDVMPYLKEKQ"

BASE COUNT      883 a    665 c    741 g    852 t
ORIGIN      1 bp upstream of EcoRI site.

Database Organization

Database Management System (DBMS)

Data * Hardware * Software * Users

Database Model

• A named logical unit (record type, data item)
• Relationships among logical units
  • one to one
  • one to many
  • many to one

DNA vs. Protein searches

DNA sequence - DNA database

• larger databases
• more random hits
• simpler scoring functions
• missing hits (similar proteins encoded by different DNAs)
Database administration

- Redundancy eliminated
- Inconsistency avoided
- Data shared
- Standards enforced
- Security applied
- Integrity maintained
- Requirements balanced

Data Warehouse

- Operational data
  - Data fusion
    - Data cleansing
      - Metadata

InterPro Database