Informatica Biomedica

lezione16

Alberto*Paoluzzi Mauro*Ceccanti

www.dia.uniroma3.it/ paoluzzi/web/did/biomed/

Informatica e Automazione, "Roma Tre" — Medicina Clinica, "La Sapienza"

May 15, 2010

Informatica Biomedica: Lezione 16

Bionformatics Archives and Information Retrieval
    The archives

fonte essenziale: A- M- Lesk, Introduction to Bioinformatics, Oxford University Press, 3rd Ed- (2008)

Contents

Bioinformatics databases

Database construction in bioinformatics involves activities that can
be classified, to some extent, into

Bioinformatics databases

Database construction in bioinformatics involves activities that can be classified, to some extent, into

- archiving with the major goals of conservation and curation of facts,

Bioinformatics databases

Database construction in bioinformatics involves activities that can be classified, to some extent, into

- archiving with the major goals of conservation and curation of facts,

- interpreting

Bioinformatics databases

Database construction in bioinformatics involves activities that can be classified, to some extent, into

- archiving with the major goals of conservation and curation of facts,

- interpreting

- annotating the compilation of biological information in a form most useful to support research-

Specialized databases

Many archival databases specialize in different kinds of data:

Specialized databases

Many archival databases specialize in different kinds of data:

- nucleic acid sequences, or

for reasons in part historical and in part because of the different curatorial skills required.

Specialized databases

Many archival databases specialize in different kinds of data:

- nucleic acid sequences, or

- protein sequences, or

for reasons in part historical and in part because of the different curatorial skills required.

Specialized databases

Many archival databases specialize in different kinds of data:

- nucleic acid sequences, or

- protein sequences, or

- structures

for reasons in part historical and in part because of the different curatorial skills required.

Primary data collections related to biological macromolecules

Primary data collections related to biological macromolecules

- ► Nucleic acid sequences, including whole-genome projects

Primary data collections related to biological macromolecules

- Nucleic acid sequences, including whole-genome projects
- Amino acid sequences of proteins

Primary data collections related to biological macromolecules

- ▶ Nucleic acid sequences, including whole-genome projects
- ▶ Amino acid sequences of proteins
- ▶ Protein and nucleic acid structures

Primary data collections related to biological macromolecules

- Nucleic acid sequences, including whole-genome projects
- Amino acid sequences of proteins
- Protein and nucleic acid structures
- Small-molecule crystal structures

Primary data collections related to biological macromolecules

- ▶ Nucleic acid sequences, including whole-genome projects
- ▶ Amino acid sequences of proteins
- ▶ Protein and nucleic acid structures
- ▶ Small-molecule crystal structures
- ▶ Protein functions

Primary data collections related to biological macromolecules

- ▶ Nucleic acid sequences, including whole-genome projects
- ▶ Amino acid sequences of proteins
- ▶ Protein and nucleic acid structures
- ▶ Small-molecule crystal structures
- ▶ Protein functions
- ▶ Expression patterns of genes

Primary data collections related to biological macromolecules

- ▶ Nucleic acid sequences, including whole-genome projects
- ▶ Amino acid sequences of proteins
- ▶ Protein and nucleic acid structures
- ▶ Small-molecule crystal structures
- ▶ Protein functions
- ▶ Expression patterns of genes
- ▶ Networks:

Primary data collections related to biological macromolecules

- ▶ Nucleic acid sequences, including whole-genome projects
- ▶ Amino acid sequences of proteins
- ▶ Protein and nucleic acid structures
- ▶ Small-molecule crystal structures
- ▶ Protein functions
- ▶ Expression patterns of genes
- ▶ Networks:
- ▶ of metabolic pathways

Primary data collections related to biological macromolecules

- ▶ Nucleic acid sequences, including whole-genome projects
- ▶ Amino acid sequences of proteins
- ▶ Protein and nucleic acid structures
- ▶ Small-molecule crystal structures
- ▶ Protein functions
- ▶ Expression patterns of genes
- ▶ Networks:
- ▶ of metabolic pathways
- ▶ of gene and protein interactions, and

Primary data collections related to biological macromolecules

- ▶ Nucleic acid sequences, including whole-genome projects
- ▶ Amino acid sequences of proteins
- ▶ Protein and nucleic acid structures
- ▶ Small-molecule crystal structures
- ▶ Protein functions
- ▶ Expression patterns of genes
- ▶ Networks:
- ▶ of metabolic pathways
- ▶ of gene and protein interactions, and
- ▶ of control cascades

Primary data collections related to biological macromolecules

- ► Nucleic acid sequences, including whole-genome projects
- ► Amino acid sequences of proteins
- ► Protein and nucleic acid structures
- ► Small-molecule crystal structures
- ► Protein functions
- ► Expression patterns of genes
- ► Networks:
- ► of metabolic pathways
- ► of gene and protein interactions, and
- ► of control cascades
- ► Publications

Nucleic acid sequence databases

World-wide nucleic-acid sequence archive is an international collaboration among:

Nucleic acid sequence databases

World-wide nucleic-acid sequence archive is an international collaboration among:

- ▶ EMBL Nucleotide Sequence Database (Europe)

Each of the three groups collects a portion of the total sequence data reported worldwide, and all new and updated database entries are exchanged between the groups on a daily basis.

Nucleic acid sequence databases

World-wide nucleic-acid sequence archive is an international collaboration among:

- EMBL Nucleotide Sequence Database (Europe)

- GenBank (USA)

Each of the three groups collects a portion of the total sequence data reported worldwide, and all new and updated database entries are exchanged between the groups on a daily basis.

Nucleic acid sequence databases

World-wide nucleic-acid sequence archive is an international collaboration among:

- EMBL Nucleotide Sequence Database (Europe)

- GenBank (USA)

- DNA Database of Japan (DDBJ)

Each of the three groups collects a portion of the total sequence data reported worldwide, and all new and updated database entries are exchanged between the groups on a daily basis.

Nucleic acid sequence databases

The nucleic acid sequence databases, as distributed, are collections of entries.

Each entry has the form of a text file containing data and annotations for a single contiguous sequence-

Many entries are assembled from several published papers reporting overlapping fragments of a complete sequence.
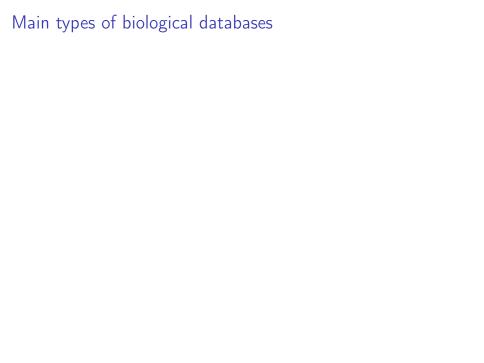
Entries have a life history

Because of the desire on the part of the user community for rapid access to data, new entries are made available before completion of annotation and checking-

Entries mature through the classes:

Unannotated → Preliminary → Unreviewed → Standard

Rarely, an entry dies; a few have been removed when they are determined to be erroneous.

Main types of biological databases

Main types of biological databases

- Genome databases and genome browsers

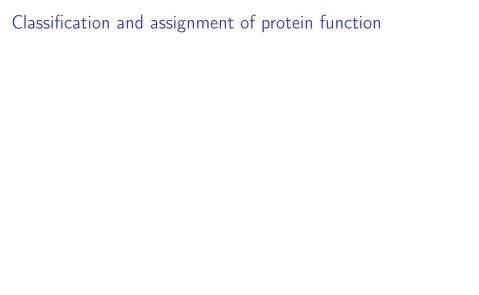# Main types of biological databases

- Genome databases and genome browsers

- Protein sequence databases

Main types of biological databases

- Genome databases and genome browsers

- Protein sequence databases

- Databases of protein families

Main types of biological databases

- Genome databases and genome browsers

- Protein sequence databases

- Databases of protein families

- Databases of structures

Main types of biological databases

- Genome databases and genome browsers

- Protein sequence databases

- Databases of protein families

- Databases of structures

- Classifications of protein structures

Main types of biological databases

- Genome databases and genome browsers

- Protein sequence databases

- Databases of protein families

- Databases of structures

- Classifications of protein structures

- Accuracy and precision of protein structure determinations

Classification and assignment of protein function

Classification and assignment of protein function

- ► The Enzyme Commission

Classification and assignment of protein function

- ▶ The Enzyme Commission

- ▶ The Gene Ontology Consortium protein function classification

Classification and assignment of protein function

- ▶ The Enzyme Commission

- ▶ The Gene Ontology Consortium protein function classification

- ▶ Specialized, or *boutique* databases

Classification and assignment of protein function

- ▶ The Enzyme Commission

- ▶ The Gene Ontology Consortium protein function classification

- ▶ Specialized, or *boutique* databases

- ▶ Expression and proteomics databases

Classification and assignment of protein function

- The Enzyme Commission

- The Gene Ontology Consortium protein function classification

- Specialized, or *boutique* databases

- Expression and proteomics databases

- Databases of metabolic pathways

Classification and assignment of protein function

- ▶ The Enzyme Commission

- ▶ The Gene Ontology Consortium protein function classification

- ▶ Specialized, or *boutique* databases

- ▶ Expression and proteomics databases

- ▶ Databases of metabolic pathways

- ▶ Bibliographic databases

Classification and assignment of protein function

- ▶ The Enzyme Commission

- ▶ The Gene Ontology Consortium protein function classification

- ▶ Specialized, or *boutique* databases

- ▶ Expression and proteomics databases

- ▶ Databases of metabolic pathways

- ▶ Bibliographic databases

- ▶ Surveys of molecular biology databases and servers

Gateways to archives

Access to databases in molecular biology

Gateways to archives

Access to databases in molecular biology

- ENTREZ

Gateways to archives

Access to databases in molecular biology

- ENTREZ

- The Sequence Retrieval Syscem (SRS)

Gateways to archives

Access to databases in molecular biology

- ENTREZ

- The Sequence Retrieval Syscem (SRS)

- The Protein Identification Resource (PIR)

Gateways to archives

Access to databases in molecular biology

- ENTREZ

- The Sequence Retrieval Syscem (SRS)

- The Protein Identification Resource (PIR)

- ExPASy-Expert Protein Analysis System