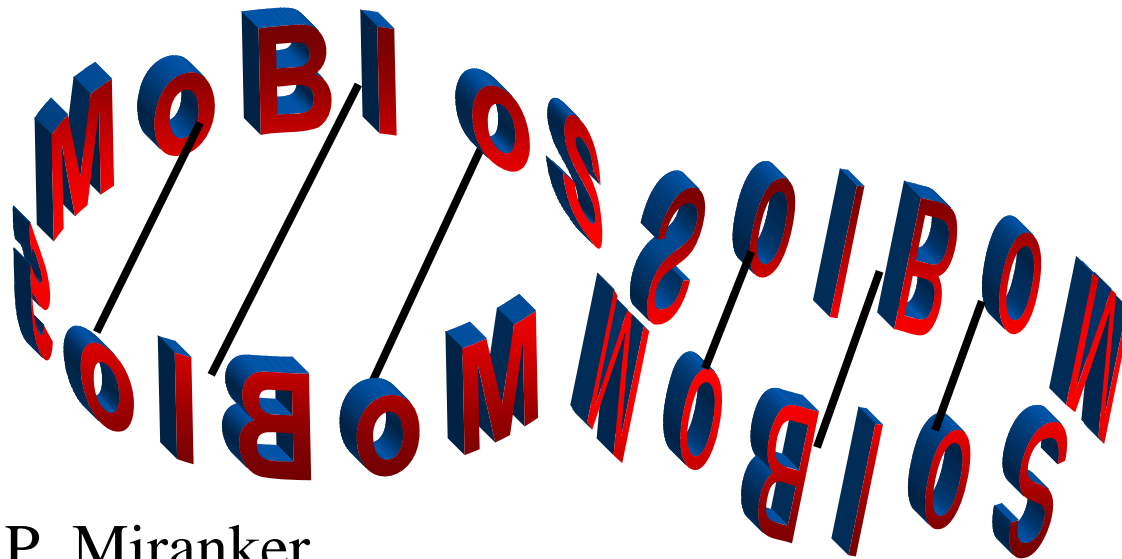


# *The MoBioS Project*

## *Molecular Biological Information System*



Daniel P. Miranker

Dept. of Computer Sciences &

Center for Computational Biology and  
Bioinformatics

University of Texas

**Weijia Xu, Rui Mao, Will**

Briggs, Smriti

Ramakrishnan, Shu Wang,

Shulin Ni, Kai Yan, Ving Lei

# Compared to Business Databases, Biological Databases

- are not that big
  - Genbank ftp mirror download,  $\ll$  1 Terabyte
  - CMS spectrometer at CERN,  $>$  Petabyte/year
- *but,*  
data management in biology is a big problem
- There must be another problem...


# You Can't Sort

- Sequences:
  - DNA, RNA, Protein databases
- Mass Spectra
  - proteomics
- Small Molecules & Protein Structure
  - Protein interaction
  - Rational drug design
- Pathways (graphs)
- Phylogenies (graphs, trees in particular)

# In Life-Sciences Database Management Systems are Souped Up File Systems

- Primary data is stored in text or blob fields

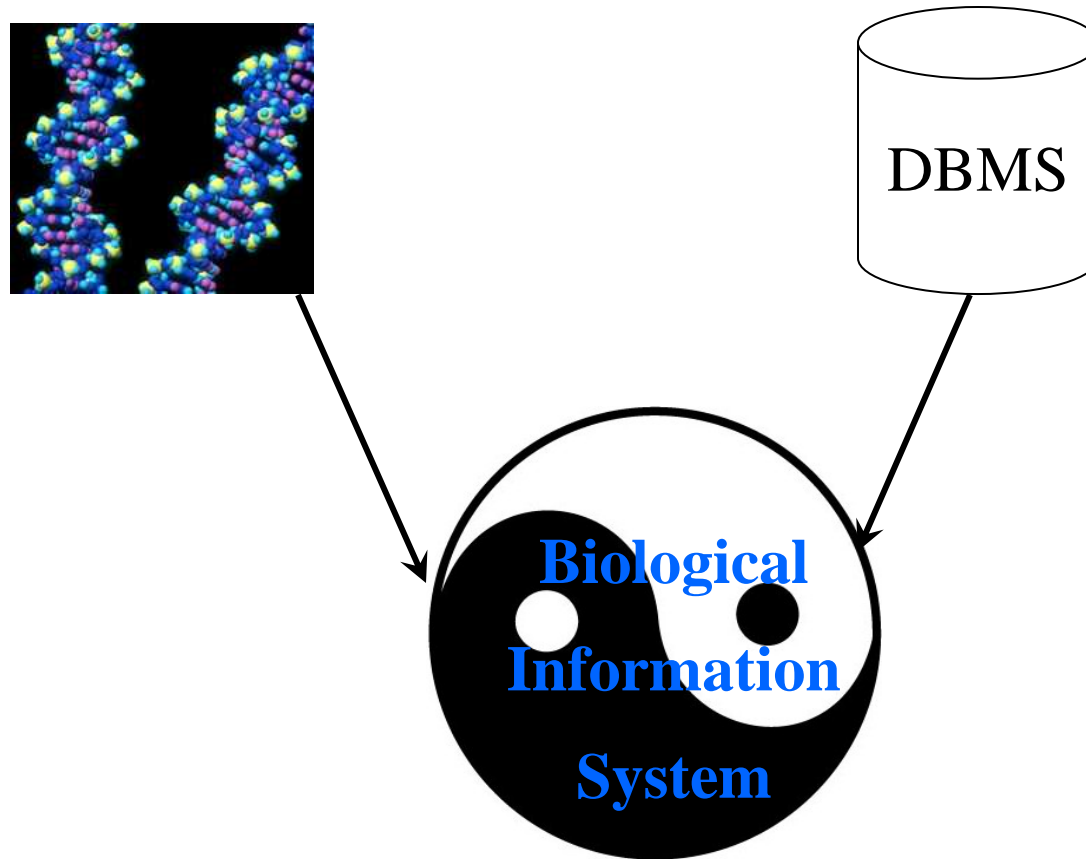
Annotations may be relational



Accession #	Gene name	Organism	Protein sequence
NM_024026	Mitochondrial ribosomal protein (MRP63)	Homo sapiens	<b>MFLTAVLLRGRIP</b>
NM_003973	Ribosomal protein L14(RPL14)	Homo sapiens	<b>VFRRFVEVGRVAY</b>
NM_026401	Mitochondrial protein	Mus musculus	<b>MFLTALLWRGRIP</b>

- Data retrieval
  - Filter DB, sequential dump,  $O(n)$ , to utilities
    - E.g. BLAST,

# Scope: To Find Common Ground Both Biology and DBMS' Have to Move



*Metric-Space Database as the Common Ground*

# Metric Space is

➤ a pair,  $M=(D,d)$ ,

where

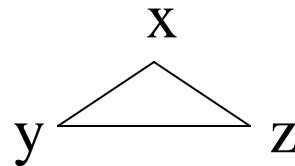
➤  $D$  is a set of points

➤  $d$  is [metric] distance function with the following properties:

➤  $d(x,y) = d(y,x)$  (symmetry)

➤  $d(x,y) > 0, d(x,x) = 0$  (non negativity)

➤  $d(x,z) \leq d(x,y) + d(y,z)$  (triangle inequality)



# Definition - By Analogy

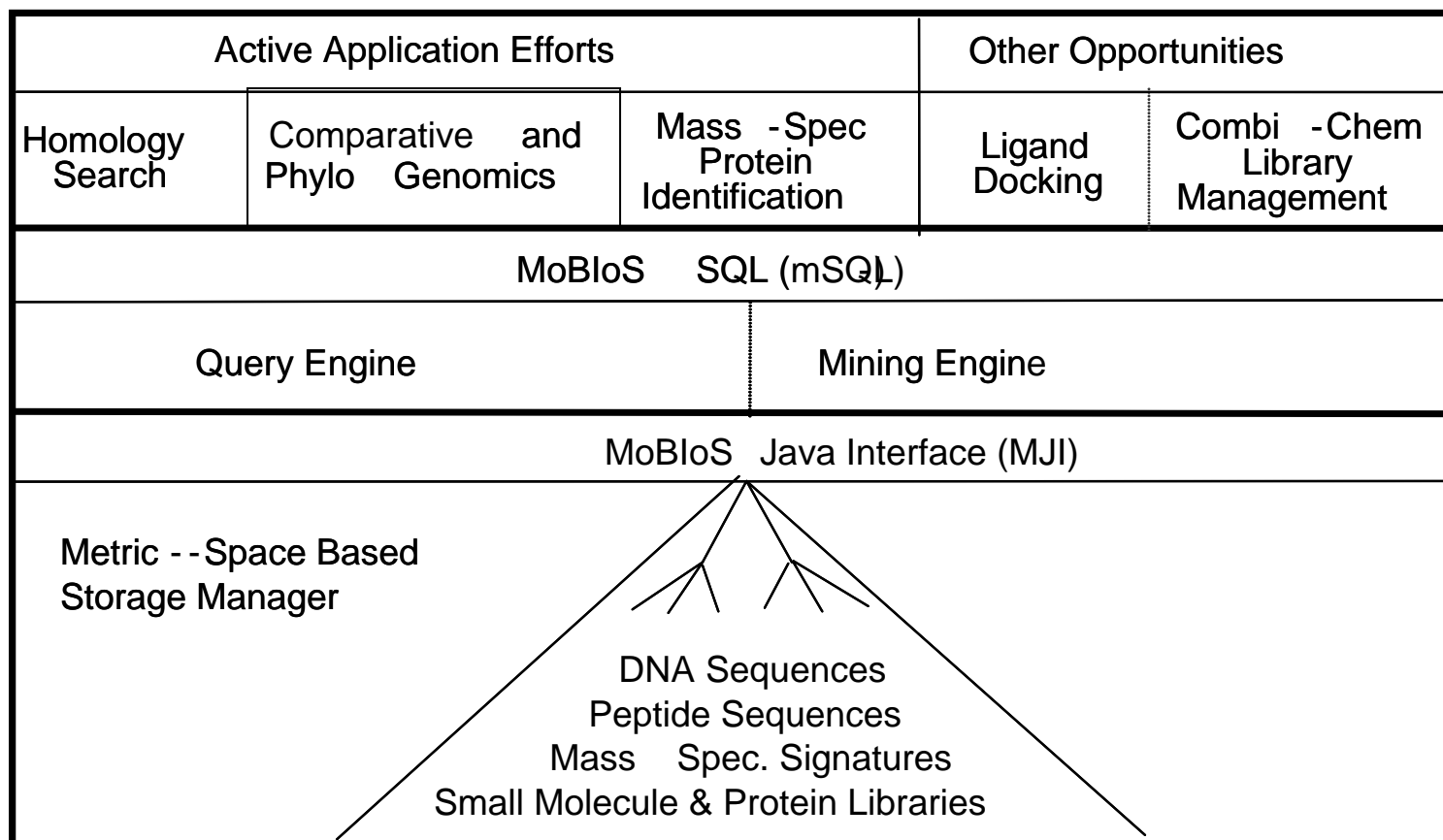
## A Spatial Database Management System:

- Extend relational DBMS
  - Special indexes for 2D and 3D data; k-d and R-trees
  - New data types
- Geographic information systems
  - Topographic maps
  - Buildings and the like

## A Metric-Space Database Management System:

- Extend Relational DBMS
  - Special indexes for metric-spaces
  - New data types
- Biological information system
  - Life science data types

# MoBioS System Overview



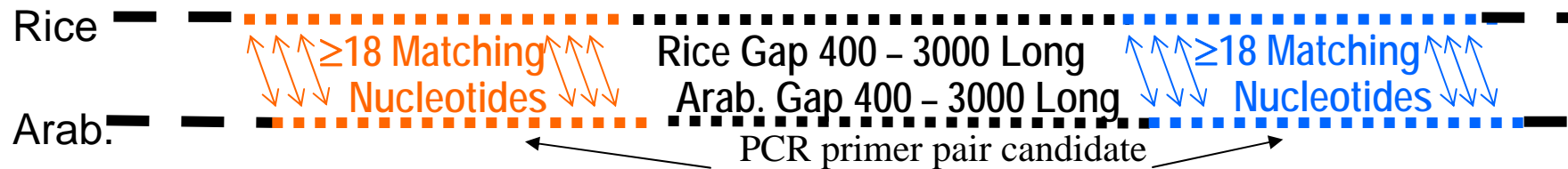
# Results to date:

- Developed and Validated Metrics for
  - protein sequence homology
  - protein mass-spectra
  - volumetric (3-d) models of protein electrostatics
- Developed Applications
  - Comparative study Rice vs. Arab. (8 CPU days)
  - Protein Identification by MS and MS/MS Database lookup

# mSQL: Enables Comparative Genomic Studies, including Secondary Structure

Compare Arabidopsis Genome X Rice Genome

1. Locate nucleotide patterns of form



2. Matching  $\equiv$   $< 5\%$  mismatches
3. Eliminate non-unique primer candidates

```
SELECT merge(R1.fragment, A1.fragment)
FROM Rice_sview R1, Rice_sview R2, Arab_sview A1, Arab_sview A2
WHERE
distance('HAMMINGDISTANCE', R1.fragment, A1.fragment) <= 1.0 AND
distance('HAMMINGDISTANCE', R2.fragment, A2.fragment) <= 1.0 AND
(FRAGOFFSET(R2.fragment)-FRAGOFFSET(R1.fragment)) >= 400 AND
(FRAGOFFSET(R2.fragment)-FRAGOFFSET(R1.fragment)) <= 3000 AND
(FRAGOFFSET(A2.fragment)-FRAGOFFSET(A1.fragment)) >= 400 AND
(FRAGOFFSET(A2.fragment)-FRAGOFFSET(A1.fragment)) <= 3000
GROUP BY R1.fragment, A1.fragment;
```