Web-based Molecular Modeling Using Java/Swarm, J2EE and RDBMS Technologies

Yingping Huang, Gregory Madey
Xiaorong Xiang, Eric Chanowich
University of Notre Dame

Partially supported by NFS-ITR

Research Area and Results

- The domain
    - Scientific simulation
        - Natural organic matter (NOM)
        - Environmental biocomplexity
- The results: A simulation model
    - Agent-based using SWARM
    - Stochastic
    - Web-based: J2EE, XML & Oracle
    - Load-balancing and fail-over enabled
    - Data warehousing & data mining features included

Motivation

- IT: A fourth paradigm of scientific study? (J. Gray, et al, 2002; Fox, 2002)
  - Three previous approaches to scientific research:
    - Observation & theory
    - Hypothesis & experiment
    - Computational X & simulation
  - Information technologies
    - J2EE & middleware & XML
    - Databases & Data Warehouses
    - Data Mining
    - Visualization
    - Statistical analysis
- Natural organic matter (NOM)

Natural Organic Matter

- NOM is ubiquitous in terrestrial, aquatic and marine ecosystems
  - Results from breakdown of animal & plant material in the environment
- Important role in processes such as
  - compositional evolution and fertility of soil
  - mobility and transport of pollutants
  - availability of nutrients for microorganisms and plant communities
  - growth and dissolution of minerals
- Important to drinking water systems
  - Impacts drinking water treatment
  - Impacts quality of well water

Background

- Compositional evolution of NOM is an interesting problem
- Important aspect of predictive environmental modeling
- Prior modeling work is often
  - too simplistic to represent the heterogeneous structure of NOM and its complex behaviors in ecosystems (e.g., carbon cycling models)
  - too compute-intensive to be useful for large-scale environmental simulations (e.g., molecular models employing connectivity maps or electron densities)
- Hence, a Middle Computational Approach is taken …
  - Agent-based & stochastic

Modeling

- Object oriented: Molecules and microbes are objects
  - Molecules and microbes have attributes
    - Heterogeneous mixture: different attributes
  - Molecules have behaviors (physical & chemical processes)
    - Behaviors are stochastically determined
    - Dependent on the:
      - Attributes (intrinsic parameters)
      - Environment (extrinsic parameters)

Modeling (cont)

- Objects of interest
  - Macromolecular precursors: large molecules
    - Cellulose
    - Proteins
    - Lignin
  - Micromolecules: smaller molecules
    - Sugars
    - Amino acids
  - Microbes
    - Bacteria
    - Fungi

Modeling (cont)

- Attributes
  - Elemental composition
    - Number of C, H, O, N, S and P atoms in molecule
  - Functional group counts
    - Double-bonds
    - Ring structures
    - Phenyl groups
    - Alcohols
    - Phenols, ethers, esters, ketones, aldehydes, acids, aryl acids, amines, amides, thioethers, thiols, phosphoesters, phosphates
  - The time the molecule entered the system
  - Precursor type of molecule
    - Cellulose, protein, lignin, etc

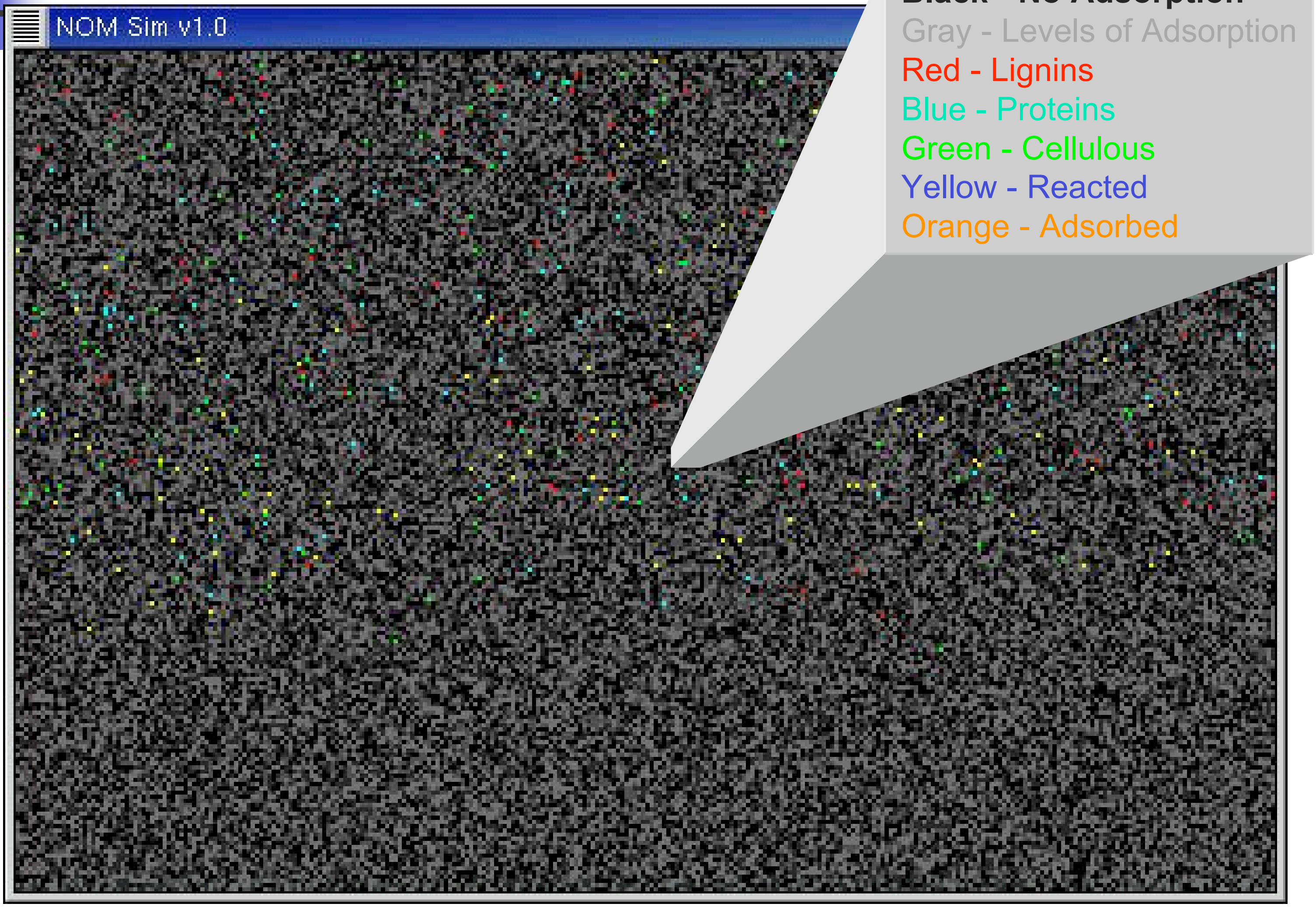
Modeling (cont)

- Behaviors (reactions and processes)
  - Physical processes
    - Adsorption (stick) to mineral surfaces
    - Aggregation/micelle formation
    - Transport downstream (surface water)
    - Transport through porous media
  - Chemical reactions
    - Abiotic bulk reactions: free molecules
    - Abiotic surface reactions: adsorbed molecules
    - Extracellular enzyme reactions on large molecules
    - Microbial uptake by small molecules

Modeling (cont)

- Environmental parameters
  - Temperature
  - pH
  - Light intensity
  - Simulation time
  - Microbial activity
  - Water flow rate/pressure gradient
  - Oxygen density

GUI Animation



Black - No Adsorption

Gray - Levels of Adsorption

Red - Lignins

Blue - Proteins

Green - Cellulous

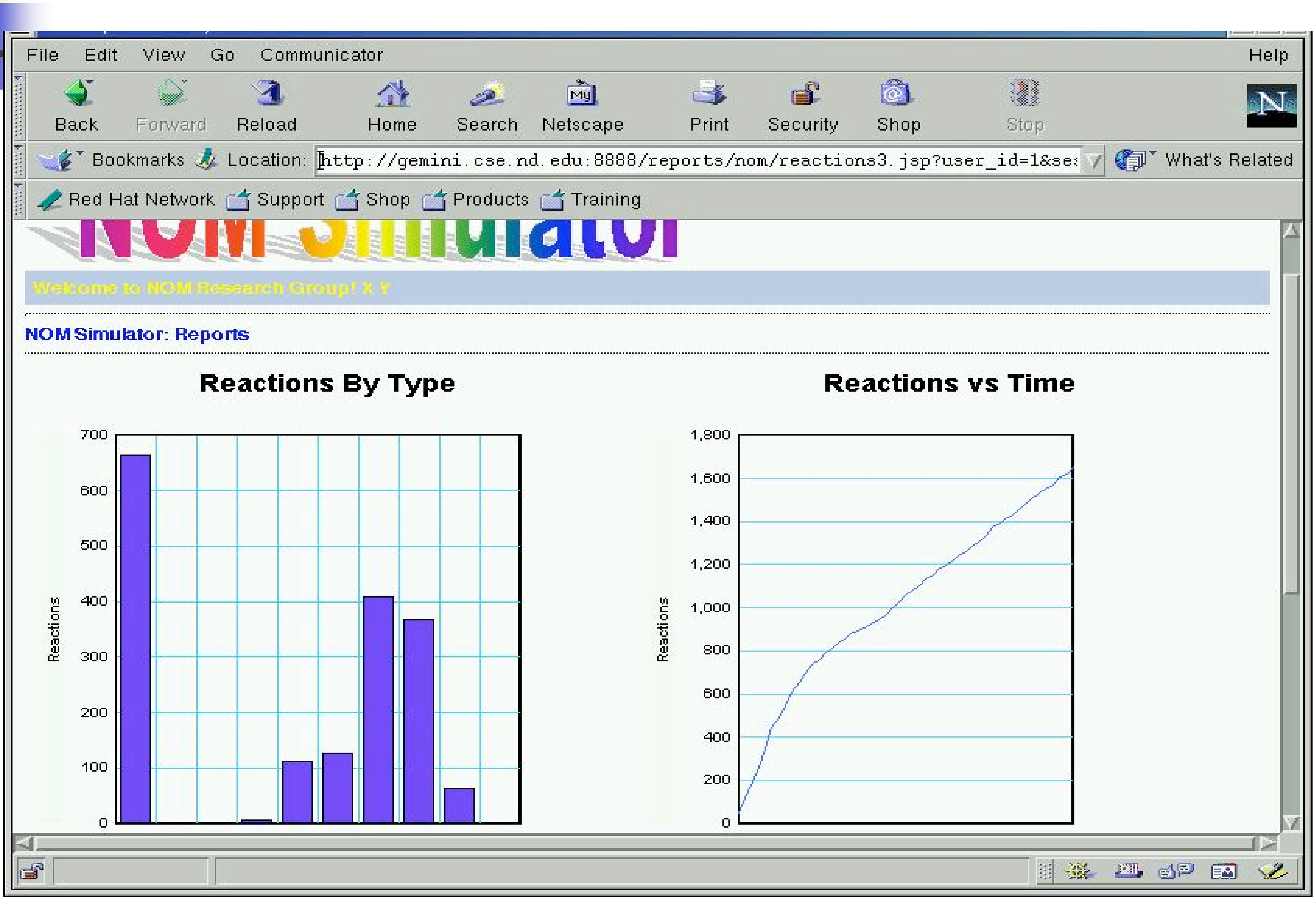
Yellow - Reacted

Orange - Adsorbed

NOM 1.0

- Loosely coupled distributed systems
  - 5Application servers (OC4J Servers)
  - 3 Database servers (Oracle: Data Warehouse, Standby Database)
  - Reports server (OC4J Server/Reports Server)
- Load balancing (implemented by JMS, AQ and MDB)
  - application servers
- Fail over
  - application servers & database servers
  - Multi-master replication of important tables
- Why fail-over (Assume down probability p for each machine)
  - No fail-over
    - Simulation system down probability: $1-(1-p)^2 = 2p-p^2$
  - With fail-over
    - Simulation system down probability: $1-(1-p^5)(1-p^2) = p^2 + p^5 - p^7$
  - Improvement:
    - $2/p = 200$ if $p=0.01$ (the smaller p, the larger improvement)

Sample Reports



Location: http://gemini.cse.nd.edu:8888/reports/nom/reactions3.jsp?user_id=1&se:

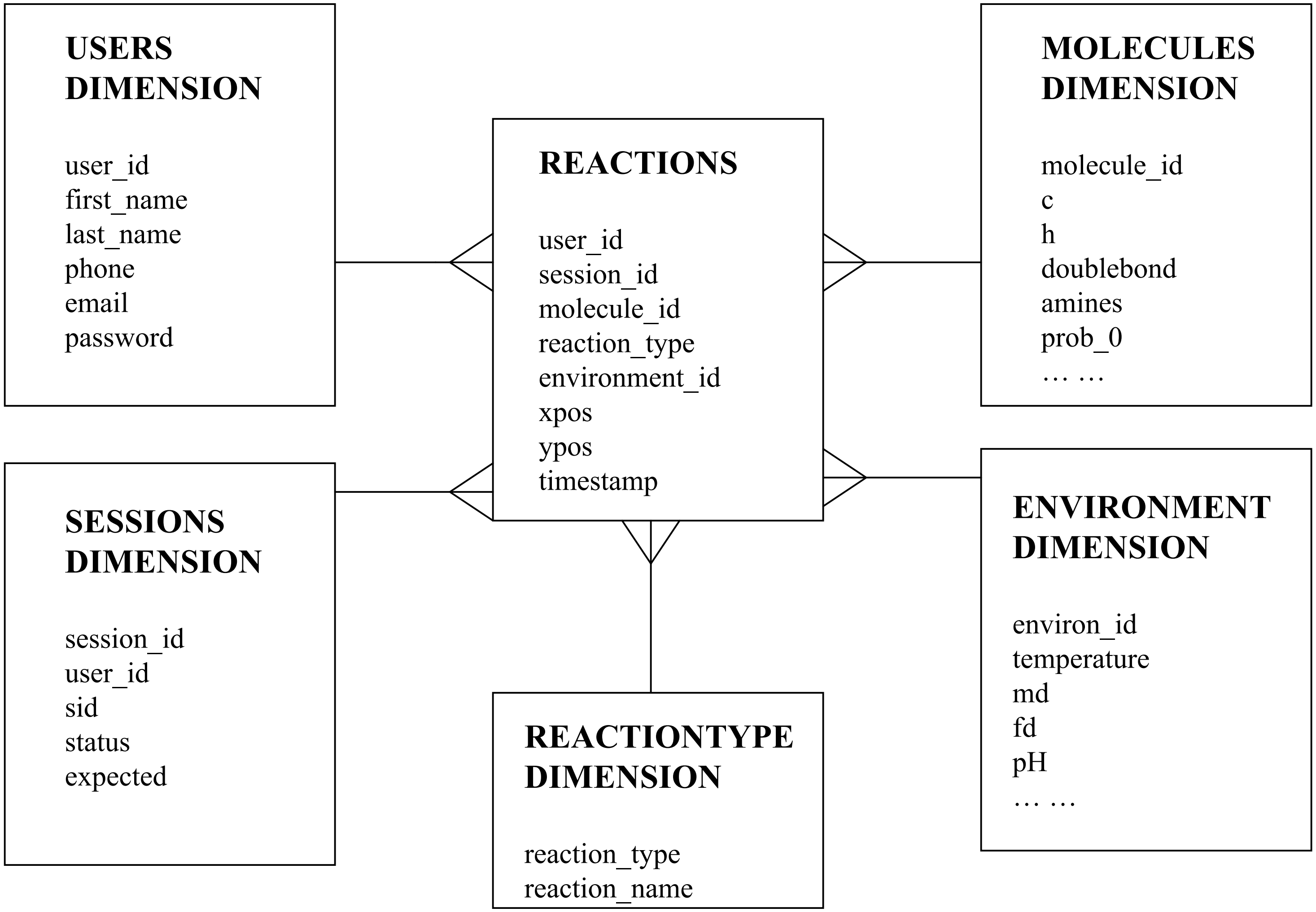
Welcome to NOM Research Group! X Y

NOM Simulator: Reports

Reactions By Type

Reactions vs Time

Data Warehousing: Star Schema



USERS DIMENSION

user_id
first_name
last_name
phone
email
password

REACTIONS

user_id
session_id
molecule_id
reaction_type
environment_id
xpos
ypos
timestamp

MOLECULES DIMENSION

molecule_id
c
h
doublebond
amines
prob_0
… …

SESSIONS DIMENSION

session_id
user_id
sid
status
expected

REACTIONTYPE DIMENSION

reaction_type
reaction_name

ENVIRONMENT DIMENSION

environ_id
temperature
md
fd
pH
… …

Data Mining: Applying Clustering

- Model-build data format
  - A table POINTS with attributes x & y
    - Points are chosen from the data warehouse
    - Standardized: x & y are in [0,1)
    - 16 million records
- Clusters explanation
  - Dense areas in soil or solution
  - Emerging behavior of random molecules (e.g. Micelles)

Summary

- Contributions are
  - New models which treats NOM as a heterogeneous mixture using SWARM
  - Simulation system with advanced web & database tools: J2EE, XML &Oracle
  - System aspects of implementation of load-balancing and fail-over using JMS, AQ, MDB, JTA, etc.
  - Data warehousing for simulation data and experimental data
  - Applying data mining to simulation data and experimental data