

Online Collaboratory for NOM Research: Agent-based Simulations, Data-Mining, and Knowledge-Discovery

Madey, G.R., Cabaniss, S.E.,
Maurice, P.A., Xiang, X., Arthurs,
L., Kennedy, R., Huang, Y

University of Notre Dame
University of New Mexico

ASLO 2005

February 24, 2005

Overview

- Project Background
- E-Science Background
- The NOM e-Science Collaboratory
- Invitation to Test, Contribute, Participate
- Summary

Background

- Small NSF-ITR involving Computer Scientists and Environmental Scientists
- Focus: “Stochastic Synthesis: Simulating the environmental transformations of NOM”
- IT Focus: e-Science, Web-Based Science, Agent-Based Simulation, Data Mining

E-Science Background

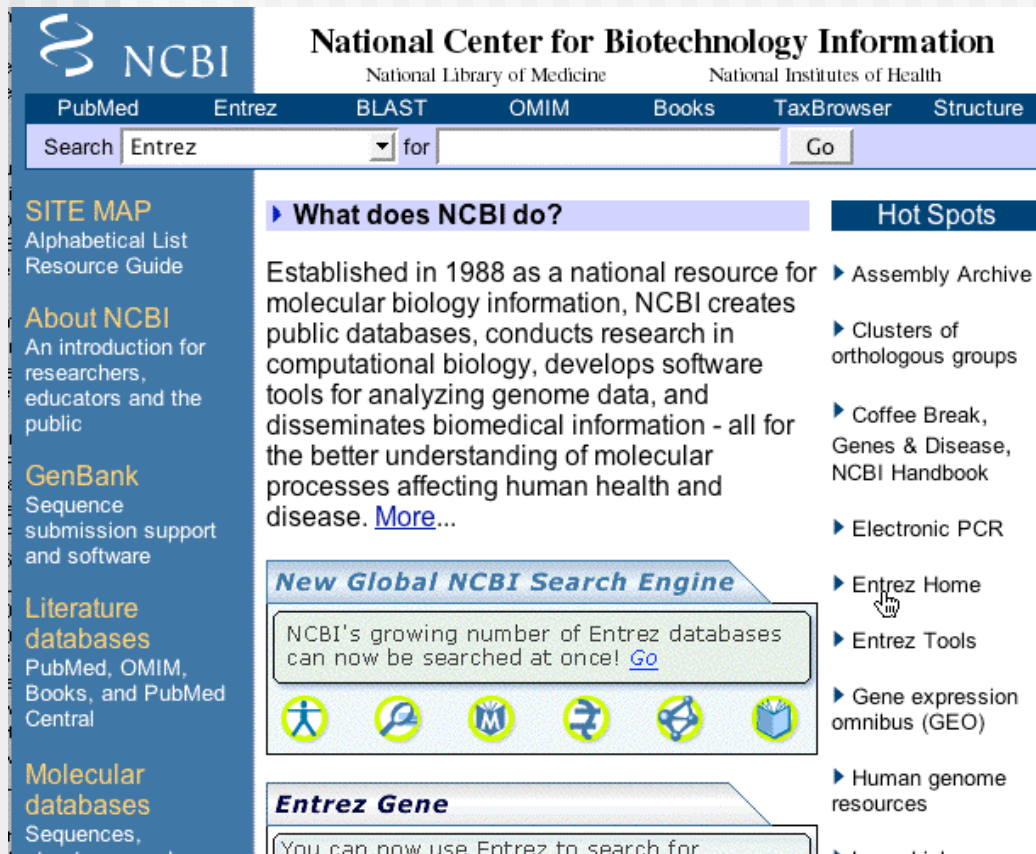
- NSF Cyberinfrastructure Program
 - Sensor networks
 - Large amounts of data => discovery through datamining
 - Online linked data repositories
 - Online analysis programs: search, extraction, matching, simulation, visualizations, etc.

E-Science Background (cont)

- WWW Telescope
- Virtual Observatories
 - NASA Skyview
 - US National Virtual Observatory
 - International Virtual Observatory Alliance
- More and more research done without every using a telescope => use of distributed data already collected

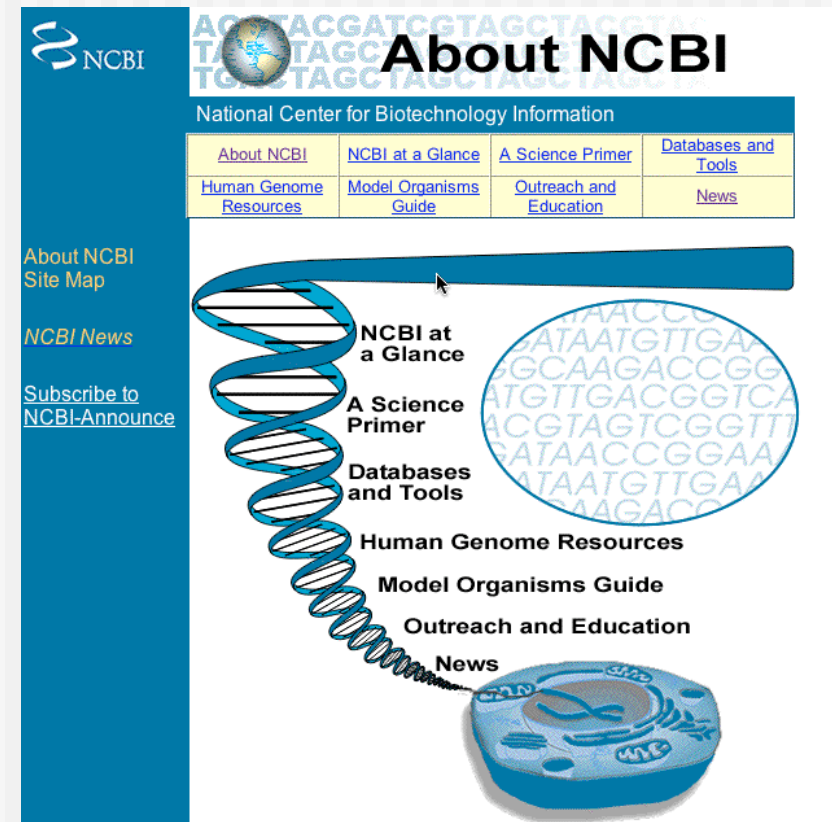
E-Science Background (cont)

■ NCBI (National Center for Biotechnology Information)



The screenshot shows the NCBI homepage with the following elements:

- Header:** NCBI logo, National Library of Medicine, National Institutes of Health.
- Navigation:** PubMed, Entrez, BLAST, OMIM, Books, TaxBrowser, Structure.
- Search:** Search Entrez for [] Go
- SITE MAP:** Alphabetical List, Resource Guide.
- About NCBI:** An introduction for researchers, educators and the public.
- GenBank:** Sequence submission support and software.
- Literature databases:** PubMed, OMIM, Books, and PubMed Central.
- Molecular databases:** Sequences.
- What does NCBI do?:** Established in 1988 as a national resource for molecular biology information, NCBI creates public databases, conducts research in computational biology, develops software tools for analyzing genome data, and disseminates biomedical information - all for the better understanding of molecular processes affecting human health and disease. [More...](#)
- Hot Spots:**
 - Assembly Archive
 - Clusters of orthologous groups
 - Coffee Break, Genes & Disease, NCBI Handbook
 - Electronic PCR
 - Entrez Home
 - Entrez Tools
 - Gene expression omnibus (GEO)
 - Human genome resources
- New Global NCBI Search Engine:** NCBI's growing number of Entrez databases can now be searched at once! [Go](#)
- Entrez Gene:** You can now use Entrez to search for



The diagram illustrates the 'About NCBI' page structure with a DNA double helix on the left and a circular graphic on the right containing DNA sequence motifs. The central text lists the following resources:

- NCBI at a Glance
- A Science Primer
- Databases and Tools
- Human Genome Resources
- Model Organisms Guide
- Outreach and Education
- News

Navigation links on the left include: About NCBI Site Map, NCBI News, and Subscribe to NCBI-Announce.

About NCBI	NCBI at a Glance	A Science Primer	Databases and Tools
Human Genome Resources	Model Organisms Guide	Outreach and Education	News








Search across databases



GO

CLEAR

Help

Welcome to the new Entrez cross-database search page

 PubMed: biomedical literature citations and abstracts ?	 Books: online books ?
 PubMed Central: free, full text journal articles ?	 OMIM: online Mendelian Inheritance in Man ?
	 Site Search: NCBI web and FTP sites ?

 Nucleotide: sequence database (GenBank) ?	 UniGene: gene-oriented clusters of transcript sequences ?
 Protein: sequence database ?	 CDD: conserved protein domain database ?
 Genome: whole genome sequences ?	 3D Domains: domains from Entrez Structure ?
 Structure: three-dimensional macromolecular structures ?	 UniSTS: markers and mapping data ?
 Taxonomy: organisms in GenBank ?	 PopSet: population study data sets ?
 SNP: single nucleotide polymorphism ?	 GEO Profiles: expression and molecular abundance profiles ?
 Gene: gene-centered information ?	 GEO DataSets: experimental sets of GEO data ?
 HomoloGene: eukaryotic homology groups ?	 Cancer Chromosomes: cytogenetic databases ?
 PubChem Compound: small molecule chemical structures ?	 PubChem BioAssay: bioactivity screens of chemical substances ?
 PubChem Substance: chemical substances screened for bioactivity ?	 GENSAT: gene expression atlas of mouse central nervous system ?

NEW 15 Nov 2004 Download the [BLAST poster](#) from [SC2004!](#)

About BLAST

- News
- Mailing list
- References
- NCBI Contributors

BLAST Services

- FAQs
- Program selection guide
- Web service interface

BLAST Software

- Databases
- Documentation
- Errata
- Executables
- Source code

Support

- Contact us

Nucleotide

- Quickly search for highly similar sequences (megablast)
- Quickly search for divergent sequences (discontiguous megablast)
- Nucleotide-nucleotide BLAST (blastn)
- Search for short, nearly exact matches
- Search trace archives with megablast or discontiguous megablast

Translated

- Translated query vs. protein database (blastx)
- Protein query vs. translated database (tblastn)
- Translated query vs. translated database (tblastx)

Special

- Search for gene expression data (GEO BLAST)
- Align two sequences (bl2seq)
- Screen for vector contamination (VecScreen)
- Immunoglobulin BLAST (IgBlast)
- SNP BLAST **NEW**

Protein

- Protein-protein BLAST (blastp)
- PHI- and PSI-BLAST
- Search for short, nearly exact matches
- Search the conserved domain database (rpsblast)
- Search by domain architecture (cdart)

Genomes

- Chicken, cow, pig, dog, sheep, cat
- Environmental samples
- Human, mouse, rat
- Fugu rubripes, zebrafish
- Insects, nematodes, plants, fungi, malaria
- Microbial genomes, other eukaryotic genomes

Meta

- Retrieve results by RID

Model Organism e-Science Sites

- FlyBase: *Drosophila*
- WormBase: *C. elegans*
- VectorBase: Mosquitos
- Mouse Genome
- DictyBase
- Etc.
- So many => GMOD

Questions?

- Has this research community begun to participate in the e-Science initiatives?
- Would this community benefit from e-Science initiatives?
- Is this community interested in an e-Science initiatives?
- Is this community willing to experiment?

The NOM e-Science Collaboratory

- Web-based
- Back-end database
- A cluster of simulation servers
- Shared simulation results
- Shared simulation configurations
- Other collaboratory features

Stochastic Synthesis: Simulating the Environmental Transformations of Natural Organic Matter

[\(Project Overview - slides\)](#)

Principal Investigators

Steve Cabaniss
Chemistry
University of New Mexico
cabaniss@unm.edu

Jerry Leenheer
US Geological Survey
Denver, CO
leenheer@usgs.gov

Laura Leff
Biology
Kent State University
lleff@kent.edu

[Greg Madey](#)
Computer Science & Engineering
University of Notre Dame
gmadey@nd.edu

[Patricia Maurice](#)
Civil Engineering & Geological Sciences
[Center for Environmental Science & Technology](#)
University of Notre Dame
pmaurice@nd.edu

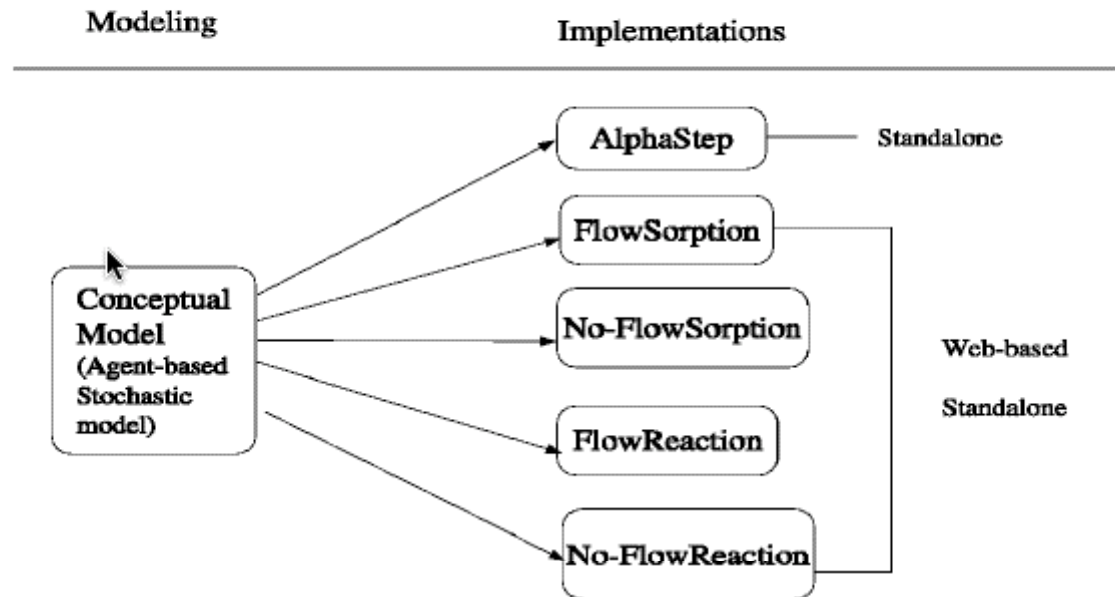
Robert Wershaw
US Geological Survey
Denver, CO
rwershaw@usgs.gov

Robert Wetzel
Biology

<http://www.nd.edu/~nom>

Nom Simulators

- Home
- Overview
- Papers
- Software**
- People
- Animations
- Photos



AlphaStep

AlphaStep is a reference implementation that is coded in Delphi 6 and runs under Windows. It is a demonstration of the NOM conceptual model that doesn't have web and collaboration features. AlphaStep simulates a variety of chemical and biological transformations, but does not simulate any type of transport and does not represent the spatial properties of NOM. AlphaStep is intended as a stand-alone application to allow ecologists, geochemists and environmental scientists to explore possible routes of NOM transformation. AlphaStep can be downloaded below:

- [AlphaStep.exe](#) (version 12/2003)
- [AlphaStep Users Guide](#)
- [AlphaStep FAQ](#)

Web-Based Simulations

The other four implementations are coded using Java programming language (Sun JDK 1.4.2) and Swarm and Repast software. Swarm is a software package for simulating complex systems that was developed at the Santa Fe Institute. It is a set of libraries that

Software download/online

Simulation HomePage: <http://tobit.cse.nd.edu/>

NOM Simulation



[Home](#)



[My NOM](#)



[Administrator](#)



[Sign In](#)



[Help](#)



[Contact Us](#)

Welcome to NOM Simulation

Available simulation models:

SorptionFlowModel

This is the sorption flow model. Administrator can update the description with detailed information

[>New simulation](#)

SorptionBatchModel

This is the sorption batch model

[>New simulation](#)

ReactionFlowModel

This is the reaction flow model

[>New simulation](#)

ReactionBatchModel

This is the reaction batch model

[>New simulation](#)

Welcome to NOM Simulation, where we can offer scientists convenient online simulations of natural organic matter (NOM).

We use the state-of-the-art agent-based stochastic simulation methods to model the behavior of natural organic matter.

These simulations also employ autonomic computing technology for self-management.

Please refer to the [project home page](#) for more information.

Existing Users

Enter your username/ password here

Username

Password

[Login](#)

New User ? Sign up [here](#)

Acknowledgement: The material presented at this web site is based in part upon work supported by the National Science Foundation, Information Technology Research/(ITR/AP-DEB), under Grant No. 0112820. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

Reaction Batch Model

The screenshot shows a web browser window titled "untitled" with the address bar containing the URL <http://tobit.cse.nd.edu/nom5/newsim.do?modelName=Reacti>. The browser's navigation bar includes back, forward, refresh, stop, and home buttons. The main content area features the "NOM Simulation" logo in a stylized font. To the right of the logo is a navigation menu with icons and labels for "Home", "My NOM", "Administrator", "Sign Out", "Help", and "Contact Us". Below the logo, a dark blue banner displays the text "Welcome Greg Madey to NOM Simulation". Underneath this banner is a progress indicator consisting of four circles connected by a horizontal line; the first circle is filled, while the others are empty. Below the progress indicator, the steps "Introduction", "Environment", "Molecules", and "Confirmation" are listed. A blue text block reads: "Welcome to the ReactionBatchModel simulation. This wizard will guide you to provide inputs for the new simulation. A new simulation ID will be used to identify your new simulation. Do you want to continue?". At the bottom right, there are two buttons: "Cancel" and "Continue", with the text "step 1 of 4" positioned between them.

untitled

<http://tobit.cse.nd.edu/nom5/newsim.do?modelName=Reacti>

NOM Simulation

[Home](#) [My NOM](#) [Administrator](#) [Sign Out](#) [Help](#) [Contact Us](#)

Welcome Greg Madey to NOM Simulation

Introduction Environment Molecules Confirmation

Welcome to the ReactionBatchModel simulation. This wizard will guide you to provide inputs for the new simulation. A new simulation ID will be used to identify your new simulation. Do you want to continue?

Cancel step 1 of 4 Continue

NOM Simulation



Home



My NOM



Administrator



Sign Out



Help



Contact Us

Welcome Greg Madey to NOM Simulation



A new simulation ID 3610 has been assigned to identify your new ReactionBatchModel simulation. In this page, you need to specify environmental parameters for your simulation.

Physical/Chemical Conditions:

pH	<input type="text"/>	I (E m ⁻²)	<input type="text"/>
O ₂ (mM)	<input type="text"/>	Celsius T	<input type="text"/>
Water	<input type="text"/>		

Bacterial Conditions:

Bacterial Density	<input type="text"/>	Protease	<input type="text"/>
Oxidase	<input type="text"/>	Decarboxylase	<input type="text"/>

Batch Information:

Simulation Time (hours)	<input type="text"/>	Time Step (hours: delta-T)	<input type="text"/>
Sample Interval (steps)	<input type="text"/>	Random Seed	<input type="text"/>

Step 2 of 4

NOM Simulation



[Home](#)



[My NOM](#)



[Administrator](#)



[Sign Out](#)



[Help](#)



[Contact Us](#)

Welcome Greg Madey to NOM Simulation



You have provided environment parameters for your new simulation in previous page. In this page, you need to specify molecules and their percentages. To create a new molecule type, press the **Create** button. To see the definition of a molecule type, move your mouse over the corresponding . Remember, your simulation ID is 3610.

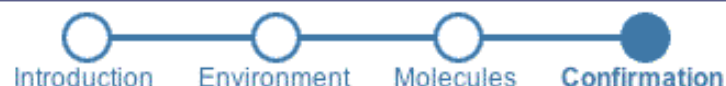
Select	Molecule Type ID	Molecule Name	Percentage
<input checked="" type="checkbox"/>	13	Cellulose	100
<input type="checkbox"/>	14	Lignin	
<input type="checkbox"/>	15	Protein	
<input type="checkbox"/>	17	Zero	
<input type="checkbox"/>	25	Cellulose2	
<input type="checkbox"/>	26	Lignin2	
<input type="checkbox"/>	27	Protein2	
<input type="checkbox"/>	28	Greg	
<input type="checkbox"/>	44	Terpene2	

Note: Click a checkbox and the focus will automatically move to the corresponding text field for percentage. The text field is not editable if the corresponding checkbox is not checked. Uncheck a checkbox will make the corresponding text field empty. The sum of the percentages must be exactly 100.

NOM Simulation

[Home](#)[My NOM](#)[Administrator](#)[Sign Out](#)[Help](#)[Contact Us](#)

Welcome Greg Madey to NOM Simulation



You have specified all necessary inputs for your new **ReactionBatchModel** simulation with ID 3610. Please press the finish button to confirm your submission. If you want to cancel your simulation, press the cancel button.

[Cancel](#) step 4 of 4 [Finish](#)

NOM Simulation

[Home](#)[My NOM](#)[Administrator](#)[Sign Out](#)[Help](#)[Contact Us](#)

Welcome Greg Madey to NOM Simulation

Dear Greg Madey:

Your new **ReactionBatchModel** simulation with simulation ID **3610** has been submitted at **02/24/2005 01:19:37**. Please save the simulation ID and check back again.

NOM Simulation

[Home](#)[My NOM](#)[Administrator](#)[Sign Out](#)[Help](#)[Contact Us](#)

Welcome Greg Madey to NOM Simulation

Simulation Inputs

Environment Parameters for ReactionBatchModel

	Ph	I	O2	CelsiusT	Water	BacterialDensity	Protease	Oxidase	Decarboxylase	ReactionTime	DeltaT	SampleInterval	UseSeed	S
3603	7	0.0001	0.0001	298	1	0.1	0.1	0.1	0.1	1000.25	0.25	1	1	

Molecule Parameters for ReactionBatchModel

SimulationId	MoleculeId	Name	Percentage	C	H	N	O	S	P	Doublebond	Rings	Phenyl	Alcohols	Phenols	Ethers	Esters	Ketones	A
3603	25	Cellulose2	34	360	602	0	301	0	0	0	60	0	182	0	119	0	0	
3603	26	Lignin2	33	400	402	0	81	0	0	160	40	40	2	1	79	0	0	
3603	27	Protein2	33	240	382	60	76	0	0	15	5	5	10	0	0	0	0	

Simulation Reports

Graphical Reports: Built by JFreeChart

[Go to the graphical reports page](#)

NOM Simulation



Welcome Greg Madey to NOM Simulation

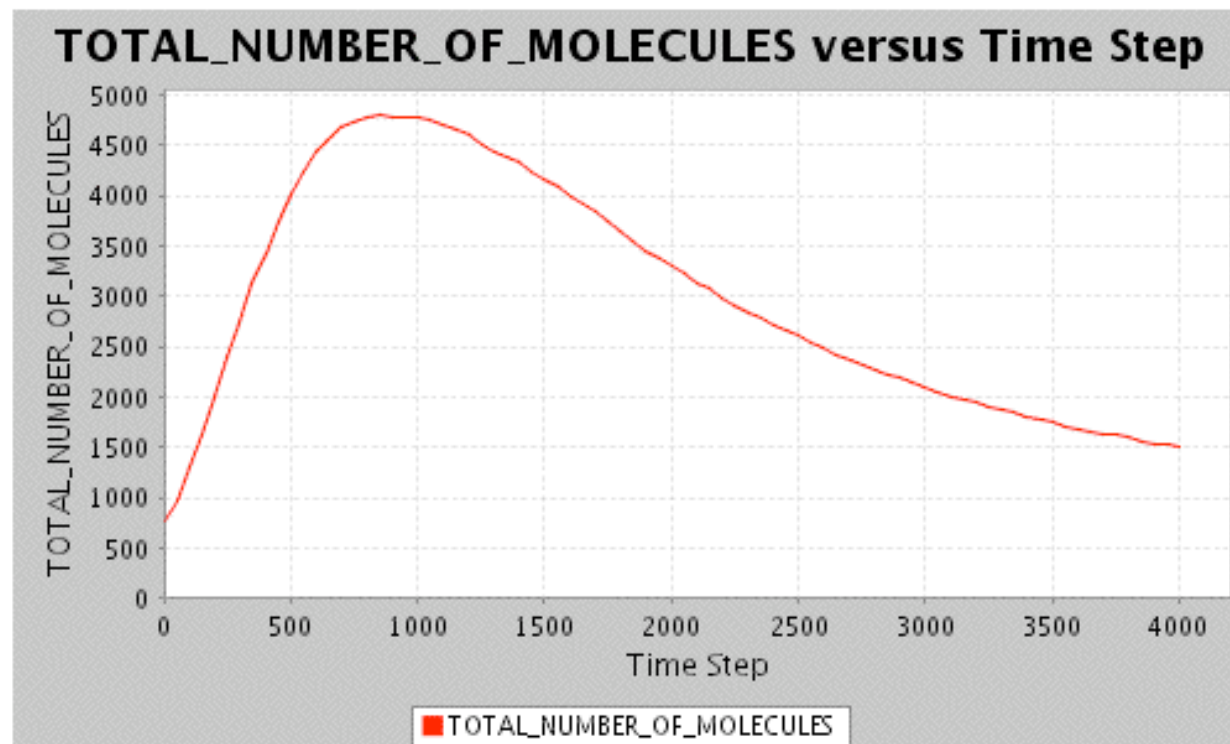
Please choose a report name, specify sample interval and click the Get Report button. For response, sample interval should be reasonably large.

TOTAL_NUMBER_OF_MOLECULES vs TimeStep ▾

Sample Interval: 50

Get Report

Here comes the Report:



NOM Simulation



[Home](#)



[My NOM](#)



[Administrator](#)



[Sign out](#)

Welcome Greg Madey to NOM Simulation

Please choose a report name, specify sample interval and click the Get Report button. For response, sample interval should be reasonably large.

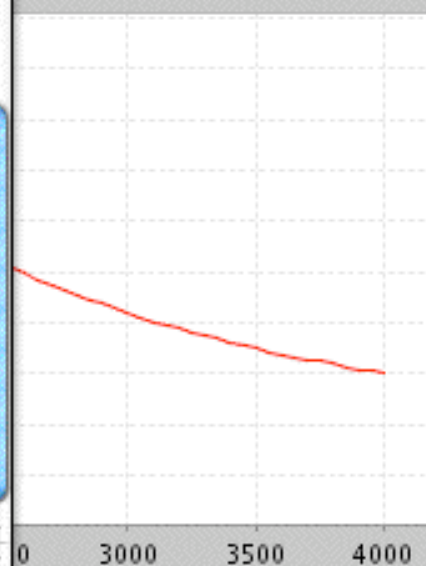
TOTAL_NUMBER_OF_MOLECULES vs TimeStep

Sample Interval:

[Get Report](#)

- TOTAL_MASS_S vs TimeStep
- TOTAL_MASS_P vs TimeStep
- PERCENT_C vs TimeStep
- PERCENT_H vs TimeStep
- PERCENT_N vs TimeStep
- PERCENT_O vs TimeStep
- PERCENT_S vs TimeStep
- PERCENT_P vs TimeStep
- ESTER_CONDENSATION vs TimeStep
- ESTER_HYDROLYSIS vs TimeStep
- AMIDE_HYDROLYSIS vs TimeStep
- MICROBIAL_UPTAKE vs TimeStep
- DEHYDRATION vs TimeStep
- CC_STRONG_OXIDATION vs TimeStep
- CC_WEAK_OXIDATION vs TimeStep
- ALCOHOL_OXIDATION vs TimeStep
- ALDEHYDE_OXIDATION vs TimeStep
- DECARBOXYLATION vs TimeStep
- HYDRATION vs TimeStep
- ALDOL_CONDENSATION vs TimeStep

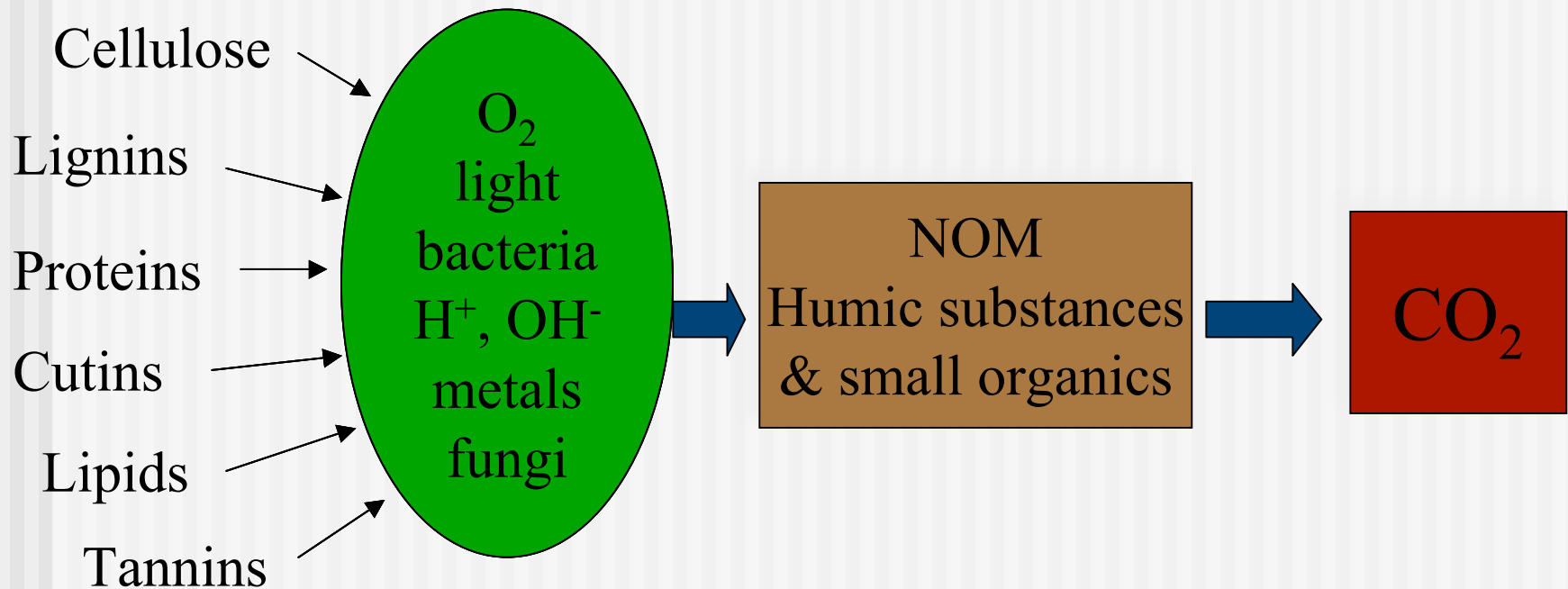
versus Time Step



Time Step

■ TOTAL_NUMBER_OF_MOLECULES

Agent-Based Modeling of NOM



Goal: A widely available, testable, mechanistic model of NOM evolution in the environment.

Data model

Pseudo-Molecule

```
graph TD; A[Pseudo-Molecule] --> B[Elemental Functional Structural Composition]; A --> C[Calculated Chemical Properties and Reactivity]; A --> D[Location Origin State];
```

**Elemental
Functional
Structural
*Composition***

**Calculated
Chemical
Properties
and *Reactivity***

**Location
Origin
State**

Environmental Parameters

Physical:
Temperature
Light Intensity

Biological:
Bacterial Density
Oxidase Activity
Protease Activity
Decarboxylase Activity

Chemical:
Water
pH
[O₂]

Invitation to Test, Contribute, Participate

- [HTTP://www.nd.edu/~nom/](http://www.nd.edu/~nom/)
- [HTTP://tobit.cse.nd.edu/](http://tobit.cse.nd.edu/)

Summary

- Growing phenomenon of e-Science based research
- One small “environmental science” e-Science site: <http://tobit.cse.nd.edu>
- Invitation to test, download, evaluate, contribute, build your own, etc.
- Description of Agent-based NOM Simulator (downloadable and online) by Steve Cabaniss next!