

**02 INFORMATION ABOUT PRINCIPAL INVESTIGATORS/PROJECT DIRECTORS(PI/PD) and  
co-PRINCIPAL INVESTIGATORS/co-PROJECT DIRECTORS**

---

Submit only ONE copy of this form for each PI/PD and co-PI/PD identified on the proposal. The form(s) should be attached to the original proposal as specified in GPG Section II.C.a. Submission of this information is voluntary and is not a precondition of award. This information will not be disclosed to external peer reviewers. **DO NOT INCLUDE THIS FORM WITH ANY OF THE OTHER COPIES OF YOUR PROPOSAL AS THIS MAY COMPROMISE THE CONFIDENTIALITY OF THE INFORMATION.**

---

**PI/PD Name:** Dennis E Shasha

**Gender:**  Male  Female  
**Ethnicity:** (Choose one response)  Hispanic or Latino  Not Hispanic or Latino

**Race:**  
(Select one or more)  
 American Indian or Alaska Native  
 Asian  
 Black or African American  
 Native Hawaiian or Other Pacific Islander  
 White

**Disability Status:**  
(Select one or more)  
 Hearing Impairment  
 Visual Impairment  
 Mobility/Orthopedic Impairment  
 Other  
 None

**Citizenship:** (Choose one)  U.S. Citizen  Permanent Resident  Other non-U.S. Citizen

**Check here if you do not wish to provide any or all of the above information (excluding PI/PD name):**

**REQUIRED: Check here if you are currently serving (or have previously served) as a PI, co-PI or PD on any federally funded project**

---

**Ethnicity Definition:**

**Hispanic or Latino.** A person of Mexican, Puerto Rican, Cuban, South or Central American, or other Spanish culture or origin, regardless of race.

**Race Definitions:**

**American Indian or Alaska Native.** A person having origins in any of the original peoples of North and South America (including Central America), and who maintains tribal affiliation or community attachment.

**Asian.** A person having origins in any of the original peoples of the Far East, Southeast Asia, or the Indian subcontinent including, for example, Cambodia, China, India, Japan, Korea, Malaysia, Pakistan, the Philippine Islands, Thailand, and Vietnam.

**Black or African American.** A person having origins in any of the black racial groups of Africa.

**Native Hawaiian or Other Pacific Islander.** A person having origins in any of the original peoples of Hawaii, Guam, Samoa, or other Pacific Islands.

**White.** A person having origins in any of the original peoples of Europe, the Middle East, or North Africa.

---

**WHY THIS INFORMATION IS BEING REQUESTED:**

The Federal Government has a continuing commitment to monitor the operation of its review and award processes to identify and address any inequities based on gender, race, ethnicity, or disability of its proposed PIs/PDs. To gather information needed for this important task, the proposer should submit a single copy of this form for each identified PI/PD with each proposal. Submission of the requested information is voluntary and will not affect the organization's eligibility for an award. However, information not submitted will seriously undermine the statistical validity, and therefore the usefulness, of information received from others. Any individual not wishing to submit some or all the information should check the box provided for this purpose. (The exceptions are the PI/PD name and the information about prior Federal support, the last question above.)

Collection of this information is authorized by the NSF Act of 1950, as amended, 42 U.S.C. 1861, et seq. Demographic data allows NSF to gauge whether our programs and other opportunities in science and technology are fairly reaching and benefiting everyone regardless of demographic category; to ensure that those in under-represented groups have the same knowledge of and access to programs and other research and educational opportunities; and to assess involvement of international investigators in work supported by NSF. The information may be disclosed to government contractors, experts, volunteers and researchers to complete assigned work; and to other government agencies in order to coordinate and assess programs. The information may be added to the Reviewer file and used to select potential candidates to serve as peer reviewers or advisory committee members. See Systems of Records, NSF-50, "Principal Investigator/Proposal File and Associated Records", 63 Federal Register 267 (January 5, 1998), and NSF-51, "Reviewer/Proposal File and Associated Records", 63 Federal Register 268 (January 5, 1998).

**02 INFORMATION ABOUT PRINCIPAL INVESTIGATORS/PROJECT DIRECTORS(PI/PD) and  
co-PRINCIPAL INVESTIGATORS/co-PROJECT DIRECTORS**

---

Submit only ONE copy of this form for each PI/PD and co-PI/PD identified on the proposal. The form(s) should be attached to the original proposal as specified in GPG Section II.C.a. Submission of this information is voluntary and is not a precondition of award. This information will not be disclosed to external peer reviewers. **DO NOT INCLUDE THIS FORM WITH ANY OF THE OTHER COPIES OF YOUR PROPOSAL AS THIS MAY COMPROMISE THE CONFIDENTIALITY OF THE INFORMATION.**

---

**PI/PD Name:** Gloria M Coruzzi

**Gender:**  Male  Female  
**Ethnicity:** (Choose one response)  Hispanic or Latino  Not Hispanic or Latino

**Race:**  
(Select one or more)  
 American Indian or Alaska Native  
 Asian  
 Black or African American  
 Native Hawaiian or Other Pacific Islander  
 White

**Disability Status:**  
(Select one or more)  
 Hearing Impairment  
 Visual Impairment  
 Mobility/Orthopedic Impairment  
 Other  
 None

**Citizenship:** (Choose one)  U.S. Citizen  Permanent Resident  Other non-U.S. Citizen

**Check here if you do not wish to provide any or all of the above information (excluding PI/PD name):**

**REQUIRED: Check here if you are currently serving (or have previously served) as a PI, co-PI or PD on any federally funded project**

---

**Ethnicity Definition:**

**Hispanic or Latino.** A person of Mexican, Puerto Rican, Cuban, South or Central American, or other Spanish culture or origin, regardless of race.

**Race Definitions:**

**American Indian or Alaska Native.** A person having origins in any of the original peoples of North and South America (including Central America), and who maintains tribal affiliation or community attachment.

**Asian.** A person having origins in any of the original peoples of the Far East, Southeast Asia, or the Indian subcontinent including, for example, Cambodia, China, India, Japan, Korea, Malaysia, Pakistan, the Philippine Islands, Thailand, and Vietnam.

**Black or African American.** A person having origins in any of the black racial groups of Africa.

**Native Hawaiian or Other Pacific Islander.** A person having origins in any of the original peoples of Hawaii, Guam, Samoa, or other Pacific Islands.

**White.** A person having origins in any of the original peoples of Europe, the Middle East, or North Africa.

---

**WHY THIS INFORMATION IS BEING REQUESTED:**

The Federal Government has a continuing commitment to monitor the operation of its review and award processes to identify and address any inequities based on gender, race, ethnicity, or disability of its proposed PIs/PDs. To gather information needed for this important task, the proposer should submit a single copy of this form for each identified PI/PD with each proposal. Submission of the requested information is voluntary and will not affect the organization's eligibility for an award. However, information not submitted will seriously undermine the statistical validity, and therefore the usefulness, of information received from others. Any individual not wishing to submit some or all the information should check the box provided for this purpose. (The exceptions are the PI/PD name and the information about prior Federal support, the last question above.)

Collection of this information is authorized by the NSF Act of 1950, as amended, 42 U.S.C. 1861, et seq. Demographic data allows NSF to gauge whether our programs and other opportunities in science and technology are fairly reaching and benefiting everyone regardless of demographic category; to ensure that those in under-represented groups have the same knowledge of and access to programs and other research and educational opportunities; and to assess involvement of international investigators in work supported by NSF. The information may be disclosed to government contractors, experts, volunteers and researchers to complete assigned work; and to other government agencies in order to coordinate and assess programs. The information may be added to the Reviewer file and used to select potential candidates to serve as peer reviewers or advisory committee members. See Systems of Records, NSF-50, "Principal Investigator/Proposal File and Associated Records", 63 Federal Register 267 (January 5, 1998), and NSF-51, "Reviewer/Proposal File and Associated Records", 63 Federal Register 268 (January 5, 1998).

**02 INFORMATION ABOUT PRINCIPAL INVESTIGATORS/PROJECT DIRECTORS(PI/PD) and  
co-PRINCIPAL INVESTIGATORS/co-PROJECT DIRECTORS**

---

Submit only ONE copy of this form for each PI/PD and co-PI/PD identified on the proposal. The form(s) should be attached to the original proposal as specified in GPG Section II.C.a. Submission of this information is voluntary and is not a precondition of award. This information will not be disclosed to external peer reviewers. **DO NOT INCLUDE THIS FORM WITH ANY OF THE OTHER COPIES OF YOUR PROPOSAL AS THIS MAY COMPROMISE THE CONFIDENTIALITY OF THE INFORMATION.**

---

**PI/PD Name:** Manpreet Katari

**Gender:**  Male  Female  
**Ethnicity:** (Choose one response)  Hispanic or Latino  Not Hispanic or Latino

**Race:**  
(Select one or more)  
 American Indian or Alaska Native  
 Asian  
 Black or African American  
 Native Hawaiian or Other Pacific Islander  
 White

**Disability Status:**  
(Select one or more)  
 Hearing Impairment  
 Visual Impairment  
 Mobility/Orthopedic Impairment  
 Other  
 None

**Citizenship:** (Choose one)  U.S. Citizen  Permanent Resident  Other non-U.S. Citizen

**Check here if you do not wish to provide any or all of the above information (excluding PI/PD name):**

**REQUIRED: Check here if you are currently serving (or have previously served) as a PI, co-PI or PD on any federally funded project**

---

**Ethnicity Definition:**

**Hispanic or Latino.** A person of Mexican, Puerto Rican, Cuban, South or Central American, or other Spanish culture or origin, regardless of race.

**Race Definitions:**

**American Indian or Alaska Native.** A person having origins in any of the original peoples of North and South America (including Central America), and who maintains tribal affiliation or community attachment.

**Asian.** A person having origins in any of the original peoples of the Far East, Southeast Asia, or the Indian subcontinent including, for example, Cambodia, China, India, Japan, Korea, Malaysia, Pakistan, the Philippine Islands, Thailand, and Vietnam.

**Black or African American.** A person having origins in any of the black racial groups of Africa.

**Native Hawaiian or Other Pacific Islander.** A person having origins in any of the original peoples of Hawaii, Guam, Samoa, or other Pacific Islands.

**White.** A person having origins in any of the original peoples of Europe, the Middle East, or North Africa.

---

**WHY THIS INFORMATION IS BEING REQUESTED:**

The Federal Government has a continuing commitment to monitor the operation of its review and award processes to identify and address any inequities based on gender, race, ethnicity, or disability of its proposed PIs/PDs. To gather information needed for this important task, the proposer should submit a single copy of this form for each identified PI/PD with each proposal. Submission of the requested information is voluntary and will not affect the organization's eligibility for an award. However, information not submitted will seriously undermine the statistical validity, and therefore the usefulness, of information received from others. Any individual not wishing to submit some or all the information should check the box provided for this purpose. (The exceptions are the PI/PD name and the information about prior Federal support, the last question above.)

Collection of this information is authorized by the NSF Act of 1950, as amended, 42 U.S.C. 1861, et seq. Demographic data allows NSF to gauge whether our programs and other opportunities in science and technology are fairly reaching and benefiting everyone regardless of demographic category; to ensure that those in under-represented groups have the same knowledge of and access to programs and other research and educational opportunities; and to assess involvement of international investigators in work supported by NSF. The information may be disclosed to government contractors, experts, volunteers and researchers to complete assigned work; and to other government agencies in order to coordinate and assess programs. The information may be added to the Reviewer file and used to select potential candidates to serve as peer reviewers or advisory committee members. See Systems of Records, NSF-50, "Principal Investigator/Proposal File and Associated Records", 63 Federal Register 267 (January 5, 1998), and NSF-51, "Reviewer/Proposal File and Associated Records", 63 Federal Register 268 (January 5, 1998).

## List of Suggested Reviewers or Reviewers Not To Include (optional)

---

### **SUGGESTED REVIEWERS:**

Not Listed

### **REVIEWERS NOT TO INCLUDE:**

Not Listed

---

## Conflict of Interest Document

Last Name	First Name	MI	Conflict Type
Aceituno	Felipe	F	Co-Author/Collaborator
Amos	Martyn		Co-Author/Collaborator
Bader	Gary		Co-Author/Collaborator
Barboza	Nora		Advisor/Advisee
Bender	Judith		Co-Author/Collaborator
Benfey	Philip		Co-Author/Collaborator
Bergmann	Dominique		Co-Author/Collaborator
Birnbaum	Kenneth		Co-Author/Collaborator
Blakeslee	Joshua		Co Author/Collaborator
Bonnet	Philippe		Co-Author/Collaborator
Bouganim	Luc		Co-Author/Collaborator
Borevitz	Justin		Co-Author/Collaborator
Borthwick	Andrew		Co-Author/Collaborator
Brenner	Eric		Co-Author/Collaborator
Buff	Robert		Advisor/Advisee
Burga	Alejandro	R	Co-Author/Collaborator
Cabello	Juan	M	Co-Author/Collaborator
Chiaromonte	Francesca		Co-Author/Collaborator
Chiu	Joanna		Advisor/Advisee
Cibrian-Jaramillo	Angelica		Co-Author/Collaborator
Colbourn	Charles	J.	Co-Author/Collaborator
Coruzzi	Gloria	m	Co-Author/Collaborator
Crawford	Nigel	M	Co-Author/Collaborator
Cruikshank	Alexis		Co-Author/Collaborator
Dangl	Jeff		Co-Author/Collaborator
Davidson	Rebecca	S	Co-Author/Collaborator
de la Torre	Eduardo		Advisor/Advisee
Dean	Alexis		Co-Author/Collaborator
Dean	Caroline		Co-Author/Collaborator
Dean	Alexis		Co-Author/Collaborator
DeSalle	Robert		Co-Author/Collaborator
Dudareva	Natalia		Advisor/Advisee
Ecker	Joseph		Co-Author/Collaborator
Egan	Mary		Co-Author/Collaborator
Ehrenreich	Ian	M	Co-Author/Collaborator
Engelmann	Katherine	E	Co-Author/Collaborator
Estelle	Mark		Co-Author/Collaborator
Fekete	Alan		Co-Author/Collaborator
Feinburg	Philip		Advisor/Advisee
Ferro	Alfredo		Co-Author/Collaborator
Fridman	Eyal		Co Author/Collaborator

Last Name	First Name	MI	Conflict Type
Gallaher	R	N	Co Author/Collaborator
Gallaher	K		Co Author/Collaborator
Gifford	Miriam		Co-Author/Collaborator
Giugno	Rosalba		Advisor/Advisee
Glazebrook	Jane		Co-Author/Collaborator
Goldberg	Arthur	P	Co-Author/Collaborator
Goodman	Nathan		Advisor/Advisee
Grant	Sarah		Co-Author/Collaborator
Green	Pamela		Co-Author/Collaborator
Guerinot	Mary Lou		Co-Author/Collaborator
Gutierrez	Rodrigo		Co-Author/Collaborator
Hai Chua	Nam		Advisor/Advisee
Hanzawa	Yoshie		Co-Author/Collaborator
Holmes	Todd		Co-Author/Collaborator
Jefferson	David		Advisor/Advisee
Katari	Manpreet		Co-Author/Collaborator
Kelfer	Jonathan		Co-Author/Collaborator
Kieber	Joseph		Co-Author/Collaborator
Kleinrock	Leonard		Advisor/Advisee
Kolokotronis	Sergios	O	Co-Author/Collaborator
Kouranov	Andrei		Advisor/Advisee
Krouk	Gabriel		Co-Author/Collaborator
Lam	Hong-Ming		Co-Author/Collaborator
Lancien	Muriel		Advisor/Advisee
Lee	Ernest		Co-Author/Collaborator
Lejay	Laurence		Co-Author/Collaborator
Lerner	Alberto		Advisor/Advisee
Li	Ming		Co-Author/Collaborator
Little	Damon	P	Co-Author/Collaborator
Long	Jeff		Co-Author/Collaborator
Maeda	H		Co Author/Collaborator
Marshall	A	C	Co Author/Collaborator
Marshall-Colon	Amy	J	Advisor/Advisee
Martienssen	Robert		Co-Author/Collaborator
McClung	C Robertson		Co-Author/Collaborator
McCombie	Richard		Co-Author/Collaborator
McCombie	Richard		Advisor/Advisee
McSorley	R		Co Author/Collaborator
Meyers	Blake		Co-Author/Collaborator
Miesak	Barbara		Advisor/Advisee
Mishra	Bud		Co-Author/Collaborator
Morgan	John	A	Co Author/Collaborator
Mukherjee	Indrani		Co-Author/Collaborator

Last Name	First Name	MI	Conflict Type
Mullen	Gary		Co-Author/Collaborator
Murphy	Angus		Co Author/Collaborator
Nero	Damion		Advisor/Advisee
Nero	Damion		Co-Author/Collaborator
Neylon	Tyler		Advisor/Advisee
Nordborg	Magnus		Co-Author/Collaborator
Nowicki	Steven	D	Co-Author/Collaborator
Obertello	Mariana		Co-Author/Collaborator
Orlova	Irina		Co Author/Collaborator
Ott	Michael		Co-Author/Collaborator
Palenchar	Peter		Advisor/Advisee
Paley	Bradford		Co-Author/Collaborator
Paley	Bradford		Co-Author/Collaborator
Parker	D. Stott		Advisor/Advisee
Peer	Wendy	A	Co Author/Collaborator
Pevzner	Ilya		Advisor/Advisee
Pichersky	eran		Co Author/Collaborator
Piel	William		Co Author/Collaborator
Poethig	Scott		Co-Author/Collaborator
Poultney	Christopher		Co-Author/Collaborator
Pucheral	Philippe		Co-Author/Collaborator
Purugganan	Micheal	D	Co-Author/Collaborator
Rabin	Michael		Co-Author/Collaborator
Raikhel	Natasha		Co-Author/Collaborator
Rhodes	David		Co Author/Collaborator
Richards	Christina	L	Co-Author/Collaborator
Rigoutsos	Isidore		Co-Author/Collaborator
Ristova	Daniela		Advisor/Advisee
Ruffel	Sandrine		Co-Author/Collaborator
Runko	Suzan		Co-Author/Collaborator
Sarkar	Neil		Co-Author/Collaborator
Schmitt	Johanna		Co-Author/Collaborator
Schnepp	Jennifer		Co Author/Collaborator
Schnittger	Arp		Co-Author/Collaborator
Sengupta	N		Co Author/Collaborator
Shasha	Dennis		Co-Author/Collaborator
Shin	Michael		Advisor/Advisee
Stamatakis	Alexandros		Co-Author/Collaborator
Stephanopoulos	Gregory		Co-Author/Collaborator
Stevenson	Dennis		Co-Author/Collaborator
Stokes	Trevor		Co-Author/Collaborator
Tanurdzic	Milos		Co-Author/Collaborator
Thompson	Lee	P	Co-Author/Collaborator

Last Name	First Name	MI	Conflict Type
Thum	Karen		Co-Author/Collaborator
Tranchina	Daniel		Co-Author/Collaborator
Tsirigos	Aristotle		Advisor/Advisee
Tzagoloff	Alexander		Advisor/Advisee
Varbanova	Marina		Co Author/Collaborator
Vidal	Marc		Co-Author/Collaborator
Wang	Jason	T.L.	Co-Author/Collaborator
Wang	Rongchen		Co-Author/Collaborator
Wang	K	H	Co Author/Collaborator
Wang	Zhihua		Advisor/Advisee
Wilson	Manda		Co Author/Collaborator
Wood	Barbara		Co Author/Collaborator
Xu	Xiangqun		Co-Author/Collaborator
Zhang	Xin		Advisor/Advisee
Zhao	Xiaojian		Advisor/Advisee
Zhu	Yunyue		Advisor/Advisee





## CERTIFICATION PAGE

### Certification for Authorized Organizational Representative or Individual Applicant:

By signing and submitting this proposal, the Authorized Organizational Representative or Individual Applicant is: (1) certifying that statements made herein are true and complete to the best of his/her knowledge; and (2) agreeing to accept the obligation to comply with NSF award terms and conditions if an award is made as a result of this application. Further, the applicant is hereby providing certifications regarding debarment and suspension, drug-free workplace, lobbying activities (see below), responsible conduct of research, nondiscrimination, and flood hazard insurance (when applicable) as set forth in the NSF Proposal & Award Policies & Procedures Guide, Part I: the Grant Proposal Guide (GPG) (NSF 10-1). Willful provision of false information in this application and its supporting documents or in reports required under an ensuing award is a criminal offense (U. S. Code, Title 18, Section 1001).

### Conflict of Interest Certification

In addition, if the applicant institution employs more than fifty persons, by electronically signing the NSF Proposal Cover Sheet, the Authorized Organizational Representative of the applicant institution is certifying that the institution has implemented a written and enforced conflict of interest policy that is consistent with the provisions of the NSF Proposal & Award Policies & Procedures Guide, Part II, Award & Administration Guide (AAG) Chapter IV.A; that to the best of his/her knowledge, all financial disclosures required by that conflict of interest policy have been made; and that all identified conflicts of interest will have been satisfactorily managed, reduced or eliminated prior to the institution's expenditure of any funds under the award, in accordance with the institution's conflict of interest policy. Conflicts which cannot be satisfactorily managed, reduced or eliminated must be disclosed to NSF.

### Drug Free Work Place Certification

By electronically signing the NSF Proposal Cover Sheet, the Authorized Organizational Representative or Individual Applicant is providing the Drug Free Work Place Certification contained in Exhibit II-3 of the Grant Proposal Guide.

### Debarment and Suspension Certification

(If answer "yes", please provide explanation.)

Is the organization or its principals presently debarred, suspended, proposed for debarment, declared ineligible, or voluntarily excluded from covered transactions by any Federal department or agency?

Yes

No

By electronically signing the NSF Proposal Cover Sheet, the Authorized Organizational Representative or Individual Applicant is providing the Debarment and Suspension Certification contained in Exhibit II-4 of the Grant Proposal Guide.

### Certification Regarding Lobbying

The following certification is required for an award of a Federal contract, grant, or cooperative agreement exceeding \$100,000 and for an award of a Federal loan or a commitment providing for the United States to insure or guarantee a loan exceeding \$150,000.

### Certification for Contracts, Grants, Loans and Cooperative Agreements

The undersigned certifies, to the best of his or her knowledge and belief, that:

- (1) No federal appropriated funds have been paid or will be paid, by or on behalf of the undersigned, to any person for influencing or attempting to influence an officer or employee of any agency, a Member of Congress, an officer or employee of Congress, or an employee of a Member of Congress in connection with the awarding of any federal contract, the making of any Federal grant, the making of any Federal loan, the entering into of any cooperative agreement, and the extension, continuation, renewal, amendment, or modification of any Federal contract, grant, loan, or cooperative agreement.
- (2) If any funds other than Federal appropriated funds have been paid or will be paid to any person for influencing or attempting to influence an officer or employee of any agency, a Member of Congress, an officer or employee of Congress, or an employee of a Member of Congress in connection with this Federal contract, grant, loan, or cooperative agreement, the undersigned shall complete and submit Standard Form-LLL, "Disclosure of Lobbying Activities," in accordance with its instructions.
- (3) The undersigned shall require that the language of this certification be included in the award documents for all subawards at all tiers including subcontracts, subgrants, and contracts under grants, loans, and cooperative agreements and that all subrecipients shall certify and disclose accordingly.

This certification is a material representation of fact upon which reliance was placed when this transaction was made or entered into. Submission of this certification is a prerequisite for making or entering into this transaction imposed by section 1352, Title 31, U.S. Code. Any person who fails to file the required certification shall be subject to a civil penalty of not less than \$10,000 and not more than \$100,000 for each such failure.

### Certification Regarding Nondiscrimination

By electronically signing the NSF Proposal Cover Sheet, the Authorized Organizational Representative is providing the Certification Regarding Nondiscrimination contained in Exhibit II-6 of the Grant Proposal Guide.

### Certification Regarding Flood Hazard Insurance

Two sections of the National Flood Insurance Act of 1968 (42 USC §4012a and §4106) bar Federal agencies from giving financial assistance for acquisition or construction purposes in any area identified by the Federal Emergency Management Agency (FEMA) as having special flood hazards unless the:

- (1) community in which that area is located participates in the national flood insurance program; and
- (2) building (and any related equipment) is covered by adequate flood insurance.

By electronically signing the NSF Proposal Cover Sheet, the Authorized Organizational Representative or Individual Applicant located in FEMA-designated special flood hazard areas is certifying that adequate flood insurance has been or will be obtained in the following situations:

- (1) for NSF grants for the construction of a building or facility, regardless of the dollar amount of the grant; and
- (2) for other NSF Grants when more than \$25,000 has been budgeted in the proposal for repair, alteration or improvement (construction) of a building or facility.

### Certification Regarding Responsible Conduct of Research (RCR)

**(This certification is not applicable to proposals for conferences, symposia, and workshops.)**

By electronically signing the NSF Proposal Cover Sheet, the Authorized Organizational Representative of the applicant institution is certifying that, in accordance with the NSF Proposal & Award Policies & Procedures Guide, Part II, Award & Administration Guide (AAG) Chapter IV.B., the institution has a plan in place to provide appropriate training and oversight in the responsible and ethical conduct of research to undergraduates, graduate students and postdoctoral researchers who will be supported by NSF to conduct research.

The undersigned shall require that the language of this certification be included in any award documents for all subawards at all tiers.

AUTHORIZED ORGANIZATIONAL REPRESENTATIVE		SIGNATURE		DATE
NAME		<b>Electronic Signature</b>		<b>Jan 26 2010 4:39PM</b>
<b>Richard L Louth</b>				
TELEPHONE NUMBER	ELECTRONIC MAIL ADDRESS	FAX NUMBER		
<b>212-998-2121</b>	<b>osp.agency@nyu.edu</b>	<b>212-995-4029</b>		

\* EAGER - EARly-concept Grants for Exploratory Research

\*\* RAPID - Grants for Rapid Response Research

**Directorate for Biological Sciences  
Division of Integrative Organismal Sys  
Plant Genome Research Project**

**Proposal Classification Form  
PI: / Proposal Number: 1025989**

**CATEGORY I: INVESTIGATOR STATUS (Select ONE)**

- Beginning Investigator - No previous Federal support as PI or Co-PI, excluding fellowships, dissertations, planning grants, etc.
- Prior Federal support only
- Current Federal support only
- Current & prior Federal support

**CATEGORY II: FIELDS OF SCIENCE OTHER THAN BIOLOGY INVOLVED IN THIS RESEARCH (Select 1 to 3)**

- |  |   |  |
|--|---|--|
| <input type="checkbox"/> Astronomy                   | <input type="checkbox"/> Engineering            | <input type="checkbox"/> Psychology        |
| <input type="checkbox"/> Chemistry                   | <input checked="" type="checkbox"/> Mathematics | <input type="checkbox"/> Social Sciences   |
| <input checked="" type="checkbox"/> Computer Science | <input type="checkbox"/> Physics                | <input type="checkbox"/> None of the Above |
| <input type="checkbox"/> Earth Science               |   |  |

**CATEGORY III: SUBSTANTIVE AREA (Select 1 to 4)**

- |  |   |   |
|--|---|---|
| <input type="checkbox"/> BEHAVIORAL STUDIES                    | <input type="checkbox"/> Keystone Species   | <input type="checkbox"/> Agricultural Ecology   |
| <input type="checkbox"/> BIOENGINEERING                        | <input type="checkbox"/> COMPARATIVE APPROACHES                                     | <input type="checkbox"/> ENDOCRINE DISRUPTORS/<br>ENVIRONMENTAL<br>ENDOCRINOLOGY          |
| <input type="checkbox"/> BIOGEOGRAPHY                          | <input checked="" type="checkbox"/> COMPUTATIONAL BIOLOGY                           | <input type="checkbox"/> EPIGENETICS  |
| <input type="checkbox"/> Island Biogeography                   | <input type="checkbox"/> CONSERVATION & RESTORATION<br>BIOLOGY                      | <input type="checkbox"/> EXTREMOPHILES  |
| <input type="checkbox"/> Historical/ Evolutionary Biogeography | <input type="checkbox"/> CORAL REEFS  | <input checked="" type="checkbox"/> GENOMICS (Genome sequence,<br>organization, function) |
| <input type="checkbox"/> Phylogeography                        | <input type="checkbox"/> CURATION   | <input type="checkbox"/> Viral  |
| <input type="checkbox"/> Methods/Theory                        | <input type="checkbox"/> DATABASES  | <input type="checkbox"/> Microbial  |
| <input type="checkbox"/> BIOMATERIALS                          | <input type="checkbox"/> DEVELOPMENTAL BIOLOGY                                      | <input type="checkbox"/> Fungal   |
| <input type="checkbox"/> BIOTECHNOLOGY                         | <input type="checkbox"/> ECOSYSTEMS LEVEL   | <input type="checkbox"/> Plant  |
| <input type="checkbox"/> Animal Biotechnology                  | <input type="checkbox"/> Physical Structure   | <input type="checkbox"/> Animal   |
| <input type="checkbox"/> Plant Biotechnology                   | <input type="checkbox"/> Decomposition  | <input type="checkbox"/> HUMAN NUTRITION  |
| <input type="checkbox"/> Environmental Biotechnology           | <input type="checkbox"/> Biogeochemistry  | <input type="checkbox"/> INFORMATICS  |
| <input type="checkbox"/> Marine Biotechnology                  | <input type="checkbox"/> Limnology/Hydrology  | <input type="checkbox"/> MARINE MAMMALS   |
| <input type="checkbox"/> Metabolic Engineering                 | <input type="checkbox"/> Climate/Microclimate                                       | <input type="checkbox"/> MOLECULAR APPROACHES   |
| <input type="checkbox"/> CHRONOBIOLOGY                         | <input type="checkbox"/> Whole-System Analysis                                      | <input type="checkbox"/> Molecular Evolution  |
| <input type="checkbox"/> COGNITIVE NEUROSCIENCE                | <input type="checkbox"/> Productivity/Biomass                                       | <input type="checkbox"/> NANOSCIENCE  |
| <input type="checkbox"/> COMMUNITY ECOLOGY                     | <input type="checkbox"/> System Energetics  | <input type="checkbox"/> ORGANISMAL SYSTEMS   |
| <input type="checkbox"/> Community Analysis                    | <input type="checkbox"/> Landscape Dynamics   | <input type="checkbox"/> Physiological Approaches   |
| <input type="checkbox"/> Community Structure                   | <input type="checkbox"/> Chemical & Biochemical Control                             | <input type="checkbox"/> Metabolic Processes  |
| <input type="checkbox"/> Community Stability                   | <input type="checkbox"/> Global Change  | <input type="checkbox"/> Hormonal Regulation/ Integration                                 |
| <input type="checkbox"/> Succession                            | <input type="checkbox"/> Climate Change   | <input type="checkbox"/> Stress Responses   |
| <input type="checkbox"/> Experimental Microcosms/ Mesocosms    | <input type="checkbox"/> Regional Studies   | <input type="checkbox"/> Sensory Biology  |
| <input type="checkbox"/> Disturbance                           | <input type="checkbox"/> Global Studies   | <input type="checkbox"/> Movement Studies   |
| <input type="checkbox"/> Deforestation                         | <input type="checkbox"/> Forestry   | <input type="checkbox"/> PALEONTOLOGY   |
| <input type="checkbox"/> Patch Dynamics                        | <input type="checkbox"/> Resource Management (Wildlife,<br>Fisheries, Range, Other) | <input type="checkbox"/> Floristic  |
| <input type="checkbox"/> Food Webs/ Trophic Structure          |   |   |

<input type="checkbox"/> Faunistic <input type="checkbox"/> Paleoecology <input type="checkbox"/> Biostratigraphy <input type="checkbox"/> Palynology <input type="checkbox"/> Micropaleontology <input type="checkbox"/> Paleoclimatology <input type="checkbox"/> Archeozoic <input type="checkbox"/> Paleozoic <input type="checkbox"/> Mesozoic <input type="checkbox"/> Cenozoic <input type="checkbox"/> PHOTOSYNTHESIS <input checked="" type="checkbox"/> PLANT BIOLOGY <input checked="" type="checkbox"/> Arabidopsis-Related Plant Research <input type="checkbox"/> POPULATION DYNAMICS & LIFE HISTORY <input type="checkbox"/> Demography/ Life History <input type="checkbox"/> Population Cycles <input type="checkbox"/> Distribution/Patchiness/ Marginal Populations <input type="checkbox"/> Population Regulation <input type="checkbox"/> Intraspecific Competition <input type="checkbox"/> Reproductive Strategies <input type="checkbox"/> Gender Allocation <input type="checkbox"/> Metapopulations <input type="checkbox"/> Extinction <input type="checkbox"/> POPULATION GENETICS & BREEDING SYSTEMS	<input type="checkbox"/> Variation <input type="checkbox"/> Microevolution <input type="checkbox"/> Speciation <input type="checkbox"/> Hybridization <input type="checkbox"/> Inbreeding/Outbreeding <input type="checkbox"/> Gene Flow Measurement <input type="checkbox"/> Inheritance/Heritability <input type="checkbox"/> Quantitative Genetics/ QTL Analysis <input type="checkbox"/> Ecological Genetics <input type="checkbox"/> Gender Ratios <input type="checkbox"/> Apomixis/ Parthenogenesis <input type="checkbox"/> Vegetative Reproduction <input type="checkbox"/> REPRODUCTIVE ANIMAL BIOLOGY <input type="checkbox"/> SPECIES INTERACTIONS <input type="checkbox"/> Predation <input type="checkbox"/> Herbivory <input type="checkbox"/> Omnivory <input type="checkbox"/> Interspecific Competition <input type="checkbox"/> Niche Relationships/ Resource Partitioning <input type="checkbox"/> Pollination/ Seed Dispersal <input type="checkbox"/> Parasitism <input type="checkbox"/> Mutualism/ Commensalism <input type="checkbox"/> Plant/Fungal/ Microbial Interactions <input type="checkbox"/> Mimicry	<input type="checkbox"/> Animal Pathology <input type="checkbox"/> Plant Pathology <input type="checkbox"/> Coevolution <input type="checkbox"/> Biological Control <input type="checkbox"/> SPINAL CORD/ NERVE REGENERATION <input type="checkbox"/> STATISTICS & MODELING <input type="checkbox"/> Methods/ Instrumentation/ Software <input type="checkbox"/> Modeling (general) <input type="checkbox"/> Modeling of Biological or Molecular Systems <input type="checkbox"/> Computational Modeling <input type="checkbox"/> Statistics (general) <input type="checkbox"/> Multivariate Methods <input type="checkbox"/> Spatial Statistics & Spatial Modeling <input type="checkbox"/> Sampling Design & Analysis <input type="checkbox"/> Experimental Design & Analysis <input type="checkbox"/> STRUCTURAL BIOLOGY <input type="checkbox"/> SYSTEMATICS <input type="checkbox"/> Taxonomy/Classification <input type="checkbox"/> Nomenclature <input type="checkbox"/> Monograph/Revision <input type="checkbox"/> Phylogenetics <input type="checkbox"/> Phenetics/Cladistics/ Numerical Taxonomy <input type="checkbox"/> Macroevolution <input type="checkbox"/> NONE OF THE ABOVE
--	---	--

**CATEGORY IV: INFRASTRUCTURE (Select 1 to 3)**

<b>COLLECTIONS/STOCK CULTURES</b> <input type="checkbox"/> Collection Enhancement <input type="checkbox"/> Collection Refurbishment <input type="checkbox"/> Living Organism Stock Cultures <input type="checkbox"/> Natural History Collections <b>DATABASES</b> <input type="checkbox"/> Database Initiation <input checked="" type="checkbox"/> Database Enhancement <input type="checkbox"/> Database Maintenance & Curation <input type="checkbox"/> Database Methods <b>FACILITIES</b> <input type="checkbox"/> Controlled Environment Facilities	<input type="checkbox"/> Field Stations <input type="checkbox"/> Field Facility Structure <input type="checkbox"/> Field Facility Equipment <input type="checkbox"/> LTER Site <input type="checkbox"/> GENOME SEQUENCING <input type="checkbox"/> Arabidopsis Genome Sequencing <input type="checkbox"/> Other Plant Genome Sequencing <input type="checkbox"/> INDUSTRY PARTICIPATION <b>INSTRUMENTATION</b> <input type="checkbox"/> Instrument Development <input type="checkbox"/> Instrument Acquisition <input type="checkbox"/> Computational Hardware Development/Acquisition	<b>TOOLS DEVELOPMENT</b> <input checked="" type="checkbox"/> Analytical Algorithm Development <input type="checkbox"/> Other Software Development <input checked="" type="checkbox"/> Informatics Tool Development <input type="checkbox"/> Technique Development <b>TRACKING SYSTEMS</b> <input type="checkbox"/> Geographic Information Systems <input type="checkbox"/> Remote Sensing <b>TRAINING</b> <input type="checkbox"/> Multi-, Cross-, Interdisciplinary Training <input type="checkbox"/> NONE OF THE ABOVE
--	---	--

**CATEGORY V: HABITAT (Select 1 to 2)**

<b>TERRESTRIAL HABITATS</b>		
<input type="checkbox"/> GENERAL TERRESTRIAL <input type="checkbox"/> TUNDRA <input type="checkbox"/> BOREAL FOREST <input type="checkbox"/> TEMPERATE	<input type="checkbox"/> Deciduous Forest <input type="checkbox"/> Coniferous Forest <input type="checkbox"/> Rain Forest <input type="checkbox"/> Mixed Forest <input type="checkbox"/> Prairie/Grasslands	<input type="checkbox"/> Desert <input type="checkbox"/> SUBTROPICAL <input type="checkbox"/> Rain Forest <input type="checkbox"/> Seasonal Forest <input type="checkbox"/> Savanna

<input type="checkbox"/> Thornwoods <input type="checkbox"/> Deciduous Forest <input type="checkbox"/> Coniferous Forest <input type="checkbox"/> Desert <input type="checkbox"/> TROPICAL <input type="checkbox"/> Rain Forest <input type="checkbox"/> Seasonal Forest <input type="checkbox"/> Savanna <input type="checkbox"/> Thornwoods	<input type="checkbox"/> Deciduous Forest <input type="checkbox"/> Coniferous Forest <input type="checkbox"/> Desert <input type="checkbox"/> CHAPPARAL/ SCLEROPHYLL/ SHRUBLANDS <input type="checkbox"/> ALPINE <input type="checkbox"/> MONTANE <input type="checkbox"/> CLOUD FOREST <input type="checkbox"/> RIPARIAN ZONES	<input type="checkbox"/> ISLANDS (except Barrier Islands) <input type="checkbox"/> BEACHES/ DUNES/ SHORES/ BARRIER ISLANDS <input type="checkbox"/> CAVES/ ROCK OUTCROPS/ CLIFFS <input type="checkbox"/> CROPLANDS/ FALLOW FIELDS/ PASTURES <input type="checkbox"/> URBAN/SUBURBAN <input type="checkbox"/> SUBTERRANEAN/ SOIL/ SEDIMENTS <input type="checkbox"/> EXTREME TERRESTRIAL ENVIRONMENT <input type="checkbox"/> AERIAL
---	--	---

**AQUATIC HABITATS**

<input type="checkbox"/> GENERAL AQUATIC <input type="checkbox"/> FRESHWATER <input type="checkbox"/> Wetlands/Bogs/Swamps <input type="checkbox"/> Lakes/Ponds <input type="checkbox"/> Rivers/Streams <input type="checkbox"/> Reservoirs <input type="checkbox"/> MARINE	<input type="checkbox"/> Open Ocean/Continental Shelf <input type="checkbox"/> Bathyal <input type="checkbox"/> Abyssal <input type="checkbox"/> Estuarine <input type="checkbox"/> Intertidal/Tidal/Coastal <input type="checkbox"/> Coral Reef <input type="checkbox"/> HYPERSALINE	<input type="checkbox"/> EXTREME AQUATIC ENVIRONMENT <input type="checkbox"/> CAVES/ ROCK OUTCROPS/ CLIFFS <input type="checkbox"/> MANGROVES <input type="checkbox"/> SUBSURFACE WATERS/ SPRINGS <input type="checkbox"/> EPHEMERAL POOLS & STREAMS <input type="checkbox"/> MICROPOOLS (Pitcher Plants, Tree Holes, Other)
---	---	---

**MAN-MADE ENVIRONMENTS**

<input type="checkbox"/> CELL/TISSUE CULTURE (In Vitro) <input type="checkbox"/> In Silico	<input type="checkbox"/> THEORETICAL SYSTEMS	<input type="checkbox"/> OTHER ARTIFICIAL SYSTEMS
---	--	---

**NOT APPLICABLE**

<input checked="" type="checkbox"/> NOT APPLICABLE	
--	--

**CATEGORY VI: GEOGRAPHIC AREA OF THE RESEARCH (Select 1 to 2)**

<input type="checkbox"/> WORLDWIDE <input type="checkbox"/> NORTH AMERICA <input checked="" type="checkbox"/> United States <input type="checkbox"/> Northeast US (CT, MA, ME, NH, NJ, NY, PA, RI, VT) <input type="checkbox"/> Northcentral US (IA, IL, IN, MI, MN, ND, NE, OH, SD, WI) <input type="checkbox"/> Northwest US (ID, MT, OR, WA, WY) <input type="checkbox"/> Southeast US (DC, DE, FL, GA, MD, NC, SC, WV, VA) <input type="checkbox"/> Southcentral US (AL, AR, KS, KY, LA, MO, MS, OK, TN, TX) <input type="checkbox"/> Southwest US (AZ, CA, CO, NM, NV, UT) <input type="checkbox"/> Alaska <input type="checkbox"/> Hawaii <input type="checkbox"/> Puerto Rico <input type="checkbox"/> Canada <input type="checkbox"/> Mexico <input type="checkbox"/> CENTRAL AMERICA (Mainland) <input type="checkbox"/> Caribbean Islands <input type="checkbox"/> Bermuda/Bahamas <input type="checkbox"/> SOUTH AMERICA	<input type="checkbox"/> Eastern South America (Guyana, Fr. Guiana, Suriname, Brazil) <input type="checkbox"/> Northern South America (Colombia, Venezuela) <input type="checkbox"/> Southern South America (Chile, Argentina, Uruguay, Paraguay) <input type="checkbox"/> Western South America (Ecuador, Peru, Bolivia) <input type="checkbox"/> EUROPE <input type="checkbox"/> Eastern Europe <input type="checkbox"/> Russia <input type="checkbox"/> Scandinavia <input type="checkbox"/> Western Europe <input type="checkbox"/> ASIA <input type="checkbox"/> Central Asia <input type="checkbox"/> Far East <input type="checkbox"/> Middle East <input type="checkbox"/> Siberia <input type="checkbox"/> South Asia <input type="checkbox"/> Southeast Asia <input type="checkbox"/> AFRICA	<input type="checkbox"/> North Africa <input type="checkbox"/> African South of the Sahara <input type="checkbox"/> East Africa <input type="checkbox"/> Madagascar <input type="checkbox"/> South Africa <input type="checkbox"/> West Africa <input type="checkbox"/> AUSTRALASIA <input type="checkbox"/> Australia <input type="checkbox"/> New Zealand <input type="checkbox"/> Pacific Islands <input type="checkbox"/> ANTARCTICA <input type="checkbox"/> ARCTIC <input type="checkbox"/> ATLANTIC OCEAN <input type="checkbox"/> PACIFIC OCEAN <input type="checkbox"/> INDIAN OCEAN <input type="checkbox"/> OTHER REGIONS (Not defined) <input type="checkbox"/> NOT APPLICABLE
--	--	--

**CATEGORY VII: CLASSIFICATION OF ORGANISMS (Select 1 to 4)**

<input type="checkbox"/> VIRUSES <input type="checkbox"/> Bacterial	<input type="checkbox"/> Plant <input type="checkbox"/> Animal	<input type="checkbox"/> PROKARYOTES <input type="checkbox"/> Archaeobacteria
--	---	--

<input type="checkbox"/> Cyanobacteria	<input type="checkbox"/> Fabaceae (Leguminosae)	<input type="checkbox"/> Merostomata (Horseshoe Crabs)
<input type="checkbox"/> Eubacteria	<input type="checkbox"/> Lamiaceae (Labiatae)	<input type="checkbox"/> Pycnogonida (Sea Spiders)
<input type="checkbox"/> <b>PROTISTA (PROTOZOA)</b>	<input type="checkbox"/> Rosaceae	<input type="checkbox"/> Scorpionida (Scorpions)
<input type="checkbox"/> Amoeboae	<input type="checkbox"/> Solanaceae	<input type="checkbox"/> Araneae (True Spiders)
<input type="checkbox"/> Apicomplexa	<input type="checkbox"/> <b>ANIMALS</b>	<input type="checkbox"/> Pseudoscorpionida (Pseudoscorpions)
<input type="checkbox"/> Ciliophora	<input type="checkbox"/> INVERTEBRATES	<input type="checkbox"/> Acarina (Free-living Mites)
<input type="checkbox"/> Flagellates	<input type="checkbox"/> MESOZOA/PLACOZOA	<input type="checkbox"/> Parasitiformes (Parasitic Ticks & Mites)
<input type="checkbox"/> Foraminifera	<input type="checkbox"/> PORIFERA (Sponges)	<input type="checkbox"/> Crustacea
<input type="checkbox"/> Microspora	<input type="checkbox"/> CNIDARIA	<input type="checkbox"/> Branchiopoda (Fairy Shrimp, Water Flea)
<input type="checkbox"/> Radiolaria	<input type="checkbox"/> Hydrozoa (Hydra, etc.)	<input type="checkbox"/> Ostracoda (Sea Lice)
<input type="checkbox"/> <b>FUNGI</b>	<input type="checkbox"/> Scyphozoa (Jellyfish)	<input type="checkbox"/> Copepoda
<input type="checkbox"/> Ascomycota	<input type="checkbox"/> Anthozoa (Corals, Sea Anemones)	<input type="checkbox"/> Cirripedia (Barnacles)
<input type="checkbox"/> Basidiomycota	<input type="checkbox"/> CTENOPHORA (Comb Jellies)	<input type="checkbox"/> Amphipoda (Skeleton Shrimp, Whale Lice, Freshwater Shrimp)
<input type="checkbox"/> Chytridiomycota	<input type="checkbox"/> PLATYHELMINTHES (Flatworms)	<input type="checkbox"/> Isopoda (Wood Lice, Pillbugs)
<input type="checkbox"/> Mitosporic Fungi	<input type="checkbox"/> Turbellaria (Planarians)	<input type="checkbox"/> Decapoda (Lobster, Crayfish, Crabs, Shrimp)
<input type="checkbox"/> Oomycota	<input type="checkbox"/> Trematoda (Flukes)	<input type="checkbox"/> Hexapoda (Insecta) (Insects)
<input type="checkbox"/> Yeasts	<input type="checkbox"/> Cestoda (Tapeworms)	<input type="checkbox"/> Apterygota (Springtails, Silverfish, etc.)
<input type="checkbox"/> Zygomycota	<input type="checkbox"/> Monogenea (Flukes)	<input type="checkbox"/> Odonata (Dragonflies, Damselflies)
<input type="checkbox"/> <b>LICHENS</b>	<input type="checkbox"/> GNATHOSTOMULIDA	<input type="checkbox"/> Ephemeroptera (Mayflies)
<input type="checkbox"/> <b>SLIME MOLDS</b>	<input type="checkbox"/> NEMERTINEA (Rynchocoela) (Ribbon Worms)	<input type="checkbox"/> Orthoptera (Grasshoppers, Crickets)
<input type="checkbox"/> <b>ALGAE</b>	<input type="checkbox"/> ENTOPROCTA (Bryozoa) (Plant-like Animals)	<input type="checkbox"/> Dictyoptera (Cockroaches, Mantids, Phasmids)
<input type="checkbox"/> Bacillariophyta (Diatoms)	<input type="checkbox"/> ASCHELMINTHES	<input type="checkbox"/> Isoptera (Termites)
<input type="checkbox"/> Charophyta	<input type="checkbox"/> Gastrotricha	<input type="checkbox"/> Plecoptera (Stoneflies)
<input type="checkbox"/> Chlorophyta	<input type="checkbox"/> Kinorhyncha	<input type="checkbox"/> Phthiraptera (Mallophaga & Anoplura) (Lice)
<input type="checkbox"/> Chrysophyta	<input type="checkbox"/> Loricifera	<input type="checkbox"/> Hemiptera (including Heteroptera) (True Bugs)
<input type="checkbox"/> Dinoflagellata	<input type="checkbox"/> Nematoda (Roundworms)	<input type="checkbox"/> Homoptera (Cicadas, Scale Insects, Leafhoppers)
<input type="checkbox"/> Euglenoids	<input type="checkbox"/> Nematomorpha (Horsehair Worms)	<input type="checkbox"/> Thysanoptera (Thrips)
<input type="checkbox"/> Phaeophyta	<input type="checkbox"/> Rotifera (Rotatoria)	<input type="checkbox"/> Neuroptera (Lacewings, Dobsonflies, Snakeflies)
<input type="checkbox"/> Rhodophyta	<input type="checkbox"/> ACANTHOCEPHALA (Spiny-headed Worms)	<input type="checkbox"/> Trichoptera (Caddisflies)
<input checked="" type="checkbox"/> <b>PLANTS</b>	<input type="checkbox"/> PRIAPULOIDEA	<input type="checkbox"/> Lepidoptera (Moths, Butterflies)
<input type="checkbox"/> NON-VASCULAR PLANTS	<input type="checkbox"/> BRYOZOA (Ectoprocta) (Plant-like Animals)	<input type="checkbox"/> Diptera (Flies, Mosquitoes)
<input type="checkbox"/> BRYOPHYTA	<input type="checkbox"/> PHORONIDEA (Lophophorates)	<input type="checkbox"/> Siphonaptera (Fleas)
<input type="checkbox"/> Anthocerotae (Hornworts)	<input type="checkbox"/> BRACHIOPODA (Lamp Shells)	<input type="checkbox"/> Coleoptera (Beetles)
<input type="checkbox"/> Hepaticae (Liverworts)	<input type="checkbox"/> MOLLUSCA	<input type="checkbox"/> Hymenoptera (Ants, Bees, Wasps, Sawflies)
<input type="checkbox"/> Musci (Mosses)	<input type="checkbox"/> Monoplacophora	<input type="checkbox"/> Chilopoda (Centipedes)
<input type="checkbox"/> VASCULAR PLANTS	<input type="checkbox"/> Aplacophora (Solenogasters)	<input type="checkbox"/> Diplopoda (Millipedes)
<input type="checkbox"/> FERNS & FERN ALLIES	<input type="checkbox"/> Polyplacophora (Chitons)	<input type="checkbox"/> Pauropoda
<input type="checkbox"/> GYMNOSPERMS	<input type="checkbox"/> Scaphopoda (Tooth Shells)	<input type="checkbox"/> Symphyta (Symphyta)
<input type="checkbox"/> Coniferales (Conifers)	<input type="checkbox"/> Gastropoda (Snails, Slugs, Limpets)	<input type="checkbox"/> PENTASTOMIDA (Linguatulida) (Tongue Worms)
<input type="checkbox"/> Cycadales (Cycads)	<input type="checkbox"/> Pelecypoda (Bivalvia) (Clams, Mussels, Oysters, Scallops)	<input type="checkbox"/> TARDIGRADA (Tardigrades, Water Bears)
<input type="checkbox"/> Ginkgoales (Ginkgo)	<input type="checkbox"/> Cephalopoda (Squid, Octopus, Nautilus)	<input type="checkbox"/> ONYCHOPHORA (Peripatus)
<input type="checkbox"/> Gnetales (Gnetophytes)	<input type="checkbox"/> ANNELIDA (Segmented Worms)	<input type="checkbox"/> CHAETOGNATHA (Arrow Worms)
<input type="checkbox"/> ANGIOSPERMS	<input type="checkbox"/> Polychaeta (Parapodial Worms)	<input type="checkbox"/> ECHINODERMATA
<input type="checkbox"/> Monocots	<input type="checkbox"/> Oligochaeta (Earthworms)	<input type="checkbox"/> Crinoidea (Sea Lilies, Feather Stars)
<input type="checkbox"/> Areaceae (Palmae)	<input type="checkbox"/> Hirudinida (Leeches)	<input type="checkbox"/> Asteroidea (Starfish, Sea Stars)
<input type="checkbox"/> Cyperaceae	<input type="checkbox"/> POGONOPHORA (Beard Worms)	
<input type="checkbox"/> Liliaceae	<input type="checkbox"/> SIPUNCULOIDEA (Peanut Worms)	
<input type="checkbox"/> Orchidaceae	<input type="checkbox"/> ECHIUROIDEA (Spoon Worms)	
<input type="checkbox"/> Poaceae (Graminae)	<input type="checkbox"/> ARTHROPODA	
<input type="checkbox"/> Dicots	<input type="checkbox"/> Cheliceriformes	
<input type="checkbox"/> Apiaceae (Umbelliferae)		
<input type="checkbox"/> Asteraceae (Compositae)		
<input type="checkbox"/> Brassicaceae (Cruciferae)		

<input type="checkbox"/> Ophiuroidea (Brittle Stars, Serpent Stars)	<input type="checkbox"/> AVES (Birds)	<input type="checkbox"/> Insectivora (Hedgehogs, Moles, Shrews, Tenrec, etc.)
<input type="checkbox"/> Echinoidea (Sea Urchins, Sand Dollars)	<input type="checkbox"/> Paleognathae (Ratites)	<input type="checkbox"/> Chiroptera (Bats)
<input type="checkbox"/> Holothuroidea (Sea Cucumbers)	<input type="checkbox"/> Sphenisciformes (Penguins)	<input type="checkbox"/> Edentata (Anteaters, Sloths, Armadillos)
<input type="checkbox"/> HEMICHORDATA (Acorn Worms, Pterobranchs)	<input type="checkbox"/> Procellariiformes (Albatrosses, Petrels, Fulmars)	<input type="checkbox"/> Primates
<input type="checkbox"/> UROCHORDATA (Tunicata) (Tunicates, Sea Squirts, Salps, Ascideans)	<input type="checkbox"/> Pelecaniformes (Pelicans, Gannets, Boobies, Tropicbirds)	<input type="checkbox"/> Monkeys
<input type="checkbox"/> CEPHALOCHORDATA (Amphioxus/Lancelet)	<input type="checkbox"/> Ciconiiformes (Herons, Bitterns, Egrets, Storks, Ibis, Flamingo)	<input type="checkbox"/> Apes (Gibbons, Orang-utan, Gorilla, Chimpanzee)
<input type="checkbox"/> VERTEBRATES	<input type="checkbox"/> Anseriformes (Ducks, Geese, Screamers)	<input type="checkbox"/> Humans
<input type="checkbox"/> AGNATHA (Hagfish, Lamprey)	<input type="checkbox"/> Falconiformes (Vultures, Hawks, Eagles, Condors, Kites, Falcons)	<input type="checkbox"/> Rodentia
<input type="checkbox"/> FISHES	<input type="checkbox"/> Galliformes (Megapodes, Turkeys, Quail, Pheasants, Peafowl, etc.)	<input type="checkbox"/> Laboratory Rodents (Rat, Mouse, Guinea Pig, Hamster)
<input type="checkbox"/> Chondrichthyes (Cartilaginous Fishes) (Sharks, Rays, Ratfish)	<input type="checkbox"/> Gruiformes (Cranes, Rails, Gallinules, Coots, Bustards, Crakes)	<input type="checkbox"/> Non-Laboratory Rodents
<input type="checkbox"/> Osteichthyes (Bony Fishes)	<input type="checkbox"/> Charadriiformes (Terns, Gulls, Stilts, Avocets, Plovers, Puffins, etc.)	<input type="checkbox"/> Lagomorphs (Rabbits, Hares, Pikas)
<input type="checkbox"/> Sarcopterygia (Lobe-finned Fishes) (Coelacanth, Lungfish)	<input type="checkbox"/> Columbiformes (Pigeons, Doves)	<input type="checkbox"/> Tubulidenata (Aardvarks)
<input type="checkbox"/> Actinopterygia (Ray-finned Fishes)	<input type="checkbox"/> Psittaciformes (Parrots, Lories, Cockatoos, Kakapo, Conures, etc.)	<input type="checkbox"/> Carnivora (Bears, Canids, Felids, Mustelids, Viverrids, Hyena, Procyonids)
<input type="checkbox"/> AMPHIBIA	<input type="checkbox"/> Cuculiformes (Cuckoos, Turacos, Anis, Coucal, Roadrunner, etc.)	<input type="checkbox"/> Ungulates
<input type="checkbox"/> Anura (Frogs, Toads)	<input type="checkbox"/> Strigiformes (Owls)	<input type="checkbox"/> Perissodactyla (Odd-toed Ungulates) (Horses, Rhinos, Tapirs, etc.)
<input type="checkbox"/> Urodela (Salamanders, Newts)	<input type="checkbox"/> Apodiformes (Hummingbirds, Swifts, Thornbills)	<input type="checkbox"/> Artiodactyla (Even-toed Ungulates) (Cattle, Sheep, Deer, Pigs, etc.)
<input type="checkbox"/> Gymnophiona (Apoda) (Caecilians)	<input type="checkbox"/> Coraciformes (Kingfishers, Todies, Bee-Eaters, Rollers, Hornbills, etc.)	<input type="checkbox"/> Sirenia (Manatees, Dugongs)
<input type="checkbox"/> REPTILIA	<input type="checkbox"/> Piciformes (Woodpeckers, Toucans, Jacamars, Barbets, Honeyguides)	<input type="checkbox"/> Proboscidea (Elephants)
<input type="checkbox"/> Chelonia (Turtles, Tortoises)	<input type="checkbox"/> Passeriformes (Passerines)	<input type="checkbox"/> Marine Mammals (Seals, Walrus, Whales, Otters, Dolphins, Porpoises)
<input type="checkbox"/> Serpentes (Snakes)	<input type="checkbox"/> MAMMALIA	<input type="checkbox"/> TRANSGENIC ORGANISMS
<input type="checkbox"/> Sauria (Lizards)	<input type="checkbox"/> Monotremata (Platypus, Echidna)	<input type="checkbox"/> FOSSIL OR EXTINCT ORGANISMS
<input type="checkbox"/> Crocodylia (Crocodilians)	<input type="checkbox"/> Marsupialia (Marsupials)	<input type="checkbox"/> NO ORGANISMS
<input type="checkbox"/> Rhynchocephalia (Tuatara)	<input type="checkbox"/> Eutheria (Placentals)	

### CATEGORY VIII: MODEL ORGANISM (Select ONE)

<input type="checkbox"/> NO MODEL ORGANISM	<input checked="" type="checkbox"/> Mouse-Ear Cress ( <i>Arabidopsis thaliana</i> )	<input type="checkbox"/> Crayfish ( <i>Procambarus</i> , <i>Astacus</i> , etc.)
MODEL ORGANISM (Choose from the list or input up to 9 characters)	<input type="checkbox"/> Ice Plant ( <i>Mesembryanthemum</i> spp.)	<input type="checkbox"/> Dragonfly ( <i>Aeschna</i> , etc.)
VIRUS/BACTERIA	<input type="checkbox"/> Barley ( <i>Hordeum vulgare</i> )	<input type="checkbox"/> Grasshopper/Locust ( <i>Schistocerca</i> , etc.)
<input type="checkbox"/> Lambda Phage	<input type="checkbox"/> Corn ( <i>Zea mays</i> )	<input type="checkbox"/> Cockroach ( <i>Periplaneta</i> , <i>Blatta</i> , <i>Blatella</i> , etc.)
<input type="checkbox"/> Rhizobacterium	<input type="checkbox"/> Pea ( <i>Pisum sativum</i> )	<input type="checkbox"/> Mantis (Mantis, <i>Parasphendale</i> , etc.)
<input type="checkbox"/> Escherichia coli	<input type="checkbox"/> Tobacco ( <i>Nicotiana</i> spp.)	<input type="checkbox"/> Six-Lined Hawk Moth ( <i>Manduca sexta</i> )
<input type="checkbox"/> Bacillus subtilis	<input type="checkbox"/> Spinach ( <i>Spinacia oleracea</i> )	<input type="checkbox"/> Fruitfly ( <i>Drosophila melanogaster</i> )
<input type="checkbox"/> Cyanobacteria ( <i>Selenococcus/Selenobacter</i> )	<input type="checkbox"/> Alfalfa ( <i>Medicago</i> spp.)	<input type="checkbox"/> Syrphid Fly ( <i>Syrphidae</i> )
PROTISTA	<input type="checkbox"/> Tomato ( <i>Lycopersicon</i> spp.)	<input type="checkbox"/> Apple Maggot ( <i>Rhagoletis</i> spp.)
<input type="checkbox"/> Acetabularia acetabulum	ANIMAL	<input type="checkbox"/> Mosquito ( <i>Culex</i> , <i>Aedes</i> , <i>Anopheles</i> , etc.)
<input type="checkbox"/> Chlamydomonas reinhardtii	<input type="checkbox"/> Nematode ( <i>Caenorhabditis elegans</i> )	<input type="checkbox"/> Flour Beetle ( <i>Tenebrio</i> spp./ <i>Tribolium</i> spp.)
<input type="checkbox"/> Paramecium	<input type="checkbox"/> Sea Slug ( <i>Aplysia californica</i> )	<input type="checkbox"/> Honeybee ( <i>Apis mellifera</i> )
<input type="checkbox"/> Tetrahymena	<input type="checkbox"/> Sea Slug ( <i>Hermisenda</i> spp.)	<input type="checkbox"/> Parasitic Wasp (Braconids, Pteromalids, etc.)
FUNGI	<input type="checkbox"/> Pond Snail ( <i>Lymnaea</i> spp.)	<input type="checkbox"/> Sea Urchin ( <i>Diadema</i> , <i>Mellita</i> , etc.)
<input type="checkbox"/> Dictyostelium	<input type="checkbox"/> Terrestrial Snail ( <i>Helix</i> spp.)	<input type="checkbox"/> Ascidian ( <i>Boltenia</i> , <i>Molgula</i> , etc.)
<input type="checkbox"/> Neurospora	<input type="checkbox"/> Squid/Cuttlefish ( <i>Loligo</i> , <i>Sepia</i> , etc.)	<input type="checkbox"/> Lancelet ( <i>Amphioxus</i> spp.)
<input type="checkbox"/> Saccharomyces cerevisiae	<input type="checkbox"/> Octopus ( <i>Octopus</i> spp.)	<input type="checkbox"/> Lamprey ( <i>Petromyzon</i> spp.)
<input type="checkbox"/> Schizosaccharomyces pombe	<input type="checkbox"/> Leech ( <i>Hirudo medicinalis</i> )	<input type="checkbox"/> Skate ( <i>Raja</i> , <i>Myliobatis</i> , etc.)
PLANT	<input type="checkbox"/> Horseshoe Crab ( <i>Limulus</i> spp.)	<input type="checkbox"/> Croaker ( <i>Sciaenid</i> Fishes)
	<input type="checkbox"/> Brine Shrimp ( <i>Artemia</i> spp.)	<input type="checkbox"/> Electric Fish ( <i>Eigenmannia</i> , <i>Sternopygus</i> , etc.)
	<input type="checkbox"/> Lobster ( <i>Homarus</i> , <i>Panilurus</i> , etc.)	

<input type="checkbox"/> Goldfish ( <i>Carassius auratus</i> , etc.) <input type="checkbox"/> Perch ( <i>Perca</i> spp.) <input type="checkbox"/> Zebrafish ( <i>Danio (Brachydanio) rerio</i> ) <input type="checkbox"/> Axolotl ( <i>Ambystoma mexicanum</i> ) <input type="checkbox"/> Mudpuppy ( <i>Necturus</i> spp.) <input type="checkbox"/> African Clawed Frog ( <i>Xenopus laevis</i> ) <input type="checkbox"/> Bullfrog ( <i>Rana catesbeiana</i> ) <input type="checkbox"/> Grass Frog ( <i>Rana pipiens</i> ) <input type="checkbox"/> Marine Toad ( <i>Bufo marinus</i> ) <input type="checkbox"/> Turtle ( <i>Chrysemys</i> , <i>Pseudemys</i> , etc.) <input type="checkbox"/> Quail ( <i>Coturnix</i> spp.) <input type="checkbox"/> Chicken Embryo ( <i>Gallus domesticus</i> )	<input type="checkbox"/> House Sparrow ( <i>Passer domesticus</i> ) <input type="checkbox"/> White-Crowned Sparrow ( <i>Zonotrichia leucophrys</i> ) <input type="checkbox"/> Zebra Finch ( <i>Poephila guttata</i> ) <input type="checkbox"/> Opossum ( <i>Monodelphis</i> , <i>Didelphis</i> , etc.) <input type="checkbox"/> Bat ( <i>Antrozous</i> , <i>Eptesicus</i> , etc.) <input type="checkbox"/> Owl Monkey ( <i>Aotus</i> spp.) <input type="checkbox"/> Rhesus Monkey ( <i>Macaca mulatta</i> ) <input type="checkbox"/> Tamarin ( <i>Sanguinus</i> , <i>Leontopithecus</i> spp.) <input type="checkbox"/> Chimpanzee ( <i>Pan troglodytes</i> ) <input type="checkbox"/> Human ( <i>Homo sapiens</i> ) <input type="checkbox"/> Chinchilla ( <i>Chinchilla laniger</i> ) <input type="checkbox"/> Deer Mouse ( <i>Peromyscus</i> spp.)	<input type="checkbox"/> Guinea Pig ( <i>Cavia porcellus</i> ) <input type="checkbox"/> Hamster ( <i>Mesocricetus</i> , <i>Phodopus</i> , etc.) <input type="checkbox"/> Kangaroo Rat ( <i>Dipodomys</i> , etc.) <input type="checkbox"/> Mouse, Laboratory <input type="checkbox"/> Rat, Laboratory <input type="checkbox"/> Vole ( <i>Microtus</i> spp.) <input type="checkbox"/> Domestic Dog ( <i>Canis domestica/familiaris</i> ) <input type="checkbox"/> Domestic Cat ( <i>Felis domestica/cattus</i> ) <input type="checkbox"/> Ferret ( <i>Mustelus</i> spp.) <input type="checkbox"/> Farm Animals (Horse, Sheep, Pigs, Cattle, Goats)  [Enter your own model organism - up to 9 characters] <input type="text"/>
---	--	---



## **PROJECT SUMMARY “TRMS: Cross Species Network Inference - From Models to Crops”**

### **1. Senior personnel**

**PI:** Dennis Shasha (NYU Courant Institute of Mathematical Sciences)

**Co-PIs:** Gloria Coruzzi & Manpreet Katari (NYU Biology, Center for Genomics & Systems Biology)

**Senior Personnel:** Arthur Goldberg (NYU Courant Institute of Mathematical Sciences)

**Collaborators:** Douglas Cook (UC Davis); Rodrigo Gutierrez, Catolica Universita de Chile.

**2. Intellectual merit of the proposed activity** We propose to develop a Cross Species Network Inference (CSNI) platform that will enable plant biologists to easily produce Plant Systems Biology studies for Crop Genomes. CSNI will enable researchers to predict how an interacting network of genes/products in crop genomes will react *as a system* in response to genetic modifications, ultimately for agricultural benefit. To implement this ambitious goal, genome-scale data acquired in a Crop Genome will be integrated into *inferred* gene networks with the aid of validated “ground truth” data (e.g. metabolic, protein interaction, etc.) from Reference Genomes. The result will generate a set of testable hypotheses about gene networks in crops, as well as suggestions for future experiments. This project will leverage the facilities of the current VirtualPlant software platform ([www.virtualplant.org](http://www.virtualplant.org)) developed under an NSF Arabidopsis 2010 Grant (DBI-0445666) including Arabidopsis multinetwork data, analysis, and manipulation tools [1]. As output, we will provide a pipeline of tools for Cross Species Network Inference to the community via a website ([www.CrossSpecies.org](http://www.CrossSpecies.org)) and the NSF *iPlant* Project (see letter). In addition, our CSNI framework will build on the infrastructure of a generic bioinformatic analysis platform engine such as Taverna [2], Kepler [3], or Galaxy [4]. These engines provide general-purpose ease-of-use and reproducible, first-class workflows. As a proof-of-principle, we will apply this Cross-Species Network Inference framework to address an economically important trait, Nitrogen (N) use efficiency in crop species (Rice and Medicago). We will then extend CSNI to other crop genomes, for which experimental data supporting network edges is not yet available. Because many crops lack a large body of experimental genomic data, our network inference approach will be useful for many pathways on many species of economic value. This work will achieve one of the main goals of Systems Biology – predicting network states under untested conditions – which should in turn enable preliminary *in silico* testing of gene manipulations in crop plants prior to testing in the field.

**We divide the work into four aims:**

**Aim 1. Inference and validation of an interaction network in Rice as a proof-of-principle.**

Using a known “ground truth” network in Arabidopsis (e.g. metabolic, protein:protein, and other validated interactions), homology between Arabidopsis and Rice genes, and transcriptome correlation data in Rice, we will create *inferred* interaction networks in Rice. To validate and refine the approach, these *inferred* Rice networks will be compared with “ground truth” networks from Rice.

**Aim 2. Inference of Regulatory Networks.** Perform a time-series N-treatment experiment on Rice to infer the early stages of the nitrogen regulatory network using the “State-Space Analysis” machine learning method. Validate the network based on its predictive accuracy on out-of-sample data.

**Aim 3. Cross-species network inference: N-regulatory networks in a nitrogen fixing Species.** Perform a time-series experiment to measure transcriptome responses of Medicago to nitrogen treatment. Perform network inference analysis on the results, as a case study of the pipeline analysis tools for cross-species network inference in Aim 4.

**Aim 4. Develop A Bioinformatic Pipeline for Cross-Species Inference (CSNI).** Provide a biologist-friendly CSNI software platform ([www.CrossSpecies.org](http://www.CrossSpecies.org)) that will infer networks in a target crop species, given “ground truth” networks in a reference species, homology information and experimental (e.g. transcriptome) data in the target species.

**3. Broader impacts of the proposed research** This project is the result of a long-standing and highly successful collaboration between biologists at NYU and elsewhere, and computer scientists at NYU's Courant Institute of Mathematical Sciences. The systems biology tools, pipelines and data resulting from this project will empower biologists to use genomic data to predict a spectrum of gene networks in biology with broad applications to agriculture and the environment. In addition to scientific results, this collaboration extends to joint training of post-docs and graduate students in Systems Biology.

## TABLE OF CONTENTS

---

For font size and page formatting specifications, see GPG section II.B.2.

	<b>Total No. of Pages</b>	<b>Page No.* (Optional)*</b>
Cover Sheet for Proposal to the National Science Foundation		
Project Summary (not to exceed 1 page)	1	_____
Table of Contents	1	_____
Project Description (Including Results from Prior NSF Support) (not to exceed 15 pages) <b>(Exceed only if allowed by a specific program announcement/solicitation or if approved in advance by the appropriate NSF Assistant Director or designee)</b>	15	_____
References Cited	3	_____
Biographical Sketches (Not to exceed 2 pages each)	8	_____
Budget (Plus up to 3 pages of budget justification)	9	_____
Current and Pending Support	5	_____
Facilities, Equipment and Other Resources	2	_____
Special Information/Supplementary Documentation	11	_____
Appendix (List below. ) <b>(Include only if allowed by a specific program announcement/ solicitation or if approved in advance by the appropriate NSF Assistant Director or designee)</b>	_____	_____
Appendix Items:		

\*Proposers may select any numbering mechanism for the proposal. The entire proposal however, must be paginated. Complete both columns only if the proposal is numbered consecutively.

---

## RELEVANCE AND JUSTIFICATION TO THE STATED GOALS OF THE PGRP

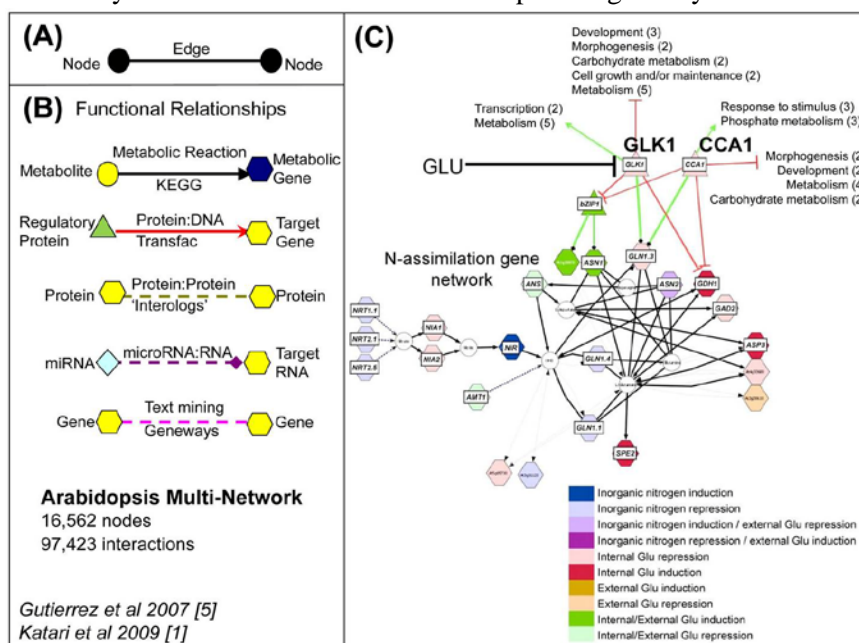
Our proposal applies to **Transferring Research from Model Systems (TRMS)** in that we use Systems Biology tools and approaches developed in the model plant *Arabidopsis* as a platform to develop Cross Species Network Inference (CSNI) approaches for crops of economic importance. We propose to apply cross-species network inference for Rice and *Medicago* using nitrogen-use efficiency as a working example of an agronomically important trait. The approach and pipeline tools we develop will be deployed on a biologist-friendly Web interface that plant scientists can use to infer regulatory networks for any crop or trait of interest. Our project addresses all five of the PGRP's goals:

1. *Advance plant systems biology*: We develop/validate systems approaches for crops (Aims 1 & 4)
2. *Translate basic discovery to field*: Derive N-regulatory networks in Rice & *Medicago* (Aims 2 & 3)
3. *Develop coordinated solutions to data access, analysis and synthesis*: Develop and deploy the Cross Species Network Inference (CSNI) Pipeline ([www.CrossSpecies.org](http://www.CrossSpecies.org)) (Aim 4)
4. *Enhance education, training and outreach*: Training in Plant Systems Biology
5. *Broaden societal impacts of Systems Biology*: Enable *in silico* predictions for modifying traits of agronomic and/or environmental value.

**RESULTS FROM PRIOR NSF SUPPORT** This NSF Plant Genome Project is most closely related to and leverages an NSF *Arabidopsis* 2010 Grant entitled “Conceptual Data Integration for the Virtual Plant” (NSF Database Activities: DBI-0445666), which is in its last year of funding (ending 11/30/10). Below, we highlight aspects of the VirtualPlant Project whose network analysis tools provide inspiration, foundation and support for this NSF Plant Genome Proposal: “**Cross Species Network Inference: From Models to Crops**”.

**I. The *Arabidopsis* Multinetwork: A systems biology tool for hypothesis generation.** Our VirtualPlant work included assembling a multinetwork for *Arabidopsis*, a first step towards a molecular wiring diagram of the plant cell [1, 5]. For a very recent review of this and other plant regulatory networks see [6].

The *Arabidopsis* multinetwork in VirtualPlant has 16,562 nodes (of which 13,960 are genes) and 97,423 interactions (Fig. 1B, Table I). The multinetwork enables us to interpret transcriptome data in the context of all known sources of interaction including protein, DNA, RNA, etc. In one example, a query against the *Arabidopsis* multinetwork with 834 N-regulated genes resulted in a sub-network of 369 genes connected by one (or more) “edges” [7]. In that example, predictions of TF→Target connections were based on significant correlation or anti-correlation ( $>0.7$  or  $<-0.7$  with  $p$ -value  $< 0.01$ ) and statistical over-representation of cis-regulatory elements (CRE) compared to the entire genome [7, 8]. This network



**Fig. 1. The VirtualPlant Multinetwork.** The *Arabidopsis* multinetwork contains genes represented as nodes (A) that are connected by edges of many types (B) including metabolic, protein-DNA, protein-protein, microRNA-RNA, and edges derived from text mining [1]. (C) shows a network neighborhood resulting from querying this multinetwork with microarray data, uncovering a regulatory hub (CCA1) involved in nitrogen signaling [7].

analysis identified potential “master” regulators of this N-responsive sub-network. At the top of the list of network TF “hubs” (with 47 connections to targets in the N-regulatory network) is the central clock control gene CCA1, a Myb family transcription factor (TF) [7]. Exploration of the network “neighborhood” surrounding this CCA1 TF hub revealed connections to target genes in N-assimilation (Fig. 1C). Using Arabidopsis lines that over-express 35S::CCA1 and by Chromatin-IP [7], we showed, using phase response curves, that distinct N-metabolites can advance or delay the circadian phase of CCA1 expression. Thus, we derived and validated the novel hypothesis that N-regulation of CCA1 mRNA expression sets the circadian clock. This is one example showing how the Arabidopsis multinet and associated software tools in VirtualPlant enabled Systems Biology approaches to derive and test new biological hypotheses. Other examples of networks derived and validated using the multinet are reported in [7, 9, 10, 13].

## II. Virtual Plant: A Software Platform for Data Integration, Analysis and Visualization.

The VirtualPlant software platform ([www.virtualplant.org](http://www.virtualplant.org)) [1] integrates genome-wide data concerning the known and predicted relationships among genes, proteins and molecules, as well as genome-scale experimental measurements. VirtualPlant also provides tools that render multivariate information into visual displays (e.g. networks) to highlight biological implications. VirtualPlant's software architecture and data model have been designed and created in a generic, species-independent manner to ease the addition of new organisms and tools.

We have previously demonstrated the use of tools embodied in the VirtualPlant system to generate hypotheses that were validated experimentally [7, 9-13].

**Software and Data Availability:** VirtualPlant is accessible via the website [www.virtualplant.org](http://www.virtualplant.org). Registered users (currently > 630) store their data sets and use many tools to analyze their genomic data such as microarray experiments. The website does not require a password and is available for free when used for non-for-profit purposes.

**VirtualPlant DB:** The VirtualPlant database contains some of the most commonly used data types including metabolic pathways from KEGG and ARACYC, protein-protein interactions from BIND and Interolog databases, and GeneOntology and Gene annotations from TAIR (see Table I for a complete listing of data sources). The database also contains processed data obtained by analyzing publicly available Microarray experiments obtained from NASC [14].

**“GeneCart” Function:** A key challenge to analyzing genomic data is the complex analysis workflow required by currently available software. VirtualPlant integrates multiple tools into a single platform that standardizes the representation of their inputs and outputs so that the output of almost any analysis can be stored in a user’s “GeneCart” and later input to any VirtualPlant analysis tool. This unique feature facilitates Systems Biology’s iterative cycles of data analysis and experimentation [15, 16]. Three working examples described in [1] illustrate the use of VirtualPlant to perform iterative data analyses that build and refine testable biological hypotheses.

**VirtualPlant User Community:** The VirtualPlant user community currently consists of 635 registered academic and commercial users from 36 countries. Among the 347 registered US users, 181 are from

Interaction	Source	# of Interactions	Reference
Biochemical Pathways	KEGG	11,197	Kanehisa et al., 2004 [20]
	ARACYC	17,498	Mueller et al., 2003 [53]
Regulatory Interactions	AGRIS	343	Davuluri et al., 2003 [24]
Protein Interactions	INTERACTOME	39,317	Geisler-Lee et al., 2007 [54]
	AIPID	24,418	Cui et al., 2008 [55]
	BIND	949	Bader et al., 2002 [21]
	MADS BOX	263	De Folter et al., 2005 [22]
	Calmodulin	755	Popescu et al., 2007 [23]
microRNA:mRNA Interactions	Collated by Dr. Pam Green's lab (mirBASE & ASRP)	582	Gustafson et al., 2005 [26] Lu et al., 2005 [27] Griffiths-Jones et al., 2006 [25]
Literature based interactions	GENEWAYS	107	Rzhetsky et al., 2004 [52]

*Katari et al 2009 [11]*

**Table I. Quantitative Information about the Edge Types of the Arabidopsis Multinet.** For detailed description of VirtualPlant and Multinet, see [1].

academia and 166 are from companies. Examples of the latter include: Monsanto, Pioneer, Ceres, Syngenta and Unilever. Other countries that have many users include: UK (78), Australia (27), Germany (24), Chile (22), France (15), Italy (11), Spain (10), Canada (9), Japan (8), Korea (8). Many anonymous users use VirtualPlant but cannot store their datasets for later analysis.

### VirtualPlant's primary tools and functions:

**BioMaps:** BioMaps takes one or more sets of genes and determines which functional terms (GO or MIPS) are statistically over-represented in each set with respect to a background population (e.g. Arabidopsis genome). The output is presented in either a tabular format which can be downloaded to Microsoft Excel or a graphical representation based on the appropriate (e.g. GO) directed acyclic graph.

**Sungear:** Sungear is a visually interactive and biologist-driven exploration of standard questions on many experiments on a genomic scale. Sungear can represent an arbitrary number of experiments/lists, all of their disjoint intersections, and their related ontological terms. The position of a circle and arrows emanating from it indicate the input lists of which it is a subset. The size of a circle is proportional to the number of genes in the intersection of those lists (see [17]). Many biologists find Sungear to be an extremely powerful and interactive tool for analyzing the interrelationships between sets of genes [57].

**Gene Network:** Gene Network analysis allows users to query our Gene Network data and displays the results in a graph using Cytoscape, an open source project for which we have built a plug-in. The tool allows users to filter interactions before displaying the graph [12].

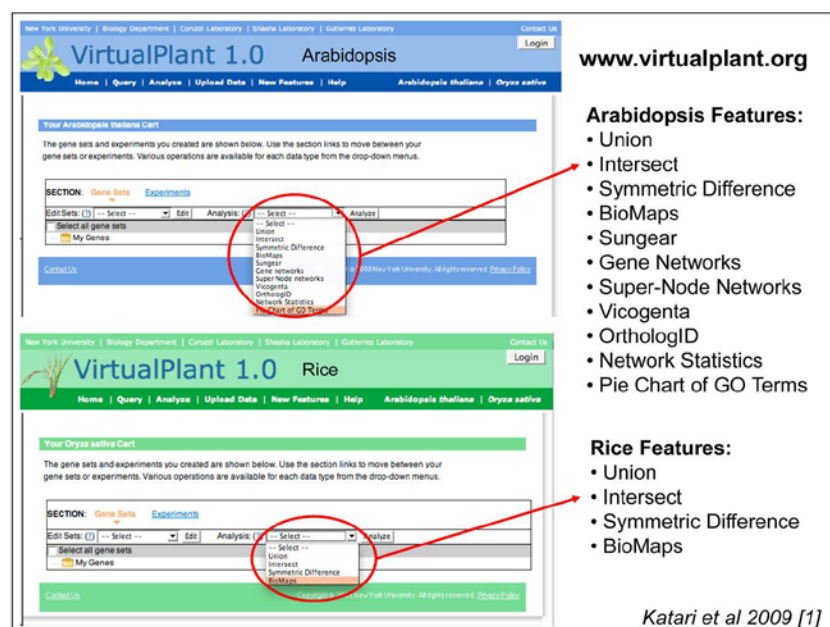
**Supernode Network:** The Supernode Network helps summarize the results of a Gene Network analysis. The genes in the gene network are grouped (Supernode) on the basis of their functional annotations and they are associated with other Supernodes with edges determined from the Gene Network data. A Supernode's size is determined by the number of genes it contains.

**NetMatch:** NetMatch, a cytoscape plug-in, finds all instances of a query graph (e.g. a network motif) in a larger graph [18]. New versions compute the statistical significance of the motifs found.

**Ongoing development of VirtualPlant:** Following an initial three years of funding of the NSF VirtualPlant Grant (DBI-0445666), we received a two-year "creativity extension" (2008-2010) to accomplish two goals. The first was to develop dynamic network modeling tools for Arabidopsis and the second to develop VirtualPlant for

Comparative Genomics. We have approached the dynamic network modeling by applying a machine learning method called "State Space" analysis to time-series data in Arabidopsis learn

regulatory networks [19, 56]. This approach is more fully described in the Research Plan (Aim 2) as it relates directly to the new Cross Species Network Inference (CSNI) approach we will develop in this NSF Plant Genome proposal. We also developed a VirtualPlant version of Rice (Fig. 2).



**Fig. 2. The VirtualPlant Arabidopsis and Rice Home Pages.** The VirtualPlant software platform ([www.virtualplant.org](http://www.virtualplant.org)) supports multiple species [1]. Shown are the two home pages for Arabidopsis and Rice. Each supports a common set of tools but is implemented on top of a separate database. An analysis within a species will not be slowed down by the addition of another species.

## **PUBLICATIONS: Peer reviewed journal articles, chapters and books:**

### **VirtualPlant: Tool development for Plant Systems Biology**

Katari M, Nowicki S, Aceituno F, Nero D, Kelfer J, Thompson L, Cabello J, Davidson R, Goldberg A, Shasha D, Coruzzi G, Gutierrez R (2009) "VirtualPlant: A software platform to support Systems Biology research". **Plant Physiol.** Dec 9 (*Epub ahead of print*).

Nero D, Kelfer J, Katari M, Tranchina D, Coruzzi G (2009) "In silico Evaluation of Predicted Regulatory Interactions in Arabidopsis thaliana". **BMC Bioinformatics.** Dec 21;10(1):435.

Poultney C, Gutierrez R, Katari M, Gifford M, Paley W, Coruzzi G and Shasha D (2007) "Sungear: Interactive visualization, exploration & functional analysis of genomic datasets". **Bioinformatics,** 23:259-61.

Ferro A, Giugno R, Pigola G, Pulvirenti A, Skripin D, Bader G, Shasha D, "NetMatch: a Cytoscape Plugin for Searching Biological Networks" **Bioinformatics,** 2007 23(7):910-912.

### **Applications of VirtualPlant: Hypothesis Generation and Testing**

Krouk G, Tranchina D, Lejay L, Cruikshank A, Shasha D, Coruzzi G and Gutierrez R (2009) "A systems approach uncovers restrictions for signal interactions regulating genome-wide responses to nutritional cues in Arabidopsis." **PLoS Comp Biol.** Mar;5(3):e1000326. (*Highly Accessed*).

Gutierrez R, Stokes T, Thum K, Xu X, Obertello M, Katari M, Tanurdzic M, Dean A, Nero D, McClung R and Coruzzi G (2008) "Systems approach identifies an organic nitrogen-responsive gene network that is regulated by the master clock control gene CCA1" **Proc. Natl Acad Sci USA** 105, 4939-4944. (*Faculty of 1000 recommended: Factor 3*)

Gutierrez R, Gifford M, Poultney C, Wang R, Shasha D, Coruzzi G, Crawford N (2007) "Insights into the genomic nitrate response using genetics and the Sungear Software System" **Journal of Experimental Botany** doi: 10.1093/jxb/erm079

Gutierrez R, Lejay L, Chiaromonte F, Shasha D, Coruzzi G (2007) "Qualitative network models and genome-wide expression data define carbon/nitrogen-responsive biomodules in Arabidopsis" **Genome Biology,** 8: R7. *Faculty 1000 (Must Read: Factor 6)*

### **Plant Systems Biology: Reviews, Books and Outreach**

Ruffel S, Krouk G, Coruzzi G (2009). "A Systems View of Responses to Nutritional Cues in Arabidopsis: Towards a Paradigm Shift for Predictive Network Modeling". **Plant Physiol.** Nov 25 (*epub ahead of print*)

Gutierrez R, Coruzzi G., Eds (2009) Book: "Plant Systems Biology", **Annual Plant Reviews;** Blackwell Publishing: Oxford, UK, 2009, Vol. 35. 360 pages.

Coruzzi GM, Burga A, Katari MS, and Gutierrez RA (2009) "Systems Biology: Principles and Applications in Plant Research". In "Plant Systems Biology", **Annual Plant Reviews;** Blackwell Publishing: Oxford, UK, 2009, Vol. 35. Pgs 3-31. *Book Chapter.*

Gifford M, Gutierrez R, and Coruzzi G (2006) "Modeling the Virtual Plant: A Systems Approach to Nitrogen-Regulatory Gene Networks". Essay 12.2 Chapter 12. Assimilation of mineral nutrients; In **A Companion to Plant Physiology,** Fourth Edition, Lincoln Taiz and Eduardo Zeiger, <http://4e.plantphys.net/article.php?ch=12&id=352>

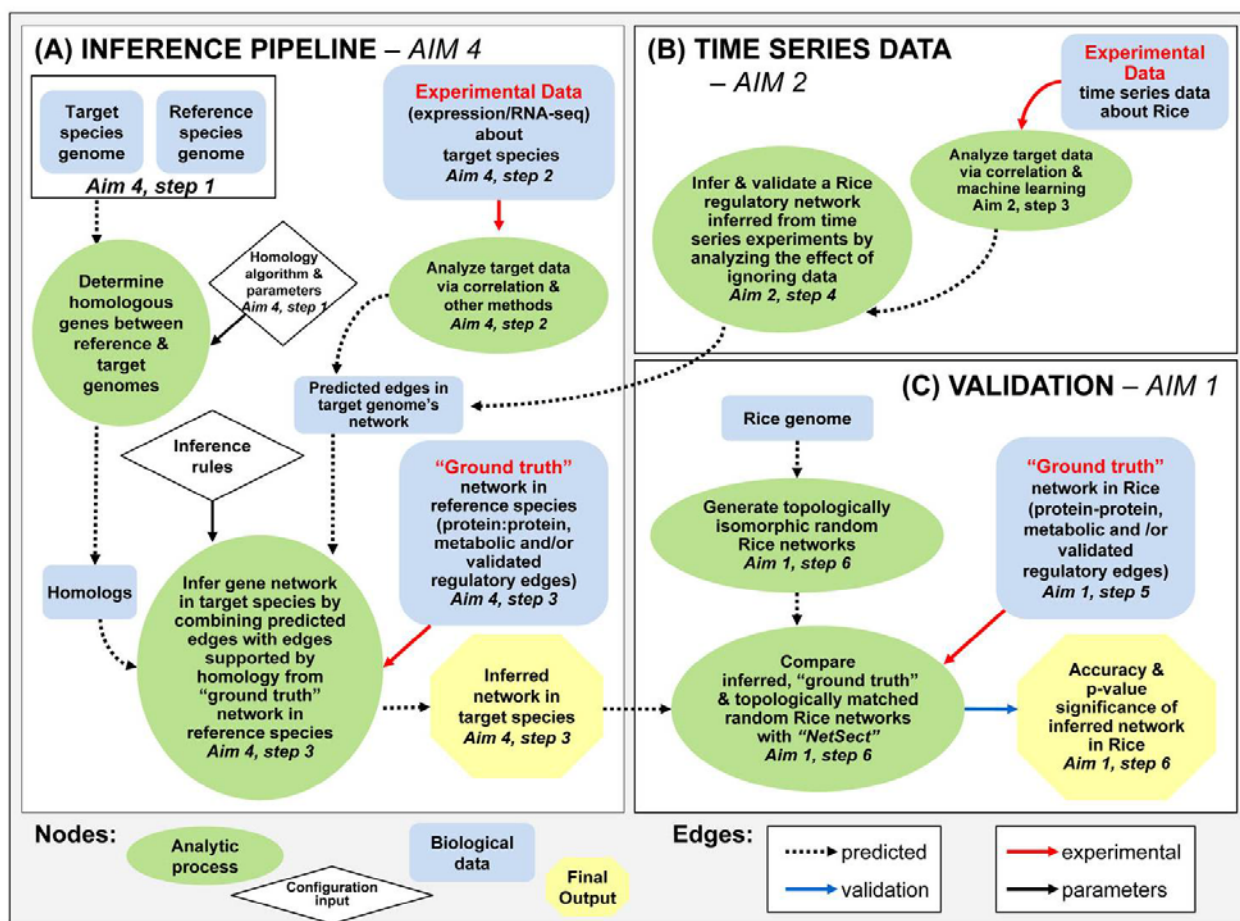
Gutierrez R, Shasha D and Coruzzi G. (2005) "Systems Biology for the Virtual Plant". **Plant Physiol.** Vol 138, pp 550-554.

**Education and Training:** The development of the Systems Biology tools and the Virtual Plant software platform has trained undergraduates (UG), MS and PhD students in Systems Biology. Students trained include **Undergraduates:** Steve Nowicki (NYU CAS), Varuni Prabhakar (Barnard College), Rebecca Davidson (BS Computer Science); **Masters Students:** Ana F. Arroja (MS student, NYU Courant), Ranjita Iyer (MS Computer Science), Jonathan Kelfer (MS Computer Science), Lee Parnell (MS Computer Science), (Jarod Wang, MS Computer Science); **PhD Students:** Chris Poultney (PhD student, NYU Courant), Jason Reisman (PhD student, NYU Courant), Saurabh Kumar (PhD student, NYU Courant). These students have gone on to PhD programs (Prabhakar, Parnell) as well as to industry (Kelfer, Bloomberg).

## RESEARCH DESIGN

### Aim 1: Inference and validation of interaction networks in Rice as a proof-of-principle.

**Rationale:** This aim will use the data-rich genomic resources of Arabidopsis and Rice to evaluate our Cross Species Network Inference (CSNI) approach by inferring an interaction network in Rice and then validating the network against Rice “ground truth” data. The data used in the network inference will be: 1) Arabidopsis “ground truth” data that includes metabolic pathway data obtained from KEGG [20], protein interaction data obtained from the Biomolecular Interaction Network Database (BIND) [21] plus other experimentally determined protein interactions [22, 23], 2) Rice-Arabidopsis gene homology, and 3) Rice microarray expression data. This *inferred* Rice network will be compared to the *known* “ground truth” data for Rice including metabolic data from KEGG and protein:protein interaction data from BIND. Our goal is to use the well-known gene interaction datasets for Arabidopsis and Rice to develop and validate a methodology for inferring networks for other species, and later apply the CSNI pipeline to crop species that have little interaction information (e.g. Medicago). See Fig. 3 for overall plan.



**Fig. 3. Network inference and Validation Pipelines for Cross Species Network Inference (CSNI).** The cross species network inference suite will contain: (Panel A) a CSNI inference pipeline, (Panel B) Time-series data collection, analysis and prediction and (Panel C) Validation. **Panel A:** The depicted CSNI pipeline will combine inference based on homology from a reference species (Aim 4, steps 1 & 2) and experimental data in a target species (Aim 4, step 2) to obtain an *inferred* network in the target species (Aim 4, step 3). **Panel B:** Aim 2 provides a method to infer a network using closely spaced time series data (e.g. for Rice and Medicago, see Aims 2 & 3) by using a machine learning technique called “State Space” modeling [56]. Validation is based on prediction on unsampled data. **Panel C:** When a validated “ground truth” network is available in the target species, the inferred network can be evaluated by using the pipeline from Aim 1 (Steps 5 & 6). Such validations will lead to improvements to homology parameters (Aim 4, step 1) and inference rules (Aim 4, small diamond).

We will also use the Arabidopsis and Rice datasets to test analogous techniques for inferring regulatory networks using other interactions including TF→Target (AGRIS) [24] and miRNA-RNA interactions [25-27]. Combining these types of inferred edges will eventually lead to a method for creating *inferred* multi-networks for any crop species.

The first test of our CSNI approach uses 1) metabolic and protein:protein interaction data from Arabidopsis as reference species “ground truth”, 2) Arabidopsis-Rice homology (using BLAST), and 3) expression correlation data (from Rice). Of the available metabolic pathway databases (e.g. KEGG, AraCyc, MetaCyc, etc.) [28, 29], we selected KEGG for these initial studies and will evaluate the others in our parameter optimization testing, described below. Compared to the metabolic data, the protein:protein interaction data is very partial and prone to false positives. However, we included it in our analysis because: i) previous studies have shown that proteins that interact are more likely to be co-expressed [15, 30]; ii) the “Interolog” approach showed that protein interactions could be predicted using BLAST scores [31], and iii) the protein:protein interaction dataset will grow significantly under the NSF Plant Genome Arabidopsis Interactome Project (<http://signal.salk.edu/interactome.html>). Studies have also shown that major enzymes in metabolic pathways are co-expressed in different conditions [32].

**Step 1. Obtain a reference “ground truth” Arabidopsis interaction network based on experimentally supported data.** For our “ground-truth” Arabidopsis networks, we have assembled metabolic interactions (KEGG; 19,688 interactions) [20], protein:protein interaction data from BIND (949 interactions) [21], protein-chip interaction data for MADS box (272 interactions) [22] and protein chip interactions for Calmodulin (755 interactions) [23]. Although we refer to the metabolic and protein interactions data as “ground truth”, many of the pathways in the KEGG and AraCyc databases are based on computational predictions, while 25% are validated in the literature [28, 29].

**Step 2. Identify Rice homologs of Arabidopsis interaction pairs.** Connect every gene in the Arabidopsis interaction network with its Rice homologs. This technique can employ various homology methods, distance or parsimony based. In our preliminary analysis (Table II), we analyzed one-to-one homology by obtaining reciprocal top BLAST pairs. We also used distance-based BLAST with an e-value cutoff of  $E^{-20}$  to capture one-to-many homology relationships [33] which captures the gene duplication events prevalent in plant genomes [34].

**Step 3. Build a Rice correlation network based on publicly available Rice microarray expression experiments.** We downloaded 32 Rice microarray experiments from GEO [35], log transformed the MAS5 [36] normalized values, and Pearson correlated all pairs of the genes whose measurement variances (after normalization) lie in the upper 20% of all genes. We inferred a correlation edge between gene pairs whose expression vectors were significantly correlated (p-value  $< 0.05$ , meaning less than a 5% chance of a non-zero correlation by chance) and correlation value  $> 0.5$  or  $> 0.7$  (Table II).

**Step 4. Build an *inferred* Rice network.** A pair of Rice genes in our *inferred* Rice network may be connected by no edge, by only an expression correlation edge, by only a homology edge from Arabidopsis, or by both of these types of edges. We focus on Rice gene pairs connected by both types of edges – homology and correlation. This network is called the *inferred* Rice network.

**Step 5. Obtain a reference “ground truth” Rice network that contains edges representing known interactions.** Our initial Rice “ground truth” network was constructed from 10,976 metabolic interactions and 334 protein-protein interactions for Rice from KEGG [20] and BIND [21], respectively.

**Step 6. Evaluate *Inferred* Rice Network:** This step computes the similarity and p-value (significance) between the *inferred* and “ground truth” Rice networks by using a network intersection tool called *NetSect* which is described below. We evaluated the quality of each subset of edge types in the *inferred* network.

***NetSect*: Evaluating the Accuracy of the *Inferred* Network.** Given networks  $N$  (“inferred”) and  $M$  (“ground truth”), with edges  $E(N)$  and  $E(M)$  respectively, one can measure their similarity by computing



$size( intersection( E(N), E(M) ) ) / size(union( E(N), E(M) ) )$ , which equals 1 when  $E(N)$  and  $E(M)$  are identical and zero when they are disjoint. We will also compute the recall and precision of the *inferred* network's ability to predict edges in the reference "ground truth" network. To compute a p-value for the *inferred* network's reconstruction of the reference network, *NetSect* computes the similarity of the inferred and ground truth networks and then computes a p-value by comparing the sample similarity with the similarity of a collection of random networks having the same topology (i.e. isomorphic) as the inferred network, with vertices drawn from the entire genome. This use of randomness corresponds to the null hypothesis that the inferred network is no better than a random choice of edges.

**Step 7. Expand "ground truth" and network inference into a "multinetwork" containing multiple edge types.** We will use techniques analogous to Steps 1-6 to infer networks based on other edge types. For example, we will add regulatory interactions including protein→DNA (AGRIS: 343 interactions) [24] and miRNA:RNA interactions [25-27]. Expanding the "ground truth" networks to include these datasets will enable us to create an inferred multinetwork that includes: protein:protein, Protein:DNA, miRNA:RNA and Metabolic edges.

Metabolic Interactions		Homology			
		Reciprocal Best Hit		Blast e-value < 1e-20	
		Inferred	Validated	Inferred	Validated
		3594	1883 (52.4%)	1,268,094	7045 (0.56%)
Correlation	> 0.5	850	547* (64.4%)	23,686	1464* (6.2%)
	> 0.7	387	275* (71.1%)	8,745	649* (7.4%)

**Table II. Validation of Network Inference using Arabidopsis (reference) and Rice (target).** Inferring interaction relationships in Rice based on homology alone (to Arabidopsis) data (using Rice expression data) yields high precision relative to the "ground truth" network (of Rice). Combining the two (homology and correlation) gives even greater precision at some cost in recall. The use of the asterisk (\*) connotes statistically significant improvement in precision

hypothesize that many of the remaining 29% of predicted edges may represent true interactions that are currently missing from the Rice KEGG metabolic database.

By contrast, due to the small number of "ground-truth" protein:protein interaction edges in both Rice (334 interactions) and Arabidopsis (1,594 interactions), the inferred Rice network barely overlaps the known "ground truth" Rice protein-protein interaction network (not shown). Of the 106,944 inferred protein interactions based on homology defined by BLAST e-value < 1E-20 only 42 (0.04%) are found in the Rice "ground truth" protein interaction data (Table not shown). Addition of correlation values enhances the predictions slightly (though statistically insignificantly, p-value = 0.27): out of 2,485 inferred predictions based on intersect of homology (e-value < 1e-20 and correlation > 0.7) only 2 (0.08%) were present in the Rice "ground truth" protein interaction network (not shown).

**Summarizing our preliminary results,** the combination of homology and correlation predict metabolic edges more accurately than they predict protein-protein edges, based on current (limited) datasets for the

We suggest two main conclusions from our preliminary analysis of Cross Species Network Inference (Steps 1-6 above) shown in Table II. *First*, homology alone does an excellent job of creating an inferred network for metabolic edges. Of the 3,594 edges in the Rice metabolic network inferred via reciprocal top hits, 52.4% or 1,883 are validated in the Rice "ground truth" KEGG metabolic interactions. *Second*, correlation significantly enhances the prediction of the inferred Rice network. Of the 387 inferred Rice metabolic interactions predicted with the

intersection of homology (reciprocal top hit) and correlation (>0.7), 275 inferred interactions (or 71.1%) are validated by the Rice metabolic "ground truth" network, which is a statistically significant improvement in precision (p-value < 0.001) (Table II). Based on *NetSect* analysis, the predictive power of the reciprocal top-hit inferred Rice metabolic network is significant (p-value < 0.001), with or without expression correlation data.

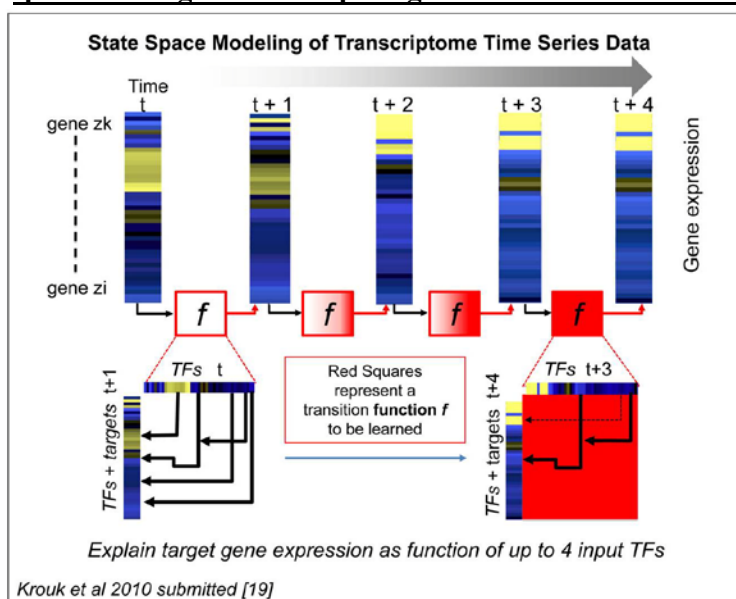
The precision of this prediction is so high, that we

latter. The homology reciprocal top BLAST hits produces the highest precision. Homology provides higher precision than correlation, for both metabolic and protein:protein edges (not shown).

**Parameter optimization.** As one would expect, the choice of data sources and homology algorithms and parameters for CSNI, greatly influences the accuracy of the inferred Rice networks. To simplify the selection of these parameters for biologists, we will systematically explore the space of these inputs, with the objective of maximizing the accuracy of our network inference predictions. A well-known technique for finding globally optimal parameters is “*simulated annealing*”, which is a probabilistic heuristic for finding global minima in large search spaces [37]. Potentially, the techniques we develop to infer networks for Rice can also be applied to the other fully sequenced crop species. Ideally, the experiments used for gene expression correlation will include many different developmental stages, different organs, and different biotic and abiotic treatments such as the ones just recently released for Rice on GEO NCBI [38].

**Expected Outcomes of Aim 1:** We will expand this cross species network inference and validation analysis to include other homology methods (e.g. parsimony-based homology [39] and other methods like COGS [40], InParanoid [41], for example). We will also expand our data sources to larger datasets for expression and protein interaction as they become available. This Aim provides a testing ground and validation for the CSNI pipeline approach that we will automate in Aim 4.

**Aim 2: Inference of regulatory networks: Develop time series expression data in two different species having well developed “ground truth” networks to infer regulatory networks.**



**Fig. 4. State-Space modeling: A machine learning approach to network inference.** State-space modeling fueled by regulatory data (transcriptome depicted as heat map) at closely spaced time points seeks to explain the expression of a target gene  $X$ , as a function of the expression of one or more other regulatory genes (e.g. transcription factors, TFs) as a fixed relationship ( $f$ ) between genes. Even though  $f$  is fixed, gene expressions can vary in value because their input genes (e.g. TFs) as well as signals to which they respond (e.g. nitrogen signals) can change over time. The function  $f$  is “simple” (or “regularized”) in the sense that each target gene in the model is forced to depend on no more than three or four input TFs. Function  $f$  is computed through a cyclic series of steps of the form guess, compute error, and then refine  $f$ , using the time-series regulatory data [19, 56].

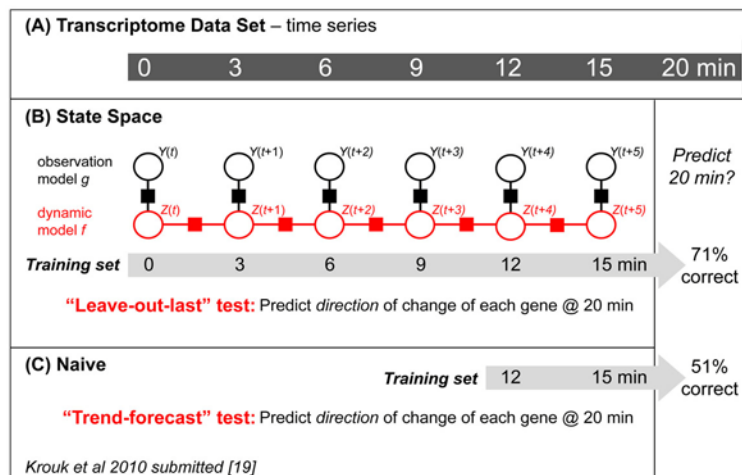
**Rationale:** In this aim, we adopt the nitrogen-treatment time-series and machine learning approach we have already used with Arabidopsis [19, 56] to infer regulatory networks for Rice. We first discuss our time-series approach in Arabidopsis, next discuss the experimental technique we will use in Rice, and finally describe the analytical framework for inferring Rice regulatory networks along with a validation strategy using experimental data.

**Predicting Arabidopsis regulatory networks using time series data and “State Space” analysis a machine learning approach.** The goal of our Arabidopsis time-series work was to

uncover gene relationships and consequently to build core regulatory networks involved in Arabidopsis root adaptation to  $\text{NO}_3^-$  provision [19]. The experimental approach was to monitor transcriptome responses to  $\text{NO}_3^-$  treatment at 0, 3, 6, 9, 12, 15 and 20 min, using ATH1 chips. This high-resolution time course analysis demonstrated that the previously known “primary nitrate response” of 20 min (e.g. genes in the nitrate reduction pathway) [11], is

actually preceded by very fast (within 3 min) regulation of genes including TFs [19]. In order to build a regulatory network that could predict these TF→target interactions, we used a machine learning method, “State-Space” modeling to generate predictions for regulatory networks [56]. The State-Space model synthesizes Bayesian and Markovian approaches (in which each gene’s expression value at a time  $t$  is assumed to depend directly only on the state of potentially all the genes at the previous time point and indirectly on values from previous time points) and entails the construction of a “trellis” indicating causalities, as shown in Fig. 5B [42, 56]. In the spirit of the close collaboration between biologists and computer scientists in this project, we depict the state space model from a biologist’s view (Fig. 4) and a computer scientist point of view (Fig. 5).

In the “State Space” model depicted in Fig. 5B, each node represents the values of all gene expressions at a particular time point. Typical values of all gene expressions are depicted as a heat map in Fig. 4. The goal of this approach, is to *learn* the function that determines the change in expression of a target gene  $z_j$ , as a linear (or if needed non-linear) combination of the expression of a relatively small number of transcription factors (typically up to three or four) (Fig. 4). As applied to our problem, the set of all genes at time  $t$  is modeled by a “latent” variable (denoted  $Z(t)$ ) from which noisy and sometimes missing observations  $Y(t)$  are made. Latent variables are represented by large red circles, and observed variables by large black circles in Fig. 5B. The relationship between latent and observed variables is the identity function  $h$  with added Gaussian noise (represented by a black square in Fig. 5B). An unknown *function*  $f$  (represented by a red square in Fig. 5B) relates the values of latent variables  $Z(t)$  and  $Z(t+1)$  (for all  $t$ ) corresponding to consecutive time measurements. The *dynamical function*  $f$  factors in both transcription factors and their target genes (e.g. other TFs or target genes), as shown in Fig. 4. The “State Space” learning algorithm iteratively infers the function of latent values of transcription factors that determines the changes to target genes. Learning the *function*  $f$  corresponds to finding parameters of  $f$  that minimize the prediction error while penalizing functions that are excessively complex (i.e. require many transcription factors to determine the change in expression of a target).



**Fig. 5. Prediction and Validation Using a Network Constructed from State-Space Modeling.** In the state space model depicted in (A), each node represents the values of all gene expressions at a particular time point. The set of all gene expressions at time  $t$  is modeled by a “latent” (i.e., hidden) variable (denoted  $Z(t)$ ) from which noisy and sometimes missing observations  $Y(t)$  are made. An unknown *function*  $f$  (represented by red square) relates the values of latent variables  $Z(t)$  and  $Z(t+1)$  (for all  $t$ ) corresponding to consecutive time measurements. In (B) Validation of the predictive modeling is tested by the ability to accurately predict the direction of change of each gene at 20 min for state space modeling (B) vs. trend-forecast test (C) [19, 56].

To test the ability of the “State Space” approach to generate a *predictive* regulatory network, we built a regulatory network using the Arabidopsis time-series data up to 15 minutes (training set: 0, 3, 6, 9, 12, 15 min) and used the resulting network to *predict* the direction of gene change (up regulation or down regulation) from 15 min to 20 min (Fig. 5). Our State Space predictions of gene regulation were correct for 70.7% of the genes (Fig. 5B). As a basis for comparison, the “naive trend forecast” that predicted the direction of change from 15 to 20 min to be in the same direction as the movement from 12 to 15 min, was correct for only 51.9% of the genes, just slightly better than random (Fig. 5C). Thus, our “State-Space” model does significantly better ( $p$  value= 0.006) than the “naive trend forecast” test based on a binomial test on a coin that is biased to be correct 51.9% of the time. This “State Space” model can also be

used to predict the “most influential TFs” in the network (e.g. the one that is predicted to influence the most genes in the network), and to generate a time-dependent regulatory network model for the control of N-assimilatory pathway genes.

**Creating a state-space regulatory network for Rice.** We will use a similar time series-based approach to examine nitrogen regulatory networks in Rice. For this, we will use similar growth and N-treatments in Rice that we have used in Arabidopsis, and for which we have shown Nitrogen-regulation of gene expression in Rice. Rice seedlings (approximately 14 days old) will be grown in nitrate-free sterile hydroponic media containing 0.5 mM NH<sub>4</sub><sup>+</sup> succinate. Plants will subsequently be N-deprived for 24 hours, then treated with 1 mM KNO<sub>3</sub>- and then shoot and root tissue will be harvested at 0, 3, 6, 9, 12, 15, 20, 25, 35, 45, 60, 120, 180 and 240 min. Controls are "T0" (harvest time zero, before treatment) and 1 mM KCl at each of these time-points. To select time-points for transcriptome analysis, "sentinel" NO<sub>3</sub>-regulated genes involved in early steps of the nitrate response (e.g. nitrate uptake or reduction) will be monitored by Q-PCR [11, 43, 44]. Based on these Q-PCR results, an initial set of RNA time points will be analyzed using transcriptomics in biological duplicates. We will prepare mRNA samples for RNA-seq (Illumina) using protocols we have previously used for Arabidopsis. Briefly, polyA mRNA will be sheared and sized to ~50-60 bp and ligated to adaptors at 5' and 3' ends for reverse transcription and PCR amplification. 14 cycles will be used for enrichment, to minimize PCR bias, but produce enough material for Illumina sequencing (see McCombie letter, CHSL).

**Step 1. Identify the “regulated” genes in the target species (Rice).** To find the genes that are N-regulated in Rice in our time-series data, we will use criteria similar to those used for Arabidopsis to minimize the False Discovery Rate (FDR). As we did for Arabidopsis time-series data [19], we will run an ANOVA (aov() function) over the data set where the signal of a probe  $i$  is  $P_i \sim \mu + \alpha N + \beta T + \gamma T*N + \epsilon$  where  $N$  is the effect of the Nitrate treatment,  $T$  is the effect of time, and  $T*N$  is the effect of their interaction,  $\mu$  is the mean signal over the data set, and  $\epsilon$  the unexplained variance. We determine for each gene the particular time point that showed the most marked effect of nitrate. We call a probe “regulated” if it has a positive call (FDR <5%) of interaction with nitrate at that particular time point. This will lead us to the identification of a nitrogen-regulated network having regulators such as transcription factors (TFs) and targets of regulation (all genes – both targets and TFs).

**Step 2. Find inferred regulatory and homology-supported correlation edges between Arabidopsis and Rice.** As a starting point for further analysis, we will next find edges between Rice genes (e.g. R1.1, R2.2) having strong Pearson correlation, where Rice gene R1.1 is a transcription factor, for which binding site over-representation also gives support [7, 8], and such that homologous genes in Arabidopsis (A1, A2) also comprise a putative edge in the transcription factor network generated from the time series data. We seek Pearson correlation between Rice genes having p-value < 0.01 (that is a 0.01 chance that the correlation is due to chance). We will use these inferred TF→Target regulatory edges among Rice genes to “prime the pump” in the “State-Space Modeling” analysis of the next step.

**Step 3. Infer the regulatory network using “State Space” analysis.** Using the homology-aided correlation edges from Step 2, we set the initial weights of the TF→target pairs for the State-Space modeling approach [56]. Specifically, gene pairs that are positively or negatively correlated in the analysis of Step 2, will have large (positive or negative, respectively) weights. As State-Space modeling is an iterative algorithm, those weights will be adjusted in the course of error reduction. Better initial edge weights can improve performance. For example, in our Arabidopsis study, the knowledge that certain genes were transcription factors improved the prediction accuracy of State Space analysis by 5% (from 67% to nearly 71%) (Fig. 5B) [19].

**Step 4. Validate the *inferred* Rice regulatory network and plan new experiments.** Validation here will mean prediction of Rice regulatory networks under untested conditions, and validation of these predictions using experimental data. Just as we did for Arabidopsis, we will build a network with the Rice time series transcriptome data, but leaving out certain time points, and then predict the values at those

missing time points and compare the predicted values with the actual ones. If the network predictions are not statistically significantly better than random based on p-value analysis, we will analyze more time-series experiments. To decide which new time series data points to use, we will make use of the following simple heuristic: perform a new measurement of a sample that is most similar to the most useful measurement just done as judged by prediction accuracy. For example, we can determine which of the existing samples are most valuable by removing them and computing prediction accuracy. In the case of our Arabidopsis time-course study [19], removing two replicate experiments from two different time-points prior to 15 minutes is less harmful to the accuracy of the prediction of the network state at 20 min than removing both replicates from a single time-point prior to 15 min. This suggests that measurements at different time-points are more valuable than replicates.

**Expected outcomes of Aim 2:** The results of this aim will generate an inferred nitrogen regulatory network for Rice that will identify the “most influential” transcription factors affecting the entire N-response “system”. From this, we can identify a core set of conserved regulatory networks (e.g. conserved between Rice and Arabidopsis). If that core set is large, Arabidopsis could be used as a true reference species to test the role of these TFs in nitrogen-use efficiency (using T-DNA mutants and overexpressors), followed by translational applied studies in Rice. That is, conserved factors shown to affect N-responses in Arabidopsis will set the stage for future translational studies in mutant Rice available from the Oryza Tag Line collection (<http://urgi.versailles.inra.fr/OryzaTagLine/>).

**Aim 3. Cross-species network inference: N-regulatory networks in a Nitrogen Fixing Species.**

**Rationale:** The experiments in this aim constitute a time-series study for Cross Species Network Inference that probes the unique nitrogen biology of legumes, i.e., the ability of legumes (such as Medicago) to switch between assimilation of external N and symbiotic N. This time-series data from Medicago will feed a Cross Species Network Inference pipeline compared to Arabidopsis, as described in Aim 4, which is based on the approaches validated for Arabidopsis and Rice, as discussed in Aims 1 & 2.

**Plant Growth conditions and N-treatments:** Medicago plants will be grown (by Doug Cook, UC Davis, see letter) in aeroponic chambers, which offer the advantage of easy access to root tissue under conditions that promote rapid and highly uniform responses to symbiotic Rhizobium.

**Aim 3A. Analyze a time-series of N-responses in “naïve” (i.e., non-symbiotic) plants.** In an initial set of experiments, we will analyze the nitrate response of “naïve” (non-symbiotic) plants. Prior nitrogen status will be established by supplementing plants with ammonium succinate for 14 days, followed by 24 hrs of N-deprivation, and treatment with 1 mM nitrate or control KCl. This again (as in Aim 2) parallels time-series N-treatments conducted in Arabidopsis and are conditions shown to affect the regulation of sentinel N-regulated genes involved in N-assimilation [19]. The time-series for nitrogen transcriptional responses will be 0, 3, 6, 9, 12, 15, 20, 25, 35, 45, 60, 120, 180 and 240 min. Transcript levels for sentinel nitrate-responsive genes will be assayed in both root and foliar tissue samples. We will follow the methods used described in Aim 2, to analyze RNA transcript levels using Illumina deep-seq.

**Aim 3B. Analyze the N-responses of plants involved in active symbiosis.** In a subsequent time series experiment, we will analyze the transcriptional response of symbiotically active Medicago truncatula to exogenous nitrate. It is well established that exogenous nitrate is inhibitory to nodule function [45]. However, it is uncertain how the immediate transcriptional response to nitrate varies between tissues of nodulated plants (i.e., foliar tissue, nodules and roots), and how this compares to nitrate responses in naive legumes and to that of non-legumes. This experiment will follow the protocol of the naive plant experiment (Aim 3A), with the following exceptions: (1) prior nitrogen status will be established by the action of bacterial nitrogenase to generate ammonia within nodules, and (2) transcription responses in roots will be assayed by partitioning the symbiotic roots into a "root" fraction (where the N-response would be activation of N-assimilation in response to external N cues) and a "nodule" fraction (where the N-response would be inhibition of N-fixation in response to external N cues).

We note that previous studies have measured the transcriptional status of legumes to varying nitrogen treatments (e.g., [46-48]). However, those studies have focused on long-term, multi-day treatments, and often on biological features or experimental designs that are unique to legumes. Thus, these prior studies report the status of plants acclimated to different nitrogen sources. They have not described nitrogen responses per se, nor have they provided data sets that permit direct comparison to the extensive N-response time-series data sets we have already generated in Arabidopsis [19] and those that we will generate in Rice (in Aim 2).

**Expected outcomes of Aim 3.** The experiments in this aim probe the unique biology of legumes, i.e., the ability of legumes (such as Medicago) to switch between assimilation of external N and symbiotic N. Even though Medicago genomics (e.g. annotation and interaction data) lags behind Rice, Medicago has the advantage of tagged mutant lines (TMT1 mutant lines from the Noble foundation (<http://bioinfo4.noble.org/mutant/>) that can be used routinely to test candidate genes. This will contribute to hypothesis testing of “highly influential” TFs in the inferred N-regulatory networks in Medicago in our collaboration with D. Cook (see letter of collaboration).

#### **Aim 4: A Bioinformatic Pipeline for Cross-Species Network Inference (CSNI).**

**Rationale:** In this aim, we will build a publicly available, production quality, Cross-Species Network Inference (CSNI) pipeline that will provide the plant scientist community (especially those with no informatics training) with a biologist-friendly tool for inferring gene networks in crops. CSNI employs data about two species, 1) the crop – which we call the *target* species, and 2) a species that has been deeply studied, which we call the *reference* species. The basic idea of CSNI is that the larger data set from the reference species will be mapped by homology into the target species, and combined with data about the target species to infer a network for the target species.

Figure 3A illustrates the CSNI pipeline. It involves three primary processes: 1) obtain homologs between the reference and target species, 2) obtain and analyze experimental data for the target species, and 3) use these data and a “ground truth” (experimentally validated) gene network for the reference species to infer a putative gene network for the target species.

The following steps describe the operation of CSNI in *Inference*, or prediction, mode. In this mode, a plant scientist who wants to infer a gene network for a target species will set the free parameters that determine the homology and inference methods of CSNI. These include i) the pair of reference and target species chosen, ii) the data sets selected from these species, iii) the homology mechanism and its parameters (such as BLAST and E-value thresholds if distance-based homology is desired), and iv) the *Inference* rules which combine these data into the target species' inferred network. As described in Aim 1, Step 7, these parameters can sometimes be set automatically based on a “ground truth” network in the target species and optimization methods.

#### **Operation of Cross-Species Network Inference (CSNI) in *Inference* mode:**

**Step 1. Choose a target (crop), reference species and Homology algorithm.** On the web site (see a mockup of the CSNI GUI in Fig. 6), the plant scientist selects the reference species and the target species. Any species can conceivably serve as the reference species provided: i) it is phylogenetically close to the target species (even as distant as Rice-Arabidopsis); and ii) it has data to support the construction of a “ground truth” network. The plant scientist next chooses a homology algorithm and its parameters, such as BLAST and its E-value cutoff, or a parsimony mechanism, and CSNI generates a set of homologous reference-to-target gene pairs. Preliminary tests in Aim 1, showed that reverse top hits gave high precision for Arabidopsis (as reference) and Rice (as target). Further work will determine whether this is the best distanced-based homology strategy and how this compares to parsimony based homology methods. In addition, we will allow users to upload any cross-species homologous gene pairs determined using their preferred method (e.g. COG [40], InParanoid [41] OrthologID [39], or homology pairs generated using analysis platforms such as Taverna [2], Kepler [3], or Galaxy [4], for example).

In our working example, we use Arabidopsis as our reference species for Medicago, because the Arabidopsis genome contains a large number of nodulin-like genes [49]. Moreover, our analysis showed that of a list of 1,458 potential Arabidopsis nodulation gene homologs (collated from Allometra database (<http://allometra.com/nodprots.shtml>)), 36% of these genes were N-regulated; the majority were N-depressed, consistent with externally supplied N inhibiting nodulation in legumes. As early nodulation events are inhibited by external nitrogen application, potentially homologous genes in Arabidopsis and Medicago and their associated N-regulated gene networks could provide insights into N-regulated root development and to the events involved in N-repression of nodulation and N-fixation in Medicago.

**Step 2. Obtain and analyze experimental data in the target species.** In this step, the plant scientist gathers and/or conducts experiments about the target plant (e.g. transcriptome). For purposes of analysis, the pipeline will offer standard tools such as correlation, linear regression, as well as machine learning tools such as “State Space” analysis (Aim 2) to identify such relationships.

**Step 3. Infer a network in the target species.** The plant scientist next uses the data obtained in Steps 1 and 2, plus the CSNI inference engine and its parameters to infer a putative network in the target species, as depicted in Fig. 3A, as follows. First, the biologist chooses a “ground truth” network in the reference species. The biologist also selects which types of edges should be included in the reference species’ “ground truth” network (e.g. protein:protein, metabolic, etc), as well as some parameter settings or Inference rules that are generated using an optimization technique such as simulated annealing as discussed in Aim 1, Step 7. For example, a combination rule might infer a regulatory edge in the target species if the edge’s genes were connected by an expression edge with correlation > 0.7 and the edge had homologous genes connected by a regulatory edge in the reference species’ ground truth network. CSNI next takes 1) the homology mapping between the reference and target species chosen in Step 1, and 2) the

**Fig. 6. The Cross Species Network Inference (CSNI) User Interface.** A biologist user selects a target and reference species. Next the user selects a homology technique and its parameters. Finally, the biologist selects a “ground truth” data set from the reference species, as well as experimental data in the target species to aid in the prediction of inferred gene regulatory networks in the target (e.g. crop) species. See the workflow in Fig. 3, panel A.

“ground truth” network in the reference species, and 3) the set of experimental data about the target species chosen. In our case, we would use the time series data for Medicago from Aim 3, as well as already existing transcriptome data in Medicago (<http://bioinfo.noble.org/gene-atlas/v2/>). Given this data, CSNI infers a gene network for the target species.

We will implement CSNI as a general-purpose tool to be used by plant scientists. We plan to deploy CSNI ([www.CrossSpecies.org](http://www.CrossSpecies.org)) on several additional platforms, first on our VirtualPlant website ([www.virtualplant.org](http://www.virtualplant.org)), and second on *iPlant* (see S. Goff letter). We will also deploy CSNI on one of the widely-used bioinformatic workflow engines: Taverna [2], Kepler [3] or Galaxy [4]. Implementing the CSNI pipeline on top of one (or more) of these bioinformatic workflow engines is important because they provide increasingly popular platforms for developing computational genetic analyses, and provide generic support for reproducible bioinformatic analyses.

While these three steps comprise the basic production operation of CSNI in Inference mode, we will also investigate three related enhancements.

#### **Enhancements to the Operation of CSNI in *Inference* mode:**

**A. Select multiple reference species.** Some target species may be phylogenetically close to multiple possible reference species. In this situation, the plant scientist may execute CSNI repeatedly with different reference species and then combine the resulting inferred networks. The networks may be combined based on set union or intersection functions or based on more sophisticated weighting functions determined by biological intuition or learning as in Aim 1, Step 7. These functions will be enhancements to the CSNI pipeline.

**B. Assemble “ground truth” networks for potential target species.** We will assemble “ground truth” networks for Grape and Corn (for which KEGG pathways exist) and in Medicago we will infer metabolic pathways, until a KEGG version is available. These “ground truth” networks will help users validate inferred networks, and help set CSNI parameters through optimization studies in these crop species based on similarity scores and p-values computed using *NetSect*, as described in Aim 1. Time and resources permitting, we will also enable “ground truth” networks for additional crop species, in consultation with iPlant.

**C. Suggest the next experiment.** As discussed in Aim 2, Step 4, in State-Space modeling a simple but powerful method for determining the most useful previous experiment or set of experiments is to suppose those experiments didn’t exist and then measure how much that degrades accuracy predictions. Using that information can help a biologist determine the next measurement to use (e.g. a new time point in a time series rather than an additional replicate). This is meant as an aid rather than as a substitute for biological insight.

**Expected outcomes of Aim 4.** The CSNI pipeline analysis constructed in Aim 4, and made available to the community as a biologist-friendly interface, will empower plant biologists to use network approaches to derive testable hypothesis for gene functions in crop species for which limited genomic information is available. Identifying networks conserved between reference and crop species will also enable researchers to focus their translational studies from models to crops.

#### **Timeline:**

**Year 1:** Aim 1. Extend cross species network inference using ground truth protein:protein and metabolic interaction networks for Rice and Arabidopsis to other homology methods. Extend network inference analysis beyond protein-protein interaction to validated regulatory (AGRIS) edges. Aim 2 and 3: Perform the time series experiments in Rice and Medicago to complete the Arabidopsis series and generate the RNA-seq data. Aim 4: Build and/or assemble high-performance network inference software for time series data (e.g. parallelize State-Space modeling) and compare it with others (especially other Bayesian methods [50, 51]). Assemble “ground truth” networks in the 3-5 target crop species beginning with Medicago, Corn, Grape. Select bioinformatic workflow platform on which we will deploy.

**Years 2-3:** Aim 2 and 3. Apply the State Space analysis to Medicago and Rice and test targeted regulatory genes in N-use pathway. Aim 4. Deploy the first version of the CSNI analysis pipeline for cross species network inference to collaborators (D. Cook, U Davis; R. Gutierrez, Chile).

**Years 4-5:** Apply the computational pipeline to infer networks in several crop species for example corn and grape. Deploy the full computational CSNI pipeline for cross-species network inference to plant community via CSNI ([www.CrossSpecies.org](http://www.CrossSpecies.org)) linked to VirtualPlant, iPlant and a selected workflow platform (e.g. Galaxy).



## **Plan to integrate Research and Education.**

**Cross training of Biologists and Computer Scientist in Systems Biology.** This project is the result of a long-standing and highly successful collaboration between biologists at NYU Center for Genomics and Systems Biology and computer scientists at NYU's Courant Institute of Mathematical Sciences. The development of Systems Biology tools in this project has and will involve biologists teaching computer scientists about topics like genetics, experimental genomics, and the computational challenges of analyzing genomic data. The computer scientists in turn teach the biologists about the computational tools and mathematical concepts behind them. We do this informally at our weekly joint lab meetings at which graduate students and post docs from NYU Biology and NYU Courant each present their work to the group. This project involves a team of three resident full time computer scientists working within a biology lab, interacting closely with wet bench biologists. The senior computer scientists (Shasha, Katari and Goldberg) are also involved in training and engaging computer scientist students at all levels in the emerging field of Systems Biology. In the last six months, they have trained one PhD student, two interns and two MS students from Courant working in this environment. For a complete listing of students trained in the past 4.5 years, see Education and Training section in Results from Prior support.

**Workshops and Classroom Training in Genomics and Systems Biology:** We also provide formal training in the form of workshops and classes to enable Systems Biology. Examples of this include a weekly software workshop in “R”, which aims to teach biologists how to analyze their own genomic data. This workshop has been taught two times, once by Jonathan Kelfer, a MS student working on the project and most recently by Manrpeet Katari, co-PI. Dr. Katari is especially suited to this role, as he is a trained geneticist (PhD) who is also a computer scientist. In addition to students and post-docs in “R” this class has also included several faculty on sabbatical at NYU including most recently: MaryLou Guerinot and Rob McClung of Dartmouth. Students will be exposed to Genomics and Systems Biology also through a series of formal courses offered by faculty at NYU’s Center for Genomics and Systems Biology including: G23.1128 Systems Biology; G23.1130 Applied Genomics: Introduction to Bioinformatics & Network Modeling; G23.1127 Bioinformatics & Genomes

**PhD practical training.** PhD students will also present their work annually in the weekly PhD seminar series hosted by the Biology Department. Computational students will be involved in constructing the pipeline and making it perform through the use of parallelization. Such students will also help to develop and test optimization and machine learning algorithms for network inference. Biological students will engage in experimental work including preparation of RNA for Illumina sequencing, as well as the analysis of data using the VirtualPlant platform.

**Training Postdocs as educators.** In this project, Post-Docs are paired up with graduate students, undergraduate students, and technicians in the laboratory to practice mentoring skills in a research context. At NYU, post-docs are also afforded the opportunity to teach and are mentored by faculty advisors. Dr. Katari, is currently co-teaching an undergraduate course “Introduction to Genomics & Bioinformatics” with a faculty mentor. Post-Docs also receive counseling from their co-mentors and practice presentation skills during regular group-lab meetings, through a Post-Doc seminar series, and at annual poster sessions at NYU.

## REFERENCES CITED

1. \* Katari, M.S., et al., *VirtualPlant: A software platform to support Systems Biology research*. Plant Physiol, 2009.
2. Oinn, T., et al., *Taverna: a tool for the composition and enactment of bioinformatics workflows*. Bioinformatics, 2004. **20**(17): p. 3045-54.
3. Altintas, I., et al., *Kepler: an extensible system for design and execution of scientific workflows*. Proceedings. 16th International Conference on Scientific and Statistical Database Management, 2004: p. 423--424.
4. Blankenberg, D., et al., *Galaxy: a web-based genome analysis tool for experimentalists*. Curr Protoc Mol Biol. **Chapter 19**: p. Unit 19 10 1-21.
5. Gutierrez, R.A., et al., *Qualitative network models and genome-wide expression data define carbon/nitrogen-responsive molecular machines in Arabidopsis*. Genome Biol, 2007. **8**: p. R7.
6. Moreno-Risueno, M.A., W. Busch, and P.N. Benfey, *Omics meet networks-using systems approaches to infer regulatory networks in plants*. Curr Opin Plant Biol, 2009.
7. Gutierrez, R.A., et al., *Systems approach identifies an organic nitrogen-responsive gene network that is regulated by the master clock control gene CCA1*. Proc Natl Acad Sci U S A, 2008. **105**(12): p. 4939-44.
8. \* Nero, D., et al., *In silico evaluation of predicted regulatory interactions in Arabidopsis thaliana*. BMC Bioinformatics, 2009. **10**: p. 435.
9. Gifford, M.L., et al., *Cell-specific nitrogen responses mediate developmental plasticity*. Proc Natl Acad Sci U S A, 2008. **105**(2): p. 803-8.
10. \* Thum, K.E., et al., *An integrated genetic, genomic and systems approach defines gene networks regulated by the interaction of light and carbon signaling pathways in Arabidopsis*. BMC Syst Biol, 2008. **2**: p. 31.
11. Wang, R., et al., *Genomic analysis of the nitrate response using a nitrate reductase-null mutant of Arabidopsis*. Plant Physiol, 2004. **136**(1): p. 2512-22.
12. \* Gutierrez, R.A., et al., *Qualitative network models and genome-wide expression data define carbon/nitrogen-responsive molecular machines in Arabidopsis*. Genome Biol, 2007. **8**(1): p. R7.
13. Nero, D., et al., *A system biology approach highlights a hormonal enhancer effect on regulation of genes in a nitrate responsive "biomodule"*. BMC Syst Biol, 2009. **3**: p. 59.
14. Craigon, D.J., et al., *NASCArrays: a repository for microarray data generated by NASC's transcriptomics service*. Nucleic Acids Res, 2004. **32**(Database issue): p. D575-7.
15. Ideker, T., et al., *Integrated Genomic and Proteomic Analyses of a Systematically Perturbed Metabolic Network*. Science, 2001. **292**(5518): p. 929-934.
16. Gutierrez, R.A., D.E. Shasha, and G.M. Coruzzi, *Systems Biology for the Virtual Plant*. Plant Physiol., 2005. **138**(2): p. 550-554.
17. \* Poultney, C.S., et al., *Sungear: interactive visualization and functional analysis of genomic datasets*. Bioinformatics, 2007. **23**(2): p. 259-61.
18. Ferro, A., et al., *NetMatch: a Cytoscape plugin for searching biological networks*. Bioinformatics, 2007. **23**(7): p. 910-2.
19. Krouk, G., et al., *High resolution dynamic transcriptome of Arabidopsis roots in response to nitrate: Molecular physiology and predictive modeling*. submitted, 2010.
20. Kanehisa, M., et al., *The KEGG resource for deciphering the genome*. Nucleic Acids Res, 2004. **32 Database issue**: p. D277-80.
21. Bader, G., D. Betel, and C. Hogue, *BIND: the Biomolecular Interaction Network Database*. Nucleic Acids Res., 2002. **31**: p. 248.
22. de Folter, S., et al., *Comprehensive interaction map of the Arabidopsis MADS Box transcription factors*. Plant Cell, 2005. **17**(5): p. 1424-33.

23. Popescu, S.C., et al., *Differential binding of calmodulin-related proteins to their targets revealed through high-density Arabidopsis protein microarrays*. Proc Natl Acad Sci U S A, 2007. **104**(11): p. 4730-5.
24. Davuluri, R., et al., *AGRIS: Arabidopsis Gene Regulatory Information Server, an information resource of Arabidopsis cis-regulatory elements and transcription factors*. BMC Bioinformatics, 2003. **4**(1): p. 25.
25. Griffiths-Jones, S., et al., *miRBase: microRNA sequences, targets and gene nomenclature*. Nucleic Acids Res, 2006. **34**(Database issue): p. D140-4.
26. Gustafson, A.M., et al., *ASRP: the Arabidopsis Small RNA Project Database*. Nucleic Acids Res, 2005. **33**(Database issue): p. D637-40.
27. Lu, C., et al., *Elucidation of the small RNA component of the transcriptome*. Science, 2005. **309**: p. 1525.
28. Masoudi-Nejad, A., et al., *EGENES: transcriptome-based plant database of genes with metabolic pathway information and expressed sequence tag indices in KEGG*. Plant Physiol, 2007. **144**(2): p. 857-66.
29. Zhang, P., et al., *MetaCyc and AraCyc. Metabolic pathway databases for plant research*. Plant Physiol, 2005. **138**(1): p. 27-37.
30. Gunsalus, K.C., et al., *Predictive models of molecular machines involved in Caenorhabditis elegans early embryogenesis*. Nature, 2005. **436**(7052): p. 861-5.
31. Matthews, L.R., et al., *Identification of potential interaction networks using sequence-based searches for conserved protein-protein interactions or "interologs"*. Genome Res, 2001. **11**(12): p. 2120-6.
32. Gachon, C.M., et al., *Transcriptional co-regulation of secondary metabolism enzymes in Arabidopsis: functional and evolutionary implications*. Plant Mol Biol, 2005. **58**(2): p. 229-45.
33. Tatusov, R.L., E.V. Koonin, and D.J. Lipman, *A genomic perspective on protein families*. Science, 1997. **278**(5338): p. 631-7.
34. Zhang, J., *Evolution by gene duplication: an update*. TRENDS in Ecology and Evolution, 2003. **18**(6): p. 292-298.
35. Barrett, T. and R. Edgar, *Mining microarray data at NCBI's Gene Expression Omnibus (GEO)\**. Methods Mol Biol, 2006. **338**: p. 175-90.
36. Pepper, S.D., et al., *The utility of MAS5 expression summary and detection call algorithms*. BMC Bioinformatics, 2007. **8**: p. 273.
37. Michalewicz, Z. and D. Fogel, *How to Solve It: Modern Heuristics*. 2004: Springer Verlag.
38. Wang, L., et al., *A dynamic gene expression atlas covering the entire life cycle of rice*. Plant J, 2009.
39. \* Chiu, J.C., et al., *OrthologID: automation of genome-scale ortholog identification within a parsimony framework*. Bioinformatics, 2006. **22**(6): p. 699-707.
40. Tatusov, R.L., et al., *The COG database: a tool for genome-scale analysis of protein functions and evolution*. Nucleic Acids Res, 2000. **28**(1): p. 33-6.
41. Berglund, A.C., et al., *InParanoid 6: eukaryotic ortholog clusters with inparalogs*. Nucleic Acids Res, 2008. **36**(Database issue): p. D263-6.
42. Murphy, K. and S. Mian, *Modelling Gene Expression Data using Dynamic Bayesian Networks*, in *Technical report, Computer Science Division*. 1999, University of California and Life Sciences Division, Lawrence Berkeley National Laboratory.
43. Wan, T.F., et al., *Correlation between ASI gene expression and seed protein contents in different soybean (Glycine max [L.] Merr.) cultivars*. Plant Biol (Stuttg), 2006. **8**(2): p. 271-6.
44. Wang, R., et al., *Microarray analysis of the nitrate response in Arabidopsis roots and shoots reveals over 1,000 rapidly responding genes and new linkages to glucose, trehalose-6-phosphate, iron, and sulfate metabolism*. Plant Physiol, 2003. **132**(2): p. 556-67.
45. Gage, D.J., *Infection and invasion of roots by symbiotic, nitrogen-fixing rhizobia during nodulation of temperate legumes*. Microbiol Mol Biol Rev, 2004. **68**(2): p. 280-300.

46. Barbulova, A., et al., *Differential effects of combined N sources on early steps of the Nod factor-dependent transduction pathway in Lotus japonicus*. Mol Plant Microbe Interact, 2007. **20**(8): p. 994-1003.
47. Jeudy, C., et al., *Adaptation of Medicago truncatula to nitrogen limitation is modulated via local and systemic nodule developmental responses*. New Phytol, 2009.
48. Ruffel, S., et al., *Systemic signaling of the plant nitrogen status triggers specific transcriptome responses depending on the nitrogen source in Medicago truncatula*. Plant Physiol, 2008. **146**(4): p. 2020-35.
49. Gresshoff, P.M., *Post-genomic insights into plant nodulation symbioses*. Genome Biol, 2003. **4**(1): p. 201.
50. Bonneau, R., et al., *The Inferelator: an algorithm for learning parsimonious regulatory networks from systems-biology data sets de novo*. Genome Biol, 2006. **7**(5): p. R36.
51. Barber, D., *Advances in Neural Information Processing Systems*. 2003, MIT Press, Cambridge MA. p. 729-736.
52. Rzhetsky, A., et al., *GeneWays: a system for extracting, analyzing, visualizing, and integrating molecular pathway data*. Journal of Biomedical Informatics, 2004. **37**(1): p. 43-53.
53. Mueller, L.A., P. Zhang, and S.Y. Rhee, *AraCyc: A Biochemical Pathway Database for Arabidopsis*. Plant Physiol, 2003. **132**(2): p. 453.
54. Geisler-Lee, J., et al., *A predicted interactome for Arabidopsis*. Plant Physiol, 2007. **145**(2): p. 317-29.
55. Cui, J., et al., *AtPID: Arabidopsis thaliana protein interactome database--an integrative platform for plant systems biology*. Nucleic Acids Res, 2008. **36**(Database issue): p. D999-1008.
56. Mirowski P & LeCun Y (2009) Dynamical Factor Graphs for Time Series Modeling. Lecture Note in Artificial Intelligence 5782:128-142
57. \* Gutierrez R, Gifford M, Poultney C, Wang R, Shasha D, Coruzzi G, Crawford N (2007) "Insights into the genomic nitrate response using genetics and the SunGear Software System" **Journal of Experimental Botany** doi: 10.1093/jxb/erm079

**Dennis E. Shasha**  
**Courant Institute of Mathematical Sciences, New York University**  
**shasha@courant.nyu.edu**

**Education**

Yale University	B.S., Engineering and Applied Science	1977
Syracuse University	M.S., Computer Science	1980
Harvard University	Ph.D., Applied Mathematics	1984

**Positions and Honors**

1995-Present	Professor of Computer Science, Courant Institute of Mathematical Sciences, New York University.
1990-1995	Associate Professor of Computer Science, Courant Institute of Mathematical Sciences, New York University.
1984-1990	Assistant Professor of Computer Science, Courant Institute of Mathematical Sciences, New York University.
1998-1999	Invited Professor at INRIA

**Selected Publications**

Manpreet S. Katari, Steve D. Nowicki, Felipe F. Aceituno, Damion Nero, Jonathan Kelfer, Lee Parnell Thompson, Juan M. Cabello, Rebecca S. Davidson, Arthur P. Goldberg, Dennis E. Shasha, Gloria M. Coruzzi, and Rodrigo A. Gutierrez, "VirtualPlant: a software platform to support system biology research" **Plant Physiology** 2009 : pp.109.147025v1

Dennis Shasha and Philippe Bonnet. "Database Tuning: Principles Experiments and Troubleshooting Techniques" (2002) *Morgan Kaufmann Publishers*, June 2002, ISBN 1-55860-753-6, Paper, 464 Pages.

Mitchell Levesque, Dennis Shasha, Wook Kim, Michael G Surette, and Philip N Benfey (2003) "Trait-To-Gene: A Computational Method for Predicting the Function of Uncharacterized Genes." **Current Biology**, vol. 13, 129-133.

Kenneth Birnbaum, Dennis E. Shasha, Jean Y. Wang, Jee W. Jung, Georgina M. Lambert, David W. Galbraith, and Philip N. Benfey (2003) "A gene expression map of the Arabidopsis root". **Science**, Dec 12 2003:1956-1960.

Rodrigo Gutierrez, Dennis Shasha, and Gloria Coruzzi (2005) "Systems Biology for the Virtual Plant" **Plant Physiology**, vol. 38, pp. 550-554.

Christopher S. Poultney, Rodrigo A. Gutierrez, Manpreet S. Katari, Miriam L. Gifford, W. Bradford Paley, Gloria M. Coruzzi and Dennis E. Shasha (2006) "Sungear: Interactive visualization and functional analysis of genomic datasets" **Bioinformatics**, doi:10.1093/bioinformatics/btl496.

Diego Regorgiato, Rodrigo Gutierrez, and Dennis Shasha(2008) "GraphClust: A Method for Clustering Databases of Graphs" **Journal of Information and Knowledge Management (JIKM)**, vol. 7, Issue: 4, pp. 231 – 241.

A. Lagana, S. Forte, A. Giudice, M. R. Arena, P. L. Puglisi, R. Giugno, A. Pulvirenti, D. Shasha, A. Ferro "miRo: a miRNA knowledge base" **Database: The Journal of Biological Databases and Curation**, Oxford University Press, 2009

Gabriel Krouk, Daniel Tranchina, Laurence Lejay, Alexis A. Cruikshank, Dennis Shasha, Gloria M. Coruzzi, Rodrigo A. Guitierrez, "A Systems Approach Uncovers Restrictions for Signal Interactions Regulating Genome-wide Responses to Nutritional Cues in Arabidopsis" **PLOS Computational Biology** March 2009, volume 5, issue 3

Bonnici V, Di Natale R, Ferro A, Giugno R, Mongiovi M, Pigola G, Pulvirenti A, Shasha D "Enhancing Graph Database Indexing By Suffix Tree Structure" Bioinformatics Italian Society Symposium 2009

### **Synergistic Activities**

Database tuning and design consulting for Wall Street companies, Bell Labs, ecommerce, and biotech companies (drug discovery). 1991-Present

Monthly puzzle column for editor Scientific American (www.sciam.com). 2001-Present

Distinguished Science Advisor, New York Hall of Science, one of 20. 2003-Present

### **List of Collaborators or Potential Collaborators**

I have collaborated with the following people during the last 48 months (or may collaborate with them due to a series editorship at Oxford for Genomics and Bioinformatics) in addition to those listed in the publications list: Michael Ashburner, David Botstein, Charles Cantor, Lee Hood, Minoru Kanehisa, Raju Kucherlapati, Gary Bader, Isidore Rigoutsos, Gregory Stephanopoulos, Martyn Amos, and Michael Rabin.

### **Names of Graduate and Post-Graduate Advisors and Advisees**

Dissertation advisor: Nathan Goodman (Harvard University)

Advisees in the last 5 years (current affiliation is in parentheses):

Rosalba Giugno (Assistant Prof, Univ of Catania, Bioinformatics)

Alberto Lerner (Google Research)

Yunyue Zhu (Finance)

Aristotle Tsirigos (Bioinformatics, IBM)

Xiaojian Zhao (Finance)

Zhihua Wang (Ask.com)

Tyler Neylon (Google research)

Xin Zhang (Finance)

**GLORIA M. CORUZZI, Ph.D.**  
New York University

**EDUCATION:**

Fordham University	B.S. Biology, cum Laude, in Cursu Honorum	1976
New York University Medical School	M.S.-Ph.D. Cell & Molecular Biology,	1979
Rockefeller University	NIH Postdoctoral Fellow, Plant Molecular Biology	1983

**APPOINTMENTS:**

2003-Present	Biology Department Chair, Carroll & Milton Petrie Professor, New York University
1991-Present	Carroll & Milton Petrie Professor, NYU Department of Biology
1990 - 1991	Associate Professor & Associate Dean of Postdoctoral Fellows Rockefeller University, Laboratory of Plant Molecular Biology
1983 - 1989	Assistant Professor, Rockefeller University
1980 - 1983	NIH Postdoctoral Fellow, Rockefeller University
1979 - 1980	Postdoctoral Research Associate, Columbia University
1976 - 1979	NIH Predoctoral Fellow, New York University Medical School

**5 PUBLICATIONS RELATED TO THIS PROPOSAL:**

- Katari MS, Nowicki SD, Aceituno FF, Nero D, Kelfer J, Thompson LP, Cabello JM, Davidson RS, Goldberg AP, Shasha DE, Coruzzi GM, Gutierrez RA (2009) VirtualPlant: A software platform to support Systems Biology research. **Plant Physiol.** Dec [Epub ahead of print <http://www.plantphysiol.org/cgi/rapidpdf/pp.109.147025v1>]
- Nero D, Kelfer J, Katari MS, Tranchina D, Coruzzi GM (2009) In Silico Evaluation of Predicted Regulatory Interactions in Arabidopsis thaliana. **BMC Bioinformatics.** Dec 21;10(1):435.
- Krouk G, Tranchina D, Lejay L, Cruikshank A, Shasha D, Coruzzi G, and Gutierrez R (2009) A systems approach uncovers restrictions for signal interactions regulating genome-wide responses to nutritional cues in Arabidopsis. **PloS Comp Biol.** Mar;5(3):e1000326.
- Poultney C, Gutiérrez RA, Katari MS, Gifford ML, Paley WB, Coruzzi GM and Shasha DE (2007) Sungear: Interactive visualization, exploration and functional analysis of genomic datasets. **Bioinformatics**, 23:259-61.
- Gutierrez R, Gifford ML, Poultney C, Wang R, Shasha DE, Coruzzi GM, Crawford NM (2007) Insights into the genomic nitrate response using genetics and the Sungear Software System. **Journal of Experimental Botany** doi: 10.1093/jxb/erm079

**5 OTHER SIGNIFICANT PUBLICATIONS:**

- Ruffel S, Krouk G, Coruzzi GM (2009) A Systems View of Responses to Nutritional Cues in Arabidopsis: Towards a Paradigm Shift for Predictive Network Modeling. *Plant Physiol.* 2009 Nov 25. [Epub ahead of print]
- Coruzzi GM, Burga A, Katari MS, and Gutierrez RA (2009) Systems Biology: Principles and Applications in Plant Research. In *Plant Systems Biology*, Annual Plant Reviews; Blackwell Publishing: Oxford, UK, 2009, Vol. 35. Pgs 3-31.
- Thum KE, Shin MJ, Gutierrez R, Katari M, Nero D, Shasha D, Coruzzi GM (2008) An integrated genetic, genomic and systems approach defines gene networks regulated by the interaction of light and carbon signaling pathways. **BMC Systems Biology** 4; 2 (1): 31
- Gutiérrez RA, Lejay L, Chiaromonte F, Shasha DE, Coruzzi GM (2007) Qualitative network models and genome-wide expression data define carbon/nitrogen-responsive biomodules in Arabidopsis. **Genome Biology**, 8: R7.
- Gutierrez R, Shasha S and Coruzzi G (2005) Systems Biology for the Virtual Plant. **Plant Physiol.** Vol 138, pp 550-554.

**SYNERGISTIC ACTIVITIES:**

Editorial Board: Trends in Plant Science, 2004-Present  
New York Botanical Garden, Botanical Science Committee, 1995-Present  
NSF Bio Advisory Panel, Systems Biology Presentation, April 19<sup>th</sup>, 2007  
NSF US-EU Taskforce on Plant Biotechnology: Panelist, June 21-22, 2005  
FRN Faculty Resource Network Summer Workshop on Integrative Approaches to Teaching Life Sciences, Lecture: Plant Genomics and Systems Biology, 2005.

**COLLABORATORS:**

Philip Benfey, Duke University	Ernest Lee, New York University
Kenneth Birnbaum, New York University	Robert Martienssen, CSHL
Hiu-Ki Chan, University of Hong Kong	Richard McCombie, CSHL
Mingsheng Chen, Chinese Acad. of Science	Walter Moss, NY Botanical Garden
Francesca Chiaromonte, Penn State University	Bradford Paley, Digital Image Design
Joanna Chiu, Rutgers University	Christopher Poultney, New York University
Michael Chou, Harvard	Steven Rudd, NY Botanical Garden
Alexis Cruikshank, New York University	Suzan Runko, New York University
Alexis Dean, New York University	Neil Sarkar, AMNH
Robert DeSalle, AMNH	Dennis Shasha, New York University
Andrew Douglas, NY Botanical Garden	Giulia Stellari, Cornell
Mary Egan, NYBG	Dennis Stevenson, NY Botanical Garden
Pamela Green, U. Delaware	Milos Tanurdzic, Cold Spring Harbor Lab
Maren Hoffman, U. Gottingen	Rudolf Tischner, U. Gottingen
Todd Holmes, UC-Irvine	Daniel Tranchina, New York University
Joseph Kieber, U. North Carolina	Robert Twigg, NY Botanical Garden
C Robertson McClung, Dartmouth	Rongchen Wang, Univeristy of California-SD
Blake Meyers, U. Delaware	Hon-Kit Wong, U. Hong Kong
Hong-Ming Lam, U. Hong Kong	Xiujuan Xing, Univeristy of California-SD
Bud Mishra, New York University	Xiangqun Xu, Zhejiang Sci-Tech University

**GRADUATE ADVISOR AND POSTDOCTORAL SPONSOR:**

PhD thesis: Dr. Alexander Tzagoloff, Columbia University, NY  
Postdoctoral: Dr. Nam Hai Chua, Rockefeller University, NY

**THESIS ADVISOR AND POSTGRADUATE-SCHOLAR SPONSOR:**

Nora Barboza, Memorial Sloan Kettering, NY	Karen Thum, (Texas A&M) former
Joanna Chiu, Rutgers University, NJ	Manpreet Katari, (CSHL) Current
Barbara Miesak, Rutgers University, NJ	Gabriel Krouk (Agro-M, France) Current
Michael Shin, Messiah College, PA	Indrani Mukherjee (U. S. Carolina) Current
Philip Feinburg, Cornell Med, NY	Sandrine Ruffel (INRA, France) Current
Damion Nero, Programmer FOJP Service Corp	Amy Marshall-Colon (Purdue) Current
Daniela Ristova, Current	
Eduardo de la Torre, Baruch College	Postdoctoral Scholars-15, Graduate Students-12
Eric Brenner, New York University	
Rodrigo Gutierrez, Catholic U. of Chile	
Andrei Kouranov, Rutgers University, NJ	
Muriel Lancien, Lancaster U, UK	
Laurence Lejay, INRA, France	
Peter Palenchar, Rutgers University, NJ	
Trevor Stokes, Times Daily Reporter, AL	
Miriam Gifford, Warwick University, UK	
Mariana Obertello, Ingebi-Conicet, Argentina	



## MANPREET S. KATARI, Ph.D.

New York University

### EDUCATION:

State University of New York at Buffalo	B.S. Biochemistry	1996
State University of New York at Stonybrook	Ph.D. Genetics	2004

### APPOINTMENTS:

Summer 1995 Laboratory Assistant, VA Medical Center, Buffalo NY.  
1997-1998 Laboratory Technician, Cold Spring Harbor Laboratory, Cold Spring Harbor, NY.  
1999-2004 Graduate Student, Cold Spring Harbor Laboratory, Cold Spring Harbor, NY.  
2004-Present Post Doctoral Fellow, New York University, New York, NY.

### 5 PUBLICATIONS RELATED TO THIS PROPOSAL:

Katari MS, Nowicki SD, Aceituno FF, Nero D, Kelfer J, Thompson LP, Cabello JM, Davidson RS, Goldberg AP, Shasha DE, Coruzzi GM, Gutierrez RA (2009) VirtualPlant: A software platform to support Systems Biology research. **Plant Physiol.** Dec [Epub ahead of print <http://www.plantphysiol.org/cgi/rapidpdf/pp.109.147025v1>]

Poultney C, Gutiérrez RA, Katari MS, Gifford ML, Paley WB, Coruzzi GM and Shasha DE (2007) Sungear: Interactive visualization, exploration and functional analysis of genomic datasets. **Bioinformatics**, 23:259-61.

de la Torre-Barcelona, JE., Egan, MG., Katari, MS., Brenner, ED, Stevenson, DW, Coruzzi, GM., DeSalle, R. (2006) ESTimating plant phylogeny: lessons from partitioning. **BMC Evol Biol.** Jun 15;6(1):48

Brenner ED, Stevenson DW, McCombie RW, Katari MS, Rudd SA, Mayer KF, Palenchar PM, Runko SJ, Twigg RW, Dai G, Martienssen RA, Benfey PN, Coruzzi GM. (2003) Expressed sequence tag analysis in *Cycas*, the most primitive living seed plant. **Genome Biol.** ;4(12):R78.

Katari, MS., Balija V, Wilson RK, Martienssen RA, McCombie WR. (2005) Comparing low coverage random shotgun sequence data from *Brassica oleracea* and *Oryza sativa* genome sequence for their ability to add to the annotation of *Arabidopsis thaliana*. **Genome Res.** Apr;15(4):496-504.

### 5 OTHER SIGNIFICANT PUBLICATIONS:

Nero D, Kelfer J, Katari MS, Tranchina D, Coruzzi GM (2009) In Silico Evaluation of Predicted Regulatory Interactions in *Arabidopsis thaliana*. **BMC Bioinformatics.** Dec 21;10(1):435.

de la Torre-Barcelona, JE., Kolokotronis, SO., Lee, EL., Stevenson, DW., Brenner, ED., Katari, MS., Coruzzi, GM., Desalle, R. (2009) The impact of outgroup choice and missing data on major seed plant phylogenetics using genome-wide EST data. **Plos One.**, Jun 3; 4(6):e5764.

Thum KE, Shin MJ, Gutiérrez RA, Mukherjee I, Katari MS, Nero D, Shasha D, Coruzzi GM. (2008) An integrated genetic, genomic and systems approach defines gene networks regulated by the interaction of light and carbon signaling pathways in *Arabidopsis*. **BMC Syst Biol** Apr 4;2:31

Gutiérrez R, Stokes T, Thum K, Xu X, Obertello M, Katari M, Tanurdzic M, Dean A, Nero D, McClung CR & Coruzzi G (2008). Systems approach identifies an organic nitrogen-responsive gene network that is regulated by the master clock gene CCA1. **PNAS**, Mar 25;105(12):4939-44.

Brenner ED, Katari MS., Stevenson DW, Rudd SA, Douglas AW, Moss WN, Twigg RW, Runko SJ, Stellari GM, Richard MW, Coruzzi GM. (2005) EST analysis in *Ginkgo biloba*: an assessment of conserved developmental regulators and gymnosperm specific genes. **BMC Genomics.** Oct 15;6(1):143

### SYNERGISTIC ACTIVITIES:

N/A

**COLLABORATORS:**

Felipe F. Aceituno, P. Univ. Catolica de Chile  
Vivekenand Balija, Cold Spring Harbor Lab  
Phillip Benfey, Duke University  
Eric Brenner, New York University  
Juan M. Cabello, P. Univ. Catolica de Chile  
Gloria M. Coruzzi, New York University  
Rebecca Davidson, New York University  
Rob Desalle, AMNH  
Alexis Dean, New York Public School District  
Jose Eduardo de la Torre, Baruch College  
Mary Egan, Montclair State University  
Miriam Gifford, Warwick University  
Arthur P. Goldberg, New York University  
Rodrigo Gutierrez, P. Univ. Catolica de Chile  
Jonathan Kelfer, Bloomberg  
Sergios O. Kolokotronis, AMNH  
Ernest Lee, AMNH  
Rob Martienssen, CSHL  
Rob McClung, Dartmouth College  
Indrani Mukherjee, New York University

Damion Nero, FOJP Service Corp  
Steve D. Nowicki, New York University  
Mariana Obertello, Ingebi-Conicet  
Peter M. Palenchar, Rutgers University  
Bradford Paley, DiDi Design  
Chris Poultney, New York University  
Stephen A. Rudd, Munich Information Center  
for Protein Sequences  
Suzan Runko, New York University  
Dennis E. Shasha, New York University  
Michael Shin, Messiah College  
Dennis Stevenson, New York Botanical Garden  
Trevor Stokes, Times Daily Reporter  
Milos Tanurdzic, CSHL  
Lee P. Thomson, University of Texas at Austin  
Karen Thum, New York University  
Dan Tranchina, New York University  
Richard W. Twigg, Duke University  
Rick Wilson, Washington University

**GRADUATE ADVISOR AND POSTDOCTORAL SPONSOR:**

PhD thesis: Dr. W.R. McCombie (Cold Spring Harbor Laboratory)  
Postdoctoral: Dr. Gloria M. Coruzzi (New York University)

**THESIS ADVISOR AND POSTGRADUATE-SCHOLAR SPONSOR:**

N/A

## ARTHUR GOLDBERG, PhD

New York University

### EDUCATION:

Harvard College	B.A. Astronomy and Astrophysics	1977
UCLA	M.S. Computer Science	1985
UCLA	Ph.D. Computer Science	1991

### APPOINTMENTS:

2009 – Present	Research Scientist in Bioinformatics, Plant Systems Biology Laboratory, NYU
2006 – 2008	CEO, ChoiceMaker Technologies, New York, NY
2000 – 2006	Co-founder and Computer Scientist, ChoiceMaker Technologies, New York, NY
1998 – 2006	Director, Masters of Science in Information Systems program, NYU,
1994 – 2006	Clinical Associate Professor of Computer Science, Computer Science Dept, Courant Institute of Mathematics, New York University
1989 – 1994	Research Scientist, IBM T.J. Watson Research Center

### 5 PUBLICATIONS RELATED TO THIS PROPOSAL:

- Katari MS, Nowicki SD, Aceituno FF, Nero D, Kelfer J, Thompson LP, Cabello JM, Davidson RS, Goldberg AP, Shasha DE, Coruzzi GM, Gutierrez RA (2009) VirtualPlant: A software platform to support Systems Biology research. *Plant Physiol.* Dec [Epub ahead of print  
<http://www.plantphysiol.org/cgi/rapidpdf/pp.109.147025v1>]
- Borthwick A, Buechi M Goldberg A. *Automated Database Blocking and Record Matching*. U.S. Patent .  
*Automated Database Blocking and Record Matching*. U.S. Patent #7,152,060. Awarded  
December 19, 2006. <http://www.google.com/patents?id=9kx-AAAAEBAJ&dq=7,152,060>
- Goldberg A, Borthwick A. *Batch Automated Blocking and Record Matching*, pending patent, filed  
November 2005.
- Buechi M, Borthwick A, Winkel A, Goldberg A, *ClueMaker: A Language for Approximate Record  
Matching*, Massachusetts Institute of Technology's Eighth International Conference on  
Information Quality (MIT ICIQ), Cambridge, MA. August 27, 2003.  
[http://www.cs.nyu.edu/artg/publications/choicemaker\\_cluemaker.pdf](http://www.cs.nyu.edu/artg/publications/choicemaker_cluemaker.pdf)

### 5 OTHER SIGNIFICANT PUBLICATIONS:

- Goldberg A, Pevzner I, Buff R, *Caching Characteristics of Internet and Intranet Web Proxy Traces*, the  
Computer Measurement Group Conference, December 1998.  
[http://www.cs.nyu.edu/artg/publications/Goldberg\\_CACHING\\_CHARACTERISTICS\\_OF\\_INTERNET\\_AND\\_INTRANET\\_WEB\\_PROXY\\_TRACES\\_1998.pdf](http://www.cs.nyu.edu/artg/publications/Goldberg_CACHING_CHARACTERISTICS_OF_INTERNET_AND_INTRANET_WEB_PROXY_TRACES_1998.pdf)
- Goldberg A, Buff R, Schmitt A, *Secure Web Server Performance Dramatically Improved By Caching  
SSL Session Keys*, Workshop on Internet Server Performance, held in conjunction with  
SIGMETRICS'98, June, 1998. <http://www.cs.nyu.edu/artg/research/ssl/ssl.doc>
- Strom R, Bacon D, Goldberg A, Lowry A, Yellin D, Yemini SA *Hermes: A Language for Distributed  
Computing*, Prentice Hall, 1991, 285 pages, cited by 109.
- Korfhage W, Goldberg A *Hermes Language Experiences, Software—Practice and Experience, Vol  
25(4).1995.*
- Goldberg A, et. al., *Restoring consistent global states of distributed computations*. Workshop on Parallel  
and Distributed Debugging, 1991, and ACM SIGPLAN Notices, vol. 26, no. 12, 1991.  
<http://citeseer.ist.psu.edu/korfhage95hermes.html>, cited by 45.

**SYNERGISTIC ACTIVITIES:**

At ChoiceMaker Technologies:

Strategic and operational co-leader of a 12-person firm that invented, built and sold mission-critical database de-duplication and approximate record matching software to 15 government agencies and firms.

Quality software management: Co-supervised construction and deployment of software by a six-person development team that produced 120+ KLOC of Java; lead creation of processes for requirements gathering and analysis, effort estimation and project scheduling, and quality assurance.

Initiated and co-led efforts that obtained \$1M in NSF SBIR “A Machine Learning Approach To Approximate Record Matching” grant.

At Computer Science Dept., New York University:

Created, promoted, ran and taught an internship-based information technology project management course which taught NYU 300 graduate students and placed them in internships at major NYC firms.

Raised over \$600,000 from firms sponsoring internships. Created fund-raising effort from scratch; all funds exceeded expectations. Coordinated almost 100 projects with numerous project clients including the AMEX, The Blackstone Group, Bank of America, Citigroup, Deutsche Bank, Lehman Bros., MetLife, Morgan Stanley, UBS, IBI, Vindigo, Wiley and HBO.

**COLLABORATORS:**

Manpreet S. Katari, PhD, Manager of Bioinformatics, Plant Systems Biology Laboratory, New York University.

Dennis Shasha, Professor, Computer Science Dept., Department of Computer Science, Courant Institute of Mathematical Sciences, New York University.

Dr. Gloria Coruzzi, Chair of Biology, Carroll & Milton Petrie Professor, Center for Genomics & Systems Biology, Department of Biology, New York University.

Andrew Borthwick, Principal Scientist at Spock.

**GRADUATE ADVISOR AND POSTDOCTORAL SPONSOR:**

Graduate Advisors:

Dissertation chair: David Jefferson, PhD, Computer scientist in the Center for Applied Scientific Computing, Lawrence Livermore National Laboratory.

Leonard Kleinrock, Distinguished Professor of Computer Science, UCLA Computer Science Department.

D. Stott Parker, Professor of Computer Science, UCLA Computer Science Department.

**THESIS ADVISOR AND POSTGRADUATE-SCHOLAR SPONSOR:**

Advised

Ilya Pevzner, dissertation: *A Probabilistic Learning Approach to Attribute Value Inconsistency Resolution*, graduated in 2006, Senior Programmer JP Morgan.

Sponsored:

Robert Buff: graduated in 2002, Senior Programmer JP Morgan.

Postdoctoral Scholars-1, Graduate Students-1

# SUMMARY PROPOSAL BUDGET

YEAR 1

ORGANIZATION <b>New York University</b>				FOR NSF USE ONLY			
				PROPOSAL NO.	DURATION (months)		
PRINCIPAL INVESTIGATOR / PROJECT DIRECTOR <b>Dennis E Shasha</b>				AWARD NO.	Proposed	Granted	
A. SENIOR PERSONNEL: PI/PI, Co-PI's, Faculty and Other Senior Associates (List each separately with title, A.7. show number in brackets)				NSF Funded Person-months		Funds Requested By proposer	Funds granted by NSF (if different)
				CAL	ACAD	SUMR	
1.	<b>Dennis E Shasha - PI</b>			0.00	0.00	0.40	\$ 7,119
2.	<b>Gloria M Coruzzi - Co-PI</b>			0.00	0.00	0.75	23,559
3.	<b>Arthur Goldberg - Senior Personnel</b>			12.00	0.00	0.00	90,653
4.	<b>Manpreet Katari - Co-PI</b>			8.00	0.00	0.00	61,583
5.							
6.	( 0 ) OTHERS (LIST INDIVIDUALLY ON BUDGET JUSTIFICATION PAGE)			0.00	0.00	0.00	0
7.	( 4 ) TOTAL SENIOR PERSONNEL (1 - 6)			20.00	0.00	1.15	182,914
B. OTHER PERSONNEL (SHOW NUMBERS IN BRACKETS)							
1.	( 1 ) POST DOCTORAL SCHOLARS			12.00	0.00	0.00	41,096
2.	( 2 ) OTHER PROFESSIONALS (TECHNICIAN, PROGRAMMER, ETC.)			18.00	0.00	0.00	79,572
3.	( 1 ) GRADUATE STUDENTS						39,332
4.	( 0 ) UNDERGRADUATE STUDENTS						0
5.	( 0 ) SECRETARIAL - CLERICAL (IF CHARGED DIRECTLY)						0
6.	( 0 ) OTHER						0
TOTAL SALARIES AND WAGES (A + B)							342,914
C. FRINGE BENEFITS (IF CHARGED AS DIRECT COSTS)							83,486
TOTAL SALARIES, WAGES AND FRINGE BENEFITS (A + B + C)							426,400
D. EQUIPMENT (LIST ITEM AND DOLLAR AMOUNT FOR EACH ITEM EXCEEDING \$5,000.)							
<b>Capital Equipment</b>							\$ 5,000
TOTAL EQUIPMENT							5,000
E. TRAVEL 1. DOMESTIC (INCL. CANADA, MEXICO AND U.S. POSSESSIONS)							2,000
2. FOREIGN							0
F. PARTICIPANT SUPPORT COSTS							
1.	STIPENDS \$ _____						0
2.	TRAVEL _____						0
3.	SUBSISTENCE _____						0
4.	OTHER _____						0
TOTAL NUMBER OF PARTICIPANTS ( 0 ) TOTAL PARTICIPANT COSTS							0
G. OTHER DIRECT COSTS							
1.	MATERIALS AND SUPPLIES						39,131
2.	PUBLICATION COSTS/DOCUMENTATION/DISSEMINATION						2,000
3.	CONSULTANT SERVICES						20,000
4.	COMPUTER SERVICES						0
5.	SUBAWARDS						0
6.	OTHER						14,553
TOTAL OTHER DIRECT COSTS							75,684
H. TOTAL DIRECT COSTS (A THROUGH G)							509,084
I. INDIRECT COSTS (F&A)(SPECIFY RATE AND BASE)							
<b>Modified Total Direct Cost (Rate: 54.0000, Base: 489531)</b>							
TOTAL INDIRECT COSTS (F&A)							264,347
J. TOTAL DIRECT AND INDIRECT COSTS (H + I)							773,431
K. RESIDUAL FUNDS							0
L. AMOUNT OF THIS REQUEST (J) OR (J MINUS K)							\$ 773,431
M. COST SHARING PROPOSED LEVEL \$ 0				AGREED LEVEL IF DIFFERENT \$			
PI/PI NAME <b>Dennis E Shasha</b>				FOR NSF USE ONLY			
ORG. REP. NAME* <b>Richard Louth</b>				INDIRECT COST RATE VERIFICATION			
		Date Checked	Date Of Rate Sheet	Initials - ORG			

# SUMMARY PROPOSAL BUDGET

YEAR 2

ORGANIZATION <b>New York University</b>				FOR NSF USE ONLY			
				PROPOSAL NO.	DURATION (months)		
PRINCIPAL INVESTIGATOR / PROJECT DIRECTOR <b>Dennis E Shasha</b>				AWARD NO.	Proposed	Granted	
A. SENIOR PERSONNEL: PI/PI, Co-PI's, Faculty and Other Senior Associates (List each separately with title, A.7. show number in brackets)				NSF Funded Person-months		Funds Requested By proposer	Funds granted by NSF (if different)
				CAL	ACAD	SUMR	
1.	<b>Dennis E Shasha - PI</b>			0.00	0.00	0.40	\$ <b>7,326</b> \$
2.	<b>Gloria M Coruzzi - Co-PI</b>			0.00	0.00	0.75	<b>24,242</b>
3.	<b>Arthur Goldberg - Senior Personnel</b>			12.00	0.00	0.00	<b>93,282</b>
4.	<b>Manpreet Katari - Co-PI</b>			8.00	0.00	0.00	<b>63,370</b>
5.							
6.	( 0 ) OTHERS (LIST INDIVIDUALLY ON BUDGET JUSTIFICATION PAGE)			0.00	0.00	0.00	<b>0</b>
7.	( 4 ) TOTAL SENIOR PERSONNEL (1 - 6)			20.00	0.00	1.15	<b>188,220</b>
B. OTHER PERSONNEL (SHOW NUMBERS IN BRACKETS)							
1.	( 1 ) POST DOCTORAL SCHOLARS			12.00	0.00	0.00	<b>42,287</b>
2.	( 2 ) OTHER PROFESSIONALS (TECHNICIAN, PROGRAMMER, ETC.)			18.00	0.00	0.00	<b>81,879</b>
3.	( 1 ) GRADUATE STUDENTS						<b>40,905</b>
4.	( 0 ) UNDERGRADUATE STUDENTS						<b>0</b>
5.	( 0 ) SECRETARIAL - CLERICAL (IF CHARGED DIRECTLY)						<b>0</b>
6.	( 0 ) OTHER						<b>0</b>
TOTAL SALARIES AND WAGES (A + B)							<b>353,291</b>
C. FRINGE BENEFITS (IF CHARGED AS DIRECT COSTS)							<b>85,906</b>
TOTAL SALARIES, WAGES AND FRINGE BENEFITS (A + B + C)							<b>439,197</b>
D. EQUIPMENT (LIST ITEM AND DOLLAR AMOUNT FOR EACH ITEM EXCEEDING \$5,000.)							
<b>Capital Equipment</b>						<b>\$ 5,000</b>	
TOTAL EQUIPMENT							<b>5,000</b>
E. TRAVEL 1. DOMESTIC (INCL. CANADA, MEXICO AND U.S. POSSESSIONS)							<b>2,080</b>
2. FOREIGN							<b>0</b>
F. PARTICIPANT SUPPORT COSTS							
1.	STIPENDS	\$	<b>0</b>				
2.	TRAVEL		<b>0</b>				
3.	SUBSISTENCE		<b>0</b>				
4.	OTHER		<b>0</b>				
TOTAL NUMBER OF PARTICIPANTS ( 0 )				TOTAL PARTICIPANT COSTS			<b>0</b>
G. OTHER DIRECT COSTS							
1.	MATERIALS AND SUPPLIES						<b>39,872</b>
2.	PUBLICATION COSTS/DOCUMENTATION/DISSEMINATION						<b>2,080</b>
3.	CONSULTANT SERVICES						<b>20,800</b>
4.	COMPUTER SERVICES						<b>0</b>
5.	SUBAWARDS						<b>0</b>
6.	OTHER						<b>15,135</b>
TOTAL OTHER DIRECT COSTS							<b>77,887</b>
H. TOTAL DIRECT COSTS (A THROUGH G)							<b>524,164</b>
I. INDIRECT COSTS (F&A)(SPECIFY RATE AND BASE)							
<b>Modified Total Direct Cost (Rate: 54.0000, Base: 504029)</b>							
TOTAL INDIRECT COSTS (F&A)							<b>272,176</b>
J. TOTAL DIRECT AND INDIRECT COSTS (H + I)							<b>796,340</b>
K. RESIDUAL FUNDS							<b>0</b>
L. AMOUNT OF THIS REQUEST (J) OR (J MINUS K)							\$ <b>796,340</b> \$
M. COST SHARING PROPOSED LEVEL \$ <b>0</b>				AGREED LEVEL IF DIFFERENT \$			
PI/PI NAME <b>Dennis E Shasha</b>				FOR NSF USE ONLY			
ORG. REP. NAME* <b>Richard Louth</b>				INDIRECT COST RATE VERIFICATION			
		Date Checked	Date Of Rate Sheet	Initials - ORG			

# SUMMARY PROPOSAL BUDGET

YEAR 3

ORGANIZATION <b>New York University</b>				FOR NSF USE ONLY			
				PROPOSAL NO.	DURATION (months)		
PRINCIPAL INVESTIGATOR / PROJECT DIRECTOR <b>Dennis E Shasha</b>				AWARD NO.	Proposed	Granted	
A. SENIOR PERSONNEL: PI/PI, Co-PI's, Faculty and Other Senior Associates (List each separately with title, A.7. show number in brackets)				NSF Funded Person-months		Funds Requested By proposer	Funds granted by NSF (if different)
				CAL	ACAD	SUMR	
1.	<b>Dennis E Shasha - PI</b>			0.00	0.00	0.40	\$ 7,538
2.	<b>Gloria M Coruzzi - Co-PI</b>			0.00	0.00	0.75	24,945
3.	<b>Arthur Goldberg - Senior Personnel</b>			12.00	0.00	0.00	95,987
4.	<b>Manpreet Katari - Co-PI</b>			8.00	0.00	0.00	65,208
5.							
6.	( 0 ) OTHERS (LIST INDIVIDUALLY ON BUDGET JUSTIFICATION PAGE)			0.00	0.00	0.00	0
7.	( 4 ) TOTAL SENIOR PERSONNEL (1 - 6)			20.00	0.00	1.15	193,678
B. OTHER PERSONNEL (SHOW NUMBERS IN BRACKETS)							
1.	( 1 ) POST DOCTORAL SCHOLARS			12.00	0.00	0.00	43,515
2.	( 2 ) OTHER PROFESSIONALS (TECHNICIAN, PROGRAMMER, ETC.)			18.00	0.00	0.00	84,254
3.	( 1 ) GRADUATE STUDENTS						42,542
4.	( 0 ) UNDERGRADUATE STUDENTS						0
5.	( 0 ) SECRETARIAL - CLERICAL (IF CHARGED DIRECTLY)						0
6.	( 0 ) OTHER						0
TOTAL SALARIES AND WAGES (A + B)							363,989
C. FRINGE BENEFITS (IF CHARGED AS DIRECT COSTS)							88,767
TOTAL SALARIES, WAGES AND FRINGE BENEFITS (A + B + C)							452,756
D. EQUIPMENT (LIST ITEM AND DOLLAR AMOUNT FOR EACH ITEM EXCEEDING \$5,000.)							
<b>Capital Equipment</b>							\$ 5,000
TOTAL EQUIPMENT							5,000
E. TRAVEL 1. DOMESTIC (INCL. CANADA, MEXICO AND U.S. POSSESSIONS)							2,163
2. FOREIGN							0
F. PARTICIPANT SUPPORT COSTS							
1.	STIPENDS \$ _____						0
2.	TRAVEL _____						0
3.	SUBSISTENCE _____						0
4.	OTHER _____						0
TOTAL NUMBER OF PARTICIPANTS ( 0 )							TOTAL PARTICIPANT COSTS
							0
G. OTHER DIRECT COSTS							
1. MATERIALS AND SUPPLIES							40,643
2. PUBLICATION COSTS/DOCUMENTATION/DISSEMINATION							2,163
3. CONSULTANT SERVICES							21,632
4. COMPUTER SERVICES							15,741
5. SUBAWARDS							0
6. OTHER							0
TOTAL OTHER DIRECT COSTS							80,179
H. TOTAL DIRECT COSTS (A THROUGH G)							540,098
I. INDIRECT COSTS (F&A)(SPECIFY RATE AND BASE)							
<b>Modified Total Direct Cost (Rate: 54.0000, Base: 519357)</b>							
TOTAL INDIRECT COSTS (F&A)							280,453
J. TOTAL DIRECT AND INDIRECT COSTS (H + I)							820,551
K. RESIDUAL FUNDS							0
L. AMOUNT OF THIS REQUEST (J) OR (J MINUS K)							\$ 820,551
M. COST SHARING PROPOSED LEVEL \$ 0				AGREED LEVEL IF DIFFERENT \$			
PI/PI NAME <b>Dennis E Shasha</b>				FOR NSF USE ONLY			
ORG. REP. NAME* <b>Richard Louth</b>				INDIRECT COST RATE VERIFICATION			
		Date Checked		Date Of Rate Sheet		Initials - ORG	

# SUMMARY PROPOSAL BUDGET

YEAR 4

ORGANIZATION <b>New York University</b>				FOR NSF USE ONLY			
				PROPOSAL NO.	DURATION (months)		
PRINCIPAL INVESTIGATOR / PROJECT DIRECTOR <b>Dennis E Shasha</b>				AWARD NO.	Proposed	Granted	
A. SENIOR PERSONNEL: PI/PI, Co-PI's, Faculty and Other Senior Associates (List each separately with title, A.7. show number in brackets)				NSF Funded Person-months		Funds Requested By proposer	Funds granted by NSF (if different)
				CAL	ACAD	SUMR	
1.	<b>Dennis E Shasha - PI</b>			0.00	0.00	0.40	\$ 7,757
2.	<b>Gloria M Coruzzi - Co-PI</b>			0.00	0.00	0.75	25,668
3.	<b>Arthur Goldberg - Senior Personnel</b>			12.00	0.00	0.00	98,771
4.	<b>Manpreet Katari - Co-PI</b>			8.00	0.00	0.00	67,099
5.							
6.	( 0 ) OTHERS (LIST INDIVIDUALLY ON BUDGET JUSTIFICATION PAGE)			0.00	0.00	0.00	0
7.	( 4 ) TOTAL SENIOR PERSONNEL (1 - 6)			20.00	0.00	1.15	199,295
B. OTHER PERSONNEL (SHOW NUMBERS IN BRACKETS)							
1.	( 1 ) POST DOCTORAL SCHOLARS			12.00	0.00	0.00	44,777
2.	( 2 ) OTHER PROFESSIONALS (TECHNICIAN, PROGRAMMER, ETC.)			18.00	0.00	0.00	86,697
3.	( 1 ) GRADUATE STUDENTS						44,243
4.	( 0 ) UNDERGRADUATE STUDENTS						0
5.	( 0 ) SECRETARIAL - CLERICAL (IF CHARGED DIRECTLY)						0
6.	( 0 ) OTHER						0
TOTAL SALARIES AND WAGES (A + B)							375,012
C. FRINGE BENEFITS (IF CHARGED AS DIRECT COSTS)							92,616
TOTAL SALARIES, WAGES AND FRINGE BENEFITS (A + B + C)							467,628
D. EQUIPMENT (LIST ITEM AND DOLLAR AMOUNT FOR EACH ITEM EXCEEDING \$5,000.)							
<b>Capital Equipment</b>							\$ 5,000
TOTAL EQUIPMENT							5,000
E. TRAVEL							2,250
1. DOMESTIC (INCL. CANADA, MEXICO AND U.S. POSSESSIONS)							
2. FOREIGN							0
F. PARTICIPANT SUPPORT COSTS							
1.	STIPENDS \$ _____						0
2.	TRAVEL _____						0
3.	SUBSISTENCE _____						0
4.	OTHER _____						0
TOTAL NUMBER OF PARTICIPANTS ( 0 )							TOTAL PARTICIPANT COSTS
							0
G. OTHER DIRECT COSTS							
1. MATERIALS AND SUPPLIES							41,447
2. PUBLICATION COSTS/DOCUMENTATION/DISSEMINATION							2,250
3. CONSULTANT SERVICES							22,497
4. COMPUTER SERVICES							16,370
5. SUBAWARDS							0
6. OTHER							0
TOTAL OTHER DIRECT COSTS							82,564
H. TOTAL DIRECT COSTS (A THROUGH G)							557,442
I. INDIRECT COSTS (F&A)(SPECIFY RATE AND BASE)							
<b>Modified Total Direct Cost (Rate: 54.0000, Base: 536072)</b>							
TOTAL INDIRECT COSTS (F&A)							289,479
J. TOTAL DIRECT AND INDIRECT COSTS (H + I)							846,921
K. RESIDUAL FUNDS							0
L. AMOUNT OF THIS REQUEST (J) OR (J MINUS K)							\$ 846,921
M. COST SHARING PROPOSED LEVEL \$ 0				AGREED LEVEL IF DIFFERENT \$			
PI/PI NAME <b>Dennis E Shasha</b>				FOR NSF USE ONLY			
ORG. REP. NAME* <b>Richard Louth</b>				INDIRECT COST RATE VERIFICATION			
		Date Checked	Date Of Rate Sheet	Initials - ORG			



# SUMMARY PROPOSAL BUDGET

YEAR 5

ORGANIZATION <b>New York University</b>				FOR NSF USE ONLY			
				PROPOSAL NO.	DURATION (months)		
PRINCIPAL INVESTIGATOR / PROJECT DIRECTOR <b>Dennis E Shasha</b>				AWARD NO.	Proposed	Granted	
A. SENIOR PERSONNEL: PI/PP, Co-PI's, Faculty and Other Senior Associates (List each separately with title, A.7. show number in brackets)				NSF Funded Person-months		Funds Requested By proposer	Funds granted by NSF (if different)
				CAL	ACAD	SUMR	
1.	<b>Dennis E Shasha - PI</b>			0.00	0.00	0.40	\$ 7,982
2.	<b>Gloria M Coruzzi - Co-PI</b>			0.00	0.00	0.75	26,413
3.	<b>Arthur Goldberg - Senior Personnel</b>			12.00	0.00	0.00	101,636
4.	<b>Manpreet Katari - Co-PI</b>			8.00	0.00	0.00	69,045
5.							
6.	( 0 ) OTHERS (LIST INDIVIDUALLY ON BUDGET JUSTIFICATION PAGE)			0.00	0.00	0.00	0
7.	( 4 ) TOTAL SENIOR PERSONNEL (1 - 6)			20.00	0.00	1.15	205,076
B. OTHER PERSONNEL (SHOW NUMBERS IN BRACKETS)							
1.	( 1 ) POST DOCTORAL SCHOLARS			12.00	0.00	0.00	46,075
2.	( 2 ) OTHER PROFESSIONALS (TECHNICIAN, PROGRAMMER, ETC.)			18.00	0.00	0.00	89,212
3.	( 1 ) GRADUATE STUDENTS						46,011
4.	( 0 ) UNDERGRADUATE STUDENTS						0
5.	( 0 ) SECRETARIAL - CLERICAL (IF CHARGED DIRECTLY)						0
6.	( 0 ) OTHER						0
TOTAL SALARIES AND WAGES (A + B)							386,374
C. FRINGE BENEFITS (IF CHARGED AS DIRECT COSTS)							95,302
TOTAL SALARIES, WAGES AND FRINGE BENEFITS (A + B + C)							481,676
D. EQUIPMENT (LIST ITEM AND DOLLAR AMOUNT FOR EACH ITEM EXCEEDING \$5,000.)							
<b>Capital Equipment</b>							\$ 5,000
TOTAL EQUIPMENT							5,000
E. TRAVEL 1. DOMESTIC (INCL. CANADA, MEXICO AND U.S. POSSESSIONS)							2,340
2. FOREIGN							0
F. PARTICIPANT SUPPORT COSTS							
1.	STIPENDS \$ _____						0
2.	TRAVEL _____						0
3.	SUBSISTENCE _____						0
4.	OTHER _____						0
TOTAL NUMBER OF PARTICIPANTS ( 0 )							TOTAL PARTICIPANT COSTS
							0
G. OTHER DIRECT COSTS							
1.	MATERIALS AND SUPPLIES						42,283
2.	PUBLICATION COSTS/DOCUMENTATION/DISSEMINATION						2,340
3.	CONSULTANT SERVICES						23,397
4.	COMPUTER SERVICES						17,024
5.	SUBAWARDS						0
6.	OTHER						0
TOTAL OTHER DIRECT COSTS							85,044
H. TOTAL DIRECT COSTS (A THROUGH G)							574,060
I. INDIRECT COSTS (F&A)(SPECIFY RATE AND BASE)							
<b>Modified Total Direct Cost (Rate: 54.0000, Base: 552036)</b>							
TOTAL INDIRECT COSTS (F&A)							298,099
J. TOTAL DIRECT AND INDIRECT COSTS (H + I)							872,159
K. RESIDUAL FUNDS							0
L. AMOUNT OF THIS REQUEST (J) OR (J MINUS K)							\$ 872,159
M. COST SHARING PROPOSED LEVEL \$ 0				AGREED LEVEL IF DIFFERENT \$			
PI/PP NAME <b>Dennis E Shasha</b>				FOR NSF USE ONLY			
ORG. REP. NAME* <b>Richard Louth</b>				INDIRECT COST RATE VERIFICATION			
		Date Checked		Date Of Rate Sheet		Initials - ORG	

# SUMMARY PROPOSAL BUDGET

Cumulative

ORGANIZATION <b>New York University</b>		FOR NSF USE ONLY			
		PROPOSAL NO.	DURATION (months)		
PRINCIPAL INVESTIGATOR / PROJECT DIRECTOR <b>Dennis E Shasha</b>		AWARD NO.	Proposed	Granted	
A. SENIOR PERSONNEL: PI/PP, Co-PI's, Faculty and Other Senior Associates (List each separately with title, A.7. show number in brackets)		NSF Funded Person-months		Funds Requested By proposer	Funds granted by NSF (if different)
		CAL	ACAD	SUMR	
1.	<b>Dennis E Shasha - PI</b>	0.00	0.00	2.00	\$ 37,722
2.	<b>Gloria M Coruzzi - Co-PI</b>	0.00	0.00	3.75	124,827
3.	<b>Arthur Goldberg - Senior Personnel</b>	60.00	0.00	0.00	480,329
4.	<b>Manpreet Katari - Co-PI</b>	40.00	0.00	0.00	326,305
5.					
6.	( ) OTHERS (LIST INDIVIDUALLY ON BUDGET JUSTIFICATION PAGE)	0.00	0.00	0.00	0
7.	( <b>4</b> ) TOTAL SENIOR PERSONNEL (1 - 6)	100.00	0.00	5.75	969,183
B. OTHER PERSONNEL (SHOW NUMBERS IN BRACKETS)					
1.	( <b>5</b> ) POST DOCTORAL SCHOLARS	60.00	0.00	0.00	217,750
2.	( <b>10</b> ) OTHER PROFESSIONALS (TECHNICIAN, PROGRAMMER, ETC.)	90.00	0.00	0.00	421,614
3.	( <b>5</b> ) GRADUATE STUDENTS				213,033
4.	( <b>0</b> ) UNDERGRADUATE STUDENTS				0
5.	( <b>0</b> ) SECRETARIAL - CLERICAL (IF CHARGED DIRECTLY)				0
6.	( <b>0</b> ) OTHER				0
TOTAL SALARIES AND WAGES (A + B)					1,821,580
C. FRINGE BENEFITS (IF CHARGED AS DIRECT COSTS)					446,077
TOTAL SALARIES, WAGES AND FRINGE BENEFITS (A + B + C)					2,267,657
D. EQUIPMENT (LIST ITEM AND DOLLAR AMOUNT FOR EACH ITEM EXCEEDING \$5,000.)					
	\$ 25,000				
TOTAL EQUIPMENT					25,000
E. TRAVEL 1. DOMESTIC (INCL. CANADA, MEXICO AND U.S. POSSESSIONS)					10,833
2. FOREIGN					0
F. PARTICIPANT SUPPORT COSTS					
1.	STIPENDS \$ _____ <b>0</b>				
2.	TRAVEL _____ <b>0</b>				
3.	SUBSISTENCE _____ <b>0</b>				
4.	OTHER _____ <b>0</b>				
TOTAL NUMBER OF PARTICIPANTS ( <b>0</b> ) TOTAL PARTICIPANT COSTS					0
G. OTHER DIRECT COSTS					
1.	MATERIALS AND SUPPLIES				203,376
2.	PUBLICATION COSTS/DOCUMENTATION/DISSEMINATION				10,833
3.	CONSULTANT SERVICES				108,326
4.	COMPUTER SERVICES				49,135
5.	SUBAWARDS				0
6.	OTHER				29,688
TOTAL OTHER DIRECT COSTS					401,358
H. TOTAL DIRECT COSTS (A THROUGH G)					2,704,848
I. INDIRECT COSTS (F&A)(SPECIFY RATE AND BASE)					
TOTAL INDIRECT COSTS (F&A)					1,404,554
J. TOTAL DIRECT AND INDIRECT COSTS (H + I)					4,109,402
K. RESIDUAL FUNDS					0
L. AMOUNT OF THIS REQUEST (J) OR (J MINUS K)					\$ 4,109,402
M. COST SHARING PROPOSED LEVEL \$ <b>0</b>		AGREED LEVEL IF DIFFERENT \$			
PI/PP NAME <b>Dennis E Shasha</b>		FOR NSF USE ONLY			
ORG. REP. NAME* <b>Richard Louth</b>		INDIRECT COST RATE VERIFICATION			
		Date Checked	Date Of Rate Sheet	Initials - ORG	

C \*ELECTRONIC SIGNATURES REQUIRED FOR REVISED BUDGET

## **BUDGET JUSTIFICATION NEW YORK UNIVERSITY**

### **A. Senior Personnel**

**PI:** Dr. Dennis Shasha will be the lead PI on this grant. He will be responsible for planning and directing the mathematic and computational aspects in this project. In particular, he will oversee the development of the time series machine learning tools and optimization algorithms for network inference that form the core analytic portions of the grant. He will also design the network visualization tools that will display the networks and simulate their changes through time. Dr. Shasha will devote 0.4 summer month of effort to this project. Salary is calculated on a 9-month base.

**Co-PI:** Dr. Gloria Coruzzi will be a co PI on this grant. She will be responsible for the overall scientific direction of the research in this project. In particular, she will be actively involved in the planning and execution of the biological aspects of the project. Dr. Coruzzi will devote 0.75 summer month of effort to this project. Salary is calculated on a 9-month base.

**Co-PI:** Dr. Manpreet Katari will be a co PI on this grant. He will play a primary role managing the Bioinformatics aspects of the project. He will primarily be involved in the analysis of the Illumina reads as described in Aims 2 and 3, including assembly of Illumina reads into contigs and analysis of expression levels. He will also be in charge of the assembling of networks in the target species. Dr. Katari will devote 8.0 calendar months of effort to this project. Salary is calculated on a 12-month base.

**Senior Personnel:** Dr. Arthur Goldberg will be senior personnel on this grant. He will manage the development of new software for the VirtualPlant System which will support the different species and inference, and also new pipelines for cross species analysis, especially as they relate to crop species. Dr. Goldberg will dedicate 12.0 calendar months effort to this project. Salary is calculated on a 12-month base.

### **B. Other Personnel**

**Post Doctoral Researcher:** Amy Marshall-Colon will be the post doctoral researcher working on this grant full-time (12 calendar months). She will be responsible for generating the tissue for Illumina sequencing. Salary is calculated on a 12-month base.

**Programmer:** Funds are requested to support one full-time (12 calendar months, Rebecca Davidson) and one part-time (6 calendar months, TBD) professional programmer. They will be responsible for coding the core of the VirtualPlant system and specifically for implementing the novel design concepts of dynamic visualization of the new aspects: dynamic networks and comparative genomic networks. The programmers will implement the different Views, the Model and the Controller of the VirtualPlant system. In addition, he/she will contribute to the design and implementation of the local database system and data analysis and prediction modules. The programmers will also be in charge of the bug-tracking facilities, moderating the VirtualPlant discussion list and the VirtualPlant system web pages. Salary is calculated on a 12-month base.

**Graduate Student:** Funds are requested to support one graduate research assistant. The graduate student will be trained by Dr. Shasha (NYU Courant) in the computational aspects of the project, especially focusing on the machine learning aspects of the project as described in Aims 2 and 4. In each year of the proposal, the graduate student will be appointed as a full-time Research Assistant for a total of 12 months: 9 months during the academic year (salary is calculated on nine-month base); plus 3 months during the summer (one summer month is equivalent to 1/6<sup>th</sup> of the academic year base).

**C. Fringe benefits**

All personnel except Graduate Students:  
 27.5% through 8/31/13  
 28.0% for the period starting 9/1/13  
 28.5% for the period starting 9/1/15

**D. Equipment**

**\$5,000/year:** We expect our warehouse to grow as new genomic data types for Arabidopsis are made available. Funds are requested to cope with the increasing demand on data storage as well as processing power to handle this data. Proposed upgrades will include additional hard drives, processors and RAM memory for the server. This equipment will be used solely and specifically for the research outlined in this proposal.

**E. Travel**

Travel budget is requested for the PI to attend 1 scientific meeting per year to present VirtualPlant progress. Travel budget will cover the cost of a domestic plane ticket and hotel stay.

**G. Other Direct Costs**

**Materials and Supplies:**

Illumina sequencing	20,580
Restriction & modification enzymes	2,550
Chemicals & biochemical	4,000
Oligonucleotide synthesis chemicals	2,000
Glass and plasticware	2,000
Plant growth supplies	2,000
Computer software	2,000
Backup Media & storage	3,000
Subscription renewals	1,000
<b>Total</b>	<b>\$39,130</b>

Summary of Illumina sequencing reactions				
RICE				
Timepoints	Treatments	Replicates	Runs	
0 min	KNO3 vs KCl	2 biological	4	
3 min	KNO3 vs KCl	2 biological	4	
6 min	KNO3 vs KCl	2 biological	4	
9 min	KNO3 vs KCl	2 biological	4	
12 min	KNO3 vs KCl	2 biological	4	
15 min	KNO3 vs KCl	2 biological	4	
20 min	KNO3 vs KCl	2 biological	4	
			TOTAL	28 RUNS
MEDICAGO - Naïve				
Timepoints	Treatments	Replicates	Runs	
0 min	KNO3 vs KCl	2 biological	4	
3 min	KNO3 vs KCl	2 biological	4	
6 min	KNO3 vs KCl	2 biological	4	
9 min	KNO3 vs KCl	2 biological	4	
12 min	KNO3 vs KCl	2 biological	4	
15 min	KNO3 vs KCl	2 biological	4	
20 min	KNO3 vs KCl	2 biological	4	
			TOTAL	28 RUNS
MEDICAGO - Rhizobium inoculated				
Timepoints	Treatments	Replicates	Runs	
0 min	KNO3 vs KCl	2 biological	4	
3 min	KNO3 vs KCl	2 biological	4	
6 min	KNO3 vs KCl	2 biological	4	
9 min	KNO3 vs KCl	2 biological	4	
12 min	KNO3 vs KCl	2 biological	4	
15 min	KNO3 vs KCl	2 biological	4	
20 min	KNO3 vs KCl	2 biological	4	
			TOTAL	28 RUNS
			<b>Total Illumina Sequencing Runs</b>	<b>84</b>
			<b>as of 01/18/2010 - \$1,225 per lane</b>	<b>\$102,900</b>
				\$20,580 per year

**Consultant Services: \$20,000/per year**

**Dr. Rodrigo Gutierrez.** Dr. Gutierrez has experience in assembling multinetworks for Arabidopsis, Rice and Grape. His primary responsibility will be to help Dr. Katari assemble networks for target crop species.

**Other Costs:** Tuition Remission in lieu of fringe benefits is calculated at 37.0% of graduate student salary.

## **I. Indirect Costs**

### **Indirect Costs:**

Overhead is calculated at a rate of 54% for the entire duration of the project per the DHHS agreement dated June 16, 2009. Overhead is not charged on the following:

- Equipment items costing \$3,000 or more
- Tuition remission for Research Assistants

### **INFLATORS**

The following increases are budgeted:

- Salary: Faculty and professional:
  - 2.9% as of 9/1/11 and thereafter
- Graduate Students: 4%
- Other Than Personnel Services: 4%

Current and Pending Support Investigator: Dennis Shasha								
TITLE	SOURCE	TOTAL AMOUNT	PERIOD	LOCATION	CAL. MONTHS	AY MONTHS	SUMMER MONTHS	STATUS
Genomics of Comparative Seed Evolution	NSF	\$5,716,000	10/1/04 - 2/28/10	NYU	0	0	1	Current
Arabidopsis 2010: Nitrogen Networks in Plants	NSF	\$2,600,000	9/1/05 - 8/31/10	NYU	0	0	1	Current
Conceptual Data Integration for the Virtual Plant	NSF	\$2,198,682	6/1/05 - 11/30/10	NYU	0	0	0.4	Current
A Systems approach to regulatory networks controlling N-assimilation	NIH	\$1,492,787	5/1/09 - 4/30/13	NYU	0	0	1	Current
Arabidopsis 2010: Nitrogen Networks in Plants (Renewal)	NSF	\$2,796,615	7/15/09 - 6/30/13	NYU	0	0	1	Current
GEPR Genomics of Comparative Seed Evolution	NSF	\$6,390,886	10/1/09 - 9/30/14	NYU	0	0	0.6	Pending
TRMS:Cross Species Network Inference: From Models to Crops	NSF	\$4,109,402	12/1/10 - 11/30/14	NYU	0	0	1	Pending

Current and Pending Support								
Investigator: Gloria Coruzzi								
TITLE	SOURCE	TOTAL AMOUNT	PERIOD	LOCATION	CAL. MONTHS	AY MONTHS	SUMMER MONTHS	STATUS
Genomics of Comparative Seed Evolution	NSF	\$4,994,583	10/1/04 - 2/28/10	NYU	0	1	0	Current
Arabidopsis 2010: Nitrogen Networks in Plants	NSF	\$2,600,000	9/1/05 - 8/31/10	NYU	0	0	0.01	Current
Conceptual Data Integration for the Virtual Plant	NSF	\$2,198,682	6/1/05 - 11/30/10	NYU	0	0	0.75	Current
The function of small RNAs in the nitrogen response	NIH-FIRCA	\$106,666	1/1/08 - 12/31/10	Catholic University of Chile	0.45	0	0	Current
Asparagine Synthetase Regulatory Network and Plant Nitrogen Metabolism	DOE	\$495,000	6/1/08 - 5/31/11	NYU	0	0	0.5	Current
A Systems Approach to Regulatory Networks Controlling N-assimilation	NIH	\$1,492,787	5/1/09 - 4/30/13	NYU	0	0	1	Current
Arabidopsis 2010: Nitrogen Networks in Plants (Renewal)	NSF	\$2,796,615	7/15/09 - 6/30/13	NYU	0	0	0.75	Current
GEPR Genomics of Comparative Seed Evolution (Renewal)	NSF	\$3,750,000	12/1/09 - 11/30/13	NYU	0	0	0.5	Pending

Current and Pending Support								
Investigator: Gloria Coruzzi								
TITLE	SOURCE	TOTAL AMOUNT	PERIOD	LOCATION	CAL. MONTHS	AY MONTHS	SUMMER MONTHS	STATUS
TRMS:Cross Species Network Inference: From Models to Crops	NSF	\$4,109,402	12/1/10 - 11/30/14	NYU	0	0	0.75	Pending



Current and Pending Support								
Investigator: Manpreet Katari								
TITLE	SOURCE	TOTAL AMOUNT	PERIOD	LOCATION	CAL. MONTHS	AY MONTHS	SUMMER MONTHS	STATUS
TRMS:Cross Species Network Inference: From Models to Crops	NSF	\$4,109,402	12/1/10 - 11/30/14	NYU	8	0	0	Pending

Current and Pending Support Investigator: Arthur Goldberg								
TITLE	SOURCE	TOTAL AMOUNT	PERIOD	LOCATION	CAL. MONTHS	AY MONTHS	SUMMER MONTHS	STATUS
TRMS:Cross Species Network Inference: From Models to Crops	NSF	\$4,109,402	12/1/10 - 11/30/14	NYU	12	0	0	Pending

## FACILITIES, EQUIPMENT AND OTHER RESOURCES

### NEW YORK UNIVERSITY

#### NYU BIOLOGY CENTER FOR GENOMICS & SYSTEMS BIOLOGY

##### **CORUZZI Laboratory:**

The co-PI (Coruzzi, NYU) has approximately 1000 sq. ft. of laboratory space in an open plan lab, with bench and desk space for 15+ researchers within the Center for Genomics and Systems Biology's 16,000 sq. ft of shared lab and facilities space. Separate areas are provided for a conference room, tissue culture room, cold room, dark room, and an autoclave and dishwashing facility. In addition, the laboratory contains bench model freezers, single-door cold boxes as well as two tissue culture hoods and two fume hoods.

**Office:** A private office of 150 sq. ft. is provided for the co-PI adjacent to the lab. The office is furnished with a computer, printer, scanner and fax machine.

**Equipment:** The CORUZZI lab is well equipped to perform the proposed research. Equipment includes a Sorval RC-5B super speed centrifuge, a Beckman L-80M Ultracentrifuge, NanoDrop ND-1000 Spectrophotometer, New Brunswick Scientific Floor Incubator/Shaker, two Revco -80° freezers, Gel Logic 200 Imaging System, Sevant ISS110 Speedvac Concentrator, Biorad Tetrad 2 thermocycler, Perkin Elmer 9600 thermocycler, Qiagen TissueLyser, water purification system as well as several bench top centrifuges, shakers, hybridization ovens, heat blocks, and other minor equipment. The lab is also fully equipped for Chromatin ImmunoPrecipitation (ChIP) experiments including a Diagenode Sonicator, Invitrogen Magnetic Particle Concentrator, and the necessary electrophoresis system.

**Genome Core Facility:** The Core Facility includes an **Imaging Facility** run by [Dr. Ignatius Tan](#) that currently includes state of the art microscopes including a Leica Spectral Confocal Microscope and Improvion Spinning Disk Confocal system for high through-put imaging. Both confocal systems have separate computer workstations configured to run Leica and Improvion's Volocity image analysis software. **The DNA Facility** includes genome expression equipment including an Affymetrix GeneScan Hybridization and Reader system, Agilent 2100 Bioanalyzer, GenePix Personal 4100A microarray scanner from Axon Instruments, a Light Cycler, and several automated DNA sequencers. **The Robotics Facility** includes an Acquarius multi-channel pipettor with autoloader by Tecan and a Tecan EVO platform configured with shaking incubator, tilting, cooling and heating carriers, a vacuum system, multi-channel pipetting head, bar code reader, and the Infinite M200 microplate reader. As a stand-alone or part of the EVO platform, the Infinite M200 reader is fully loaded incubating, shaking reader with (top/bottom read) fluorescence, luminometer, and excitation and emission monochromator options. **The Cellular & Phenotyping Facility** consists of the BD FACS Aria a high speed cell sorter with two laser lines (488, 633 nm) capable of automated dispensing into a 96 well format and sterile environment. Bacterial samples can be sorted through a forward scattering PMT.

##### **Other:**

Twelve 8x10 ft. walk-in plant growth chambers and five reach in Percival Growth Chambers are located in the basement and are accessed by private elevator.

**BIOINFORMATICS:** The Coruzzi laboratory is equipped with two Apple G4s, two Apple G5, two Dell PCs with Pentium IV processors and one HP Omnibook 6000 with a Pentium II processor (used to run the Roche LightCycler Quantitative PCR instrument and the NanoDrop spectrophotometer). The VirtualPlant team uses one 4 core and one 8 core Mac Pro, each with 4GB of memory as workstations. The

VirtualPlant web server, database server, and development server are IBM e326 servers with 4 cores and 4GB of RAM each. In addition, a cluster has been purchased from Silicon Mechanics comprised of 8 nodes, each with 8 cores and 32GB of RAM with a 5 Terabyte file system. The purpose of this cluster is to support the computationally intense calculations related to the VirtualPlant project.

**GENOME CENTER Bioinformatics Suites and Cluster:** The Bioinformatics suites in the Center for Genomics and Systems Biology contain approximately 30 workstations and associated computers. The computer cluster that serves these bioinformatics stations is presently comprised of 26 computational nodes, each with at least 4GB RAM. In addition to many essential software programs, the cluster uses the Sun Grid Engine for load balancing and process distribution among 34 Opteron 242 2.2 GHz processors. Disk storage available is currently over 10 terabytes. The Systems Administrator for the Bioinformatics cluster at NYU's Center for Genomics and Systems Biology is John Bako.

**Dennis Shasha of the Courant Institute** brings to this project the resources associated with the Computer Science department including a network of several hundred desktop workstations, file servers, multi-processor computational servers and clusters, and state-of-the-art visualization facilities, all supported by a central server infrastructure. The Courant Institute also has access to the Academic Computing Services, a unit of the NYU-wide Information Technology Services offering an additional wide range of computational resources in support of research and instruction. This includes a 4.5 TeraFlops supercomputer, consisting of a cluster of 256 IBM eServer BladeCenter JS20 dual-processor nodes and using Myrinet switch technology. It is the first supercomputer in the United States that is using IPv6, the next generation Internet Protocol.

## **(A-1) Sharing of Results and Management of Intellectual Property**

### **Results Sharing**

**1. Data Distribution:** In this NSF Plant Genome project, we will generate deep-transcriptomic (Illumina) time-series data for N-treatment of two crop species, Rice and Medicago. This project proposes to generate short read sequences for transcriptome analysis. A typical sequencer run will generate several hundred gigabytes of data. Once NCBI fully implements the equivalent of a trace archive for short read sequences, we will deposit our sequence data into that database. Once we have finished aligning the reads to the respective genome sequence, the alignment information will be stored in a database and made available to the public using the Gbrowse ([www.gbrowse.org](http://www.gbrowse.org)) interface. Normalized expression values from these experiments will be made available as supplemental data in the journal where it is accepted.

**2. Informatic Resources & Software:** The informatic analysis pipelines for Cross Species Network Inference (CSNI), discussed in Aim 4 will be made available to the community free of charge. We will implement CSNI as a general-purpose tool to be used by plant scientists. We plan to deploy CSNI as a website ([www.crossspecies.org](http://www.crossspecies.org)) linked to several additional platforms, first to VirtualPlant website ([www.virtualplant.org](http://www.virtualplant.org)), and second to *iPlant* (see S. Goff letter). We will also deploy CSNI on one of the widely-used bioinformatic workflow engines: Taverna [2], Kepler [3] or Galaxy [4]. Implementing the CSNI pipeline on top of one (or more) of these bioinformatic workflow engines is important because they provide increasingly popular platforms for developing computational genetic analyses, and provide generic support for reproducible bioinformatic analyses.

**Publications:** The results of our analysis of the data we generate will be made available through peer-reviewed literature as it is the most appropriate way to make this information available.

### **Intellectual Property**

#### **1. Invention Disclosure and Patent Management:**

**Invention Disclosures** Invention disclosures will be reported by the inventors to the office responsible for patenting and licensing at their institutions. At NYU, this is the Office of Industrial Liaison. Invention disclosure forms are available on the Office of Industrial Liaison website ([www.nyu.edu/oil](http://www.nyu.edu/oil)) under “New Invention Disclosure Form.” The principal investigator at each institution will be responsible for ensuring that inventions are disclosed to their institution’s patenting and licensing office sufficiently prior to public disclosure to allow the office to evaluate patentability and commercial potential, and to have a patent application filed if appropriate. The offices at institutions other than NYU will notify the NYU Office of Industrial Liaison of any inventions made at their institutions.

**Patenting** The institution at which the inventor is employed will be responsible for evaluating whether to retain title to inventions, filing U.S. and/or foreign patent applications, and notifying NSF. Inventions will be disclosed to the funding agency promptly upon receipt. Decisions on whether to file a patent application will be based upon an evaluation of the commercial potential of the invention by the institutional patenting and licensing office, and an evaluation of the patentability of the invention (including a search of prior art) in conjunction with an outside patent law firm. It is expected that the first patent filing will typically be a provisional filing. Decisions on whether to file non-provisional U.S. and or PCT applications will be made in the 10-11 month timeframe following the first (provisional) filing.

**Patent Reporting** NYU participates in the Interagency Edison System. NYU will notify the funding agency and grant the required non-exclusive license for government purposes for inventions for which NYU elects title. A final invention statement will be sent to NSF upon completion of the grant.

**2. Licensing and Commercialization:** The Office of Industrial Liaison at NYU shall be responsible for

seeking to make technology developed under the grant at NYU commercially available. Technology developed at the other participating institutions will be licensed by the technology licensing offices of those institutions. Research materials developed under the grant will be made available to researchers at not-for-profit institutions under a material transfer agreement with intellectual property terms consistent with NIH guidelines. The Office will also seek out appropriate industry partners to commercially develop technology. Potential arrangements will include exclusive or non-exclusive licenses. Any licenses will contain provisions requiring the licensing company to diligently develop the technology. Options may also be granted to companies considering licensing for limited time periods (e.g., 3-6 months) to allow them to conduct due diligence and evaluate their interest in a license. It is expected that technology will be licensed to existing companies, but if the technology is of sufficient breadth to justify the creation of a new start-up company, this will be considered.

**3. Inter-Institutional Agreement:** As described above, inventions made solely by one institution shall be managed by that institution. For inventions made by more than one institution, an inter-institutional agreement shall be agreed upon, under which the parties shall agree on which institution shall take the lead in managing patenting and licensing activities, and on sharing patent expenses and license revenues based on relative contributions. License agreements shall require the approval of all institutions with an ownership interest in the licensed technology.

## **(A-2) Management Plan**

**Overview:** Our proposal is a result of a highly successful collaboration between computer scientists at *NYU Courant Institute of Mathematical Sciences* and biologists at *NYU Biology's Center for Genomics and Systems Biology*. Our close proximity and ongoing successful working relationships during the "VirtualPlant" software development NSF project (DBI DBI-0445666) to enable Systems Biology analysis in Arabidopsis encourages us to further develop and expand the project to crop species. In this proposed NSF Plant Genome Grant, we will generate the data (transcriptomic) for several crop species (Rice and Medicago) as proof-of-principle, as well as informatic tools and resources and pipeline of informatic analysis tools to develop cross species network inference and machine learning approaches to discover genes associated with important economic traits. The diverse background of the participants in this project provides an opportunity to combine the expertise of investigators with backgrounds in plant biology, genomics, bioinformatics and computer science.

**A successful management plan is already in place and this plan is described below:** To coordinate and facilitate interactions between individuals, Dennis Shasha (NYU Courant) and Gloria Coruzzi (NYU Biology) will serve as Project Managers for computational and experimental aspects respectively. The role of the Project Managers is to oversee the daily operations of the project and insure that the needs and concerns of the participants are addressed on a day-to-day basis between the participants involved. This includes monitoring and facilitating the technical operations both experimental and computational ranging from plant growth and tissue collection, RNA purification and sequencing to software development, testing and implementation. The project manager will also facilitate communication between PIs, post-docs, graduate students and laboratory technicians by scheduling weekly meetings of all participants to manage immediate issues regarding research needs. We will also schedule day-long meetings twice a semester with collaborators on the experimental end (Doug Cook, UC Davis) and on the computational end (Rodrigo Gutierrez, Chile), to do evaluation of work status and long term planning.

**Bioinformatics manager:** Dr. Manpreet Katari (NYU Biology) will be in charge of the bioinformatics analysis aspects including RNA-seq analysis, databases, and multinet network generation. He will also serve as a liaison to researchers and technicians at CSHL where Illumina sequence generation of deep-seq libraries and bioinformatic analysis occurs (see letter from McCombie, CSHL). Dr. Katari will facilitate the processing of this data into contigs and expression data for Machine Learning, State Space analysis. To enable efficient information exchange of raw and processed data, a file server has been set up at the NYU to store and distribute data and its analysis among users at NYU Biology and NYU Courant. A secondary server is also established to create daily backups of the primary server in order to insure and protect data. Relevant information is also accessible in a web-based interface for authorized users. Dr. Katari will maintain the web server, database server, and the multinet network database.

**Software development manager:** Dr. Arthur Goldberg (NYU Courant) will manage the development of new software analysis tools and pipelines to enable Cross Species Network Inference (CSNI) which will support the different species and inference, and also new pipelines for cross species analysis, especially as they relate to crop species.

**Website:** We have set up a web site to house the development of Cross Species Network Inference tools and pipelines, which is accessible at: [www.CrossSpecies.org](http://www.CrossSpecies.org)

**Education & outreach:** MS and PhD students involved in this project will be co-mentored by a Biology and Courant faculty advisors, Coruzzi (NYU Biology) and Shasha (NYU Courant). The students will present their work at weekly joint lab meetings. PhD students will also present their work annually in the weekly PhD seminar series hosted by the Biology Department. Computational students will be involved

in constructing the pipeline and making it perform through the use of parallelization. Such students will also help to develop and test optimization and machine learning algorithms for network inference. Biological students will engage in experimental work including preparation of RNA for Illumina sequencing, as well as the analysis of data using the VirtualPlant platform. Students will be exposed to Genomics and Systems Biology also through a series of courses offered by faculty at NYU's Center for Genomics and Systems Biology including: G23.1128 Systems Biology; G23.1130 Applied Genomics: Introduction to Bioinformatics & Network Modeling; G23.1127 Bioinformatics & Genomes.

**Principal Investigators:** The three principal investigators will each commit a minimum of a half-day per week to this project. This will include supervision of personnel, organizational meeting attendance, and intellectual developments and contributions.

**Role of senior participants and timeline:**

Name	Institution	Role	Aim
<i>Dennis Shasha</i> PI	NYU Courant	Project Leader	Oversee Aims 1, 2, 3, 4
<i>Gloria Coruzzi</i> , Co-PI	NYU Biology	Co-leader: Experimental	Oversee Aims 2 & 3
<i>Manpreet Katari</i> Co-PI	NYU Biology	Bioinformatics Manager	Aims 1 & 4
<i>Arthur Goldberg</i> Senior Personnel	NYU Courant	Software developer	Aims 1 & 4
<i>Amy Marshall-Colon</i> Post-Doc	NYU Biology	Experimental Treatments RNA processing for Seq	Aim 2 Rice experiments
<i>Doug Cook</i> Collaborator	UC Davis	Experimental support Plant treatments	Aim 3 Medicago Experiments
<i>Rodrigo Gutierrez</i> Consultant	U Catolica, Chile	Assembling "ground truth" networks for crop	Aims 1 & 4

**Timeline:**

**Year 1:** Aim 1. Extend cross species network inference by 1) using ground truth protein:protein and metabolic interaction networks for Rice and Arabidopsis from several different sources, 2) using other tools, methods and databases for defining homology, and 3) incorporating validated regulatory (AGRIS) edges. Aim 2 & 3: Perform the time series experiments in Rice and Medicago and complete the Arabidopsis series by generating RNA-seq data. Aim 4: Build and/or assemble high-performance network inference software for time series data (e.g. parallelize State-Space modeling) and compare it with others (especially other Bayesian methods). Assemble "ground truth" networks in the 3-5 target crop species beginning with Medicago, Corn, Grape. Evaluate different bioinformatic workflow platforms and decide which we will deploy.

**Years 2-3:** Aim 2 & 3. Apply the state space analysis to Medicago and Rice and test targeted regulatory genes in N-use pathway. Aim 4. Deploy the first version of the analysis pipeline for cross species network inference to collaborators (D. Cook, U Davis; R. Gutierrez, Chile).

**Years 4-5:** Apply the computational pipeline to infer networks in several crop species for example corn and grape. Deploy the full computational CSNI pipeline for cross-species network inference to plant community via CSNI ([www.CrossSpecies.org](http://www.CrossSpecies.org)) linked to VirtualPlant, iPlant and a selected workflow platform (e.g. Galaxy).



### **(A-3) Coordination with Outside Groups**

**Please see attached letters of collaboration:**

**Doug Cook (UC Davis)** Dr. Cook, an expert in the field of Medicago genomics will perform the N-treatment time series and supply tissues for RNA-seq analysis (Aim 3), to fuel the regulatory network inference studies in Medicago.

**Rodrigo Gutierrez (U Catolica, Chile)** Dr. Gutierrez, the creator of the Arabidopsis multinetwork (Gutierrez et al 2007) will assist in the assembly of multinetworks for crop species including Vitis, Corn and Medicago.

**iPlant (see letter from iPlant Project Director, Steve Goff)** We will coordinate with iPlant to make our Cross species network inference platform (CSNI) modular, independent and accessible with and compatible with iPlant, and accessible using other annotation analysis platforms such as Galaxy and Taverna. We will also make our currently developed VirtualPlant tools accessible to iPlant, as per letter by (S. Goff).

**Richard McCombie (CSHL)** Dr. McCombie, will process our time-series RNA seq data using CSHL Genome Sequencing facility using an Illumina sequencing platform.

### **(A-5) Postdoctoral Mentoring Plan**

**Co-mentorship across disciplines and institutions:** Post-Docs and students will receive novel cross-disciplinary training across biology & genomics (NYU Biology, Center for Genomics & Systems Biology) and informatics and systems Biology (NYU Courant). The PIs with expertise spanning these disciplines will co-mentor students and post docs individually and also at the weekly meetings where postdocs & students present their results.

**Training as Educators:** In this project, Post-Docs are paired up with graduate students, undergraduate students, and technicians in the laboratory/at the computer to practice mentoring skills in a research context. Post-docs are also afforded the opportunity to teach in NYU Biology courses where they are mentored by faculty advisors. For example, Dr. Katari, is currently co-teaching an undergraduate course “Introduction to Genomics & Bioinformatics” with a faculty mentor (Kris Gunsalus).

**Career Development:** Post-Docs receive counseling from their co-mentors and practice presentation skills during regular group-lab meetings, through a special NYU Biology Post-Doc seminar series, and at annual poster sessions at the NYU Biology retreat. Funds are provided for students and Post-Docs to attend at least one meeting each year and are expected to widely disseminate their work.



COLLEGE OF AGRICULTURE AND  
ENVIRONMENTAL SCIENCES  
AGRICULTURAL EXPERIMENT STATION  
DEPARTMENT OF PLANT PATHOLOGY  
TELEPHONE: (530)752-0300  
FAX: (530)752-5674

ONE SHIELDS AVENUE  
354 HUTCHISON HALL  
DAVIS, CA 95616-8680

January 14, 2010

Dr. Gloria Coruzzi  
New York University  
Department of Biology  
Center for Genomics & Systems Biology

Dear Gloria,

The purpose of this letter is to express my enthusiasm for participating in your proposed project to the NSF Plant Genome Research Program entitled "Cross species network inference: From Models to Crops", as a non-funded collaborator. According to the experimental plan, my group would provide expertise in *Medicago* as well as tissues of *Medicago truncatula*, grown according to differing N-regimes (naïve and symbiotic) as specified in Aim 3. We would also help identify appropriate genome resources (e.g., genetic and genomic) in *Medicago truncatula* that will aid in the construction of a *Medicago* multinet, which would be useful to us, and to the rest of the *Medicago* community.

The proof-of-principle topic of nitrogen responses in legumes that you have chosen for your cross species network inference studies is of clear scientific and agronomic interest, especially as it provides a point of comparison for the detailed data sets you have developed in *Arabidopsis* and that you will also develop in rice, as per the proposal. Moreover, nitrogen-response pathways are of potential practical importance for legumes in agriculture. We know essentially nothing of how legumes respond transcriptionally to reduced nitrogen, including how they distinguish nitrogen derived from the soil environment versus symbiotic nitrogen fixation. As external nitrogen represses N-fixation, the elucidation of this regulatory network would have potential applications in agriculture.

As our own lab is developing protein:protein interaction data for *Medicago*, we would also be happy to contribute this information to aid in the analysis & validation of inferred networks in *Medicago*. Finally, my group also has recently initiated studies to examine the impact of domestication on nitrogen fixation in chickpea, one of the first domesticated legumes and a close relative of *Medicago truncatula*. This is the "nitrogen-network" in action, viewed through the domestication bottleneck, and I am excited about the potential synergy between the two projects.

Best of luck in the review process.

Sincerely,

Doug

A handwritten signature in black ink that reads "Douglas R. Cook".

Douglas Cook  
Professor of Plant Pathology  
[drcook@ucdavis.edu](mailto:drcook@ucdavis.edu); 530-754-6561; 530-304-0759



Núcleo Milenio en  
**Genómica Funcional de Plantas**



**Rodrigo A. Gutiérrez, Ph.D.**

Director Centro Núcleo Milenio en Genómica Funcional de Plantas

Profesor Asistente

Laboratorio de Biología de Sistemas

Departamento de Genética Molecular y Microbiología

Facultad de Ciencias Biológicas

P. Universidad Católica de Chile.

Casilla 114-D, Santiago.

[rgutierrez@uc.cl](mailto:rgutierrez@uc.cl)

Fax: (56-2) 222-5515

Tel : (56-2) 686-2663

**January 20, 2010**

Dr. Gloria M. Coruzzi

Chair of Biology

Carroll and Milton Petrie Professor

New York University

Department of Biology

100 Washington Square East

1009 Main Building

New York, N.Y. 10003

Dear Dr. Coruzzi,

I would like to express my interest in collaborating as a consultant to develop a Cross-species Network Inference Platform that is biologist-user friendly.

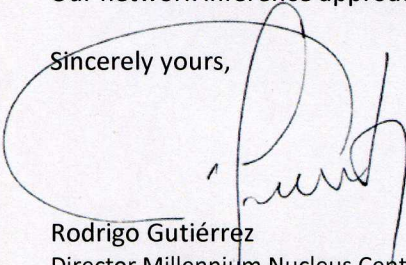
My experience in assembling multi-networks for Arabidopsis, Rice and Grape will be beneficial as the methods we have developed are species independent and could be applied to develop multi-network models for new crop species.

I understand that my primary responsibility will be to help Dr. Katari assemble networks for target crop species. I will contribute by identifying the necessary data sources and by adapting and optimizing the algorithms for each species as needed to develop the new multi-network models.

My fee of \$20,000 a year includes 3 months of work at a daily rate of \$250 as well as travel funds to attend the progress meetings twice a year. The consultant fee will cover my time as well as the time of software developers and bioinformaticians in my team to carry out the tasks outlined above.

The proposed work is valuable because many crop genomes lack a large body of experimental data. Our network inference approach will be useful for many species of economic value.

Sincerely yours,

A handwritten signature in black ink, appearing to read 'Rodrigo', written over a large, light-colored oval scribble.

Rodrigo Gutiérrez  
Director Millennium Nucleus Center for Plant Functional Genomics  
Assistant Professor  
Department of Molecular Genetics & Microbiology  
P. Universidad Católica de Chile.  
Alameda 340. Santiago. 8331010. Chile  
P: (56-2) 686-2663 | F: (56-2) 354-2185  
[rgutierrez@uc.cl](mailto:rgutierrez@uc.cl)  
<http://virtualplant.bio.puc.cl>



January 21, 2010

Dr. Gloria Coruzzi  
New York University  
Center for Genomics and Systems Biology  
Department of Biology  
1009 Silver Center  
100 Washington Square East  
New York, NY 10003-6688

Dear Gloria,

We are delighted to support your proposal "Network & Functional Inference: From Models to Crops" to develop tools and a pipeline to enable cross-species network inference for crop species based on reference species including *Arabidopsis*.

As part of the iPlant Genotype-to-Phenotype (iPG2P) Grand Challenge, the iPlant Collaborative is currently building a system in which individual existing tools will be brought in to analyze and visualize diverse sets of omics data in an integrated format, enabling the plant biologist to generate novel hypotheses and obtain new insights into regulatory pathways. Your team's user-friendly, cross-species network inference platform will mesh and integrate well with the analysis pipeline and tools that iPlant is now developing. iPlant's system is also intended to be user-friendly as well, such that the user may employ the analytical tools of his/her own choosing. Additionally, our system will be flexible enough to incorporate and integrate better and newer tools, such as the Cross-Species Network Inference Platform, as they appear.

Your goal to design tools that are modular and independent so that they can be accessed in multiple ways, such as via Galaxy, Taverna, or via iPlant's system, will complement iPlant's resources and help us achieve our mission to create a cyberinfrastructure collaborative for the plant sciences. An additional benefit to iPlant will be the availability of the VirtualPlant modular tools you have already developed for *Arabidopsis*, such as Sungear and Biomaps.

Additionally, your application of the cross-species network inference framework to address N-use efficiency in crop species will be of particular value to users of iPlant's NextGen Sequencing pipeline. The pipeline will feed directly into our visualization system, and will be ideal for processing the N-use efficiency data on *Medicago* that you anticipate generating in the course of your project. We are excited at the synergy that will benefit the plant biology research community from the tools and platforms being developed by our respective teams.

Sincerely,

A handwritten signature in blue ink, appearing to read "Stephen A. Goff", is written over a light blue horizontal line.

Stephen A. Goff  
Project Director  
iPlant Collaborative

[sgoff@iplantcollaborative.org](mailto:sgoff@iplantcollaborative.org)  
520-626-4224 (Office)  
520-301-6719 (Mobile)



Cold Spring Harbor Laboratory

P.O. Box 100, 1 Bungtown Road  
Cold Spring Harbor, New York 11724

Dr. Gloria Coruzzi  
New York University  
Center for Genomics & Systems Biology  
100 Washington Square East  
New York, NY 10003

Jan. 20, 2010

Dear Gloria:

I would be delighted to collaborate with you on your upcoming Plant Genome project on Cross Species Network Inference: From Models to Crops. We have had a number of successful collaborations over the years and very much look forward to this one.

As you have seen from our pilot study, where we compared gene expression using the same RNA samples for Affymetrix ATH1 hybridization and Illumina sequencing, using the Illumina sequencers gives a unique view into the RNA World. Many crop species lack a large body of experimental data necessary to build a molecular interaction network, which is essentially to do a systems level analysis. Your comparative network inference approach appears to be a promising method to help build such networks for the crop species where there is little interaction data. Time-series nitrogen treatment experiments on species with sequenced genomes such as Rice, and newly sequenced genomes such as Medicago, an N-fixing legume, will be useful to identify nitrogen regulatory network edges for both species. This regulatory network will help address an economically important trait: N-use efficiency in crop species.

We will certainly be happy to sequence the samples you generate in this Plant Genome project in collaboration with you. I wish you the best of luck with the proposal. Thank you.

Sincerely,

W. Richard McCombie  
Professor