MINING THE EVOLUTIONARY DYNAMICS OF PROTEIN LOOP STRUCTURE AND ITS ROLE IN BIOLOGICAL FUNCTIONS

PI:

Dr. Gustavo Caetano-Anollés
Professor of Bioinformatics (Crop Science / IGB)
University of Illinois at Urbana-Champaign

ILLINOIS
UNIVERSITY OF ILLINOIS AT URBANA-CHAMPAIGN

Presented By:
Fizza Mughal
Graduate Student (Informatics)
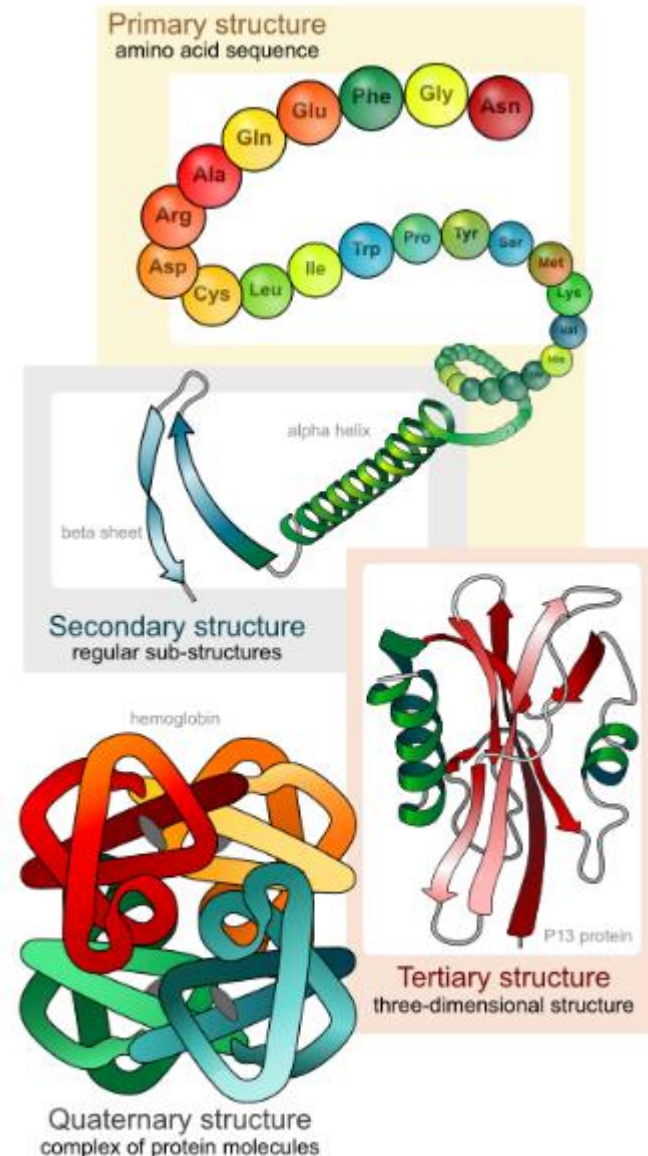University of Illinois at Urbana-Champaign

Objectives

- Flexible-unstructured regions of proteins introduce fundamental heterogeneity for molecular function

- Exploring dynamics of loops to ascertain their role in protein function

- Identify protein motions exclusive to specific functions

- Examine biophysical properties (flexibility and fluctuations) in the light of evolution
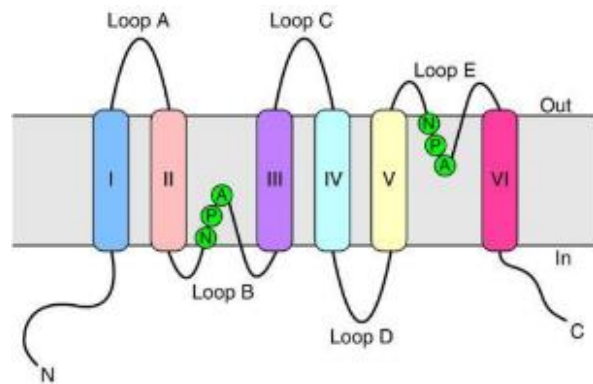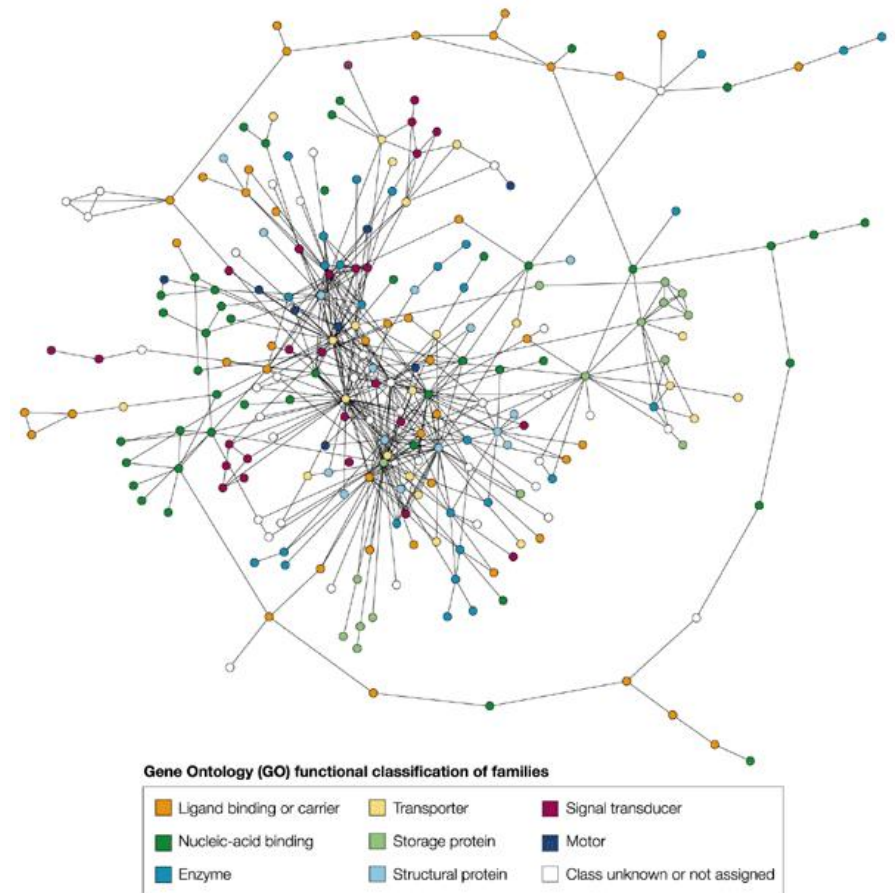
Protein Structure

- Levels
  - Primary
  - Secondary
  - Tertiary
  - Quaternary
- Domains: folded stable units
- Structural Classification Of Proteins (SCOP)
  - Fold Families: recent common ancestry
  - Fold Super Families: distant common ancestor
  - Folds: common structural topology

Primary structure
amino acid sequence

Secondary structure
regular sub-structures

Tertiary structure
three-dimensional structure

Quaternary structure
complex of protein molecules

Protein Molecular Function

- Gene Ontology (GO):
  - Cellular Component
    - intracellular or extracellular
  - Molecular function
    - Binding or catalysis
  - Biological Process
   - Operations critical to functioning of living units

Gene Ontology (GO) functional classification of families

■ Ligand binding or carrier	■ Transporter	■ Signal transducer
■ Nucleic-acid binding	■ Storage protein	■ Motor
■ Enzyme	■ Structural protein	□ Class unknown or not assigned

Source:
Ouzounis, *et al.*, 2003 *Nature Reviews Genetics*, *4*(7), 508-519.

Protein Evolution

- Assumption:
  - Most abundant = most ancient
- Phylogenomic reconstruction
  - Characters
  - Taxa
1. FF Assignment
2. Genomic Abundance calculation
3. Character states defined (N= most ancient; 0= most recent) and polarized
4. Tree construction using PAUP* (maximum parsimony)
5. Age (node distance, *nd*) calculated (0=most ancient; 1=most recent)
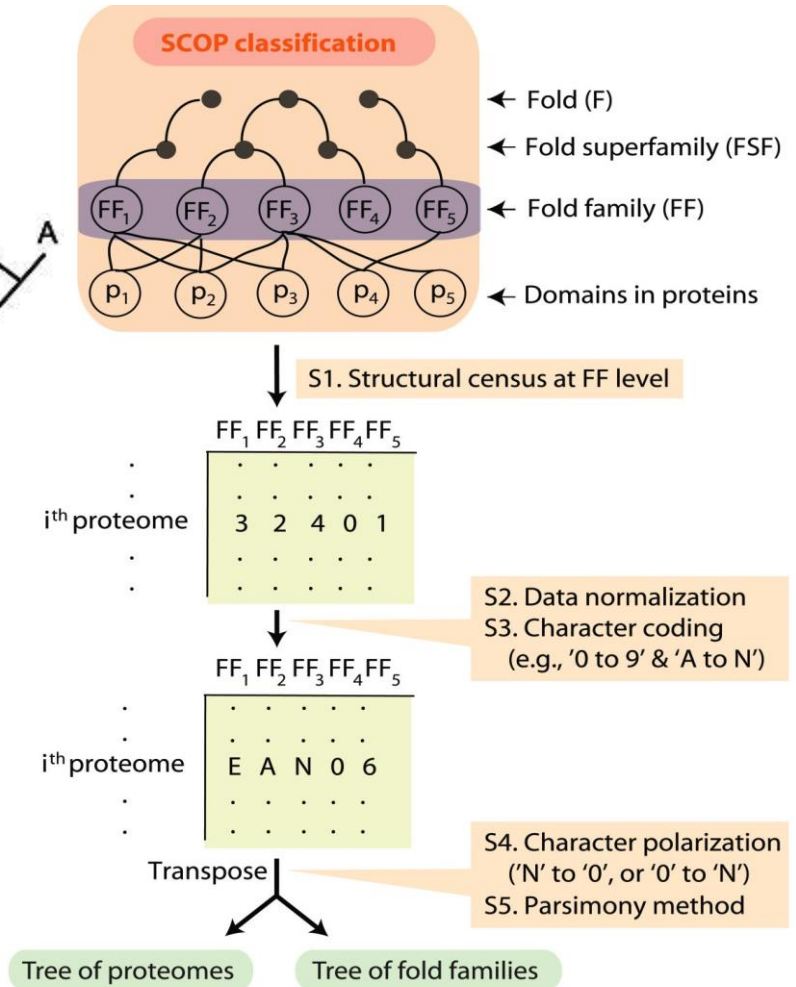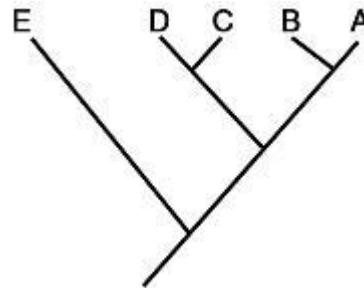
Source:
Kim & Caetano-Anollés, 2012. *BMC evolutionary biology*, *12*(1), 13.

Approach

- Dataset
  - Aminoacyl-tRNA synthetase (aaRS) domain FFs
  - ArchDB loop classification
  - Annotation with nd values
  - 87 Classifications
    - Density Search (DS)
    - Lowest p-value
    - Loop length >2 AA
    - Sec struc length ≥ 8 AA
    - Overall length < ~40 AA
- MD Simulations (NPT)
  - NAMD 2.9
  - CHARMM36
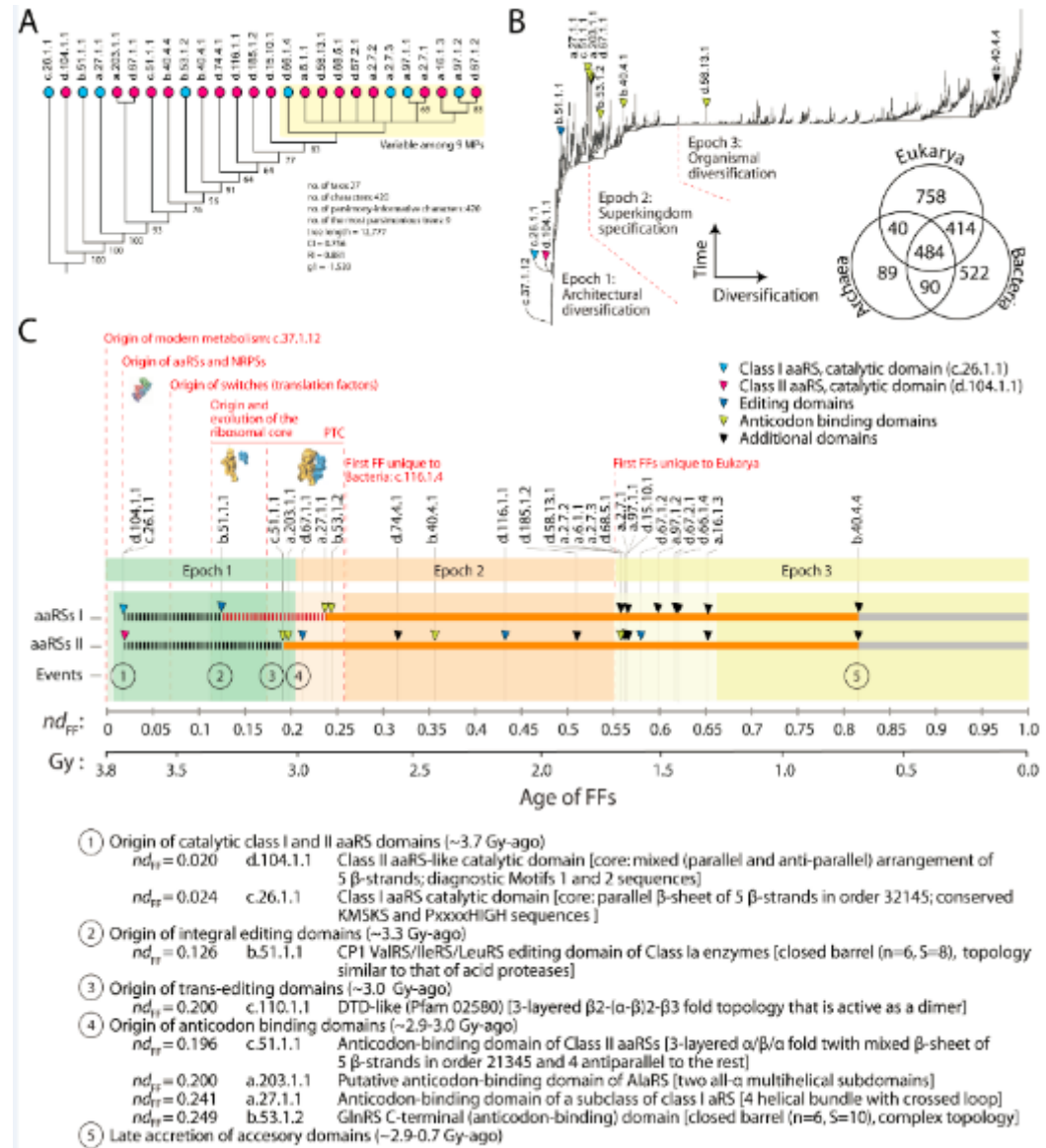
Source:
Caetano-Anollés, *et al*.,2013. *PLoS One*, *8*(8), e72225..

Why Blue Waters?

- Computing capability

- Storage of temporary files

- Impact: International Collaboration

- Key Challenge: Output File Storage

The Journey So Far …

- 73 Simulations performed
- Associated molecular functions
- Example: b.40.4.4 (MyF domain)
- Global Parameters:
  - RMSD
  - Radius of gyration

GO Term	Molecular Function	Occurrence
GO:0002161	aatRNA binding activity	4
GO:0003677	DNA binding	4
GO:0004825	metionine tRNA-ligase	4
GO:0042802	Identical protein bindir	4
GO:0000287	Mg ion binding	5
GO:0008270	Zinc ion binding	7
GO:0000049	tRNA Binding	9
GO:0046872	Metal ion binding	17
GO:0005524	ATP binding	25

Classification	FF	Loop ID	Loop Length	GO Term	Molecular Function
DS.BN.3.13.1	b.40.4.4	1JMZ_A_182	3	GO:0020037	Heme binding
				GO:0046872	Metal ion binding
				GO:0009055	Electron carrier activity
DS.BN.4.2.13	b.40.4.4	1T77_A_2080	4	None	None
DS.BN.5.2.2	b.40.4.4	1FJR_A_36	5	GO:0004930	G-coupled receptor protein activity
DS.BN.6.69.1	b.40.4.4	4MLL_B_208	6	GO:0008800	Beta-lactamase activity
				GO:0008658	Penicillin binding

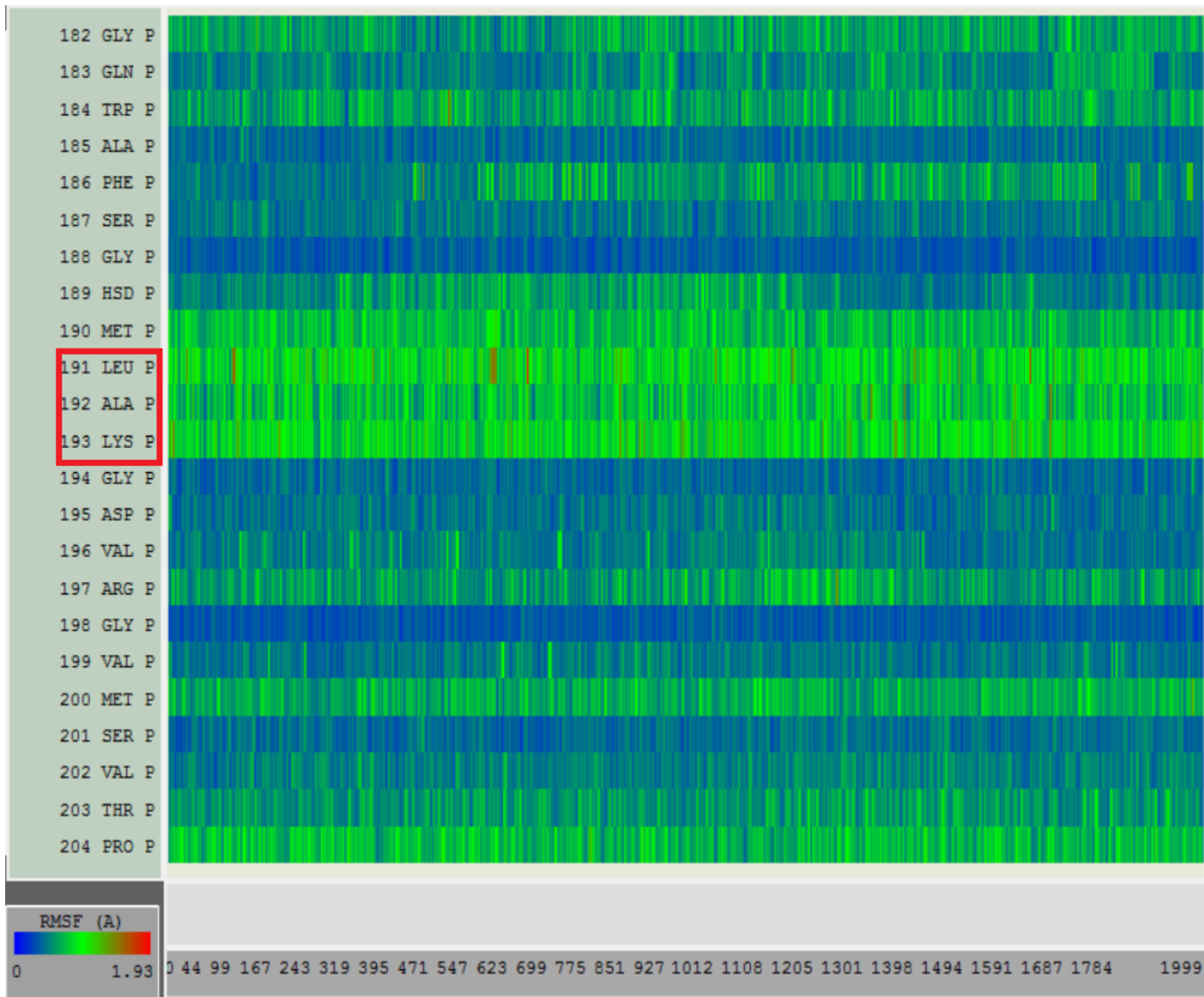**Conformational diversity (RMSD) vs. evolutionary age (nd)**
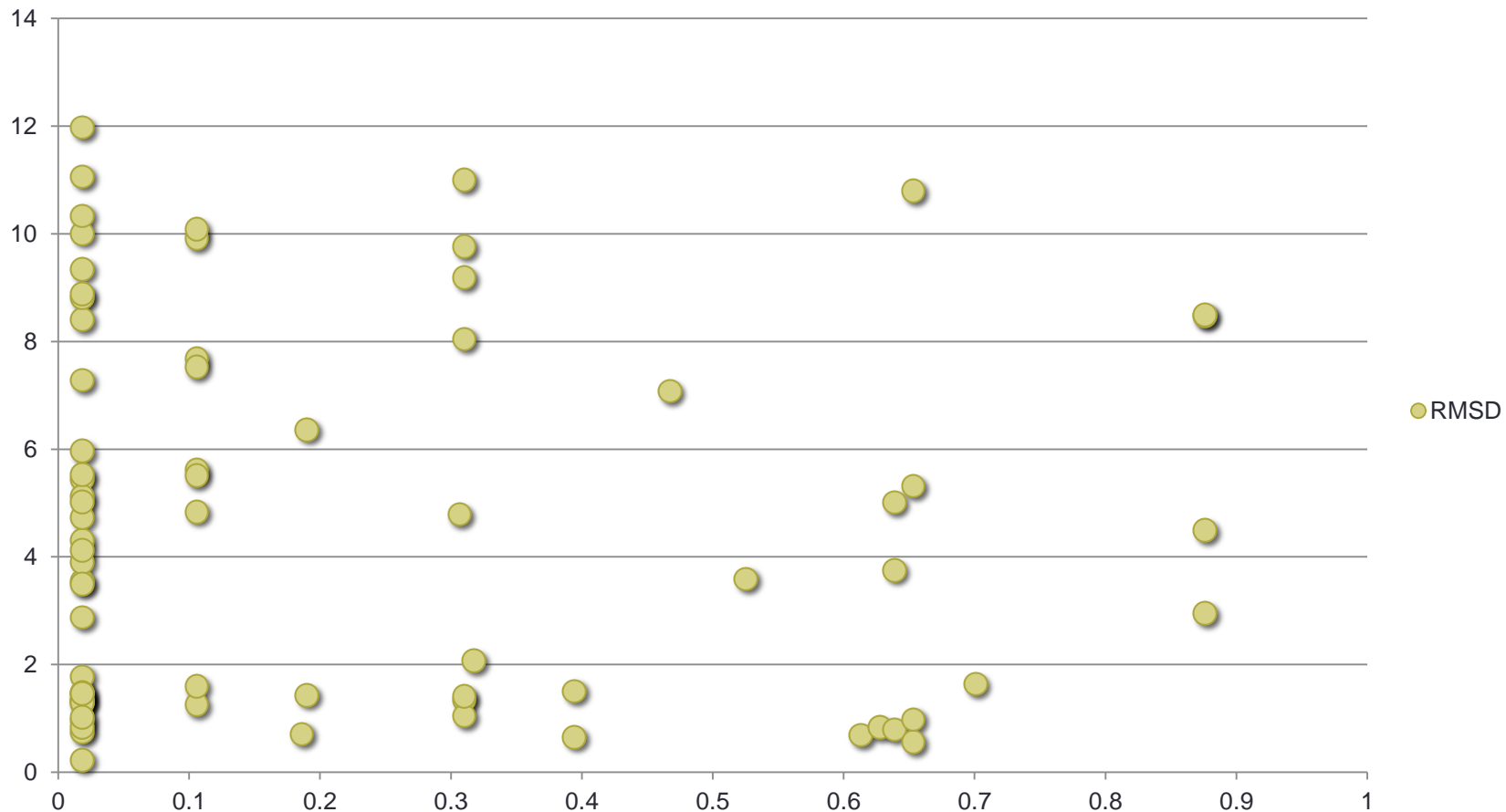
Radius of Gyration vs. age (nd)
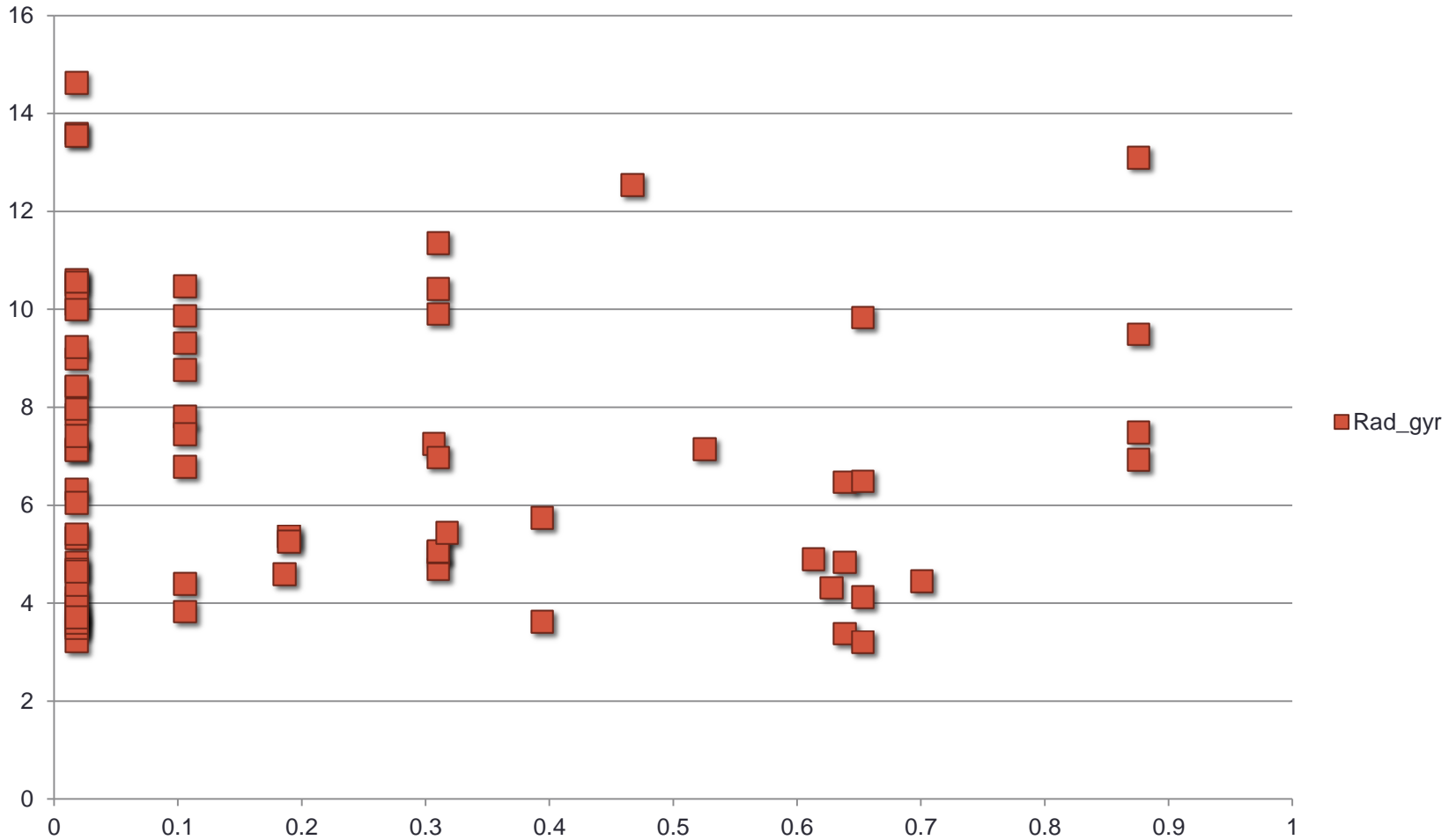
1JMZ

- 1JMZ

Conclusion & Future Directions

- Identification of fundamental principles of molecular evolution is achieved by reconstructing past events

- Advances in synthetic biology and translational medicine

- Methods to predict future "evolutionary trajectories"
  - predict evolvability of viruses
  - treatment of viral diseases with interfering agents (Wilke, 2012 *PLoS computational biology*)

- Map motions specific to classification/function based on molecular dynamic simulations

- Energy analysis

- Expand the data set!

Acknowledgements

- NCSA Blue Waters
- Illinois Research Board Grant
- Dr. Frauke Gräter (Germany)
- Evolutionary Bioinformatics Lab Members

THANK YOU!

Questions/Comments/Suggestions