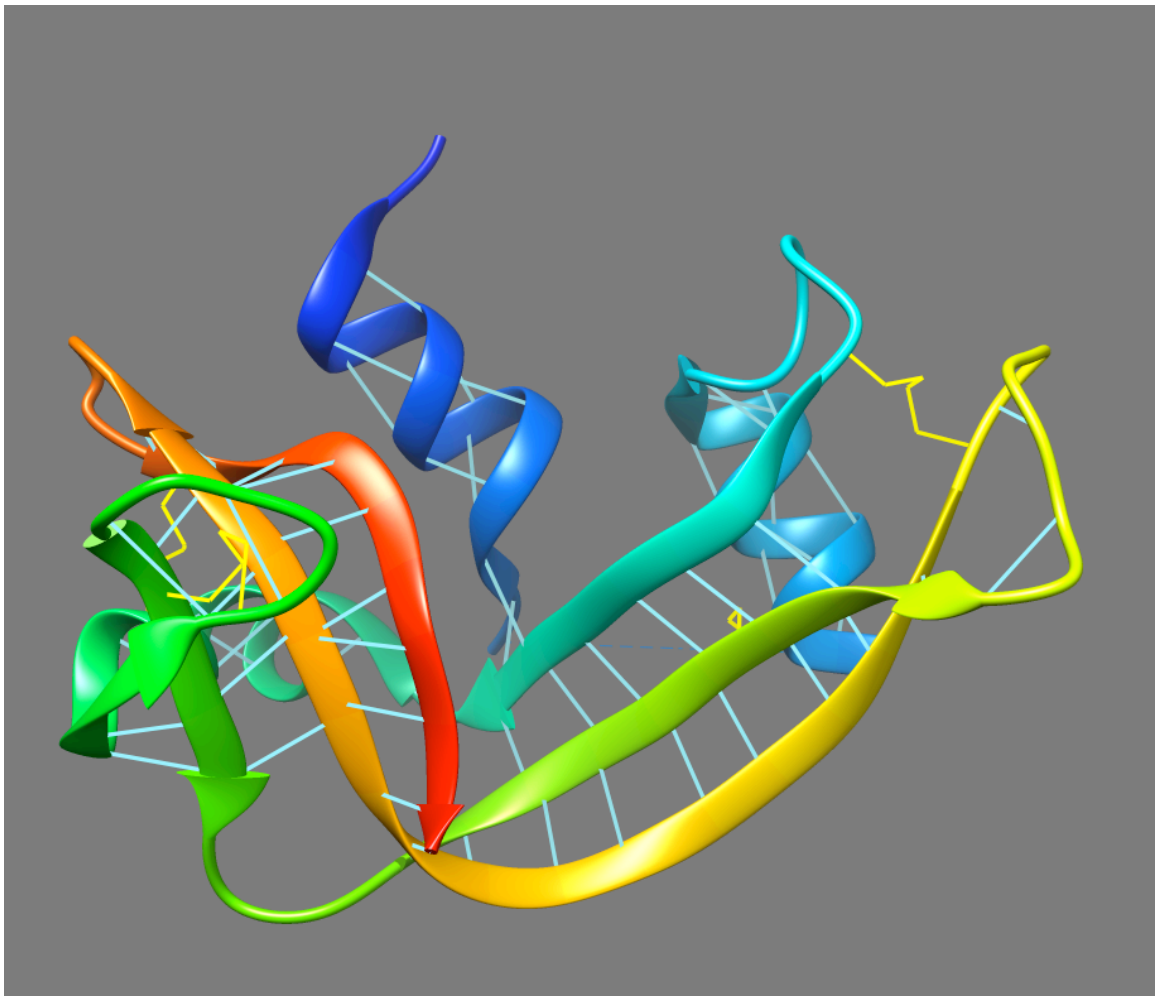
40+ Years of Protein Structure Analysis



Masur Auditorium, Building 10
National Institutes of Health
Bethesda, Maryland

November 7th, 2014

Overview

In the late 1960s, the three-dimensional structure was known for only five proteins and atomic coordinates were available for only three. The accessible surface area calculations, introduced at this time, were among the first computational analyses on protein structures. Today, there are more than 100,000 structures in the Protein Data Bank and computational analysis of protein structures is flourishing under the name of structural bioinformatics or computational structural biology. In this mini-symposium, ten world-renowned experts will speak on their current research activity in this and related areas.

Agenda

09:00-09:50 Registration

09:50-10:00 **Welcome and Opening remarks: Ira Pastan**

Chair: Stephen Bryant

10:00-10:30 **Shoshana Wodak**, *Scientific Director, Centre for Computational Biology, The Hospital for Sick Children, Toronto, Canada*
Protein Interactions: Short Stories from Then and Now

10:30-11:00 **George Rose**, *Professor, Department of Biophysics, Johns Hopkins University, Baltimore, MD*
Seeing is Deceiving

11:00-11:30 Break

11:30-12:10 **John Moult**, *Professor, Institute for Bioscience and Biotechnology Research, University of Maryland, Rockville, MD*
10 CASP Experiments: Successes, Bottlenecks, and Challenges in Protein Structure Modeling

12:10-12:40 **Robert Jernigan**, *Director, Baker Center for Bioinformatics and Biological Statistics, Iowa State University, Ames, IA*
Extracting Dynamics Information from Multiple Experimental Molecular Structures

12:40-2:00 Lunch

Chair: Ruth Nussinov

2:00-2:30 **Giuseppe Graziano**, *Associate Professor, Department of Science and Technology, University of Sannio – Benevento, Benevento, Italy*
On the Conformational Stability of Globular Proteins

2:30-3:00 **Jean Garnier**, *Editor-in-Chief, Biophysical Reviews, IUPAB, Paris, France*
Protein Secondary Structure, a Recurrent Object

3:00-3:40 **David Baker**, *Professor, Institute for Protein Design, University of Washington*
Design of Novel Protein Structures, Functions and Assemblies

3:40-4:00 Break

Chair: Susan Gottesman

4:00-4:30 **Phil Bourne**, *Associate Director for Data Science, National Institutes of Health, Bethesda, MD*
Protein Structure in the Era of Big Data

4:30-5:00 **Jongsun Jung**, CEO/CTO, *Syntekabio, Inc., Seoul, Korea*
PHR (Personal health records) & genomics in Korea

5:00-5:30 **Ira Pastan**, *Co-Chief, Laboratory of Molecular Biology, National Cancer Institute, Bethesda, MD*
Making Immunotoxins with BK Lee

6:00-6:30 Reception

6:30-9:30 Dinner

Abstracts

Shoshana Wodak[1,2,3]
[1] *Scientific Director, Centre for Computational Biology*
The Hospital for Sick Children
555 University Avenue
Toronto, Canada
[2] *Professor, Biochemistry Department, University of Toronto, Toronto, Canada*
[3] *VIB-VUB Structural Biology Research Center, Brussels, Belgium*

E-mail: shoshana@sickkids.ca

Protein interactions: short stories from then and now

I'll review some of the very early protein-ligand and protein-protein docking methods and their application to the elucidation of the quaternary structure change in hemoglobin, and how they gave rise to thriving areas of research. This will be followed by the description of a recent computational analysis of the mechanism and free energy profile of three-dimensional domain swapping reactions. In this analysis we propose that a large category of domain swapped systems form by a process that does not involve chain unfolding but proceeds by gradual exchange between intra- and intermolecular interactions, describing a minimum energy path.

George Rose
Professor, Biophysics
Johns Hopkins University
Jenkins Department of Biophysics
3400 North Charles Street
Baltimore, MD 21218

E-mail: grose@jhu.edu

Seeing is deceiving

The preceding half century of structural biology has transformed molecular biology (and journal covers). Few things are as compelling as an experimentally-derived macromolecular model in atomic detail. However, consciously or unconsciously, tangible models condition our thinking, resulting in implicit assumptions about the underlying mechanism(s) of self-assembly.

The protein folding mechanism has remained elusive because at equilibrium, small globular proteins of biophysical interest tend to partition into either unfolded or native species, with only a negligible population of intermediate species (the two-state approximation).

Consequently, analyzing proteins at equilibrium is like observing the first and last scenes of a Shakespeare play, then trying to infer the action in between. It is not hard to understand why the field remains controversial.

From molecules to skyscrapers, persisting structure is ultimately a consequence of unseen stabilizing energetics. What is a protein domain? Are there non-hydrogen-bonded backbone polar groups in the protein core? Is the peptide plane actually as rigid as Pauling thought? These are fundamentally thermodynamic questions, and attempts to answer them by analyzing X-ray structures of folded proteins is guaranteed to be misleading. Illustrative examples will be presented.

John Moult
Professor, Institute for Bioscience and Biotechnology Research
University of Maryland
Shady Grove Road
Rockville, MD

Email: moult@umbi.umd.edu

Extracting Dynamics Information from Multiple Experimental Molecular Structures

CASP (Critical Assessment of Structure Prediction) is an organization that conducts community wide experiments to measure the state of the art in modeling protein structure from amino acid sequence. The core principle of CASP is to objectively evaluate methods through bona fide blind testing of the ability to predict experimentally determined structures, and to do that through the participation of members of the modeling community. The ten CASP experiments so far span almost 20 years of the field of protein structure modeling, and there have been enormous advances in the capabilities of the methods in that period. These changes have been driven by three primary factors – improved algorithms for structure modeling, improved software engineering to implement algorithms and make methods accessible, and greatly expanded databases of protein sequences and experimental structures. As a consequence, structure modeling has moved from a rather academic pursuit of little practical relevance to a technique that is utilized by many biologists. In this talk I will review the progress of the field as seen through CASP, examining the principal innovations and the impact each had, the current accuracy and scope of the methods, and the barriers still to be overcome.

Robert L. Jernigan
Director, Baker Center for Bioinformatics and Biological Statistics
Dept. of Biochemistry, Biophysics & Molecular Biology
Iowa State University
Ames, IA 50011-3020

Email: jernigan@iastate.edu

Extracting Dynamics Information from Multiple Experimental Molecular Structures

Meaningful dynamics information can be extracted from multiple coarse-grained experimental structures of the same, or closely related, proteins or RNAs. The covariance matrix of atom positions is decomposable into its principal components, and these serve to identify the most important motions. Normally only a few principal components dominate the motions of the structures, and these usually relate to the functional dynamics. This dynamics information provides strong evidence for the plasticity of protein and RNA structures, and also suggests that these structures almost always have a highly limited repertoire of motions. Coarse-graining significantly aids in identifying these motions. In some cases, such as many enzymes the dominant motions are opening and closing over the active site. For myoglobin the changes are much smaller, reflecting in part the small changes in the loop conformations, but nonetheless they show characteristic details of the motions that depend in part upon the species. These approaches also can be used to compute the global entropies. We are using these to compute pathways for transitions on free energy landscapes between the different conformations by combining these entropies with 4-amino acid cooperative potentials.

Giuseppe Graziano
Associate Professor of Physical Chemistry
Department of Science and Technology
University of Sannio - Benevento
Via Port'Arsa 11 - 82100 Benevento, Italy

Email: graziano@unisannio.it

On the conformational stability of globular proteins

A polypeptide chain can adopt very different conformations, a fundamental distinguishing feature of which is the water accessible surface area, WASA, that is a measure of the layer around the polypeptide chain where the center of water molecules cannot physically enter, generating a solvent-excluded volume effect. The large WASA decrease associated with the folding of a globular protein leads to a large decrease in the solvent-excluded volume, and so to a large increase in configurational/translational freedom of water molecules. The latter is a quantity that depends upon temperature. Simple SPT calculations over the -30 – 150 °C temperature range, where liquid water can exist at 1 atm, show that such a gain decreases significantly on lowering the temperature below 0 °C, paralleling the decrease in liquid water density. There will be a temperature where the destabilizing contribution of the polypeptide chain conformational entropy exactly matches the stabilizing contribution of the water configurational/translational entropy, leading to cold denaturation. The same approach is able to show that the denaturing action of urea and GdmCl is due to their direct energetic interactions with the polypeptide chain.

References

G. Graziano, *Phys.Chem.Chem.Phys.*, 2010, **12**, 14245.
G. Graziano, *Phys.Chem.Chem.Phys.*, 2011, **13**, 12008.
G. Graziano, *Phys.Chem.Chem.Phys.*, 2011, **13**, 17689.
G. Graziano, *Phys.Chem.Chem.Phys.*, 2012, **14**, 13088.
G. Graziano, *Phys.Chem.Chem.Phys.*, 2014, **16**, 21755.

Jean Garnier
Editor-in-Chief, Biophysical Reviews, IUPAB
International Council for Science (ICSU)
5 rue Auguste-Vacquerie, 75116 PARIS, France

Email: jgarnier@icsu.org

Protein secondary structure, a recurrent object

In average about half of a globular protein structure is made of secondary structure, alpha helices and beta strands, where the other half is made of isolated residues in these two conformations (aperiodic or coil structure) or residues behaving like a random polymer floating in the solvent (disordered structure). The secondary structure of a protein is like its backbone which has been recently found more recurrent than expected under the form of locally similar structural pieces (LSSPs or 3D substructures) made of a small sequential number of helices and strands from the N terminal down to the C terminal amino acid sequence then following the DNA coding sequence order (Tai et al., Proteins 2011; 79:853-866). We devoted a web site to use this recurrence to assign domains in proteins and search for structural neighbors in the PDB: http://genome.jouy.inra.fr/domire/. I will comment about some aspects of these recurrences.

David Baker
Professor
Institute for Protein Design
University of Washington
Molecular Engineering and Sciences
Box 351655
Seattle, WA 98195-1655

Email: dabaker@u.washington.edu

Design of novel protein structures, functions and assemblies

I will describe the design of proteins with new structures and functions. We compute amino acid sequences predicted to fold into proteins with new structures and functions, produce synthetic genes encoding these sequences, and characterize them experimentally. I will describe the design of ultra-stable idealized proteins, flu neutralizing proteins, high affinity ligand binding proteins, and self assembling protein nanomaterials. I will also describe the contributions of the general public to these efforts through the distributed computing project Rosetta@home and the online protein folding and design game FoldIt.

Philip Bourne
Associate Director for Data Science
National Institutes of Health
Bethesda, MD 20892

Email: philip.bourne@nih.gov

Protein Structure in the Era of Big Data

The PDB, which I was proud to be associated with for many years, and which is a foundation upon which this community functions, has been described as exemplary by many stakeholders – funders, users, depositors, but does a resource like this make the most sense going forward? Such a question is not about the PDB per se, but as a reference point in a scientific data ecosystem that is in a state of rapid change. A change in which issues of integrating multiple data types to make discoveries, sustainability of data resources, reward for producing quality data, and many more aspects come into play. I will use the PDB as a reference point to begin a dialog about the future.

Jongsun Jung
CEO/CTO, Syntekabio, Inc.
Data Coordination Center in PGM21
992 Venture Town,
Korea Institute of Science and Technology,
Seoul, Korea

E-mail: jung@syntekabio.com

PHR (Personal Health Records) & genomics in Korea

Owing to the recent and widespread deployment of NGS sequencers at most R&D organizations, NGS data analytics has become increasingly important as a technique to accommodate the stunning increase in the amount of whole genome and exome data generated by these sequencers. In particular, the integration of the data from 10,000 to 100,000 whole genome & exome sequences should help enormously in the integration of PHR (personal health records) with genomics, called PHRG, by furnishing NGS-generated data to professionals in clinical institutions and life science R & D organizations. PHRG can be divided into five sub-topics; [1] high throughput simulation of protein-drug-mutation complexes, [2] direct comparison of one vs. 1,000 genomes, [3] ethnic-based human_haplo ID, [4] annotation of ethnic-based known disease variants, and [5] IT-infrastructure. The clinical phenotype data of PHR can be fed directly from EHR or EMR. This new project is similar to the eMERGE project at NIH in terms of setting, but is not the same in terms of IT-driven NGS technology where eMERGE is focused on complex diseases detected by GWAS. PHRG has broader coverage than eMERGE, and addresses rare diseases and cancer, as well as pharmacogenomics. Currently, PHRG technologies have been largely tested and vetted, and some are at the assembly-stage. Ideally, in the near future, it should be possible to access centralized PHRG databases on cloud platforms, which would considerably facilitate the exchange of information between health specialists and patients during clinical encounters by providing the health specialists with an up-to-date, rapid access to patient medical history, including drug interactions, the results of previous laboratory tests and diagnostic studies, and the genetics of the person's genome. The availability of such platforms would also provide an information hub for patient health management.

In the talk, the core science & technology of the five sub-topics outlined above as well as the technology required for their fusion will be discussed.

Ira Pastan
Co-Chief, Laboratory of Molecular Biology
National Cancer Institute
Bethesda, MD 20892

E-mail: pastani@mail.nih.gov

Making Immunotoxins with BK Lee

Recombinant Immunotoxins (RITs) are chimeric proteins designed to kill cancer cells. They originally consisted of a single chain Fv connected to a portion of a potent bacterial toxin. The Fv directs the toxin to the cancer cell, where it enters and kills the cell. First generation RITs had several difficulties. The single chain Fv was often unstable, leading to aggregation and inactivation of the protein, and the toxin was immunogenic in humans and mice. BK Lee designed, and we produced and tested RITs containing Fvs that were connected by a disulfide bond that made them very stable and eliminated the aggregation problem. Next, using the **Lee-Richards** algorithm, he identified residues likely to make up B cell epitopes that could be changed to alanine without affecting the activity of the RIT. My group introduced these mutations into the RIT and tested their properties. Using this approach we have produced a new generation of RITs that are very stable and much less immunogenic. Clinical trials with one of these agents are scheduled to begin early in 2015.