# Laboratory Directed Research and Development

**Modeling Conformational Ensembles of Proteins and Complexes**

<table>
<thead>
<tr>
<th><strong>Lead Scientist:</strong></th>
<th>Henry van den Bedem</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Phone:</strong></td>
<td>926-2314</td>
</tr>
<tr>
<td><strong>Email:</strong></td>
<td><a href="mailto:vdbedem@slac.stanford.edu">vdbedem@slac.stanford.edu</a></td>
</tr>
<tr>
<td><strong>Date:</strong></td>
<td>8/26/2011</td>
</tr>
<tr>
<td><strong>Department/Division/ Directorate:</strong></td>
<td>Structural Genomics/SSRL</td>
</tr>
</tbody>
</table>
| **Other Scientists:** | Ashley Deacon (Structural Genomics/SSRL)  
Jean-Claude Latombe (CS, Stanford University)  
Hiro Tsuruta (SMB/SSRL) |
| **Proposal Term**   | From: 10/2011  
Through: 09/2014 |

<table>
<thead>
<tr>
<th><strong>Business Manager:</strong></th>
<th>Stephanie Carlson</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Phone:</strong></td>
<td>650-926-2033</td>
</tr>
<tr>
<td><strong>Email:</strong></td>
<td><a href="mailto:steph@slac.stanford.edu">steph@slac.stanford.edu</a></td>
</tr>
</tbody>
</table>
This document and the material and data contained herein were developed under the sponsorship of the United States Government. Neither the United States nor the Department of Energy, nor the Leland Stanford Junior University, nor their employees, makes any warranty, express or implied, or assumes any liability or responsibility for accuracy, completeness or usefulness of any information, apparatus, product or process disclosed, or represents that its use will not infringe privately owned rights. Mention of any product, its manufacturer, or suppliers shall not, nor it is intended to imply approval, disapproval, or fitness for any particular use. A royalty-free, non-exclusive right to use and disseminate same for any purpose whatsoever, is expressly reserved to the United States and the University.
Abstract

We propose to develop a computational method to build challenging, high impact protein structures into medium to low resolution crystallography, SAXS and nano-crystallography data, and provide insight into conformational dynamics. The program will provide structural biology researchers with a tool of major scientific potential while enhancing the lab's mission readiness in data analysis and scientific computing.

Summary of Proposal

Description of Project

Macromolecules often exhibit significant conformational flexibility, especially when they interact with ligands or protein partners to form assemblies. Probing these conformational changes is essential to understanding their cellular function, yet these scientifically important molecules are often difficult to crystallize and diffract poorly. Such assemblies are studied by a variety of experimental techniques, including protein crystallography (PX) and Small Angle X-ray Scattering (SAXS). Nano-crystallography (nX) and single particle imaging data from LCLS are also being targeted at these systems. A ‘divide and conquer’ approach is typically used for structure determination, where individual components (domains or proteins) are determined at high resolution and then the entire complex is studied at lower resolution. Time-resolved SAXS can be further used to probe the dynamics. However, automatically fitting large molecules or assemblies to experimental data, allowing for their flexibility remains a challenge.

We propose a computational research program to model challenging systems with experimental data from multiple sources. Our approach is to extend the innovative, robotics-inspired Kino-Geometric Sampler conformational search algorithm (KGS, [4]) to fit previously determined domains or proteins to experimental data. KGS exhibits a singularly large radius of convergence and optimally reduces the number of free parameters. These unique features enable flexible 'docking' of atomic models in the data while moderating the risk of overfitting at low resolution. We aim to automatically compute 3-D models, and to provide structural insights for systems that diffract poorly or cannot be crystallized. Our method will also provide insight into functionally relevant dynamics, especially for time-resolved experiments or a series of static structures.

Expected Results

This program capitalizes on synergy between SLAC experimental capabilities. It will (1) Lead to algorithms enabling the study of macromolecular assemblies, an active research area at the frontier of current structural biology; (2) Automate low resolution model-building through atomic resolution models; (3) Capture functionally relevant dynamics through conformational substates and their fractional occupancies.
Proposal Narrative

Purpose/Goals

Structural biology is an important component of the laboratory lightsource mission. Each year, more than 1,000 users perform experiments in structural biology at SLAC, a number that is expected to grow rapidly as LCLS continues to develop new capabilities. The unique and state-of-the-art experimental facilities at SLAC provide researchers with many tools to study important macromolecular systems across size scales and resolutions, ranging from large protein complexes such as photosystem I reported in a proof-of-concept nX experiment at 8.5Å resolution to small proteins determined at sub-angstrom resolution. The Joint Center for Structural Genomics (JCSG) at SLAC aims to solve large protein and DNA complexes involved in regulating the fate of stem cells, but current algorithms are unable to automatically and optimally model (intra)-domain motion necessary to fit components of large assemblies across data sources. Indeed, the Nobel Prize (2009) winning initial characterization of the ribosome at 5.5Å for instance, required a laborious process of rigid-body docking of large fragments of a separately obtained 50S subunit structure and numerous other subunit proteins into PX data, followed by more detailed fitting [11].

For low resolution (below 3Å) PX/nX data sets conventional model building algorithms break down and only large secondary structure elements can be identified. Some progress was recently reported in modeling low resolution PX data for single proteins: MR_rosetta [3] rebuilds a homology model as part of a molecular replacement (MR) software suite, but has a limited radius of convergence. If an approximate model can be obtained, extreme care should be taken while fitting it to the experimental data at low resolution to avoid overfitting and introducing model-bias [10]. Existing algorithms for SAXS can only fit user-defined rigid domains [5, 6] to model complexes, or require user-defined constraints to flexibly fit domains by ‘trial and error’ [7]. Our purpose is to avoid these limitations and flexibly model challenging structures at low resolution using multiple data sources, and identify conformational substates and their occupancies.

Approach/Methods

KGS will be extended to fit previously determined atomic models to fit current data capitalizing on (1) KGS’s large radius of convergence, and (2) optimal reduction of the free parameter count owing to a kinematic representation of proteins. Conformational substates and their fractional occupancies will be computed from diffusive sampling.

Kino-geometric sampling of the folded state

Covalent and hydrogen bonds in a biomolecule define rigid groups that greatly reduce its modes of deformation. In a jointly NSF-funded project Professor Latombe’s group at Stanford recently developed KGS, a novel and powerful algorithm that models a folded protein by way of a kinematic linkage, with rigid groups of atoms as links and rotatable bonds as joints. The inclusion of hydrogen bonds in such a model results in a complex
network of interdependent cycles (Fig 1a,b), greatly reducing the free parameter count.

Figure 1 Left Panel. (a) A protein fragment with 3 H-bonds. (b) The kinematic model for this fragment. Each node is a rigid group of atoms, and each edge a remaining rotatable bond. Three interdependent cycles are colored in red, cyan and orange. (c) Cyanovirin-n’s two conformations (gray and blue) are distinguished by a large domain motion. Starting at the gray conformation KGS (magenta) quickly sampled more than 10Å away, close to the blue conformation. Right panel A. The crystal structure of β2GPI. B. Three manually constructed conformations using rotations between domains CCP2/CCP3 with fractional occupancies shown improve the fit to SAXS data (grey envelope) relative to a single conformation. (Image after [12])

KGS combines a novel strategy capable of sampling the folded state uniformly, with a null-space method which simultaneously deforms a large number of interdependent kinematic cycles without breaking covalent or favorable hydrogen bonds. It encodes dominant energy terms implicitly by kinematic and geometric constraints, and efficiently computes an ensemble that diffuses rapidly through the folded state. Indeed, cyanovirin-n, an HIV-inactivating protein, exists as both a monomer and a dimer in solution. The dimer conformation is distinguished by a large domain motion that KGS sampled a remarkable 10Å away from the monomer conformation (Fig 1(c)). Similar large motions have shown to be a key element in improving the data fit in solution (Fig 1, right panel). In contrast, the maximum radius of convergence reported by MR_rosetta does not exceed 3.4Å.

Placing components in medium to low resolution data

In some cases, a model will already have been placed in the data, but may be partially incomplete, incorrect or sub-optimally fit as a rigid model. In other (PX/nX) cases, MR may have failed due to a deviating template. KGS may then be used to propose new search models, or heavy atoms can be used to locate models. To fit SAXS profiles, a (parallelized) search of configurations in SE(3) will precede diffusive sampling. We will furthermore investigate the use of KGS search models in a MR technique in combination with a recently proposed phasing method for nX data [8].

Modeling single components at medium to low resolution

Initially we will develop KGS to model a single protein into low-resolution PX data using an atomic model of it in a different conformation (e.g. native vs. ligand-bound form), or a homologous or comparative model when the identical protein structure is unknown.
Test data is abundant within the JCSG, where ~20% of unsolved protein structures diffract between 3.0Å-5.0Å. For the majority, experimental phases are available but known homologues are not sufficiently accurate for successful MR. KGS will be adapted to generate conformational variability for a torsion angle Monte-Carlo Minimization (MCM) protocol where the gradient of the data is projected onto the null-space of the first-order approximation to the kinematic model, i.e. \( dq = NN^T \nabla \rho \), where \( q \) is the current conformation, \( N \) is a basis for the null space, and \( \rho \) is experimental data. This technique has proven extremely effective in our more limited kinematic optimization algorithm [12], but its performance with a vast number of kinematic cycles is unknown. Careful attention will be needed to control model bias by using cross-validation. Simulated data allows us to control key parameters and validate the algorithm.

**Low resolution assembly modeling with reduced dimensionality**

In years 2 and 3 our objective is to develop KGS to enable modeling of large assemblies, first with PX/nX and then with SAXS data. Comparative and atomic models will be combined and flexibly fit into medium to low-resolution data. A major difficulty is extending KGS's kinematic representation to include inter-protein h-bond networks. Furthermore, while kinematic cycles reduce the number of degrees of freedom and moderate the risk of overfitting, they may also prevent large domain movement. Thus, we will investigate when h-bonds are allowed to break and new ones form, requiring updating KGS's representation at each step.

**Identifying substates and functionally relevant dynamics**

Distinct experimental techniques or time-resolved experiments may capture models in different conformations (Fig 1). Conformational substates and their fractional occupancies in the crystal or solution can be obtained from a set of optimized diffusive samples using a constrained convex optimization algorithm to select a subset that collectively best explains the data, akin to our work in [13,16]. Furthermore, by flexibly linking these snapshots while ensuring a good data fit and preserving key hydrogen bonds, sampling points translate into 'clash-free' structural pathways, giving researchers access to transient intermediates. Endowing KGS with realistic long range energy potentials would open up possibilities to compute low-energy transitions.

**Specific Location of Work**

The work will take place in building 278.

**Anticipated Outcomes/Results**

We anticipate to greatly automate low-resolution model building for large, challenging structures, enabling the study of macromolecular assemblies, an increasingly important focus of leading structural biology research. Our research will provide structural insight at low resolution through atomic models or homologues, and through explicit modeling of conformational substates and functionally relevant dynamics.
References


Previous Work

The authors have previously developed and applied kinematics-based algorithms to protein model building [13,14]. Applications of their research have appeared in high profile journals [15-17].


VITA (Lead Scientist)

Henry van den Bedem

Joint Center for Structural Genomics / Stanford Synchrotron Radiation Lightsource
SLAC National Accelerator Laboratory
Ph: (650) 926-2314
Email: vbedem@slac.stanford.edu

Henry van den Bedem is a staff scientist with the Joint Center for Structural Genomics at the Stanford Synchrotron Radiation Lightsource. His research interests are in Computational Structural Biology, in particular methods and algorithms development for interpreting structure and dynamics of macromolecules from experimental data, and physics-based refinement of comparative protein models. He pioneered application of robotic motion-planning techniques to X-ray crystallography for model building and refinement.

Education

• PhD in Mathematics (1999), University of Alabama at Birmingham
• MSc in Mathematics (1995), Delft University of Technology, The Netherlands

Research and Professional Experience

• Staff Scientist (2007 - ) SLAC National Accelerator Laboratory
• Software Developer, Research (2001 - 2007) SLAC National Accelerator Laboratory

Grants

• Joint DMS/NIGMS initiative to support research in the area of Mathematical Biology (2005-2009) Co-Investigator

Selected Publications


Budget

Budget Request by Fiscal Year

To conduct the proposed studies, we expect the budget to cover the expense of

- One Principal Investigator for 3 years (1%)
- One post-doctoral fellow for 3 years (100%)
- Materials and supplies for workstation, laptop, software licenses, books $5.1K, $4.9K, $4.6K in Yrs 1, 2, and 3, respectively
- Travel to/from East Coast to/from SLAC 1 trip/year @ $2.7K each (air, lodging, and local per diem)
- Travel to/from Europe to/from SLAC 1 trip/year @ $2.6K each (air, lodging, and local per diem)
Approvals

Department/Division Administrator

[Signature]

SSRL Business Manager

Department Chair/Division Manager

[Signature]

Brett Holman

Associate Laboratory Director

[Signature]