RESEARCH STATEMENT

Macromolecular complexes form the machinery responsible for most biological processes and are relevant to understanding many diseases such as cancer and metabolic disorders. Knowledge of these structures would provide not only the mechanistic descriptions for how macromolecules act in an assembly but also clues in developing therapeutic interventions related to disease. Efforts in structural proteomics have lead to a rapid increase of the number three-dimensional (3-D) structures of individual proteins. Moreover, knowledge of networks of interactions and signaling pathways is also expanding rapidly through genomic and proteomic approaches. Yet, our picture of the structures of both stable and transient protein interactions lags behind. Efforts in crystallizing macromolecular complexes have met with some limited success, and hybrid experimental approaches, utilizing cryo-electron microscopy and crystallography or NMR to give structural details of complex assemblies are evolving and have led to a rapid increase of the number of three-dimensional (3-D) structures of Large Biomolecular Complexes (LBCs) (atomic level deposited in the Protein Data Bank (PDB) and volumetric electron density map data at the European Bioinformatics Institute (EBI)). However, along with these experimental methods, there is a growing need for efficient and robust computational approaches to predicting the structures of protein and their interactions. The goal of my research is to answer the following biological questions:

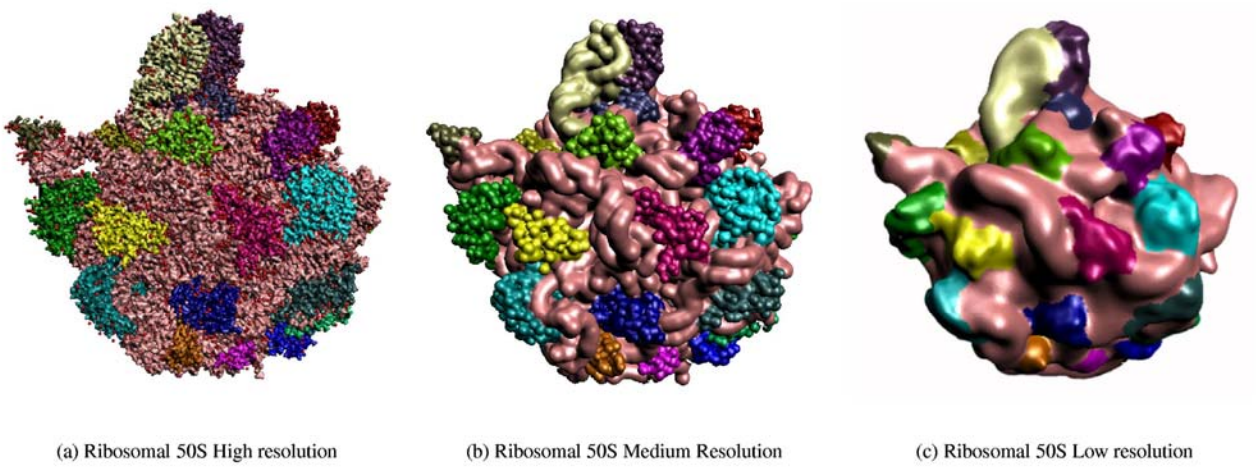
- How do we identify, represent, match and visualize **Large Biomolecular Complexes** (LBCs) structures efficiently and fast?
- How do these structures interact?

Supercomputers across the world are also now routinely deployed for computer simulations and generation of associated properties (APs) of LBCs, ranging from volumetric electrostatics potential, to accurate 3D force fields and interaction potentials, to atomic level molecular dynamics trajectories. These AP's are routinely employed to predict docking binding sites. However, due to the large complexity of these data sets, most docking codes are quite limited in the size of proteins they can dock or exhibit low accuracy. Moreover, crucial for analysis of molecular interactions is the visualization of LBCs. Indeed, a need for *interrogative* visualization is a necessary tool for validation and verification of computational results. While molecular visualization software has developed over the years, today, most tools still operate on individual molecular structures and small 3-D electron density maps with little facility to manipulate large complexes, integrate geometric and volumetric visual representations, or depict molecular flexibility. Few, if any, currently used programs allow for or enable interaction with BCs and their APs at the atomic level, or LBCs as reconstructed volumetric maps from tomographic and cryo imaging, that will become common in the next ten years.

My approach to answer these biological questions effectively involves the development of **Multi-Resolution and Radial Basis Functions (MR)** representation theory to optimize computational modeling, information processing and visualization techniques, in particular, for large-scale complex systems. More precisely, MR theory is leading to

- Feature preserving molecular representations of LBCs.
- Fast and stable extraction of suitable representations of LBCs for large scale visualization and information processing.
- Fast Biomolecular convolution codes for structure identification and docking.
- Simulation of biophysical phenomena.

. In addition to a strong mathematical computational approach to answer these biological questions appropriately, I believe a collaboration between computational biologists, computer scientists, applied mathematicians and experimental biologists is essential. To his end I form part of a network of collaborators, principally among these include



(a) Ribosomal 50S High resolution (b) Ribosomal 50S Medium Resolution (c) Ribosomal 50S Low resolution

Figure: Molecular representation at different resolution views of the Ribosomal 50S subunit molecular surface. The RNA is in pale pink and dull yellow, and the proteins are variously colored.

- Prof. **C. Bajaj** (Post. Doc Advisor, Professor in Computer Science and Head of the Center for Computational Visualization (CCV) at the University of Texas),
- Prof. **A. Olson** (Head of the Molecular Graphics Laboratory (MGL) at The Scripps Research Institute)
- Prof. **W. Chiu** (Head of the National Center for Macromolecular Imaging (NCMI)).

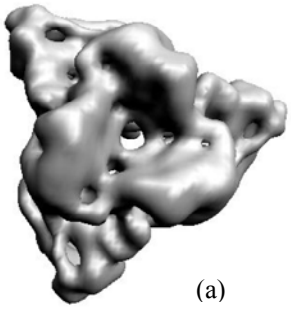
Other References include:

- Prof. **Kevin Amaratunga** (Ph.D. Advisor, Department of Civil and Environmental Engineering, MIT)
- Prof. **Gilbert Strang** (Department of Mathematics, MIT)
- Dr. **Pavel Bochev** (Sandia National Laboratories)

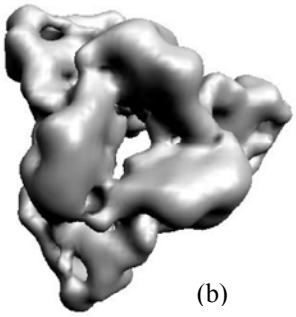
Moreover, I have also worked with experimental toxicologists (see [4]). More detailed information on my research can be found on my website at http://www.ices.utexas.edu/~julio/index.html. Note that current manuscripts will be updated at this website.

Radial Basis Functions Representations of LBCs

Due to the size of molecular density maps, most molecular processing software developed over the years still operate on individual molecular structures and small 3-D APs density maps with little facility to manipulate large complexes, integrate geometric and volumetric visual representations, or depict molecular flexibility. Few, if any, currently used programs allow for or enable interaction with LBCs and their APs at the atomic level, or LBCs as reconstructed volumetric maps from tomographic and cryo imaging, that will become common in the next ten years.



(a)



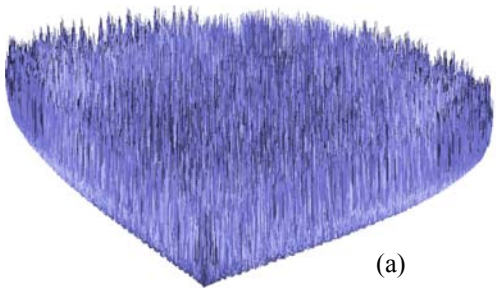
(b)

Radial Basis Functions (RBFs) interpolation methods have become increasingly popular. The ability to generate highly accurate representations of density maps without the burden of a mesh have made them attractive to the visualization community. RBFs pseudo-atomic RBFs representation has many advantages. First, they significantly reduce the dimensionality of the original density map while preserving Biomolecular structures. The figure to the left (a) shows an isosurface for the p8 trimer from Cryo-Em map from the Rice Dwarf Virus (RDV) with 128^3 voxels. The trimer on the right (b) has been reconstructed from a thin-plate RBF with only 7000 expansion points while preserving molecular features. Second, this representation is especially attractive since they couple with fast and efficient Irregular Fourier methods for fast volume rendering visualization, structural identification and docking.

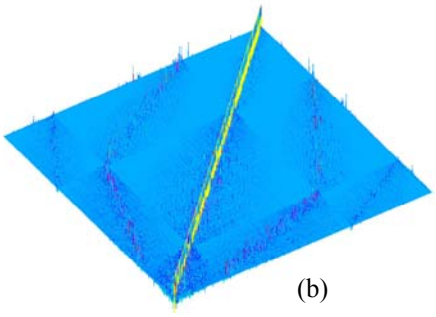
Fast and stable extraction of RBF representations

The computational cost for extracting RBFs representations can be prohibitively expensive for large amounts of interpolation points. Indeed using direct methods, for an N point interpolation problem it requires $O(N^2)$ memory and $O(N^3)$ computational steps. Moreover, this problem is badly conditioned and for relatively small problems it can cause the linear system solver to stagnate. One key aspect of MR is to derive fast, stable and memory efficient RBFs representations.

Development of algorithms for RBFs extraction has been a recent topic in the scientific computing and visualization communities. Some work has been done in this area. However, current algorithms are either based on unreliable heuristics or incomplete methods. Moreover, in many instances it is desirable to employ non-radial symmetrical (i.e. anisotropic basis representations) since it leads to a more accurate description of the molecular data. One crucial observation of the RBF interpolating problem is that it can be posed as an integral equation. This observation allows us to extend the sparsification and conditioning techniques developed for integral equations [6,8], to RBFs representations This method sparsifies and stabilizes badly conditioned dense RBF matrices. In the following figure a 7000 point RBF thin-plate spline matrix (a) (Corresponding to the p8 trimer), which is dense and unstable, is transformed into a highly sparse and well conditioned matrix (b) ([1]).



(a)



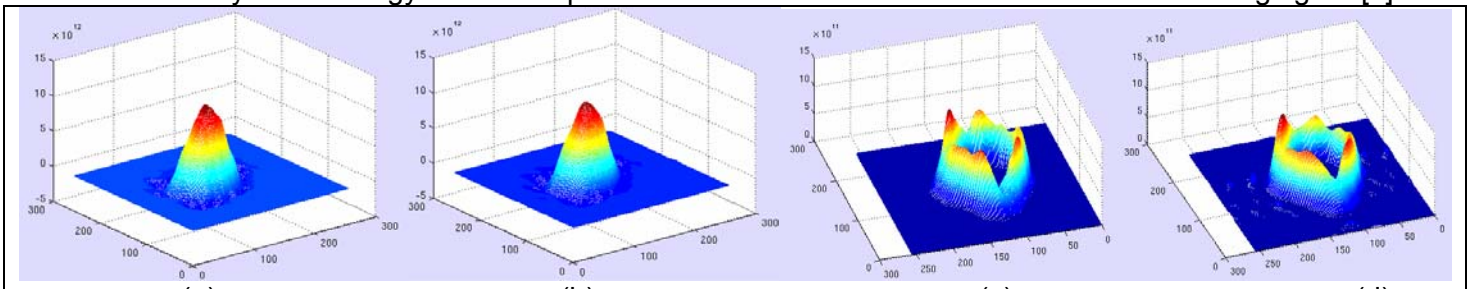
(b)

Fast Biomolecular Convolution Codes for Structure Identification and Docking.

Molecular docking and identification usually consists of two primary selections. One is the choice of goodness of fit measure (sometimes called the scoring function) while the other is the choice of the search algorithm. Both of these decisions are based on an assumed molecular model. The scoring function includes consideration for molecular properties in addition to a representation of molecular shape.

Many approaches exist to optimize the scoring function, among these graph theory, geometric approaches, spherical harmonics, and grid based Fourier methods. All of these methods suffer drawbacks in accuracy of representation or computational efficiency. However, the convolution theorem of the Discrete Fourier Transform (DFT) makes Gird based Fourier methods maintain the best balance between accuracy and computational efficiency.

Our approach significantly optimizes this balance by avoiding the construction of the high resolution N^3 Grid and constructs the Fourier spectrum directly from the atomic representation. From a RBFs representation a highly accurate convolution profile can be computed extremely fast with a Non equiDistant Fourier Transform (NDFT) algorithm. This approach only requires *O(M)* coefficients and *O(MlogM))* operations, where *M* is the number of atoms or pseudo atoms in the molecule. In practice, the number of atoms will be significantly smaller than the grid needed to represent it. Indeed, our method was tested on the mACHE and Super oxide dismutase molecule vs. a 256^3 grid FFT method. Computational speed improvements of 100 times and 50 times less memory are observed with only 0.1% energy and 1.5 % peak location detection error as shown in the following figure [2].



(a) (b) (c) (d)

Figure: Convolution profile example obtained from the FFT and NDFT. Computational speed improvements of 100 times and 50 times less memory are observed with only 0.1% energy and 1.5 % peak location detection error. (a) Slice of a convolution profile of correlating mAChE with itself obtained with the FFT grid based method on a 256^3. (b) Convolution profile obtained with our method. (c) Profiles of docking the molecular complex of superoxide dismutase with the FFT grid based method on a 256^3 where The core clash regions, which are negative, were removed for clarity . (d) Convolution profile obtained with our method.

Currently, this approach optimizes the grid-based Fourier docking method. This method will be extend to tackle Multi-Resolution flexible molecules representation

Simulation of Biophysical Phenomena

Scientific computing techniques have gain traction as a tool to describe the underlying physical phenomena of biological systems. These methods provide insights on the basic mechanisms of biomolecular dynamics and function. At CCV we have examined computational methods to describe biomolecular phenoma such as the synaptic transmission in neuromuscular junctions and the solvation/electrostatic properties of large biomolecular systems. Our goal is to develop computational efficient solvers that are well adapted to describing biophysical phenomena.

Wavelets have already been shown to be conceptually attractive for problems in computational mechanics, where one is often interested in studying local features such as shocks, stress concentrations and other discontinuities. Wavelets have led to efficient, fast, hierarchal, and adaptive algorithms for solving Partial Differential Equations (PDEs) in fields such as computational fluid dynamics, elasticity, and electrostatics. A considerable body of literature on wavelets now exists. Most of the wavelet constructions to date have been constructed over regular grids. Although they have led to very efficient algorithms, the restriction to regular grids has limited them to simple problems in a 3D scenario. This is in contrast to finite elements, which have been used in the last 30 years. They are flexible and very well suited to complex grids. This motivates the construction of wavelets representations with all of their advantages with the flexibility of finite elements.

The various challenges involved in wavelet-based physical simulation algorithms are the design operator sparsification techniques, error estimators, adaptive refinement strategies and pre-conditioners. This involves not only mathematical developments, but also efficient data structure programming techniques.

In [6] I construct continuous Multi-resolution Spatially Adaptive Multiwavelets. These constructions lead to a class of interpolating irregular wavelets of arbitrary polynomial order which combine the fast transform, decorrelation and localization properties of wavelets with the flexibility of finite elements on irregular meshes. Moreover, these constructions compression capabilities are independent of the geometry; this includes curves and surfaces that have sharp edges. This makes them an excellent basis to solve integral and differential equations.

The power of this approach was demonstrated for a model 3D potential problem over a complex skull mesh cast as an integral equation of the second kind. It is shown in [8] theoretically that an optimal convergence rate is achieved, with only $O(N \log N^p)$ entries of the discrete operator matrix, where p is a small number and N is the number of unknowns. Moreover, in practice, only $O(N \log N^p)$ entries are needed. This is in contrast to boundary elements which require a full dense matrix to achieve optimal convergence. Furthermore, since wavelet constructions only require the neighboring nodes this opens up opportunities for parallel code implementations. Moreover, this work has been extended to decouple the refinement scales on a weak formulation discretization of PDEs [5]. This implies that as we refine the mesh we do not have to solve the entire problem again, just local updates would have to be solved and the added to the courser solution. These methods have been developed in conjunction with my Ph.D. advisor, Prof. Kevin Amaratunga from MIT.

Among the many biophysical problems, the study of diffusion in biomolecular systems is particularly attractive, due the role it plays in protein-protein interactions, ligand binding and signal transmission at synaptic junctions. The Multi-resolution Spatially Adaptive Multiwavelet solver will be extended to the study diffusion in biomolecular systems using continuum mechanics equations. More precisely, to build fast and stable solvers for the steady state Smoluchowski equation to calculate ligand binding rate constants for large biomolecules.

References

[1] J. E. Castrillon-Candas, C Bajaj and J. Li, “Sparsification and Stabilization of Radial and Anisotropic Basis Functions”, *ICES Technical Report*.

[2] J. E. Castrillon-Candas, C Bajaj and V.K. Siddavanahalli, “An Adaptive Compact Fourier Representation Method for Protein-Protein Docking”, *ICES Technical Report*.

[3] C. Bajaj and J. E. Castrillon-Candas, “Hierarchical Compressed Volumetric Representations of Molecular Structures”, *ICES Technical Report*.

[4] W. Luo , J. E. Castrillon-Candas, H. Zarbl and W. G. Thilly (2004), “Inducible DNA Repair Accounts for Time Dependent resistance of Human AHH-1 Cells to Mutation by PAH”, Submitted to *Mutation Research.*

[5] S.D. Heedene, K. Amaratunga and J. E. Castrillon-Candas, “Generalized Hierarchical Bases: a Wavelet-Ritz-Galerkin Framework for Lagrangian FEM”, To appear in *Engineering Computations* 2002.

[6] J. E. Castrillon-Candas and K. Amaratunga, "Spatially Adapted Multiwavelets and Sparse Representation of Integral Equations on General Geometries", *SIAM SISC*, **24**, 5, 1530-1566, (2003).

[7] J. E. Castrillon-Candas and K. Amaratunga, "Fast Computation of Continuous Karhunen-Loeve Eigenfunctions using Wavelets*", IEEE Transactions on Signal Processing*, 50, 1, 78-86, January 2002.

[8] K. Amaratunga and J. E. Castrillon-Candas: "Surface Wavelets: A Multiresolution Signal Processing Tool for 3D Computational Modeling". *International Journal for Numerical Methods in Engineering*, 52, 3, 239-271, September 2001.

TEACHING STATEMENT

Computational Biology is a fast evolving field where its foundation has not yet been cast in stone. A single field approach to research and teaching will no longer be sufficient. I believe that computational mathematics will form an integral part of Biomolecular systems. This presents an opportunity to shape the thinking of students that are entering the field for the first time.

Scientific computing forms a solid base to study physical processes with mathematical modeling and informatics. This base touches numerical methods, signal processing, visualization and system theory. Indeed, many of these approaches have been surfacing and heavily funded in the study of biological processes on topics such as fold determination, protein-protein interaction and simulation of biophysical phenomena.

My initiatives would be to employ my substantial experience in undergraduate and graduate education to form a student body affluent in mathematical modeling and applications to biological processes. This can be achieved through a series of courses at the undergraduate and graduate levels to expose, motivate and teach students about the fundamentals and advances of Computational Mathematics and Biology.

My general philosophy to teaching is that you should prepare a class, but leave part of it open-ended thus increasing student participation and interaction. The objective is for the students to take an active role in their education. I tested this hypothesis frequently in the classroom and always achieved the same result. The students would loose any fear of me and thus participate more actively in the discussion. I had particular satisfaction when a student would ask me a question, I would wait and pretend that I did know the answer until another student would intervene, start a discussion that then led the class to solve the problem. Sometimes, a student would save me when I really did not know the answer!

In the Math and Civil Engineering Departments at MIT, I participated in the design of a research-oriented graduate course: Wavelets and Filter Banks, with Professors Gilbert Strang and Kevin Amaratunga. The multi-disciplinary approach to this course emphasized the applications of wavelets across the spectrum of fields. Indeed, we had students from many different departments, ranging from the sciences to engineering. My participation included teaching recitations and designing problem sets. However, the one-on-one research interactions with the students was of significant satisfaction as I learned so much from them on research problems in other fields, while teaching them the fundamentals of wavelets and signal processing. This class introduced me to the mentoring of students on research topics. This ability has further extended to advising Ph.D students on research topics at MIT and here at ICES.

In addition to graduate course experience, I have extensively participated in the formation of the undergraduate student body in Electrical Engineering and Computer Science at MIT as Head Teaching Assistant (TA). My experience includes teaching tutorials, labs and one-on-one education. Moreover, I learned organizational skills as a Head Teaching Assistant, where I coordinated large groups of TAs (11), who worked with more than 200 students.

At the University of Texas at Austin, I participated fully in the formation of Ph D students at the Computational Applied Mathematics program and Computer Science departments. Moreover, my responsibilities included lecturing and organizing the weekly seminars at CCV.

I have taught for several years at the graduate and undergraduate level. I enjoyed the experience so much that I even volunteered to teach recitation for the wavelet graduate course, although I already had funding. While I enjoy teaching, I also bring years of experience in staff management, administration, course design, and advising students at all levels, including Ph.D. candidates. In addition, I can be a strong member to your faculty as I have multidisciplinary experience in applied mathematics, computational sciences, biology, electrical engineering and visualization.