Workshop on

Mathematics and Computer Science in Modeling and Understanding of Structure and Dynamics of Biomolecules

Program and Abstracts

Organizers

Adam Liwo, University of Gdańsk, Poland

Gabriel del Rio Guerra, Universidad Nacional Autónoma de México, Mexico

François Major, Université de Montréal, Canada

Banff International Research Station for Mathematical Innovation and Discovery Vancouver, Canada

August 9 - 11, 2019

Workshop program

Friday, August 9

16:00		**Check-in begins (Front Desk – Professional Development Centre)**
19:30	20:15	**Robert Jernigan:** ***The importance of correlations in biology*** (opening lecture; TCPL 201; chairperson: Adam Liwo)
20:30	23:00	Informal gathering in 2nd floor lounge, Corbett Hall

Saturday, August 10

07:00	**08:45**	**Breakfast (Vistas Dining Room)**
08:45	09:00	**Welcome Talk by BIRS Staff** (TCPL 201)
Session 1		**Protein structure and function** **Chairperson: Marek Cieplak**
09:00	09:30	**Daisuke Kihara:** ***Computational protein tertiary structure modeling from cryo-EM maps of intermediate resolution*** (invited talk) (TCPL 201)
09:30	10:00	**Ilya Vakser:** ***Comparative Modeling of Protein Complexes*** (invited talk; TCPL 201)
10:00	10:30	Coffee Break (TCPL Foyer)
10:30	11:00	**Gregory Chirikjian:** ***Mathematical methods for biomolecular structure determination*** (invited talk; TCPL 201)
11:00	11:30	**Changbong Hyeon:** ***Cost-precision trade-off and transport efficiency of molecular motors*** (invited talk; TCPL 201)
11:30	13:00	Lunch (Vistas Dining Room)
13:00	13:20	Group Photo (TCPL Foyer)
Session 2		**Topology, disorder and allostery** **Chairperson: Gabriel del Rio**
13:20	14:05	**Marek Cieplak:** ***Emergence of knots in intrinsically disordered proteins*** (invited lecture; TCPL 201)
14:05	14:35	**Banu Ozkan:** ***Nature utilizes dynamic allostery for evolution*** (invited talk; TCPL 201)
14:35	15:05	**Nina Pastor:** ***Effect of phosphorylation and mutation to Asp and Glu on the conformational landscape of an intrinsically disordered region*** (invited talk; TCPL 201)
15:05	15:30	Coffee Break (TCPL Foyer)
Session 3		**Structure/function prediction** **Chairperson: Banu Ozkan**
15:30	15:50	**Andrzej Kloczkowski:** ***Modeling structure, stability and dynamics of proteins and protein aggregates*** (invited talk; TCPL 201)
15:50	16:10	**Michał Boniecki:** ***SimRNA: a coarse-grained method for RNA 3D structure modeling - new ideas accounting for on non-canonical base pairing*** (invited talk; TCPL 201)
16:10	16:30	**Agnieszka Karczyńska:** ***Structure prediction of mono- and oligomeric tarets using the physics-based coarse-grained UNRES force field and information from databases – CASP13/CAPRI46*** (contributed talk; TCPL 201)
16:30	16:50	**Marcelino Arciniega:** ***Pruning false positives cases from small molecule docking results using machine learning techniques*** (contributed talk; TCPL 201)

17:30	19:30	Dinner (Vistas Dining Room)
19:30	22:00	**Poster session + discussion** (TCPL Foyer, TPCL 201)

Sunday, August 11

07:00	**09:00**	**Breakfast (Vistas Dining Room)**
Session 4		**Models and algorithms** **Chairperson: Changbong Hyeon**
09:00	09:30	**Carlos Brizuela:** ***Limits of performance for protein side chain packers*** (invited talk; TCPL 201)
09:30	10:00	**Zhiyun Wu:** ***A global subspace optimization algorithm for minimum energy molecular cluster conformation*** (invited talk; TCPL 201)
10:00	10:30	Coffee Break (TCPL Foyer)
10:30	10:50	**Tamara Bidone:** ***Computational model of kinetochore-microtubule attachments*** (conributed talk; TCPL 201)
10:50	11:10	**Maribel Hernández Rosales:** ***Graph Theory in Orthology Detection*** (contributed talk; TCPL 201)
11:10	11:30	**Closing remarks** (TCPL 201)

Abstracts

L - Lectures

IT - Invited talks

CT - Contributed talks

P - posters

L-1. The Importance of Correlations in Biology

Robert L Jernigan
Iowa State University
Ames, IA, USA

The availability of large genome-related data provides the opportunity to extract sequence correlations. Already there have been huge advances in structure prediction and predictions of interactions by using information about the correlated pairs of amino acids in multiple sequence alignments. We have also been using this same type of information to improve sequence matching for the purpose of annotating genes and proteins for their function. We observe major gains in the specificities of function, estimated at nearly half of genes, in nearly all cases confirming the function assignments from BLAST/Blosum62 but providing significantly more useful information.

The denseness of biological systems would immediately suggest the importance of these pairs, and higher order correlations. Reliably extracting higher order correlations requires a larger set of data. Although the data is growing rapidly, success in obtaining the higher order correlations at present will require the integration of diverse sets of data.

The ultimate goal for such efforts is to understand the origin of specific phenomes and disease states, as well as the effects of mutations.

L-2. Emergence of knots in intrinsically disordered proteins

Marek Cieplak

Institute of Physics, PAS, Warsaw, Poland

Transient knotted structures are expected to arise during the volatile evolution of intrinsically disordered peptide chains. We show that this is indeed the case for sufficiently long polyglutamine tracts and α-synuclein. The polyglutamine tracts are fused within huntingtin protein that is associated with the Huntington neurodegenerative disease. We show that the presence of knots in the tracts hinders and sometimes even jams translocation, especially when the knots are deep. The knots in polyglutamine may form in tracts exceeding about 40 residues. This fact explains the existence of a similarly sized length treshold above which there is an experimentally observed toxicity at the monomeric level. We also discuss emergence of knots in α-synuclein. We show that these knots are either shallow or deep and last for about 3 – 5 μs, as inferred from an all-atom explicit-solvent 30 μs trajectories. We discuss conformational biasses that take place in α-synuclein and contact formation during aggregation of two chains of this protein. We then discuss several aspects of dynamics of knotted structured proteins as assessed within a Go-like model. In particular, we argue that folding under the nascent conditions is essential to fold to a structure that is deeply knotted.

In collaboration with: M. Chwastyk, Ł. Mioduszewski, A. Gomez-Sicilia, M. Carrion-Vazquez, P. Robustelli, and Y. Zhao.

IT-1. Computational protein tertiary structure modeling from cryo-EM maps of intermediate resolution

Daisuke Kihara[1,2], Genki Terashi[1], Sai Raghavendra Maddhuri Venkata Subramaniya[2]

[1]Department of Biological Sciences; [2]Department of Computer Science, Purdue University, West Lafayette, IN, 47907, USA

The significant progress of the cryo-EM poses a pressing need for software for structural interpretation of EM maps. Particularly, protein structure modeling tools are needed for EM maps determined at a resolution around 4 Å or lower, where finding main-chain structure and assigning the amino acid sequence into EM map is challenging. In this seminar, we discuss computational protein structure modeling tools we have been developing and future directions, opportunities, and challenges. We have developed a de novo modeling tool named MAINMAST (MAINchain Model trAcing from Spanning Tree) for EM maps for this resolution range [1, 2, 3]. MAINMAST builds main-chain traces of a protein in an EM map from a tree structure constructed by connecting high-density points without referring to known protein structures or fragments. The method has substantial advantages over the existing methods. MAINMAST showed better modeling performance than existing methods. The method is further enhanced recently to be able to model symmetric protein complexes and ligand (drug) molecules that bind to a protein in a map. Moreover, to provide structure information for maps determined at a lower resolution (5~10 Angstroms), we have recently developed a new tool, Emap2sec, which uses convolutional deep learning for detecting secondary structures of proteins [3]. Emap2sec scans an EM map with a voxel and assigns a secondary structure, i.e. alpha helix, beta strand, or coil, from density patterns of the voxel and its neighbors.

Acknowledgments: This research was partly supported by NIH (R01GM123055), NSF (DMS1614777, CMMI1825941), and Purdue Institute of Drug Discovery.

[1] G. Terashi, D. Kihara, Nature Communications, **9**, 1618, 2018
[2] G. Terashi, D. Kihara, J. Struct. Biol., **204**, 351-359, 2018.
[3] S. R. M. V. Subramanya, G. Terashi, D. Kihara, accepted, 2019.
[4] http://kiharalab.org/mainmast/

IT-2. Comparative Modeling of Protein Complexes

Ilya A. Vakser, Devlina Chakravarty, Taras Dauzhenka, Petras J. Kundrotas

Computational Biology Program and Department of Molecular Biosciences, The University of Kansas, Lawrence, KS 66047, USA

Comparative modeling of protein complexes is a technique complementary to free docking. Accounting for similarity of protein-protein interfaces in the docking algorithms is important for generating near-native docking models. Traditionally, template-based modeling of protein-protein complexes has relied on similarity of the entire proteins (targets) to known structures of protein-protein complexes (templates). However, similarity can be also inferred between the target protein surface and the template protein-protein interface. Earlier studies indicated that interface alignment can also be used to refine docking models generated by the full structure alignment. We performed a comprehensive comparative benchmarking of these two approaches (full structure alignment and interface alignment) based on 223 bound and unbound complexes from the DOCKGROUND benchmark set 4. The results showed that docking performance for the unbound structures compared to that for the bound structures decreases only slightly for both methods. Overall, the interface alignment performs marginally better than the full structure alignment in selecting templates which provide models with low i-RMSD and good quality. If a common template is selected in both interface and full structure alignments, the docking success rate increases significantly. Interface templates perform better than full structures for targets with highly specific interactions, when the full structures are different (e.g. in case of multidomain targets with a different relative orientation of domains than in the templates). Full structure alignment performs better for proteins with different binding sites for different functions. The interface-based predictions require less refinement for all high-quality models irrespective of the rank. In general, matching CATH annotations for target and template yield correct solutions. However, the success can be affected by alternative binding sites. Combining templates from sequence and structure homology increases the success rate to ~87%. Structural refinement of modeled protein-protein complexes is an essential step in protein docking. Such refinement becomes increasingly challenging when the interacting proteins undergo significant conformational changes upon binding (unbound/bound interface RMSD $\geq$ 3 Å). We developed a flexible interface refinement protocol and benchmarked it on a set of top 100 docking models, which are inside the docking funnel, generated for binary protein-protein complexes from the DOCKGROUND benchmark set 4. The refinement of the docking models was performed by a systematic local conformational search using previously developed contact potentials. The results showed that the procedure performs better than other widely used refined protocols benchmarked on the same set of protein-protein complexes.

IT-3. Mathematical Aspects of Biomolecular Structure Determination

Gregory S. Chirikjian[1],[2]

[1]National University of Singapore; [2]Johns Hopkins University

This talk discusses advances in modelling data acquisition and information fusion in biomolecular structure determination from x-ray crystallography, SAXS, and EM. Whereas the configuration space of a single rigid body is the group of orientation-preserving Euclidean motions, the configuration space of a collection of rigid bodies that move in lock step constrained by a discrete symmetry group is actually a coset space (or quotient space) of the configuration space of a single body by the discrete symmetry group. This quotient space, called a `motion space', is the configuration space in the molecular replacement method in macromolecular crystallography. Of particular importance is the characterization of those coordinated motions which place symmetry mates in collision, since only the complement of this space is physically viable. This talk also discusses how SAXS and single-particle EM provide complementary information which can be fused to obtain better electron density estimates than each one can obtain independently.

Acknowledgments: This work was funded under grants NSF CCF-1640970, NSF IIS-1619050, NIH R01GM113240. The collaborations with the co-authors on the papers listed below are greatly appreciated.

References

Chirikjian, GS. "Mathematical aspects of molecular replacement. I. Algebraic properties of motion spaces." *Acta Crystallographica Section A: Foundations of Crystallography v*67, no. 5 (2011): 435-446.

Chirikjian, GS. "Kinematics Meets Crystallography: The Concept of a Motion Space." *Journal of Computing and Information Science in Engineering* 15, no. 1 (2015): 011012.

Chirikjian, GS., and B Shiffman. "Collision-free configuration-spaces in macromolecular crystals." *Robotica* 34, no. 8 (2016): 1679-1704.

Chirikjian GS, Sajjadi S, Shiffman B, Zucker SM. Mathematical aspects of molecular replacement. IV. Measure-theoretic decompositions of motion spaces. Acta Crystallographica Section A: Foundations and Advances. 2017 Sep 1;73(5):387-402.

Kim, J.S., Afsari, B., Chirikjian, G.S., ``Cross-Validation of Data Compatibility Between SAXS and Cryo-EM," Journal of Computational Biology, 24(1):13--30, 2017.

Dong, H., Kim, J.S., Chirikjian, G.S., ``Computational Analysis of SAXS Data Acquisition," Journal of Computational Biology, 22(9): 787-805, 2015.

IT-4. Cost-precision trade-off and transport efficiency of molecular motors

Changbong Hyeon[1]

[1]Korea Institute for Advanced Study, Seoul 02455, Republic of Korea

An efficient molecular motor would deliver cargo to the target site at a high speed and in a punctual manner while consuming a minimal amount of energy. However, according to a recently formulated thermodynamic principle, known as the thermodynamic uncertainty relation, the travel distance of a motor and its variance are constrained by the free energy being consumed. Here we use the principle underlying the uncertainty relation to quantify the transport efficiency of molecular motors for varying ATP concentration ([ATP]) and applied load (*f*). Our analyses of experimental data find that transport efficiencies of the motors studied here are semi-optimized under the cellular condition. The efficiency is significantly deteriorated for a kinesin-1 mutant that has a longer neck-linker, which underscores the importance of molecular structure. It is remarkable to recognize that, among many possible directions for optimization, biological motors have evolved to optimize the transport efficiency in particular.

[1] "Energetic Costs, Precision, and Transport Efficiency of Molecular Motors" W. Hwang, C. Hyeon, *J. Phys. Chem. Lett.* (2018) 9, 513-520

[2] "Physical insight into the thermodynamic uncertainty relation using Brownian motion in tilted periodic potentials" C. Hyeon, W. Hwang, *Phys. Rev. E.* (2017) 96, 012156

[3] "Quantifying the Heat Dissipation from a Molecular Motor's Transport Properties in Nonequilibrium Steady States" W. Hwang, C. Hyeon *J. Phys. Chem. Lett.* (2017) 8, 250-256

IT-5. Nature utilizes dynamic allostery for evolution

S. Banu Ozkan[1]

[1]Center for Biological Physics Department of Physics Arizona State University, Tempe AZ

The discovery of the protein structure proved to be one of the major scientific achievements in biological sciences and marked a milestone in the field, spawning the sequence-structure-function paradigm. That is, how can the 1-dimensional set of information from the amino acid sequence give rise to a unique 3-dimensional structure which, in turn, determines the function of a given protein. However, it quickly became apparent that it was not so simple as assigning a single 3-dimensional structure to an amino acid sequence. Some proteins which bind to ligands could take on multiple, distinct conformations depending on the binding event or environmental conditions. Furthermore, there are proteins sharing the similar 3-D structure despite the significant difference in amino acid sequences or even specific domains of proteins present across protein families. Thus, all these studies gave rise to the ensemble picture of proteins. That is, proteins are not simple static objects but rather dynamic entities, that sample many conformational states even in the absence of ligand binding. Variations in amino acid sequences can alter ligand recognition, binding rates, and other biophysical, thermodynamic and kinetic properties of homologous enzymes while still maintaining similar 3-dimensional folds. Differing functions between structural homologues gave rise to a view of protein evolution which proceeds through conformational dynamics and functional promiscuity, where a relationship may exist between active site flexibility and amino acid evolvability. Here we introduce two tools, the Dynamic Flexibility Index (DFI)[1] and the Dynamic Coupling Index (DCI) [2]which can quantify structural flexibility and dynamic coupling at a site-specific, single amino acid level. We show that it is possible to relate evolutionary conservation corelates with the flexibility of a given position. Through conformational dynamics analysis of ancestral proteins, we present that conformational ensemble of a protein is modified to adopt to a new environment and/or to emerge a new function by modifying rigidity and flexibility of its positions [3]. Finally, we discuss how Nature can modulate change through allosteric mutations which alter the internal interaction network of proteins, and how changes in allosteric regulation can result in disease phenotypes[4].

Acknowledgments: This research was supported by NSF

[1] Z.N. Gerek, S. Kumar, S.B. Ozkan Evol Appl **6**, 423–433, 2013.
[2] A.Kumar, T.J. Glembo, S.B. Ozkan Biophys J **109**:1273–1281, 2015
[3] V.A. Risso, J.M. Sanchez-Ruiz, S.B. Ozkan Curr. Opp. Struct. Biol. **51**,106-115, 2018.
[4] A. Kumar, B.M. Butler, S. Kumar, S.B.Ozkan Curr. Opp. Struct. Biol. **51**,135-142, 2015

IT-6. Effect of Phosphorylation and Mutation to Asp and Glu on the Conformational Landscape of an Intrinsically Disordered Region

Nina Pastor[1] and Marco A. Ramírez-Martínez[1]

[1]Centro de Investigación en Dinámica Celular-IICBA, Universidad Autónoma del Estado de Morelos, Av. Universidad 1001, Col. Chamilpa, 62209 Cuernavaca, Morelos, México

Intrinsically disordered proteins are notorious for their conformational flexibility and capacity for interaction with different targets by acquiring distinct conformations depending on the specifics of the binding site. They can also engage in specific interactions without loosing conformational freedom, forming fuzzy complexes. The particular conformations favored by these proteins can be tuned by posttranslational modifications, such as phosphorylation. A common experimental strategy to study the effect of phosphorylation is to perform mutations of the modified residues by aspartate (D) or glutamate (E), under the assumption that the main effect of phosphorylation is the inclusion of negative charge. Whether the mutation to D or E is equivalent to phosphorylation is case dependent. In this work we explore the conformational landscape of an intrinsically disordered region at the C-terminus of adenoviral protein E1B55kDa, which is regulated by phosphorylation at three residues at the C-terminus, with the aim of establishing whether mutation to D or E is equivalent to modification by phosphorylation. In the context of the complete virus, the triple mutant with Ds produces a more efficient virus compared to wild type, and mutation to alanine (A), which cannot be phosphorylated, is equivalent to not having the full protein. We chose the last 20 residues of E1B55kDa as our reference peptide, as multiple disorder predictors consider it to be disordered. This peptide has only one cationic residue (arginine 9), and the C-terminal half is enriched in negative residues. We submitted the wild type sequence to Pepfold3, and obtained 100 different structures for it. Taking these as a reference, we built versions with three phosphorylated residues (two serines and one threonine), three Ds, three Es and three As. We placed each peptide in a water box with 0.15M NaCl using Charmm-gui, and ran it in NAMD in the NPT ensemble at 298K and 1 atm with the Charmm36m forcefield for 50 ns, achieving a total simulated time of 5 μs for each peptide variant. The distribution of the radius of gyration shows the prevalence of extended structures, with a slightly expanded ensemble for the triple D and triple E variants, and a slightly compressed one for the phosphorylated version, compared to the wild type. This is reflected in almost saturated hydration for the peptide in all its residues, except for a small decrease in hydration number for arginine 9 in the phosphorylated version. A closer look at intrapeptide hydrogen bonds reveals that there are few interactions in general, but arginine 9 engages in many more interactions with the phosphorylated residues than with the other charges in the peptide; the interaction with phosphorylated threonine 19 is preferred above all. This interaction leads to the formation of a loop that prefers to adopt disordered conformations, so we propose that it engages in fuzzy complexes with other proteins. In general, phosphorylation leads to an increase in alpha helix formation in the peptide, while substitution for Ds and Es leads to a loss of this structure. We conclude that phosphorylation and the mutation to D and E are not equivalent.

Acknowledgments: This research was supported by a CONACYT scholarship for MARM and supercomputing time at the Laboratorio Nacional de Supercómputo del Sureste (LNS), LANCAD in México City, and the Laboratorio de Dinámica de Proteínas at UAEM.

IT-7. Modeling structure, stability and dynamics of proteins and protein aggregates

Andrzej Kloczkowski

Battelle Center for Mathematical Medicine, The Research Institute at Nationwide Children's Hospital, Columbus, OH 43215

Recent progress in modeling structure and dynamics of proteins and protein aggregates will be reviewed. Recent advancements in structure prediction such as development of better potentials and force-fields (including multibody potentials), and improved modeling of free energies will be presented. We significantly improved protein structure evaluations by considering the effects of amino acid variants on protein stability, and we have shown that the outliers in stability are typically aberrant proteins. Recently, we have made significant progress in understanding protein stability and dynamics by computing protein free energies extracted from structures to account for the high packing densities in proteins, including important novel evaluations of protein entropies. Preliminary results show large improvements over previous potentials for assessing protein stabilities. Our results demonstrate that there are substantial gains in specificity from combining the sequence with structural and protein dynamic data. These developments may significantly impact the advancement of precision/personalized medicine.

IT-8. SimRNA: a coarse-grained method for RNA 3D structure modeling - new ideas accounting for on non-canonical base pairing

Michal J. Boniecki[1]

[1]Internatonal Institute of Molecular and Cell Biology in Warsaw, Poland

The molecules of the ribonucleic acid (RNA) perform a variety of vital roles in all living cells. Their biological function depends on their structure and dynamics, both of which are difficult to experimentally determine but can be theoretically inferred based on the RNA sequence. SimRNA [1] is one of the computational methods for molecular simulations of RNA 3D structure formation. The method is based on a simplified (coarse-grained) representation of nucleotide chains, a statistically derived model of interactions (statistical potential), and the Monte Carlo method as a conformational sampling scheme.

In SimRNA, the backbone of the RNA chain is represented by two atoms per nucleotide, whereas nucleotide bases are represented by three atoms each. In fact, these three atoms are used to calculate a system of local coordinates that allows for positioning of a 3D grid - the actual representation of the base. The 3D grid contains information about the interactions of the entire base moiety (not only the three atoms explicitly included in the SimRNA representation).

The current version of SimRNA (3.22) is able to predict basic topologies of RNA molecules with sizes up to about 50-70 nucleotides, based on their sequences only, and larger molecules if supplied with appropriate distance restraints. However, it should be noted that the current version of SimRNA, as well as other methods for RNA 3D structure prediction, exhibit a number of limitations, which reduce the accuracy of RNA 3D structure models obtained. One of the biggest challenges is the prediction of non-canonical base pairs, which are crucial for the formation of functional motifs in RNA structure. Current studies and developments are focused on a new version of SimRNA, which will overcome the key limitations that exist in the current version of the program, as well as general limitations in current methods for RNA 3D structure prediction. The major idea is to split all the contacts corresponding to base-base interactions into classes that describe specific types of base-base interactions (canonical and non-canonical), while derivation of the statistical potential.

Acknowledgments: This work was supported by the Polish National Science Center Poland (NCN) (grant 2016/23/B/ST6/03433 to M.J.B.)

[1] M.J. Boniecki, G. Lach, W.K. Dawson, K. Tomala, P. Lukasz, T. Soltysinski, K.M. Rother, J.M. Bujnicki, Nucleic Acids Res. 2016 Apr 20;**44**(7):e63

IT-9. Limits of performance for protein side chain packers

Carlos A. Brizuela[1] and José D. Colbes[2]

[1]Cicese Research Center, Ensenada, Baja California, Mexico; [2]Universidad Nacional de Asunción, San Lorenzo, Paraguay.

To date, it is possible to design proteins with an improved function starting from known scaffolds. This design applies to the enhancement of enzymatic capabilities, inhibitors of protein-protein interactions, among others [1]. The next generation for the design of functional proteins will be guided by the approach known as template-free design. The goal of this approach is to design a sequence of amino acids that will have a predefined function. A more conservative approach seeks to find a chain of amino acids that will fold into a predefined backbone geometry. A central challenge to the latter approach is the side chain packing problem (SCPP) that aims to find a set of rotamers that minimizes a given scoring function, for a fixed backbone geometry associated to a candidate sequence. In this talk, we will define the computational model for the SCPP, analyze the results achieved by state-of-the-art packers, and determine a lower bound for the maximum achievable accuracy of a simple rotamer library [2]. We also show that a strong limitation to reduce the gap between state-of-the-art results and the maximum attainable accuracy is the scoring function. Furthermore, we show that the limitation in the scoring function is not related to an incorrect weighting of its components nor to the constrained geometry of the crystal [3].

[1]. P.S. Huang, S.E. Boyken, and D. Baker. ¨*The coming of age of de novo protein design*¨. *Nature* 537 (7620): 320 – 327, 2016.

[2]. J. Colbes, R.I. Corona, C. Lezcano, D. Rodriguez, C.A. Brizuela. "Protein aide-chain packing problem: is there still room for improvement?." *Briefings in Bioinformatics*, doi:10.1093/bib/bbw079, 2016.

[3]. J. Colbes, S. Aguila, C.A. Brizuela. "Scoring of side-chain packings: An analysis of weight factors and molecular dynamics structures". *Journal of Chemical Information and Modeling*, 58 (2), 443-452, 2018.

IT-10. A Global Subspace Optimization Algorithm for Minimum Energy Molecular Cluster Conformation

Zhijun Wu
Department of Mathematics
Iowa State University

We consider the problem to obtain the optimal conformation of a given molecular cluster with the lowest possible potential energy. This problem has been studied as a test case for global optimization algorithms, and considered as a starting point for the study of more complicated conformational problems such as protein folding through potential energy minimization. Here we propose a global subspace optimization algorithm for the solution of the problem, with the variables of the energy function divided into subgroups and optimization performed successively in the subspaces corresponding to the subgroups of variables. The idea behind the algorithm comes from the study of group behaviors of biological populations, where species compete for resources yet find strategies to co-exist and co-evolve. We show that such behaviors can be modeled as a multi-player evolutionary game, and the potential energy minimization problem can be reduced to such a game, with each subgroup of variables considered as a strategy set to be determined by a subpopulation of species. Thus, the successive subspace minimization of an energy function proceeds like a game played among subgroups of species in a biological population. We show that a Nash-equilibrium of the game is equivalent to a KKT point of the energy minimization problem subject to a set of linear and nonnegative constraints, and an evolutionary stable equilibrium corresponds to a strict energy minimizer. We describe the implementation of the algorithm and present some preliminary test results for a small group of clusters.

CT-1. Structure prediction of mono- and oligomeric tarets using the physics-based coarse-grained UNRES force field and information from databases – CASP13/CAPRI46

A.S. Karczyńska[1], E.A Lubecka[1,2], A.G.Lipska[1], A.K. Sieradzan[1], A. Giełdoń[1], A. Liwo[1,3], and C. Czaplewski[1]

[1]Faculty of Chemistry, University of Gdańsk, Wita Stwosza 63, 80-308 Gdańsk, Poland; [2]Institute of Informatics, Faculty of Mathematics, Physics, and Informatics, University of Gdańsk, Wita Stwosza 57, 80-308 Gdańsk, Poland; [3]Korea Institute for Advanced Study, 85 Hoegiro, Dongdaemun-gu, 130-722 Seoul, Republic of Korea

The results of blind prediction of the structures of monomeric and oligomeric proteins obtained in the recent CASP/CAPRI experiment by the KIAS-Gdansk/Czaplewski groups, by using the physics-based coarse-grained UNRES force field [1] and information from databases are presented. For both monomeric and oligomeric targets, the methodology of the KIAS-Gdansk/Czaplewski groups included extensive conformational search by means of the Multiplexed Replica Exchange Molecular Dynamics (MREMD) simulations [2] with the UNRES force field [3], with geometry restraints from server models [4,5]. For the monomeric targets, the restraints were derived from fragments with similar geometry that occured in the server models (the consensus fragments), while monomer geometries except for the terminal, flexible loop, and linker regions, were restrained in oligomer simulations. The server models were selected mainly based on DeepQA score [6] ranking; when the score was below 0.5, models from the servers which performed well in previous CASP exercises: Zhang, Quark, and BAKER-ROSETTASERVER were selected. For oligomeric targers, monomers were modeled first and, subsequently, initial oligomeric structures were modeled based with the aid of the packing proposed by the HHpred server [7]; for smaller targets the monomers were oriented randomly. The results of MREMD simulations were processed by using the Weighted Histogram Analysis Method (WHAM) [8] to obtain the probabilities of conformations and subsequently subjected to cluster analysis to obtain the 5 (CASP) or 10 (CAPRI) families of conformations, from which the conformations closest to the mean conformations were, in turn, selected as candidate predictions [1], which were subsequently converted to all-atom conformations submitted to CASP/CAPRI. The obtained models were ranked solely based on the computed probabilities of the families obtained by summing up the probabilities of the constituent conformations computed by WHAM based on the UNRES effective function [1].

Acknowledgments: This research was supported by grant UMO-2017/26/M/ST4/00044 from the National Science Centre of Poland (Narodowe Centrum Nauki).

[1] A. Liwo et al., J. Mol. Model. 20 (2014): 2306.
[2] Y.M. Rhee et al., Biophys. J. 84 (2003): 775-786.
[3] C. Czaplewski et al., J. Chem. Theor. Comput. 5 (2009): 627-640.
[4] P. Krupa et al., J. Chem. Inf. Model. 55 (2015): 1271-1281.
[5] M. Mozolewska et al., J. Chem. Inf. Model. 56 (2016): 2263-2279.
[6] R. Cao et al., BMC Bioinformatics. 17 (2016): 495.
[7] L. Zimmermann et al., J Mol Biol. 430 (2018): 2237-2243.
[8] S. Kumar et al., J. Comput. Chem., (2001) 8, 1011-1021.

CT-2. Pruning false positives cases from small molecule docking results using machine learning techniques

Marcelino Arciniega-Castro[1], Fernanda Álvarez-Esquinca[1], Adrián A. Rodríguez-Pie[1]

[1]Cell Physiology institute, National Autonomous University of Mexico.

Over the last three decades, Computer Aided Drug Design (CADD) has positioned as one of the more useful approaches aiding the research at early stages of drug discovery process [1]. Particularly, small molecule docking algorithms have been employed exhaustively to identify the possible atomic interactions, between the protein target and a suggested small molecule, that support the formation of the protein-ligand complex. This evaluation is performed by employing a scoring function that relates geometric patterns of the interacting molecules to free energy values. However, the accurate and exact prediction of the binding free energy, along with the complex conformation, remains as an open problem [2]. As consequence, docking results present a high rate of false positive cases. The high complexity of the physicochemical process, together with the vast amount of structural experimental information available, renders the use of machine learning algorithms an attractive possibility [3, 4, 5]. In the present work, we briefly describe the problems associated with current docking scoring functions and posit the idea of pruning the false positive cases using machine learning algorithms. Then, we expose the limitations the docking algorithm (not only of the scoring function) of AutodockVina [6] by analyzing a set of approximately 15000 crystallographic complexes retrieved from Protein Data Bank [7]. Finally, we present preliminary results obtained with our tools designed for false positive identification. Specifically, we show how a relatively simple Bayesian Network, based on interaction fingerprints, can be used to infer the badly placed fragment molecules (with molecular weights in the range of 150-350 Da). Additionally, we present the results obtained of a Convolutional Neural Network to analyze docking poses (molecules with molecular weights in the range of 150-850 Da). Both networks show promising results by improving Receiver Operating Characteristic metrics as compared with the use of the docking protocol alone.

Acknowledgments: This project was supported by Dirección General de Asuntos del Personal Académico at Universidad Nacional Autónoma de México (PAPIIT-IA202917). The authors thank Dirección General de Cómputo y de Tecnologías de Información y Comunicación at Universidad Nacional Autónoma de México for granting the use of the supercomputer Miztli (LANCAD-UNAM-DGTIC-320).

[1] G. Sliwoski, S. Kothiwale, J. Meiler, E. W. Lowe-Jr. *Pharmacol. Rev.*, **66**, 334-395, **2014**.
[2] HA Carlson, et. al. *J. Chem. Inf. Model.*, **56**, 1063-1077, **2016**.
[3] M. Arciniega, O. F. Lange. *J. Chem. Inf. Model.*, **54**, 1401-1411, **2014**.
[4] J. C. Pereira, et. al. *J. Chem. Inf. Model.*, **56**, 2495-2506, **2016**.
[5] M. Wójcikowski, P. J. Ballester, P. Siedlecki. *Sci. Rep.* **7**, 46710, **2017**.
[6] O. Trott, A. J. Olson. *J. Compt. Chem.* **31**, 455-461, **2010**.
[7] www.rcsb.org.

CT-3. Computational model of kinetochore-microtubule attachments

Tamara C Bidone[1,2], Samuel Campbell[1], Mohammed A Amin[3], Dileep Varma[3]

[1]Department of Bioengineering, University of Utah; [2]Scientific Computing and Imaging Institute, University of Utah; [3]Department of Cell and Molecular Biology, Northwestern University

The ability of cells to separate chromosomes during mitosis is critical to several phases of their physiology. Chromosome segregation is mediated by spindle microtubules that attach to mitotic kinetochores via a dynamic protein interface, which includes Ndc80 and its accessory proteins, Ska, Cdt1 and ch-TOG [1-3]. The Ndc80 complex forms the core component of the attachment sites while Ska, Cdt1 and ch-TOG binds kinetochores via the Ndc80 complex. From prometaphase to metaphase, the kinetochore levels of Ska and Cdt1 increase in Hela cells, while that of Ndc80 remains constant. This suggests a correlation between concentration of proteins at the kinetochore-microtubule (kMT) interface and increasing amounts of load during mitosis. Interestingly, while being dynamic, the kMT interface ensures stability of the connection between chromosomes and kinetochore microtubules. How the various interface proteins interplay to ensure a dynamic yet stable connection is not known because their exact roles in this process are still elusive. An interesting hypothesis is that the Ndc80-accessory proteins Ska, Cdt1 and ch-TOG directly strengthen the kinetochore-microtubule interface by forming additional connections between kinetochore-bound Ndc80 and spindle microtubules. However, since Ska, Cdt1 and ch-TOG dynamically form and break their connections with microtubules, a synergy between them is likely to exist. Here, in order to characterize the synergy between Ska, Cdt1 and ch-TOG, we developed a new computational model, based on a kinetic Monte Carlo approach. The model allowed us to explicitly incorporate Ndc80, Ska1, Cdt1 and ch-TOG, isolate their contributions, and characterize their synergistic effects on the stability of the interface. Each protein is defined by a position along a tubulin protofilament, and exists in two states, bound or unbound, while undergoing biased diffusion, as observed in experiments. The model also incorporates tension-dependent unbinding rates for each protein, including catch bond kinetics for ch-TOG, as detected experimentally [2]. As for the output, the model evaluates: (*i*) displacement of the kMT interface along the tubulin protofilament; (*ii*) time of kMT attachment under tension; and (*iii*) kMT attachment rupture force, corresponding to the force that detaches all proteins. We find that combining Ndc80, Ska and Cdt1 enhances kMT attachment strength with respect to individual components. Ch-TOG further strengthens the complex because of its catch bond kinetics. In addition, the model shows that the rupture force, corresponding to the load under which no protein is bound, increases in proportion to the number of simulated microtubules. Taken together, our results provide important mechanistic insights into how kMT proteins coordinate with each other to withstand tension and ensure accurate chromosome segregation.

[1] D. Varma, and E. D. Salmon. J. Cell. Sci., 2013.
[2] M. P. Miller, C.L. Asbury, and S. Biggins. Cell, 2016.
[3] S. Agarwal, K.P. Smith, Y. Zhou, A. Suzuki, R.J. McKenney, and D. Varma. J. Cell Biol., 2018.

CT-4. Graph Theory in Orthology Detection

Hernández-Rosales, Maribel[1], Hellmuth, Marc[2], Stadler, Peter[3]

[1]Conacyt-Institute of Mathematics, UNAM, Mexico; [2]A University of Greifswald, Germany;
[3]University of Leipzig, Germany;

During evolution genes go throw many events, such as duplication, speciation, loss, horizontal gene transfer, among others. Two genes are said to be paralogs if they are the product of a duplication event, and orthologs if they are the product of a speciation event. The distinction of paralogs and orthologs is an important problem in comparative and evolutionary genomics. Moreover, orthology detection is a first step towards any functional annotation study.

The evolutionary history of a set of genes can be represented as a phylogenetic tree where leaves represent genes and internal nodes evolutionary events. In this work, we investigate a graph theory-based method for the prediction of large-scale orthologous genes and the reconstruction of their evolutionary history [1,2,3]. We represent genes as vertices of a graph and place an edge between two genes if their sequence similarity is high. We characterize mathematically the topological properties of this graph in order to have only valid orthology relations. Surprisingly, graphs that represent valid orthology relations are P4-free, i.e. graphs which do not contain induced paths of length four. These graphs have been studied earlier and have been characterized as cographs [4]. We further investigate a set of induced subgraphs that give us evidence of noise in the data set or of wrong orthology predictions. In order to remove those induced subgraphs, we need to come up with a solution for the cograph editing problem, which has been found to be NP-complete [5]. Here we also present a work-in-process heuristic for the cograph edting problem that will help us to induce valid orthology relations.

Acknowledgments: This research was supported by Conacyt Mexico and DAAD Germany.

[1] Marcus Lechner, Maribel Hernandez-Rosales, Daniel Doerr, Nicolas Wieseke, Annelyse Thevenin, Jens Stoye, Sonja J. Prohaska and Peter F. Stadler. Orthology Detection Combining Clustering and Synteny for Very Large Data Sets. PlosONE, 9(8):e105015, (2014).

[2] Marc Hellmuth, Maribel Hernandez-Rosales, Katharina T. Huber, Vincent Moulton, Peter F. Stadler, and Nicolas Wieseke. Orthology relations, symbolic ul- trametrics, and cographs. J. Math. Biol. 66(1-2):399-420, (2013).

[3] Maribel Hernandez-Rosales, Marc Hellmuth, Nicolas Wieseke, Katharina Huber, Vincent Moulton, and Peter F. Stadler. From event-labeled gene trees to species trees. BMC Bioinformatics 13(Suppl. 19):S6, (2012)

[4] Corneil DG, Lerchs H, Stewart Burlingham LK. Complement reducible graphs. Discrete Appl Math 3:163–174, (1981).

[5] Liu Y, Wang J, Guo J, Chen J. Cographs editing: complexity and parametrized algorithms. In: Fu B, Du DZ (eds) COCOON 2011. Lecture notes computer science, vol 6842. Springer, Berlin, pp 110–121, (2011).

CT-5. Chemical Shifts Based Structural Ensemble Generation for Intrinsically Disordered Proteins: A Case Study

Yi He[1], Laura I. Gil Pineda[1,2], Laurie Milko[1]

[1]Department of Chemistry & Chemical Biology, University of New Mexico, Albuquerque, New Mexico 87131, United States; [2]Department of Chemical Sciences, Universidad Icesi, Cali, Colombia

About a third of the proteome consists of intrinsically disordered proteins (IDPs) that fold, whether fully or partially, upon binding to their partners [1]. IDPs use their inherent flexibility to play key regulatory roles in many biological processes [2]. Such flexibility makes their structural analysis extremely challenging, being nuclear magnetic resonance (NMR) the most suitable high-resolution technique. However, conventional NMR structure determination methods, which seek to determine a single high-resolution structure [3], are inadequate for IDPs. There are several tools available for the structural analysis of IDPs using NMR data and primarily Chemical Shifts (CS) [4-6]. However, a persistent problem is how to effectively sample the extensive, but not random, conformational space of IDPs. We have implemented a novel relational database, termed Glutton, that links all existing CS data with corresponding protein 3D structures with the goal of enabling the conformational analysis of IDPs directly from their experimental CS. Glutton's uniqueness is in its focus on dihedral angle distributions consistent with a given set of CS rather than with unique structures. Such dihedral distributions define how native-like is the ensemble and lead to the effective calculation of large ensembles of structures that efficiently sample the available conformational space. With Glutton, we examined Nuclear Coactivator Binding Domain (NCBD), an IDP with NMR structure obtained using osmolyte stabilizers that is partly disordered in native conditions [7]. As means of comparison, we produced a 60μs long MD simulation of NCBD in explicit solvent starting from the NMR structure and using the CHARMM36m force field with modified TIP3P water which was suggested as a good combination to explore the conformational space of IDPs [8]. The structural ensembles obtained from Glutton are based only on geometric considerations and CS restraints, but they can be further refined using additional computational (force field) and/or experimental (distance restraints) information.

Acknowledgments: This work was supported by grants: the startup fund at the University of New Mexico, the W.M. Keck Foundation, the National Science Foundation [NSF-MCB-161759 and NSF-CREST-1547848] and the European Research Council [ERC-2012-AdG-323059].

[1] H.J. Dyson, P.E. Wright, Chem. Rev., **104**, 3607–3622, 2004
[2] M.M. Babu, Biochem. Soc. Trans., **44**, 1185–1200, 2016
[3] A.M., Gronenborn, G.M. Clore, Anal. Chem., **62**, 2–15, 1990
[4] V. Ozenne, F. Bauer, L. Salmon, J. Huang, M.R. Jensen, S. Segard, P. Bernadó, C. Charavay, M. Blackledge, Bioinformatics, **28**, 1463–1470, 2012
[5] M. Krzeminski, J.A. Marsh, C. Neale, W. Choy, J.D. Forman-Kay, Bioinformatics, **29**, 398–399, 2013
[6] D.H. Brookes, T. Head-Gordon, J. Am. Chem. Soc., **138**, 4530–4538, 2016
[7] A. Naganathan, M. Orozco, J. Am. Chem. Soc., **133**, 12154–12161, 2011
[8] J. Huang, S. Rauscher, G. Nawrocki, T. Ran, M. Feig, B. L de Groot, H. Grubmüller, A.D. MacKerell Jr, Nat Methods, **14**, 71–73, 2017

P-1. Protein translocation through a α-hemolysin biological nanopore

M. A. Shahzad

Department of Physics, University of Swat, Pakistan.

We study the phenomena of translocation of a protein through a α-hemolysin nanopore under the action of a driving force using the method of Langevin dynamics simulation (Gō-like protein model). The work is motivated by recent experiments on voltage driven transport of protein across nano-channel. In this transport mechanism of protein across a channel, the force-fields consists of stretching energy, bending energy and torsion energy. The non-bonded interaction energy is modeled by Lennard-Jones potential. We model the nanopore through which the protein is confined by a step-like soft-core repulsive potential with cylindrical symmetry which is set parallel the *x*-axis of the frame of reference to used for translocation simulations (Fig. 1). We characterized the translocation mechanism by studying the thermodynamical and kinetic properties of the process. In particular, we study the average of translocation time, the mobility, and the translocation probability as a function of pulling force *F* acting in the channel. We used the free-energy profile to built up a phenomenological one-dimensional drift-diffusion model in the reaction coordinate based on the Smoluchowski stochastic differential equation. The results obtained from the mathematical model are then used in comparison with molecular dynamics simulation to explains and reproduces the behavior of the translocation observables.

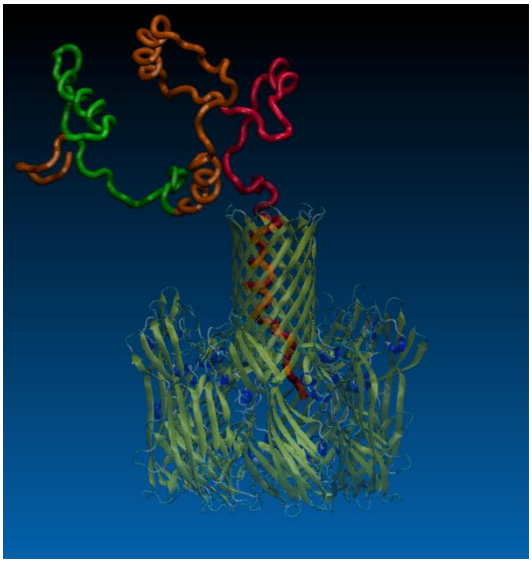

FIG. 1: Schematic view of protein pulled across a α-hemolysin nanopore.

[1] N. Gō and H. A. Scheraga, Macromolecules, 9, 535-542, 1976.
[2] N. Gō, Ann. Rev. of Biophys. Bioeng., 12, 183-210, 1983.
[3] A. Liwo, Y. He, H.A. Scheraga, Phys. Chem. Chem. Phys., 13, 16890-16901, 2011.
[4] W. Wickner and R. Schekman, Science, 310, 1452-1456, 2005.
[5] J.J. Kasianowicz, E. Brandin, D. Branton, and D.W. Deamer, Proc. Natl. Acad. Sci. USA, 93, 13770-13773, 1996.
[6] F. Cecconi, M. A. Shahzad, U. M. B. Marconi and A. Vulpiani, Phys. Chem. Chem. Phys., 19, 11260-11272, 2017.

P-2. De novo Protein Structure Prediction using RCC

Fernando Fontove-Herrera[1], Gabriel del Rio-Guerra[2]

[1]C3 Idea; [2]Department of Biochemistry and structural biology, Institute of cellular physiology, UNAM

Residue Cluster Class is a representation for proteins designed with the purpose to predict structure classification, achieving high performances [1]. The feature space if generated by building the contact graph of all the protein residues and then calculating its maximal cliques, each feature of the RCC corresponds to the number of cliques of 26 different types including sizes 3 to 6 (Figure 1).

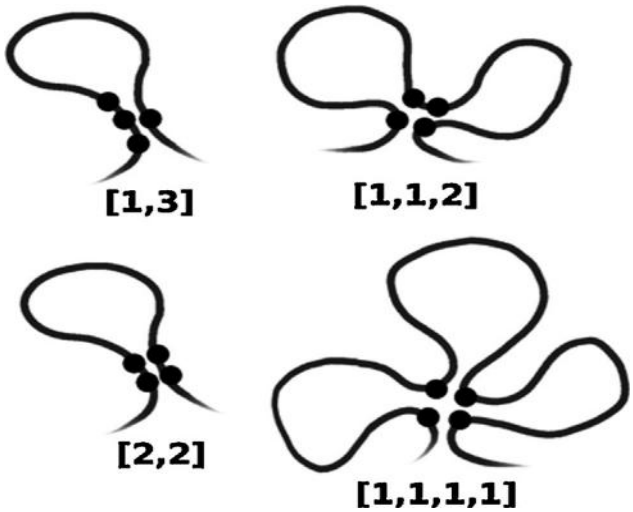

Figure 1. Four Examples of size 4 for maximal cliques. [1,3] has 3 residues contiguous in sequence and 1 separated, [2,2] has two blocks of 2 residues contiguous in sequence, [1,1,2] has three blocks, one with 2 contiguous residues and two with separated residues; [1,1,1,1] four blocks with all separated residues.

RCC have also showed great promise in the prediction of function, mainly at cellular localization and biological process level (unpublished work). Being a single representation that could be used to produce good predictions of both structure classification and function, we decided to explore its usefulness for de Novo protein structure prediction. As a proof of concept, we took random protein foldings and linearly transformed them to the real fold, analyzing how its RCC and RMSD behaved with respect to the real fold (Figure 2).

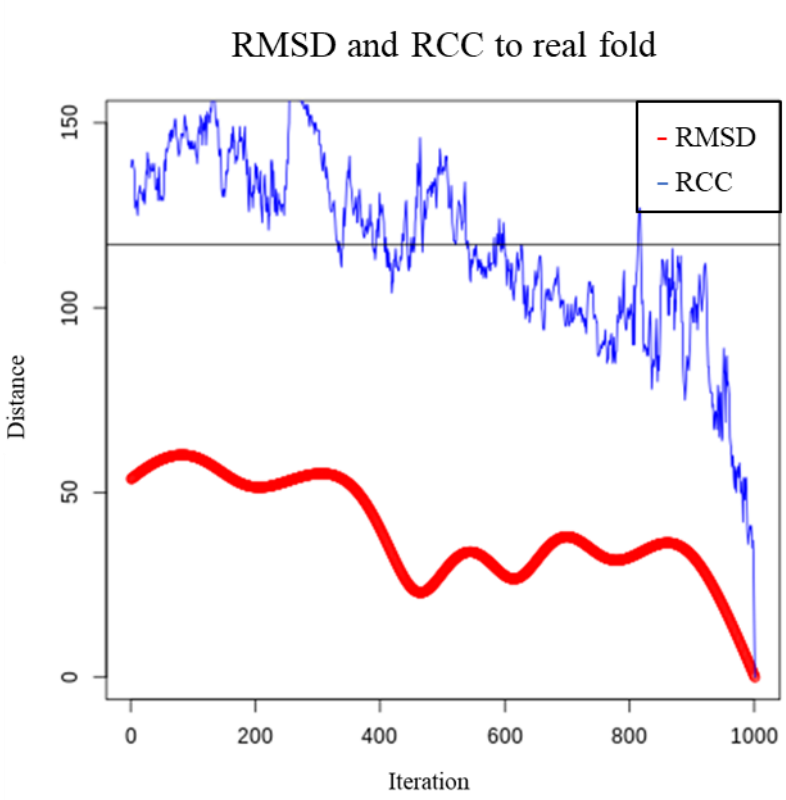

Figure 2. In red the RMSD to the real fold, in blue the distance of its RCC to the real fold.

We implemented an optimization algorithm using as energy the distance of the RCC to the real fold, but this showed us that contrary to what we thought, there are several folds that have the same RCC as the target. Our work now is on two fronts: (i) Do a smart selection of starting points for the optimization algorithm and (ii) Discard candidate folds using different criteria. For (i) we are generating a database of known protein subsequences that we can use as building blocks to jump start the optimization process. For (ii) we aim to analyze each maximal clique and penalize all the ones that are not observed in our database (i.e. not present in known proteins). According to our estimates, front (i) may solve up to 60% of the contacts; since the contact map of proteins do not present the network features required to uniquely solve the localization problem with distance information [2], we will explore by simulation to what degree this front will reduce this problem.

Acknowledgments: CONACyT (CB-252316), C3 Consensus.

[1] Corral-Corral Ricardo, et al. 2015. Machine Learnable Fold Space Representation based on Residue Cluster Classes. DOI: 10.1016/j.compbiolchem.2015.07.010

[2] James Aspnes, Tolga Eren, David K. Goldenberg, A. Stephen Morse, Walter Whiteley, Yang Richard Yang,Brian D.O. Anderson and Peter N. Belhumeur (2006) A Theory of Network Localization. https://ieeexplore.ieee.org/document/1717436.

P-3. Simulations of calcium-ion binding by an amyloid forming peptide

Lipska Agnieszka G.[1], Antoniak Anna [1], Platts Jamie[2,] Sieradzan Adam K.[1]

[1]Faculty of Chemistry, University of Gdańsk, Poland; [2]School of Chemistry, Cardiff University, United Kingdom

Amyloid formation leads to many diseases, such as the Alzheimer and Parkinson diseases, type II diabetes, and a number of systemic amyloidoses [1]. It is hypothesized that calcium cations play important promoting role in formation of deposits [2].

In this study, we performed multiplexed replica exchange molecular dynamics (MREMD) simulations of Aβ1-40 with the use of UNRES force field [3]. Calcium-amino-acid potentials [4] were introduced into the unresf90 software [5] and calculations with and without cations were performed for, a dimer, a trimer and a tetramer of Aβ1-40, respectively. The simulations were carried out in a [6], the box dimensions being 200 Å x 300 Å x 400 Å x 500 Å, in order to maintain proper cation concentration (2 mM). The temperature range in MREMD was 250-480 K. The results of MREMD simulations were processed with the use of the weighted histogram analysis method [7] and cluster analysis, in order to identify the dominant families of conformations. Additionally, the Ca^{2+}...SC and Ca^{2+}...p radial distribution functions, where SC and p denote the side-chain and peptide-group centers, respectively, were calculated. Our results suggest that the presence of cations significantly promotes the formation of associated forms.

Acknowledgments: This research was supported by Preludium 2016/21/N/ST4/03154 (to AGL) form the National Science Centre (Poland).

[1] F. Chiti, C.M. Dobson, Ann. Rev. Biochem., **86,** 27-68, 2017
[2] A. Ahmad, C.M. Staratton, J.L. Scemama, M. Muzaffar, Int. J. Biol. Macromol., **89,** 297-304, 2016
[3] A. Liwo, M. Baranowski, C. Czaplewski et al, J. Mol. Mod., **20**, 2306, 2014
[4] M. Khalili, J.A. Sunders, A. Liwo, S. Ołdziej, H. A. Scheraga, Protein Sci., **13**, 2725-2735, 2004
[5] E.A. Lubecka, A. Liwo, TASK Quaterly, **20**, 399-408, 2016
[6] A.K. Sieradzan, J. Comput. Chem., **36**, 940-946, 2015
[7] S. Kumar, D. Bouzida, R.H. Swendsen, P.A. Kollman, J.M. Rosenberg. Comput. Chem., **8**, 1011-1021, 2001.

P-4. Contact-distance restraints in a prediction of protein structures with UNRES force field

Emilia A. Lubecka[1,2], Agnieszka G. Lipska[2], Adam Liwo[2]

[1]Institute of Informatics, Faculty of Mathematics, Physics and Informatics, University of Gdańsk, Wita Stwosza 57, 80-308 Gdańsk, Poland;

[2]Faculty of Chemistry, University of Gdańsk, Wita Stwosza 63, 80-308 Gdańsk, Poland

Contact-assisted simulations, the contacts being predicted or determined experimentally, have become very important in the determination of the structures of proteins and other biological macromolecules. Recently we have implemented a modified bounded flat-bottom restraint function [1] that does not generate a gradient when a restraint cannot be satisfied [2]. The results of our study demonstrated that contact-distance restraints, even not completely accurate, can significantly improve the quality of the models of protein structures obtained with the coarse-grained UNRES force field [3,4]. We have obtained in this study GDT_TS values for the CASP11 and CASP12 targets usually higher than those obtained at the time of the respective CASP experiments [5], in which an unbounded flat-bottom quartic contact-distance penalty function was used; it should be noted that we carried out the simulations in exactly the same blind-prediction mode as in the CASP experiments. This successful application of the bounded contact-distance function suggests that it can be applied to other problems in which contact-type restraints are used, some of which can turn out to be false, e.g., in the determination of protein structure by NMR or with the aid of chemical cross-link information, etc. From our study it follows that the borderline beyond which model quality starts to improve is about 20 % of true contacts; for a lower percentage of true contacts unrestrained and restrained simulations give comparable results. Our new approach was also used with success in the CASP13 experiment, in which we participated as the wf-BAKER-UNRES group. The average GDT_TS of our CASP13 contact-assisted predictions was by over 15 units higher than those obtained in the CASP11 and CASP12 experiments, whereas, for an *ab initio* mode predictions (the UNRES group), we have obtained over 10 GDT_TS units improvement [6].

Acknowledgments: This work was supported by grant No. UMO-2017/25/B/ST4/01026 from the National Science Center of Poland (Narodowe Centrum Nauki). Computational resources were provided by (a) the Interdisciplinary Center of Mathematical and Computer Modeling (ICM) the University of Warsaw under grant No. GA71-23, (b) the Centre of Informatics – Tricity Academic Supercomputer & Network (CI TASK) in Gdańsk, (c) the Academic Computer Centre Cyfronet AGH in Krakow under grants: casp13unres, casp13unres2, and (d) our 796-processor Beowulf cluster at the Faculty of Chemistry, University of Gdańsk.

[1] A.K. Sieradzan, R. Jakubowski, J. Comput. Chem., **38**, 553-562, 2017.
[2] E.A. Lubecka, A. Liwo, J. Comput. Chem. doi:10.1002/jcc.25847.
[3] C. Czaplewski, S. Kalinowski, A. Liwo, H.A. Scheraga, J. Chem. Theor. Comput., **5**, 627-640, 2009.
[4] A. Liwo, C. Czaplewski, S. Ołdziej, et al., Simulation of protein structure and dynamics with the coarse-grained UNRES force field, in: G. Voth (Ed.), Coarse-Graining of Condensed Phase and Biomolecular Systems, CRC Press, 2008, Ch. 8, pp. 1391-1411.
[5] C. Keasar, L.J. McGuffin, E.B. Wallner, et al., Sci. Rep. **8**, 9939, 2018.
[6] E.A. Lubecka, A.S. Karczyńska, A.G. Lipska, et al., J. Mol. Graph. Model., **92**, 154-160, 2019.

P-5. Efficient classification of protein structure and function with the same representation

Gabriel Del Rio[1], Fernando Fontove[2]

[1]Department of Biochemistry and structural biology, Institute of cellular physiology, UNAM;
[2]C3 Idea

Residue Cluster Class (RCC) is a new feature space to represent proteins that is learnable by any machine-learning algorithm, rendering the best performance to identify structural neighbors [1]. If protein fold is related to protein function, then RCC should excel to represent protein function. Optimizing the efficiency to compute RCCs allowed us to identify that a residue contact graph built without lateral chains and a distance of 7 angstroms improved our previous results on protein fold classification. Preliminary results on protein function classification rendered up to 58% for Cellular localization, 48% for molecular function and 54 % for biological process (considering only proteins with single chains). This efficiency is achieved by including or not the amino acid sidechains. Our protein function prediction efficiency improved the state of the art predictions based on protein sequence, despite our training set included less examples because 3D structure is experimentally harder to obtain than sequence, hence it would seem that RCC best represent protein function than sequence-based approaches. In this regard, it should be noted that protein structure-function relationship based on 3D structural parameters has been shown to improve on sequence-based ones [2-4]. However, protein function prediction based on sequence includes many more functions than we can currently analyze with structure, so these results should not be overestimated. Thus, we provide evidence that the same representation of protein structure or protein function rendered the best classification for protein structure and function. Our results support the notion that the structure and function of proteins are related and such relationship is encoded in RCC.

Acknowledgments: This work was funded under grant CONACYT CB252316 and the Institute of cellular physiology.

References

1. Corral-Corral R, Chavez E, Del Rio G. Machine Learnable Fold Space Representation based on Residue Cluster Classes. Comput Biol Chem. 2015 59 Pt A:1-7.
2. Corral-Corral R, Beltrán JA, Brizuela CA, Del Rio G. Systematic Identification of Machine-Learning Models Aimed to Classify Critical Residues for Protein Function from Protein Structure. Molecules. 2017 Oct 9;22(10). pii: E1673.
3. Cusack MP, Thibert B, Bredesen DE, Del Rio G. Efficient identification of critical residues based only on protein structure by network analysis. PLoS One. 2007 2(5):e421.
4. Thibert B, Bredesen DE, del Rio G. Improved prediction of critical residues for protein function based on network and phylogenetic analyses. BMC Bioinformatics. 2005 Aug 26;6:213.

P-6. Extension of the force matching method to anisotropic coarse-grained transferable force fields: application to the UNRES model of proteins

Adam Liwo and Cezary Czaplewski

Faculty of Chemistry, University of Gdańsk, Wita Stwosza 63, 80-308 Gdańsk, Poland.

Coarse-grained (CG) models of biological macromolecules are nowadays extensively used in simulations because of tremendous extension of the time- and size-scale that they offer. To provide the link between coarse-grained and all-atom dynamics, the Force Matching (FM) [1] and, further, the Multi-Scale Coarse-Graining (MSCG) approaches to biomolecular systems have been developed [2], in which the forces computed at the all-atom level are averaged over the degrees of freedom omitted from a coarse-grained representation. However, because isotropic site-site interactions have been assumed, aggressive coarse graining is not possible and the resulting force fields are not transferable. We propose a new approach to FM, which overcomes both problems. Coarse-rained anisotropic forces are calculated from all-atom forces by imposing the constraint that the relative spacing of the atoms that belong to a CG site along the virtual-bond vector of this site does not change, this implying concerted motion of these atoms when one of the anchor point of the virtual-bond vector of that site moves. To achieve transferability, the factor-expansion approach developed in our earlier work [3] has been implemented, in which the PMF of a system is expanded into Kubo cluster-cumulant functions [4], termed factors, that can be identified with the respective energy terms; the mean forces are partitioned likewise. The new FM variant has been applied to our highly reduced physics-based UNRES model of polypeptide chains [5], in which only united peptide groups and united side chains are the interaction sites. The training protein was the 20-residue tryptophan cage miniprotein (PDB: 1L2Y) for which all-atom dynamics simulations were carried out with the AMBER 14 force field [6] with implicit solvent; the miniprotein folded and unfolded several times during the course of simulations whose total length was 0.1 μs. A total of 50,000,000 snapshots were collected from the trajectory and used to re-calibrate the recent NEWCT-4P scale-consistent variant of the UNRES force field [7]. The complete target function consisted of the force-matching and maximum-likelihood term [8], which drives the modeled conformational ensemble(s) close to the NMR- determined ensembles. Enegy-term weights and the radii and anisotropies of the united side chains were optimized. The resulting force field produced fewer clashes for the conformations of proteins reduced from the all-atom to the UNRES representation, while retaining the capacity of folding proteins in the ab initio mode. Optimization with the FM term alone did not produce a predictive force field.

Supported by grant UMO-2017/25/B/ST4/01026 from the National Science Centre (NCN). Computational resources were provided by the Informatics Center in Gdańsk TASK and Interdisciplinary Center of Mathematical and Computer Modeling (ICM), University of Warsaw, grant GA76-11.

1. Izvekov, S & Voth GA (2005) *J Phys Chem B* **109**, 2469-2473
2. Das, A *et al* (2012) *J Chem Phys* **136** 194115
3. Sieradzan AK *et al* (2017) *J Chem Phys* **146** 124106
4. Kubo, R (1962) *J Phys Soc Japan* **17** 1100-1120
5. Liwo, A *et al* (2014) *J Molec Model* **20** 2306
6. Case, D.A. et al. (2005) *J. Comput. Chem.* **26,** 1668-1688
7. Liwo, A. et al. (2019) J. Chem. Phys., **150**, 155104.
8. Zaborowski, B. et al. (2015) *J. Chem. Inf. Model.*, **55**, 2050–2070

Participants (alphabetic order)

#	Name	Affiliation	E-mail
1	Arciniega, Marcelino	Univesidad Autonoma de Mexico	marciniega(at)fc.unam.mx
2	Bidone, Tamara	University of Utah	tamarabidone(at)sci.utah.edu
3	Boniecki, Michał	International Institute of Molecular and Cell Biology in Warsaw	mboni(at)genesilico.pl
4	Brizuela, Carlos	Centro de Investigación Científica y de Educación Superior de Ensenada	cbrizuel(at)cicese.mx
5	Chirikjian, Gregory	National University of Singapore and Johns Hopkins University	gchirik1(at)jhu.edu
6	Cieplak, Marek	Polish Academy of Science	mc(at)ifpan.edu.pl
7	Del Rio, Gabriel	Universidad Nacional Autónoma de México	gdelrio(at)ifc.unam.mx
8	Fontove, Fernando	C3 Idea	fernando.fontove(at)c3consensus.com
9	He, Yi	University of New Mexico	yihe(at)unm.edu
10	Hernández Rosales, Maribel	Universidad Nacional Autónoma de México	maribel(at)im.unam.mx
11	Hyeon, Changbong	Korea Institute for Advanced Study	hyeoncb(at)kias.re.kr
12	Jernigan, Robert	Iowa State University	jernigan(at)iastate.edu
13	Karczynska, Agnieszka	University of Gdansk	agneska.kar(at)gmail.com
14	Kihara, Daisuke	Purdue University	dkihara(at)purdue.edu
15	Kloczkowski, Andrzej	Nationwide Children's Hospital	andrzej.kloczkowski(at)nationwidechildrens.org
16	Lipska, Agnieszka	University of Gdańsk	lypstyk(at)gmail.com
17	Liwo, Jozef Adam	University of Gdansk	adam.liwo(at)ug.edu.pl
18	Lubecka, Emilia	University of Gdańsk	emilia.lubecka(at)ug.edu.pl
19	Ozkan, Banu	Arizona State University	banu.ozkan(at)asu.edu
20	Pastor, Nina	Universidad Autonoma del Estado de Morelos	nina(at)uaem.mx
21	Shahzad, Muhammad Adnan	University of Swat	muhammad.shahzad(at)unicam.it
22	Vakser, Ilya	University of Kansas	vakser(at)ku.edu
23	Wu, Zhiyun	Iowa State University	zhijun(at)iastate.edu