BioGeometry NEWS

August
2003
biogeometry.cs.duke.edu

Education

An important component of education in computational structural biology is the development and delivery of topical courses. We present short descriptions of five of the more mature such courses.

SSB 228: Computational Structural Biology: Protein Simulation and Structure Prediction (Winter 2001, an online course in subsequent semesters; Levitt; Stanford)

Computers now play a central role in biology, in that they provide platforms to simulate and analyse complex biological systems and processes. Nature is still far ahead of what we can do with computers: for example, while protein sequences spontaneously fold into their native structures, simulation of this process is still well beyond present computer power. The central goal of the computational structural biology course SB228 is to provide an overview of the current impact of computers on our understanding of protein structure and dynamics.

The main topics discussed during the course include: I. Principles of protein structures; II. Energetics of protein structures; III. Protein simulations; IV. Protein Structure Prediction; IV. Large scale genomics projects.

Modeling softwares, modeling web servers, as well as state-of-the-art computer experiments on solving the protein structure prediction problems, are discussed.

During winter 2001, the ten plenary lectures of SB228 were filmed, and stored on video by the Stanford Center for Professional Development. Since then, SB228 has been offered as an online course. The online version of the course consists of two parts: the recorded lectures, and the online material (handouts and assignments).

http://csb.stanford.edu/class

CPS 296: Bio-Geometric Modeling (Fall 2002; Edelsbrunner; Duke)

This is a one semester graduate seminar course that teaches geometric methods used in the simulation and analysis of biomolecular systems and processes. Sample questions are the representation of protein structures, the docking of proteins, and the measuring of surface area and volume. While the course focuses on relatively recently developed combinatorial methods, the broader goal is the synthesis with classical numerical methods to form the concrete (continuous/discrete) foundations for a broad computational approach to understand life on a molecular level.

The course synthesizes concepts from algorithms, geometry and topology and includes the following main topics: I. Biomolecules; II. Geometric Models; III. Surface Meshing; IV. Connectivity; V. Shape Features; VI. Density Maps; VII. Match and Fit; VIII. Measures; IX. Derivatives.

The course material has been distilled into a set of notes that are available on the web. The notes form the basis of a future textbook on the topic.

http://www.cs.duke.edu/education/courses/fall02/cps296.1/

COMP 006: Folding: From Paper to Proteins (Fall 2002; Snoeyink; UNC)

This course, one of several First Year Seminar courses in Computer Science, is held every other year. First Year Seminar courses at UNC are designed to engage first year students in active learning, and to expose them to current research, with a focus on research methodologies as well as results.

We cover four areas of folding (paper, polyhedra, linkages, and proteins), looking both at the results in the individual areas, and at how they relate. These reflect Dr. Snoeyink's research interest in computational geometry, an area of computer science that uses geometric techniques to solve computational problems in other fields. Although the course does not teach programming, we try to teach thought processes that lead to solving such problems.

This course is designed to appeal to a wide variety of students with different backgrounds and abilities. We use diverse teaching methods, ranging from student driven discussions to lecture style teaching. In the origami unit, for example, students folded various objects (including animals and modular polyhedra), then unfolded some to observe the crease patterns formed. We attended the Southeast Origami Festival in Charlotte, NC, and some of the students taught the others what they learned. Students wrote reports on either the festival, or on an origami math research paper. We related paper folding to physics by discussing why paper airplanes fly, students designed their own airplanes. In the protein unit, we explored various interesting proteins (i.e. silk, wool and hemoglobin).

CPS 260: Algorithms in Computational Biology (Fall 2003; Agarwal; Duke)

This course, cross-listed as BGT 204 in the Bioinformatics and Genome Technology PhD program administered by Center for Bioinformatics and Computational Biology, is intended to provide a systematic introduction to the algorithmic techniques behind the most commonly used tools in computational biology. While the course will survey a wide range of methods and tools in the field, the primary emphasis will be on understanding and analyzing the algorithms behind these tools.

Topics to be covered include basic techniques in design and analysis of algorithms, dynamic programming, string matching, probabilistic techniques, hidden Markov models, geometric algorithms, and data mining. These topics will be exposed in the context of applications of sequence alignment, genome sequence assembly, gene finding, gene expression and regulation, protein structure prediction, protein dynamics.

http://www.cs.duke.edu/education/courses/fall03/cps260

Education, continued

COMP 290-079: Applied Optimization in Computational Biology (Fall 2003; Snoeyink; UNC)

This is a one credit seminar/working lab that will use case studies in specific biological problems of interest to the class members to motivate algorithmic methods (e.g. dynamic programming), software engineering (e.g., code factoring and tuning), and languages and tools (e.g., Perl, MATLAB, debuggers, profilers).

This is not a replacement for introductory courses in computer science. We assume that each student can write some sort of program, and that some students will want to learn more sophisticated techniques that apply to their particular problems, while other students will want to learn sophisticated problems that might be amenable to their particular techniques.

This course is drawing strong interest from both life sciences and computer science students.

http://www.cs.unc.edu/~snoeyink/
comp290opt/syllabus.htm

Software

Jean-Claude Labombe's group has posted two software systems on the web:

- Software to perform efficient Monte Carlo simulation of proteins is available at http://robotics.stanford.edu/~itayl/mcs/

- Stochastic Roadmap Simulation (SRS), software designed to compute ensemble properties of molecular motions (folding, binding), is available at http://robotics.stanford.edu/~apaydin/software.html

Student Projects

SHANTANU SHARMA, undergraduate student, Indian Institute of Technology, Kanpur, worked on a summer project with Jeff Roach and Charlie Carter at UNC.

Shantanu completed some initial, exploratory work on a new method of protein structure comparison based on the Delaunay tetrahedralization. He showed that the particular tetra-hedralization with which he was working can be expressed as a sequence of integers that preserves much of the structural information present in the atomic coordinates. In particular he identified subsequences that indicate which residues participate in alpha helices and beta sheets. This information has been combined into a structural alignment procedure and similarity measure that we will continue to develop.

MATTHEW MIAN, Howard Hughes Summer Undergraduate Student, worked this summer with Herbert Edelsbrunner and Johannes Rudolph at Duke.

The aim of the project was to investigate the docking of small molecules (drugs or substrates) to proteins using shape complementarity alone. The goal we set ourselves was to determine to what extent our successes for protein-protein docking can be extended to the docking of small molecules to proteins.

For our test data we chose 29 different ligand-protein complexes from the PDB. The ligand size varied from 9 to 46 heavy atoms, and the proteins belonged to a variety of classes. We began our initial search using score parameters and density of search space identical to our standard protein-protein docking values. Calculations took 30 min to 4 hours on the cluster of 80 PCs.

As shown in the table, correct docking solutions were achieved as the top-scoring complex in 18 of 29 cases (62%). In all 18 of these complexes, the ligand size was greater or equal to 28 heavy atoms and no complexes with ligand size less than 26 heavy atoms were docked correctly. This suggests that a minimum ligand size is required to achieve a shape complementarity that suffices for docking prediction. In an experiment to extend this minimum size (and perhaps the geometric complementarity), we added hydrogen atoms to five of the proteins that had failed to dock correctly. As seen in the Table, correct docking solutions were found within the top 5 scores for 4 of 5 complexes, extending correct docking solutions down to 18 heavy atoms. Thus, addition of hydrogens may be useful for smaller interaction surfaces, although it is computationally costly and comes at the expense of false positives. Experiments were also performed to systematically vary the distance parameter lambda used in the scoring function as well as the density of the rotational and translational space. Our current parameters are close to optimal for finding the correct solution for cases without hydrogens (data not shown) and further experimental variations of lambda are planned for cases with hydrogens added.

PDB	Protein	Ligand	Ligand Size	RMSD (top score)	RMSD (with H's) (top 5 scores)
1A4R	Cdc42	GDP	28	0.7	--
1A4R	Cdc42	GNH	28	1.0	--
1A2B	RhoA	GSP	32	0.7	--
1AS0	Gia1	GSP	32	0.5	--
1CLU	H-Ras	DBG	33	0.4	--
1JAH	H-Ras	GCP	32	0.7	--
1RVD	H-Ras	DBG	33	0.7	--
121P	H-Ras	GTO	32	0.6	--
1AQ1	Cdk2	STU	36	1.6	--
1CKP	Cdk2	PVB	26	20	13.1, 0.7, 0.8, 2.9, 1.0
1DI8	Cdk2	DTQ	22	20	7.3, 7.3, 1.0, 0.5, 1.9
1DM2	Cdk2	HMD	19	9.4	1.0, 1.8, 0.6, 5.3, 1.2
1E1V	Cdk2	CMG	18	13	4.7, 4.6, 4.5, 22, 4.6
1GZ8	Cdk2	MBP	18	9.0	1.6, 22, 0.8, 0.6, 13
1B39	Cdk2	ATP	31	0.9	--
1DWC	a-thrombin	MIT	31	0.8	--
1DWD	a-thrombin	MID	37	0.9	--
1TMN	thermolysin	CLT	12	25	--
2CTC	carboxypeptidase	LOF	12	28	--
3PTB	trypsin	BEN	9	21	--
1D4S	HIV protease	TPV	42	0.9	--
1HSG	HIV protease	MK1	45	1.0	--
1NPW	HIV protease	LGZ	46	0.6	--
1IEP	cAbl	STI	37	0.7	--
1FPU	cABL	PRC	29	0.7	--
1DB1	vitamin D receptor	VDX	30	1.5	--
1E3G	androgen receptor	R18	20	13	--
1ULB	phosphorylase	GUN	11	29	--
2PHH	PHBH	PHB	10	16	--

The BioGeometry project is funded by the National Science Foundation under grant CCR-00-86013. Any opinions, findings, and conclusions or recommendations expressed in this newsletter are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.
BioGeometry News is the project's monthly newsletter. For more information, please visit http://biogeometry.cs.duke.edu/newsletter.