CSE 350/450: Structural Bioinformatics

Location: Packard Lab 416, Tuesdays and Thursdays, 1:10 pm - 2:25 pm
Professor: Brian Y. Chen, Department of Computer Science and Engineering, Lehigh University

---

1. Course Description

Solving problems at the leading edge of medical and industrial technologies depends, in many cases, on improving our understanding of protein function. For example, the debilitating side effects of cancer treatment could be reduced by developing drug molecules that selectively fit the unique structures of cancer proteins. In this case, and in many others, protein shape can yield many deep insights into how proteins function. This course is an exploration, through collaborative and interdisciplinary projects, of the biological, computational, and statistical ideas developed for protein structure alignment, finding functional sites, structure-function inference, molecular surfaces, and evolution in protein structure.

2. Textbook

Structural Bioinformatics, 2nd Edition
Edited by Jenny Gu, and Philip E. Bourne.
Publisher: Wiley-Blackwell, 2009.

3. Lecture Topics and Readings

Week 1: Introduction

Lecture 1: Introduction to Structural Bioinformatics
Lecture 2: Introduction to Proteins
Gu and Bourne, Chapter 1, 2, 4. Optional: Chapter 3.

Week 2: Motif Design and Geometric Matching

Lecture 3: Motif Design
Lecture 4: Geometric Matching
Gu and Bourne, Chapter 10, 11, 21.
Algorithms for structural comparison and statistical analysis of 3D protein motifs [1].

Week 3: Modeling Matching Catalytic Sites in Protein Structure Data

Lecture 5: Match Scoring
Gu and Bourne, Chapter 8
A Statistical Model to Correct Systematic Bias Introduced by Algorithmic Thresholds in Protein Structural Comparison Algorithms [2].
An algorithm for constraint-based structural template matching: application to 3D templates with statistical analysis [3].
A Model for Statistical Significance of Local Similarities in Structure [4].
Inferring Functional Relationships of Proteins from Local Sequence and Spatial Surface Patterns [5].
A New Method to Detect Related Function Among Proteins Independent of Sequence and Fold Homology [6].

Week 4: Algorithmic Motif Design

Lecture 6: Algorithmic Motif Design

Lecture 7: Composite motifs
The MASH pipeline for protein function prediction and an algorithm for the geometric refinement of 3D motifs [7].
Cavity-aware motifs reduce false positives in protein function prediction [8].
Composite motifs integrating multiple protein structures increase sensitivity for function prediction [9].
Analysis of substructural variation in families of enzymatic proteins with applications to protein function prediction [10]

Week 5: Protein Structure Alignment

Lecture 8: Whole Structure Alignment
Lecture 9: Geometric Hashing and GRATH
Gu and Bourne, Chapters 16, 17, 18.
Mapping the Protein Universe [11]
Efficient Alignment of 3d structures with Geometric Hashing [12]
Protein structure comparison using iterated double dynamic programming [13]
Protein structure alignment by incremental combinatorial extension (CE) of the optimal path [14]
Quantifying the Similarities in Fold Space [15]

Week 6: Multiple Structure Alignment

Lecture 10: Multiple Structure Alignment
Lecture 11: Applications of Multiple Structure Alignment
Gu and Bourne, Chapter 23.
Multiple Structural Alignment by Optimal RMSD Implies that the Average Structure is a Consensus [16]
Automated Multiple Structure Alignment and Detection of a Common Substructural Motif [17]
A Method for Simultaneous Alignment of Multiple Protein Structures [18]
MC-CE: Monte Carlo Combinatorial Extension [19]
Structural Evolution of the Protein Kinase-like Superfamily [20]
Multiple Flexible Structural Alignment with POSA [21]

Week 7: Protein Electrostatics

Lecture 12: Protein Electrostatics
Gu and Bourne, Chapter 24.
On the Role of Electrostatic Interactions in the Design of Protein-Protein Interactions [22]
Focusing of Electric Fields in the Active Site of Superoxide Dismutase [23]
On the impact of desolvation versus electrostatic attraction [24]
Rapid Electrostatically assisted association of proteins [25]

Week 8: Analyzing Molecular Surfaces

Lecture 13: Alpha Shapes and Protein Cavities
Lecture 14: Molecular Surfaces and Surface Analysis
Analytical Shape Computation of Macromolecules I: alpha shapes [26]
Analytical Shape Computation of Macromolecules II: protein cavities [27]
Are Proteins Well packed? [28]
SURFNET: A program for visualizing molecular surfaces, cavities, and intermolecular interactions [29]
A method for localizing ligand binding pockets in protein structures. [30]
Protein Clefts in Molecular Recognition and function [31]
Travel Depth: A new shape descriptor for macromolecules [32]
Identifying protein binding pockets with PASS [33]

Week 9: Analyzing Molecular Volumes

Lecture 15: Volumetric Analysis of Protein Surfaces
Lecture 16: Using Machine Learning to classify functional sites
VASP: A Volumetric Analysis of Surface Properties Yields Insights into Protein-Ligand Binding Specificity file [34]
Hotpatch: A Statistical Approach to Finding Biologically Relevant Features on Protein Surfaces [35]
SCREEN: Finding Druggable Cavities with Random Forests [36]

Week 10: Protein-Protein Interactions

Lecture 17: Protein-Protein Interactions
Lecture 18: Predicting Protein Protein Interactions and Hotspots
Gu and Bourne, Chapter 26
The Atomic Structure of Protein-Protein Recognition Sites [37]
Protein-Protein Interactions: Hot Spots and Structurally Conserved Residues Often Locate in Structurally Complemented Pockets that Pre-Organized in the Unbound State: Implications for Docking [38]
Protein interface conservation across structure space [39]
Segmenting Motifs in Protein-Protein Interface Surfaces [40]

Week 11: Protein-DNA Interactions

Lecture 19: Protein-DNA Interactions
Lecture 20: Predicting Protein-DNA Interactions
Gu and Bourne, Chapter 25.
Recognition of Specific DNA sequences [41]
The Role of DNA shape in protein-DNA recognition [42]
Structure-based Prediction of C2H2 Zing-Finger binding specificity: Sensitivity to docking geometry [43]
Structural alignment of protein-DNA interfaces: insights into the determinants of binding specificity [44]
Exploring the DNA-binding specificities of zinc fingers with DNA microarrays [45]

Week 12: Molecular Simulation and Docking

Lecture 23: Molecular Simulation
Gu and Bourne, Chapter 27, 37
Molecular Dynamics Simulations in Biology [46]
Energetics of Ion Conduction through the K+ channel [47]

Week 13: Computational Drug Design

Lecture 21: Structure Based Drug Design
Gu and Bourne, Chapter 34
Evaluating the Substrate-Envelope Hypothesis: Structural analysis of Novel HIV-1 Protease Inhibitors Designed to be Robust against Drug Resistance [48]

Week 14: Quantative Structure-Activity Relationships

Lecture 22: From Hits to leads
The role of quantitative structure-activity relationships (QSAR) in biomolecular drug discovery [49]
Hit and Lead generation: beyond high throughput screening [50]

Rationalizing fragment based drug discovery for BACE1: insights from FB-QSAR, FB-QSSR, multi objective (MO-QSPR) and MIF studies [51]
Virtual screening for R-groups, including predicted pIC50 contributions, within large structural databases, using Topomer CoMFA [52]
Rethinking 3D-QSAR [53]

Week 15: Protein Structure Prediction

Lecture 24: Knowledge-based Protein Structure Prediction
Lecture 25: Ab Initio Protein Structure Prediction
Gu and Bourne, Chapters 28-32
Comparative Protein Modelling by Satisfaction of Spatial Constraints [54]
On the Role of Structural Information in Remote Homology Detection and Sequence Alignment: New Methods Using Hybrid Sequence Profiles [55]
Solving and analyzing side-chain positioning problems using linear and integer programming [56]
Assembly of Protein Tertiary Structures from Fragments with Similar Local Sequences using Simulated Annealing and Bayesian Scoring Functions [57]
Improved Recognition of Native-Like Protein Structures Using a Combination of Sequence-Dependent and Sequence-Independent Features of Proteins [58]

4. Course Structure

Semester Project (CSE350: 90%) (CSE450: 70%)

The semester project will be based on a prototype bioinformatics pipeline with biological, algorithmic and statistical modules. For their semester projects, students, working alone or in interdisciplinary groups, will retrofit the prototype with one or more modules. See the project wiki for details.

Class Participation (CSE350/450: 10%)

a) **Actively starting and/or participating in discussions** during the lecture periods that connect the topic of the lecture to a major field in the course can significantly improve your participation grade. Example: Suppose the lecture is covering protein structure alignment. Questions like "is protein structure alignment used in current biological research?", "It seems like there are an infinite number of possible structural alignments, how is an optimal alignment even possible?", "Is an optimal structural alignment even a biologically valid piece of information?" are discussion-starters that will be noted for class participation. Participating in these discussions will, likewise, be noted. There is no distinction in score between starting and participating in a discussion. Do not be concerned about interrupting the lecture, until the lecturer says it is time to move on from the discussion. For full credit, students are expected to ask one question or participate in one discussion per lecture.

b) **Attendance.** In a collaborative course like CSE 350, participation is key for both your own benefit and that of those around you. If you do not attend the course, then you deny yourself access to the knowledge of those around you, and you deny those around you a chance to learn from your experience. For this reason, attendance is very important. This is an advanced course, and, for efficiency, attendance will not be taken. If the lecturer notices that you are not in class, you will lose participation points. Obtain approval from the lecturer in advance for planned absences, which will not count against you.

c) **Cell phones and laptops** are to be set on silent mode. Sounds from any electronic device, including the vibration of a device against hard surfaces, are considered interruptions and will thus affect participation grades. However, lectures will pause for questions to be answered, and those questions may turn into discussions. Typing during lecture is considered an interruption, but during discussions, it is permissible, presumably because you are looking up information to add to the discussion and boost your participation grade. Tactful use of electronic devices respects the interests of your peers.

Review Paper (CSE450: 20%)
You will write a review of the current research relating to one of the topics covered in class. Inform the instructor of what topic you have chosen to do. See further descriptions below.

5. Prerequisites

This course draws from three primary subjects: biochemistry, algorithms, and statistics. Knowledge of all subjects is unnecessary, but a familiarity with at least one of these subjects is required. Having taken courses in one of the following groups is highly recommended, though only instructor permission is necessary to enroll. Students will be asked to state which set(s) of prerequisites they fulfill, and be thus associated with one of the primary subjects.

Group 1: BIOS 371+372 or CHM 371+372
Group 2: MATH 205, CSE 109, CSE 340
Group 3: MATH 312, MATH 334.

A background in programming is not required for the biological module and only light programming is necessary for the statistical module. Thus, programming is not required to complete the course with full marks. Likewise, a background in biology is not required for the computational or statistical modules, and a background in statistics is not required for the biological and computational modules. However, integrating the knowledge gained about other fields, from other students, the lectures, and the textbook, can provide significant extra credit.

6. Assessment

For the project, full credit is possible for completing, alone, a module of the pipeline (see course structure) in the student's declared background. Students working alone on a module outside of their declared background are eligible for up to 15% extra credit, to recognize the difficulty of working on material outside the student's background.

Groups can have at most three students, and must complete one module for each student in the group. Modules completed by a group of students cannot be of the same type - i.e. two biological modules for a group of two students is disallowed. Students integrating multiple modules as part of a group will be eligible for up to 5% extra credit for each integrated module, and thus a maximum of 15%. The instructor will divide dysfunctional groups into individuals, with individual modules, if necessary.

Additional Requirements for CSE450

Students taking CSE450 must prepare a 7-10 page (single spaced, 12 point font, 1 inch margins, including citations.) review paper on the current research relating to one of the fields discussed in class. CSE450 Reports are individual work, and not to be discussed with other students.

7. Outcomes

By completing this course, students will:

1. Understand the basic design and purpose of several major computational technologies in the field of structural bioinformatics
2. Be aware of how biological, algorithmic, and statistical concepts can be integrated to draw meaningful conclusions from multi-faceted biological data.
3. Have experience in the implementation challenges relating to these major technologies.
4. Have experience in technical communication with collaborators with technical expertise outside of their own field.

This course supports program missions to educate students that will:

1. Apply their education in computer science to the analysis and solution of scientific, business, and industrial problems.
3. Function effectively in a collaborative team and effectively communicate with members of the team.
4. Engage in continued education in their field of expertise

Students with Disabilities:

If you have a documented learning disability, and will be requesting academic accommodation for this class, please contact Dean Cheryl Ashcroft in the Office of the Dean of Students, UC 212, at x84152, or by email at caa4@lehigh.edu. She will establish the appropriate accommodations for your case.

Lehigh Student Senate Academic Integrity Statement:

We, the Lehigh University Student Senate, as the standing representative body of all undergraduates, reaffirm the duty and obligation of students to meet and uphold the highest principles and values of personal, moral and ethical conduct. As partners in our educational community, both students and faculty share the responsibility for promoting and helping ensure an environment of academic integrity. As such, each student is expected to complete all academic course work in accordance to the standards set forth by the faculty and in compliance with the university's Code of Conduct.

8. Course Readings

[1] Algorithms for structural comparison and statistical analysis of 3D protein motifs. Chen BY, Fofanov VY, Kristensen DM, Kimmel M, Lichtarge O, Kavraki LE. Pac Symp Biocomput. 2005:334-45.

[2] A Statistical Model to Correct Systematic Bias Introduced by Algorithmic Thresholds in Protein Structural Comparison Algorithms. Fofanov VY*, Chen BY*, Bryant DH, Moll M,

Lichtarge O, Kavraki LE, Kimmel M. Computational Structural Bioinformatics Workshop, IEEE International Conference on Bioinformatics and Biomedicine (BIBM 2008).

[3] An algorithm for constraint-based structural template matching: application to 3D templates with statistical analysis. Barker JA, Thornton JM. Bioinformatics. 2003 Sep 1;19(13):1644-9.

[4] A model for statistical significance of local similarities in structure. Stark A, Sunyaev S, Russell RB. J Mol Biol. 2003 Mar 7;326(5):1307-16.

[5] Inferring functional relationships of proteins from local sequence and spatial surface patterns. Binkowski TA, Adamian L, Liang J. J Mol Biol. 2003 Sep 12;332(2):505-26.

[6] A new method to detect related function among proteins independent of sequence and fold homology. Schmitt S, Kuhn D, Klebe G. J Mol Biol. 2002 Oct 18;323(2):387-406.

[7] The MASH pipeline for protein function prediction and an algorithm for the geometric refinement of 3D motifs. Chen BY, Fofanov VY, Bryant DH, Dodson BD, Kristensen DM, Lisewski AM, Kimmel M, Lichtarge O, Kavraki LE. J Comput Biol. 2007 Jul-Aug;14(6):791-816.

[8] Cavity-aware motifs reduce false positives in protein function prediction. Chen BY, Bryant DH, Fofanov VY, Kristensen DM, Cruess AE, Kimmel M, Lichtarge O, Kavraki LE. Comput Syst Bioinformatics Conf. 2006:311-23.

[9] Composite motifs integrating multiple protein structures increase sensitivity for function prediction. Chen BY, Bryant DH, Cruess AE, Bylund JH, Fofanov VY, Kristensen DM, Kimmel M, Lichtarge O, Kavraki LE. Comput Syst Bioinformatics Conf. 2007;6:343-55.

[10] Analysis of substructural variation in families of enzymatic proteins with applications to protein function prediction. Bryant DH, Moll M, Chen BY, Fofanov VY, Kavraki LE. BMC Bioinformatics. 2010 May 11;11:242.

[11] Mapping the protein universe. Holm L, Sander C. Science. 1996 Aug 2;273(5275):595-603.

[12] Efficient detection of three-dimensional structural motifs in biological macromolecules by computer vision techniques. Nussinov R, Wolfson HJ. Proc Natl Acad Sci U S A. 1991 Dec 1;88(23):10495-9.

[13] Protein structure comparison using iterated double dynamic programming. Taylor WR. Protein Sci. 1999 Mar;8(3):654-65.

[14] Protein structure alignment by incremental combinatorial extension (CE) of the optimal path. Shindyalov IN, Bourne PE. Protein Eng. 1998 Sep;11(9):739-47.

[15] Quantifying the similarities within fold space. Harrison A, Pearl F, Mott R, Thornton J, Orengo C. J Mol Biol. 2002 Nov 8;323(5):909-26.

[16] Multiple Structure Alignment by Optimal RMSD Implies that the Average Structure is a Consensus. Wang X, Snoeyink J. Comput Syst Bioinformatics Conf. 2006;5:79-87.

[17] Automated multiple structure alignment and detection of a common substructural motif. Leibowitz N, Fligelman ZY, Nussinov R, Wolfson HJ. Proteins. 2001 May 15;43(3):235-45.

[18] A method for simultaneous alignment of multiple protein structures. Shatsky M, Nussinov R, Wolfson HJ. Proteins. 2004 Jul 1;56(1):143-56.

[19] CE-MC: a multiple protein structure alignment server. Guda C, Lu S, Scheeff ED, Bourne PE, Shindyalov IN. Nucleic Acids Res. 2004 Jul 1;32(Web Server issue):W100-3.

[20] Structural evolution of the protein kinase-like superfamily. Scheeff ED, Bourne PE. PLoS Comput Biol. 2005 Oct;1(5):e49. [21] Multiple flexible structure alignment using partial order graphs. Ye Y, Godzik A. Bioinformatics. 2005 May 15;21(10):2362-9.

[22] On the role of electrostatic interactions in the design of protein-protein interfaces. Sheinerman FB, Honig B. J Mol Biol. 2002 Apr 19;318(1):161-77.
[23] Focusing of electric fields in the active site of Cu-Zn superoxide dismutase: effects of ionic strength and amino-acid modification. Klapper I, Hagstrom R, Fine R, Sharp K, Honig B. Proteins. 1986 Sep;1(1):47-59.
[24] Computation of electrostatic complements to proteins: a case of charge stabilized binding. Chong LT, Dempster SE, Hendsch ZS, Lee LP, Tidor B. Protein Sci. 1998 Jan;7(1):206-10.
[25] Rapid, electrostatically assisted association of proteins. Schreiber G, Fersht AR. Nat Struct Biol. 1996 May;3(5):427-31.
[26] Analytical shape computation of macromolecules: I. Molecular area and volume through alpha shape. Liang J, Edelsbrunner H, Fu P, Sudhakar PV, Subramaniam S. Proteins. 1998 Oct 1;33(1):1-17.
[27] Analytical shape computation of macromolecules: II. Inaccessible cavities in proteins. Liang J, Edelsbrunner H, Fu P, Sudhakar PV, Subramaniam S. Proteins. 1998 Oct 1;33(1):18-29.
[28] Are proteins well-packed? Liang J, Dill KA. Biophys J. 2001 Aug;81(2):751-66.
[29] SURFNET: a program for visualizing molecular surfaces, cavities, and intermolecular interactions. Laskowski RA. J Mol Graph. 1995 Oct;13(5):323-30, 307-8.
[30] A method for localizing ligand binding pockets in protein structures. Glaser F, Morris RJ, Najmanovich RJ, Laskowski RA, Thornton JM. Proteins. 2006 Feb 1;62(2):479-88.
[31] Protein clefts in molecular recognition and function. Laskowski RA, Luscombe NM, Swindells MB, Thornton JM. Protein Sci. 1996 Dec;5(12):2438-52.
[32] Travel depth, a new shape descriptor for macromolecules: application to ligand binding. Coleman RG, Sharp KA. J Mol Biol. 2006 Sep 22;362(3):441-58.
[33] Fast prediction and visualization of protein binding pockets with PASS. Brady GP Jr, Stouten PF. J Comput Aided Mol Des. 2000 May;14(4):383-401.
[34] VASP: a volumetric analysis of surface properties yields insights into protein-ligand binding specificity. Chen BY, Honig B. PLoS Comput Biol. 2010 Aug 12;6(8). pii: e1000881.
[35] HotPatch: a statistical approach to finding biologically relevant features on protein surfaces. Pettit FK, Bare E, Tsai A, Bowie JU. J Mol Biol. 2007 Jun 8;369(3):863-79.
[36] On the nature of cavities on protein surfaces: application to the identification of drug-binding sites. Nayal M, Honig B. Proteins. 2006 Jun 1;63(4):892-906.
[37] The atomic structure of protein-protein recognition sites. Lo Conte L, Chothia C, Janin J. J Mol Biol. 1999 Feb 5;285(5):2177-98.
[38] Protein-protein interactions: hot spots and structurally conserved residues often locate in complemented pockets that pre-organized in the unbound states: implications for docking. Li X, Keskin O, Ma B, Nussinov R, Liang J. J Mol Biol. 2004 Nov 26;344(3):781-95.
[39] Protein interface conservation across structure space. Zhang QC, Petrey D, Norel R, Honig BH. Proc Natl Acad Sci U S A. 2010 Jun 15;107(24):10896-901.
[40] Segmenting Motifs in Protein-Protein Interface Surfaces. Phillips JM, Rudolph J, Agarwal, PK. Proceedings of the 6th Workshop on Algorithms in Bioinformatics.
[41] Recognition of specific DNA sequences. Garvie CW, Wolberger C. Mol Cell. 2001 Nov;8(5):937-46.

[42] The role of DNA shape in protein-DNA recognition. Rohs R, West SM, Sosinsky A, Liu P, Mann RS, Honig B. Nature. 2009 Oct 29;461(7268):1248-53.
[43] Structure-based prediction of C2H2 zinc-finger binding specificity: sensitivity to docking geometry. Siggers TW, Honig B. Nucleic Acids Res. 2007;35(4):1085-97.
[44] Structural alignment of protein--DNA interfaces: insights into the determinants of binding specificity. Siggers TW, Silkov A, Honig B. J Mol Biol. 2005 Feb 4;345(5):1027-45.
[45] Exploring the DNA-binding specificities of zinc fingers with DNA microarrays. Bulyk ML, Huang X, Choo Y, Church GM. Proc Natl Acad Sci U S A. 2001 Jun 19;98(13):7158-63.
[46] Molecular dynamics simulations in biology. Karplus M, Petsko GA. Nature. 1990 Oct 18;347(6294):631-9.
[47] Energetics of ion conduction through the K+ channel. Bernèche S, Roux B. Nature. 2001 Nov 1;414(6859):73-7.
[48] Evaluating the substrate-envelope hypothesis: structural analysis of novel HIV-1 protease inhibitors designed to be robust against drug resistance. Nalam MN, Ali A, Altman MD, Reddy GS, Chellappan S, Kairys V, Ozen A, Cao H, Gilson MK, Tidor B, Rana TM, Schiffer CA. J Virol. 2010 May;84(10):5368-78.
[49] The role of quantitative structure--activity relationships (QSAR) in biomolecular discovery. Winkler DA. Brief Bioinform. 2002 Mar;3(1):73-86.
[50] Hit and lead generation: beyond high-throughput screening. Bleicher KH, Böhm HJ, Müller K, Alanine AI. Nat Rev Drug Discov. 2003 May;2(5):369-78.
[51] Rationalizing fragment based drug discovery for BACE1: insights from FB-QSAR, FB-QSSR, multi objective (MO-QSPR) and MIF studies. Manoharan P, Vijayan RS, Ghoshal N. J Comput Aided Mol Des. 2010 Oct;24(10):843-64.
[52] Virtual screening for R-groups, including predicted pIC50 contributions, within large structural databases, using Topomer CoMFA. Cramer RD, Cruz P, Stahl G, Curtiss WC, Campbell B, Masek BB, Soltanshahi F. J Chem Inf Model. 2008 Nov;48(11):2180-95.
[53] Rethinking 3D-QSAR. Cramer RD. J Comput Aided Mol Des. 2010 Nov 26.
[54] Comparative protein modelling by satisfaction of spatial restraints. Sali A, Blundell TL. J Mol Biol. 1993 Dec 5;234(3):779-815.
[55] On the role of structural information in remote homology detection and sequence alignment: new methods using hybrid sequence profiles. Tang CL, Xie L, Koh IY, Posy S, Alexov E, Honig B. J Mol Biol. 2003 Dec 12;334(5):1043-62.
[56] Solving and analyzing side-chain positioning problems using linear and integer programming. Kingsford CL, Chazelle B, Singh M. Bioinformatics. 2005 Apr 1;21(7):1028-36.
[57] Assembly of protein tertiary structures from fragments with similar local sequences using simulated annealing and Bayesian scoring functions. Simons KT, Kooperberg C, Huang E, Baker D. J Mol Biol. 1997 Apr 25;268(1):209-25.
[58] Improved recognition of native-like protein structures using a combination of sequence-dependent and sequence-independent features of proteins. Simons KT, Ruczinski I, Kooperberg C, Fox BA, Bystroff C, Baker D. Proteins. 1999 Jan 1;34(1):82-95.