Project and Reading List: CSC 857 Bioinformatics Computing, Spring 2013
Instructor: Dr. Rahul Singh, TA Daniel Asarnow

You should be able to access most of the papers from inside the campus. If you can't find a paper, contact the TA.

Project Area A: Structural Bioinformatics

Project A.1: Exploring the protein space with Dali and MATT

Protein structure space is defined as the "universe" of all protein structures. Structural biologists have partitioned this space into a hierarchy of structural clades, exemplified by databases such as SCOP and CATH. Such classifications are partially based on subjective judgments and have been found to disagree substantially in some cases. Thus researchers have begun to use automatic clustering methods to place proteins into existing hierarchies and even to produce entirely new classification schemes. If they are biologically meaningful, such automated approaches are appealing practically, due to their enhanced speed relative to human analysis, and theoretically, due to objectivity, uniform applicability and independence from specific experts.

It has become possible to represent the universe of all protein structures by using protein structure alignment algorithms to measure the distances between pairs of proteins. Particularly interesting are explicit coordinate representations derived from pairwise distance measurements, in which each protein corresponds to a single point. Such representations are called maps of protein structure space (MPSS) and may carry advantages over pairwise distances alone.

In this project, students will use distance data or MPSS to investigate automatic clustering and classification of protein structures. Pre-computed alignment distances and MPSS for ~4,000 protein structures will be provided by the TA.

1. N. Daniels, A. Kumar, L. Cowen, and M. Menke, "Touring Protein Space with Matt," IEEE/ACM Transactions on Computational Biology and Bioinformatics, vol. 9, no. 1, pp. 286–293, 2012.
2. V. Sam, C.-H. Tai, J. Garnier, J.-F. Gibrat, B. Lee, and P. J. Munson, "Towards an automatic classification of protein structural domains based on structural similarity," BMC Bioinformatics, vol. 9, no. 1, p. 74, Jan. 2008.
3. L. Holm and C. Sander, "Touring protein fold space with Dali/FSSP," Nucleic Acids Res., vol. 26, no. 1, pp. 316–319, Jan. 1998.

See also

1. L. L. Conte, S. E. Brenner, T. J. P. Hubbard, C. Chothia, and A. G. Murzin, "SCOP database in 2002: refinements accommodate structural genomics," Nucl. Acids Res., vol. 30, no. 1, pp. 264–267, Jan. 2002.
2. G. Murzin, S. E. Brenner, T. Hubbard, and C. Chothia, "SCOP: A structural classification of proteins database for the investigation of sequences and structures," Journal of Molecular Biology, vol. 247, no. 4, pp. 536–540, Apr. 1995.
3. J. Hou, S.-R. Jun, C. Zhang, and S.-H. Kim, "Global mapping of the protein structure space and application in structure-based inference of protein function," Proceedings of the National Academy of Sciences of the United States of America, vol. 102, no. 10, pp. 3651–3656, Mar. 2005.
4. K. Jain, "Data clustering: 50 years beyond K-means," Pattern Recognition Letters, vol. 31, no. 8, pp. 651–666, Jun. 2010.
5. PSPACE: http://pspace.info

Project A.2: Predicting functional annotations in GO with maps of protein structure space

Recently there has been an effort to massively increase the number of proteins with known molecular structure, under the heading of “structural genomics”. Due in large part to the success of this effort, there are now a large number of proteins with known structure but unknown function. Due to the difficulty and slowness of manual function annotation, automatic methods for structure-based prediction of protein function are now highly desired.

At the same time, it has become possible to represent the universe of all protein structures by using protein structure alignment algorithms to measure the distances between pairs of proteins. Particularly interesting are explicit coordinate representations derived from pairwise distance measurements, in which each protein corresponds to a single point. Such representations are called maps of protein structure space (MPSS) and may carry advantages over pairwise distances alone. Given that the function of a protein is intimately tied to its structure, it should be possible to develop function prediction methods based on the measured distances between protein structures, either in terms of pairwise alignments or low-dimensional MPSS.

In this project, students will use protein distances to predict protein function as defined in the Gene Ontology database. Pre-computed alignment distances and MPSS for ~4,000 protein structures will be provided by the TA.

1. J. Hou, S.-R. Jun, C. Zhang, and S.-H. Kim, “Global mapping of the protein structure space and application in structure-based inference of protein function,” *Proceedings of the National Academy of Sciences of the United States of America*, vol. 102, no. 10, pp. 3651–3656, Mar.2005.
2. M. Osadchy and R. Kolodny, “Maps of protein structure space reveal a fundamental relationship between protein structure and function,” *Proceedings of the National Academy of Sciences*, vol. 108, no. 30, pp. 12301–12306, Jul. 2011.
3. D. Lopez and F. Pazos, “Gene ontology functional annotations at the structural domain level,” *Proteins: Structure, Function, and Bioinformatics*, vol. 76, no. 3, pp. 598–607, 2009.
4. J. D. Watson, S. Sanderson, A. Ezersky, A. Savchenko, A. Edwards, C. Orengo, A. Joachimiak, R. A. Laskowski, and J. M. Thornton, “Towards Fully Automated Structure-based Function Prediction in Structural Genomics: A Case Study*,” Journal of Molecular Biology*, vol. 367, no. 5, pp. 1511–1522, Apr. 2007.

See also

1. I. Friedberg, “Automated protein function prediction—the genomic challenge.” *Brief Bioinform* 2006, 7:225–242.
2. H. Ma et al., “The Gene Ontology (GO) database and informatics resource” *Nucleic Acids Res* 32: D258–61.
3. M. Ashburner et al., “Gene Ontology: tool for the unification of biology,” *Nature Genetics*, vol. 25, no. 1, pp. 25–29, May 2000.
4. T. Fawcett, “ROC Graphs: Notes and Practical Considerations for Researchers,” 2004.
5. PSPACE: http://pspace.info

Project A.3: Understanding the topology and distribution of the protein space

The goal of this project is to study how protein molecules are distributed in terms of their neighborhood relationships in the protein structure space. The team will utilize 2D and 3D maps of the protein structure space and compute neighborhood graphs (Nearest neighbor graphs and relative neighborhood graphs) for these maps. The questions to be investigated are (1) how these graphs vary amongst the 2D and 3D maps (2) How do the relative neighbors compare to neighborhood relations derived from the raw alignment data, and (3) How do SCOP classifications correspond to the neighborhood graphs at the family, superfamily, and fold levels.

1. I Friedberg and A. Godzik, "Connecting the Protein Structure Universe by Using Sparse Recurring Fragments", Structure, 13, pp. 1213-1224, 2005 http://linkinghub.elsevier.com/retrieve/pii/S0969-2126(05)00216-9
2. C.A Orengo, T.P. Flores,W.R. Taylor and J.M. Thornton, "Identification and classification of protein fold families", Protein Engineering vol. 6 no. 5 pp. 485-500, 1993 http://peds.oxfordjournals.org/cgi/reprint/6/5/485

See Also

1. J. Jaromczyk and G. Toussant, "Relative Neighborhood Graphs and their Relatives", Proceedings of the IEEE, Vol 80, No. 9, 1992

Project A.4: Domain Identification in Proteins

The goal of this project is to implement the alpha-carbon based domain identification method of Feldman (Reference 1 below) and test it on the Benchmark_2 dataset from this paper. How well do your inferences agree with that of the paper. How can you improve the performance of the method further?

1. H. Feldman, “Identifying Structural Domains of Proteins Using Clustering”, BMC Bioinformatics, 2012, 13:286
2. T. Holland, S. Veretnik, I. Sindyalov and P. Bourne, “Partitioning Protein Structures in Domains: Why is it so Difficult?, Journal of Molecular Biology, 2006, Vol 361, pp. 562-590

See also

1. Emmert-Streib F, Mushegian A: A topological algorithm for identification of structural domains of proteins. BMC Bioinformatics 2007, 8:237.
2. Alden K, Veretnik S, Bourne PE: dConsensus: a tool for displaying domain assignments by multiple structure-based algorithms and for construction of a consensus assignment. BMC Bioinformatics 2010, 11:310.

Project Area B: Transcriptional Bioinformatics

Project B.1: Imputing missing values in microarray data

The goal of this project is to: (1) Implement the algorithm of Troyanskaya et al. (Reference 1) and examine its effectiveness is imputing missing values from microarrays. (2) How would you extend/improve upon the method?

1. Troyanskaya, O., Cantor, M., Sherlock, G., Brown, P., Hastie, T., Tibshirani, R., Botstein, D., Altman, R. B.: Missing value estimation methods for DNA microarrays. Bioinformatics. (2001) 520--525.
2. Brock GN, Shaffer JR, Blakesley RE, Lotz MJ, Tseng GC., Which missing value imputation method to use in expression profiles: a comparative study and two selection schemes, BMC Bioinformatics. (2008), 12
3. Scholz M, Kaplan F., Guy. CL. Kopka J., and Selbig J., Non-Linear PCA: A Missing Data Approach, Bioinformatics, 2005, 3887-3895
4. D. Wong, F. Wong, and G. Wood, A multi-stage approach to clustering and imputation of gene expression profiles, Bioinformatics 2007 23(8):998-1005 http://bioinformatics.oxfordjournals.org/cgi/reprint/23/8/998

See Also

1. Scheel I, Aldrin M., Glad I, Sorum R., Lyng H., Frigessi A., The Influence of Missing Value Imputation on Detection of Differentially Expressed Genes from Microarray Data, Bioinformatics 2005 21(23) pp. 4272-4279
2. Oba, S., Sato, M., Takemasa, M. M., Matsubara, K., Ishii, S.: A Bayesian missing value estimation method for gene expression profile data. Bioinformatics. (2003) 2088--2096.

Project B.2: Reconstructing Biological Time-Series Data from Microarrays

Consider a set of points that have a temporal ordering. However, this ordering is not known to you. Knowing just the points, can you recover the temporal ordering? In this research, the "points" correspond to the mRNA levels of different genes and reconstructing their temporal ordering can help in increasing data quality as well as tell us about the underlying relationships between these genes.

1. P. Magwene, P. Lizardi, and J. Kim, "Reconstructing the temporal ordering of biological samples using microarray data", Bioinformatics, Vol. 19, No. 7, 2003, pp. 842 – 850:
2. Qiu P, Gentles AJ, Plevritis SK, Discovering Biological Progression Underlying Microarray Samples. PLoS Comput Biol 7(4): e1001123

See Also

1. T. Dey and P. Kumar, "A simple provable algorithm for curve reconstruction", Proc. 10th Symp. Discrete Algorithms, ACM and SIAM, Jan 1999: http://www.compgeom.com/~piyush/papers/mypaper.pdf

Project B.3: Temporal Gene Expression Analysis

The research goal of this project is to analyze patterns in time-series microarray data by techniques of time-series data mining to find interesting patterns.

1. Kasturi J, Acharya R., Ramanathan M, An Information Theoretic Approach for Analyzing Temporal Patterns of Gene Expression, Bioinformatics, 2003 Mar 1;19(4):449-58 http://bioinformatics.oxfordjournals.org/cgi/reprint/19/4/449
2. Wichert S, Fokianos K, Strimmer K. Identifying periodically expressed transcripts in microarray time series data. Bioinformatics, 2004 Jan 1;20(1):5-20. http://bioinformatics.oxfordjournals.org/cgi/reprint/20/1/5
3. Andersson CR, Isaksson A, Gustafsson MG. Bayesian detection of periodic mRNA time profiles without use of training examples.BMC Bioinformatics. 2006 Feb 9;7:63. http://www.biomedcentral.com/content/pdf/1471-2105-7-63.pdf
4. Kim J, Kim H. Clustering of change patterns using Fourier coefficients. Bioinformatics. 2008 Jan 15;24(2):184-91 http://bioinformatics.oxfordjournals.org/cgi/reprint/24/2/184

See Also

1. Leng X, Müller HG. Time ordering of gene coexpression. Biostatistics. 2006 Oct;7(4):569-84. Epub 2006 Feb 22. http://biostatistics.oxfordjournals.org/cgi/reprint/7/4/569
2. Luan Y, Li H. Clustering of time-course gene expression data using a mixed-effects model with B-splines. Bioinformatics, 2003 Mar 1;19(4):474-82. http://bioinformatics.oxfordjournals.org/cgi/reprint/19/4/474
3. Yu Chuan Tai and Terence P. Speed, Statistical Analysis of Microarray Time Course Data http://www.ds.unifi.it/StatGen2005/works/day4/speed_latest.pdf
4. Boutros PC, Okey AB. Unsupervised pattern recognition: an introduction to the whys and wherefores of clustering microarray data. Brief Bioinform. 2005 Dec;6(4):331-43. http://bib.oxfordjournals.org/cgi/reprint/6/4/331

Project B.4: Graph-theoretic methods to find co-regulated gene groups

(Data courtesy of S. Lonardi and Y. Cheng)
Use the file yeastEx.txt, which is an expression matrix of 2882 yeast genes, with names available in the file yeastNames.txt, under 17 conditions (different snapshots of the yeast cell cycle. These files are available at: http://www.cs.ucr.edu/~stelo/cs234winter13/h3.html
The goal of the project is to evaluate different neighborhood graphs in terms of their ability to find co-regulated genes. Explore the use of different measures of expression vector similarity to define thresholds that can be used to construct nearest neighbor graphs, relative neighbor graphs, and Gabriel graphs on the genes. Using the number of incidental edges for each vertex in each type of graph, identify the co-regulated genes. Generate a plot of expression for each gene cluster.

1. Patrik D'haeseleer, How does gene expression clustering work? Nature Biotechnology 23, 1499 - 1501 (2005)
2. K. Y. Yeung1, D. R. Haynor and, W. L. Ruzzo, Validating clustering for gene expression data, *Bioinformatics (2001) 17 (4): 309-318*

See Also.

1. J. Jaromczyk and G. Toussant, “Relative Neighborhood Graphs and their Relatives”, Proceedings of the IEEE, Vol 80, No. 9, 1992

Project Area C: Sequence Analysis

Project C.1: Whole Genome Alignment

This project deals with the question of aligning entire genomes (rather than genes). The research goal is to understand one of such techniques in detail and apply it.

1. Darling AC, B. Mau, F. Blattner, and N. Perna, Mauve: Multiple Alignment of Conserved Genomic Sequence with Rearrangement, Genome Research 2004, 14: 1394-1403

2. Choi JH, Cho HG, Kim S. GAME: a simple and efficient whole genome alignment method using maximal exact match filtering. Comput Biol Chem. 2005 Jun;29(3):244-53.
http://www.ncbi.nlm.nih.gov/pubmed/15979044

3. Haubold B, Pierstorff N, Möller F, Wiehe T. Genome comparison without alignment using shortest unique substrings., BMC Bioinformatics, 2005 May 23;6:123.
http://www.biomedcentral.com/content/pdf/1471-2105-6-123.pdf

4. Delcher AL, Phillippy A, Carlton J, Salzberg SL., Fast algorithms for large-scale genome alignment and comparison. Nucleic Acids Res. 2002 Jun 1;30(11):2478-83
http://nar.oxfordjournals.org/cgi/reprint/30/11/2478

3. Wong PW, Lam TW, Lu N, Ting HF, Yiu SM. An efficient algorithm for optimizing whole genome alignment with noise. Bioinformatics. 2004 Nov 1;20(16):2676-84.
http://bioinformatics.oxfordjournals.org/cgi/reprint/20/16/2676

See Also

1. Kurtz S, Phillippy A, Delcher AL, Smoot M, Shumway M, Antonescu C, Salzberg SL: Versatile and open software for comparing large genomes. Genome Biology 2004, 5:R12.
http://genomebiology.com/content/pdf/gb-2004-5-2-r12.pdf

2. Batzoglou S. The many faces of sequence alignment, Brief Bioinform. 2005 Mar;6(1):6-22.
http://bib.oxfordjournals.org/cgi/reprint/6/1/6

3. Page RD. Introduction to comparing large sequence sets. Curr Protoc Bioinformatics. 2003 Feb;Chapter 10:Unit 10.1.

4. Lippert RA. Space-efficient whole genome comparisons with Burrows-Wheeler transforms. J Comput Biol. 2005 May;12(4):407-15.
http://math.mit.edu/~lippert/research/JCB-LMW-04-submission-5.pdf

5. Bellgard M, Kenworthy W. FBSA: feature-based sequence alignment technique for very large sequences. Appl Bioinformatics. 2003;2(3):145-50

6. Joseph J, Sasikumar R., Chaos game representation for comparison of whole genomes. BMC Bioinformatics. 2006 May 5;7:243
http://www.biomedcentral.com/1471-2105/7/243

Project Area D: Molecular Informatics and Computational Drug Discovery

Project D1: Molecular Matching of Small Molecules

The research question is how to compare small (drug-like) molecules. This can be used for querying molecules in large databases. The investigations should convert molecules to representations (descriptors) that can be used for rapid querying and investigate different measures to compare such descriptors.

1. BLASTing Small Molecules: Statistics and Extreme Statistics of Chemical Similarity Scores. P. Baldi and R. W. Benz., Bioinformatics, 24(13):i357-i365, (2008).
2. Speeding Up Chemical Database Searches Using a Proximity Filter Based on the Logical Exclusive-OR. P. Baldi, D. S. Hirschberg and R. J. Nasr. Journal of Chemical Information and Modeling,48, 7, 1367-1378, (2008).

See Also

1. ChemDB: A Public Database of Small Molecules and Related Chemoinformatics Resources. J. Chen, S. J. Swamidass, Y. Dou, J. Bruand, and P. Baldi. Bioinformatics, 21, 4133-4139, (2005).
2. R. Singh, “Surface Similarity-Based Molecular Query-Retrieval”, BMC Cell-Biology, Vol. 8, Suppl.1 (S6): July, 2007.
3. Wang Y, Xiao J, Suzek T, Zhang J, Wang J, Bryant S. “PubChem: a public information system for analyzing bioactivities of small molecules.” Nucleic Acids Res. 2009 July 1; 37(Web Server issue): W623–W633. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2703903/
4. Chemistry Development Kit http://sourceforge.net/apps/mediawiki/cdk/index.php?title=Main_Page
5. Daylight Theory Manual http://www.daylight.com/dayhtml/doc/theory/index.html

Project D2: Molecular Matching of Large Molecules

The research question is how to compare large molecules (proteins and enzymes).

1. Sael L, Li B, La D, Fang Y, Ramani K, Rustamov R, Kihara D., Fast protein tertiary structure retrieval based on global surface shape similarity. Proteins. 2008 Sep;72(4):1259-73.
   http://www.ncbi.nlm.nih.gov/pubmed/18361455
2. Daras, P.; Zarpalas, D.; Axenopoulos, A.; Tzovaras, D.; Strintzis, M.G., "Three-Dimensional Shape-Structure Comparison Method for Protein Classification," Computational Biology and Bioinformatics, IEEE/ACM Transactions on , vol.3, no.3, pp.193-207, July-Sept. 2006
   http://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=1668019&isnumber=34920
3. CATHEDRAL: A Fast and Effective Algorithm to Predict Folds and Domain Boundaries from Multidomain Protein Structures. Oliver C Redfern, Andrew Harrison, Tim Dallman, Frances M. G Pearl, and Christine A Orengo. PLoS Comput Biol. 2007 November; 3(11): e232.

See also

1. Prlic A, Bliven S, Rose P, Bluhm W, Bizon C, Godzik A, Bourne P. "Pre-calculated protein structure alignments at the RCSB PDB website." Bioinformatics (2010) 26 (23): 2983-2985.
   http://bioinformatics.oxfordjournals.org/content/26/23/2983.full
2. Partitioning Protein Structures into Domains: Why is it so difficult? T. Holland, S. Veretnik, I. Shindylov, P. Bourne. J. Mol. Biol. 2006 Aug 18;361(3):562-90. Epub 2006 Jun 22.
   http://www.ncbi.nlm.nih.gov/pubmed/16863650
3. Yeh JS, Chen DY, Chen BY, Ouhyoung M. "A web-based three-dimensional protein retrieval system by matching visual similarity." Bioinformatics (2005) 21 (13): 3056-3057.
   http://bioinformatics.oxfordjournals.org/content/21/13/3056.full
4. Chi PH, Scott G, Shyu CR. "A fast protein structure retrieval system using image-based distance matrices and multidimensional index." BIBE 2004 Proceedings.
   http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.112.342&rep=rep1&type=pdf
5. Shyu CR, Chi PH, Scott G, Xu D. "ProteinDBS: a real-time retrieval system for protein structure comparison." Nucleic Acids Res. 2004 July 1; 32(Web Server issue): W572–W575.
   http://www.ncbi.nlm.nih.gov/pmc/articles/PMC441574/
6. R. Singh, "Surface Similarity-Based Molecular Query-Retrieval", BMC Cell-Biology, Vol. 8, Suppl.1 (S6): July, 2007.

INSTRUCTIONS AND SUGGESTIONS FOR PAPER SELECTION AND PRESENTATIONS

0. Select a project following the procedural guidelines given in the class. Select a paper from this project. Projects and papers are selected on a first-come-first-serve basis. Make sure to intimate the TA about your selection. If you want to propose your own project, get approval from the professor. Your project selection has to be minimally of the same complexity as the class project and must use methods and ideas from bioinformatics computing as taught in the class.

1. In the week you present, PowerPoint slides must be emailed to (and received by) the TA no later than the Monday preceding the day of presentation. Failure to do so will constitute a late submission and will result in loss of points. Your slides will be made available to everyone in the class through iLearn.

2. If you need a laptop for presentation, inform the TA one week in advance

3. Each presentation will last for 30 minutes, followed by questions for 15-20 minutes. As such, you should prepare between 25 - 30 slides. Talks significantly over/under time will loose points. Use a simple slide template

4. Lastly, your presentation should convey a broad view of the paper's contribution to its field of research and the main technical details in it. So focus on the key contributions of the work and its implications rather than trying to present everything in the paper.

5. Generally, avoid making slides "text-heavy". If you are using images in slides have them in formats such as jpeg to avoid large file sizes.

6. Understanding a paper may require you to "research" the topic: - For example, you may need to read another paper by the same author to better understand what they are writing about or find a tutorial on the web that explains the problem the authors are looking at. Doing such research will help you in understanding the content and thus lead to a better presentation.

7. Access to electronic editions of most papers and to major online-libraries can be obtained through the university library – use the library resources available to you. Also many authors put their papers online.

8. Please check the iLearn page on Wednesday evening for slides of presentations for the following Thursday.

9. Everyone is expected to read the paper(s) being presented and participate in Q&A and discussions.