BI217: Introductory Bioinformatics & Biostatistics

SYLLABUS

Instructor: Maoying Wu
Email: ricket.woo@gmail.com
Webpage: http://cbb.sjtu.edu.cn/~mywu/bi217/2010

Course Description

This is a course on the fundamentals of bioinformatics geared towards second-year undergraduate students, who are interested in the principle knowledge in bioinformatics and biostatistics. The course will cover a lot of research fields in both bioinformatics and biostatistics but won't go too detail, with a number of interesting topics including biological side of bioinformatics, mathematics in bioinformatics, computer techniques for bioinformatians, biological sequence analysis, structural analysis, microarray analysis, systems biology. There will also be some additional topics such as parameter estimation, hypothesis testing, survival analysis, multivariate analysis, etc.

The course aims at guiding you into the kingdom of bioinformatics and biostatistics, with ease step and great comfort.

Prerequisites

Knowledge of Calculus and Probability is preferred but not a must.

Textbooks and References

1. R. Durbin, S. R. Eddy, A. Krogh and G. Mitchison (1998). ***Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids.*** Cambridge University Press
2. Andreas Baxevanis and Francis Ouellett (eds.), ***Bioinformatics: A Practical Guide to the Analysis of Genes and Proteins****, 3rd Edition*, Wiley & Sons, 2005.
3. Gibas, Cynthia, and Per Jambeck. ***Developing Bioinformatics Computer Skills***. O'Reilly, 2001

Calendar

#Week	*Topics*	*Assignments*
2	What is Bioinformatics?	reading assignment
3	R Language and Biostatistics Overview	R Computer Lab
4	Biological Side of Bioinformatics	assignment1.pdf
5	Probability Distribution	problem1.pdf
6	Mathematical Principles in Bioinformatics	assignment2.pdf
7	Parameter Estimation	problem2.pdf
8	Developing Your Computer Skills for Bioinformatics	assignment3.pdf
9	Hypothesis Testing	problem3.pdf
10	Pooling Algorithms for Bioinformatics	assignment4.pdf
11	Linear Models	problem4.pdf
12	Analysis of Biological Sequences, Structures and Functions	assignment5.pdf

13	Multivariate Analysis	problem5.pdf
14	Sequence Alignment & Phylogenic Reinference	assignment6.pdf
15	Survival Analysis	problem6.pdf
16	Microarray Data Analysis	assignment7.pdf
17	Preprocessing and Analysis of High-dimensional Data	problem7.pdf
18	Reading Discussion	/

Homework

- There will be 16, more or less equally spaced, problem sets, with half for bioinformatics, another half for biostatistics. They will count for 80% of the final grade.
- There will be an about-4-hour discussion on assigned readings, which will account for the remaining 20% of the final grade.
- There will be no midterm exam and final exam.

Policy on Collaboration

You may discuss with your peers when preparing your homework solutions. However, duplicating is not acceptable. If you do collaborate on homework, you must cite, in your solution, your partners. Additionally, if you 'consult' an 'expert', do not forget to cite the source in your solutions.

Recommended Readings

Sequence Alignment [1-4]

Phylogeny [5-21]

Gene Identification [22-29]

Structure Prediction [30-38]

Microarray [39-61]

Systems Biology [62-72]

Statistical Genetics [72-82]

1. Agrawal A, Brendel VP, Huang X: ***Pairwise statistical significance and empirical determination of effective gap opening penalties for protein local sequence alignment***. *Int J Comput Biol Drug Des* 2008, ***1***(4):347-367.
2. Agrawal A, Huang X: ***Pairwise statistical significance of local sequence alignment using multiple parameter sets and empirical justification of parameter set change penalty***. *BMC Bioinformatics* 2009, ***10 Suppl 3***:S1.
3. Altschul SF, Bundschuh R, Olsen R, Hwa T: ***The estimation of statistical parameters for local alignment score distributions***. *Nucleic Acids Res* 2001, ***29***(2):351-361.
4. Mitrophanov AY, Borodovsky M: ***Statistical significance in biological sequence analysis***. *Brief Bioinform* 2006, ***7***(1):2-24.
5. Suchard MA, Weiss RE, Sinsheimer JS: ***Bayesian selection of continuous-time Markov chain evolutionary models***. *Mol Biol Evol* 2001, ***18***(6):1001-1013.
6. Alexe G, Satya RV, Seiler M, Platt D, Bhanot T, Hui S, Tanaka M, Levine AJ, Bhanot G: ***PCA and clustering reveal alternate mtDNA phylogeny of N and M clades***. *J Mol Evol* 2008, ***67***(5):465-487.

7. Gambin A, Slonimski PP: ***Hierarchical clustering based upon contextual alignment of proteins: a different way to approach phylogeny***. *C R Biol* 2005, ***328***(1):11-22.
8. Cotta C, Moscato P: ***A memetic-aided approach to hierarchical clustering from distance matrices: application to gene expression clustering and phylogeny***. *Biosystems* 2003, ***72***(1-2):75-97.
9. Ninio M, Privman E, Pupko T, Friedman N: ***Phylogeny reconstruction: increasing the accuracy of pairwise distance estimation using Bayesian inference of evolutionary rates***. *Bioinformatics* 2007, ***23***(2):e136-141.
10. Wang LS, Warnow T, Moret BM, Jansen RK, Raubeson LA: ***Distance-based genome rearrangement phylogeny***. *J Mol Evol* 2006, ***63***(4):473-483.
11. Cheon S, Liang F: ***Bayesian phylogeny analysis via stochastic approximation Monte Carlo***. *Mol Phylogenet Evol* 2009, ***53***(2):394-403.
12. Winkworth RC, Bell CD, Donoghue MJ: ***Mitochondrial sequence data and Dipsacales phylogeny: mixed models, partitioned Bayesian analyses, and model selection***. *Mol Phylogenet Evol* 2008, ***46***(3):830-843.
13. Liu L, Pearl DK: ***Species trees from gene trees: reconstructing Bayesian posterior distributions of a species phylogeny using estimated gene tree distributions***. *Syst Biol* 2007, ***56***(3):504-514.
14. Clark TG, De Iorio M, Griffiths RC: ***Bayesian logistic regression using a perfect phylogeny***. *Biostatistics* 2007, ***8***(1):32-52.
15. Castoe TA, Parkinson CL: ***Bayesian mixed models and the phylogeny of pitvipers (Viperidae: Serpentes)***. *Mol Phylogenet Evol* 2006, ***39***(1):91-110.
16. Sullivan J: ***Maximum-likelihood methods for phylogeny estimation***. *Methods Enzymol* 2005, ***395***:757-779.
17. Zhang H, Gu X: ***Maximum likelihood for genome phylogeny on gene content***. *Stat Appl Genet Mol Biol* 2004, ***3***:Article31.
18. Katoh K, Kuma K, Miyata T: ***Genetic algorithm-based maximum-likelihood analysis for molecular phylogeny***. *J Mol Evol* 2001, ***53***(4-5):477-484.
19. Schadt EE, Sinsheimer JS, Lange K: ***Computational advances in maximum likelihood methods for molecular phylogeny***. *Genome Res* 1998, ***8***(3):222-233.
20. Yang Z: ***Maximum-likelihood estimation of phylogeny from DNA sequences when substitution rates differ over sites***. *Mol Biol Evol* 1993, ***10***(6):1396-1401.
21. Hasegawa M, Fujiwara M: ***Relative efficiencies of the maximum likelihood, maximum parsimony, and neighbor-joining methods for estimating protein phylogeny***. *Mol Phylogenet Evol* 1993, ***2***(1):1-5.
22. Aggarwal G, Ramaswamy R: ***Ab initio gene identification: prokaryote genome annotation with GeneScan and GLIMMER***. *J Biosci* 2002, ***27***(1 Suppl 1):7-14.
23. DeCaprio D, Vinson JP, Pearson MD, Montgomery P, Doherty M, Galagan JE: ***Conrad: gene prediction using conditional random fields***. *Genome Res* 2007, ***17***(9):1389-1398.
24. Burset M, Guigo R: ***Evaluation of gene structure prediction programs***. *Genomics* 1996, ***34***(3):353-367.
25. Stanke M, Schoffmann O, Morgenstern B, Waack S: ***Gene prediction in eukaryotes with a generalized hidden Markov model that uses hints from external sources***. *BMC Bioinformatics* 2006, ***7***:62.
26. Stanke M, Waack S: ***Gene prediction with a hidden Markov model and a new intron submodel***. *Bioinformatics* 2003, ***19 Suppl 2***:ii215-225.
27. Meyer IM, Durbin R: ***Comparative ab initio prediction of gene structures using pair HMMs***. *Bioinformatics* 2002, ***18***(10):1309-1318.
28. Ter-Hovhannisyan V, Lomsadze A, Chernoff YO, Borodovsky M: ***Gene prediction in novel fungal genomes using an ab initio algorithm with unsupervised training***. *Genome Res* 2008, ***18***(12):1979-1990.
29. Wang Z, Chen Y, Li Y: ***A brief review of computational gene prediction methods***. *Genomics Proteomics Bioinformatics* 2004, ***2***(4):216-221.

30. van Batenburg FH, Gultyaev AP, Pleij CW: ***An APL-programmed genetic algorithm for the prediction of RNA secondary structure***. *J Theor Biol* 1995, ***174***(3):269-280.
31. Juan V, Wilson C: ***RNA secondary structure prediction based on free energy and phylogenetic analysis***. *J Mol Biol* 1999, ***289***(4):935-947.
32. Mathews DH, Disney MD, Childs JL, Schroeder SJ, Zuker M, Turner DH: ***Incorporating chemical modification constraints into a dynamic programming algorithm for prediction of RNA secondary structure***. *Proc Natl Acad Sci U S A* 2004, ***101***(19):7287-7292.
33. Andronescu M, Zhang ZC, Condon A: ***Secondary structure prediction of interacting RNA molecules***. *J Mol Biol* 2005, ***345***(5):987-1001.
34. Ding Y: ***Statistical and Bayesian approaches to RNA secondary structure prediction***. *Rna* 2006, ***12***(3):323-331.
35. Mathews DH, Turner DH: ***Prediction of RNA secondary structure by free energy minimization***. *Curr Opin Struct Biol* 2006, ***16***(3):270-278.
36. Mathews DH, Turner DH, Zuker M: ***RNA secondary structure prediction***. *Curr Protoc Nucleic Acid Chem* 2007, ***Chapter 11***:Unit 11 12.
37. Zhao Y, Wang Z: ***[RNA secondary structure prediction based on support vector machine classification]***. *Sheng Wu Gong Cheng Xue Bao* 2008, ***24***(7):1140-1148.
38. Poolsap U, Kato Y, Akutsu T: ***Prediction of RNA secondary structure with pseudoknots using integer programming***. *BMC Bioinformatics* 2009, ***10 Suppl 1***:S38.
39. Chen X, Wang L, Smith JD, Zhang B: ***Supervised principal component analysis for gene set enrichment of microarray data with continuous or survival outcomes***. *Bioinformatics* 2008, ***24***(21):2474-2481.
40. Zhao Q, Sun J: ***Cox survival analysis of microarray gene expression data using correlation principal component regression***. *Stat Appl Genet Mol Biol* 2007, ***6***:Article16.
41. van der Werf MJ, Pieterse B, van Luijk N, Schuren F, van der Werff-van der Vat B, Overkamp K, Jellema RH: ***Multivariate analysis of microarray data by principal component discriminant analysis: prioritizing relevant transcripts linked to the degradation of different carbohydrates in Pseudomonas putida S12***. *Microbiology* 2006, ***152***(Pt 1):257-272.
42. Wang A, Gehan EA: ***Gene selection for microarray data analysis using principal component analysis***. *Stat Med* 2005, ***24***(13):2069-2087.
43. Iwafuchi H, Mori N, Takahashi T, Yatabe Y: ***Phenotypic composition of salivary gland tumors: an application of principal [corrected] component analysis to tissue microarray data***. *Mod Pathol* 2004, ***17***(7):803-810.
44. Crescenzi M, Giuliani A: ***The main biological determinants of tumor line taxonomy elucidated by a principal component analysis of microarray data***. *FEBS Lett* 2001, ***507***(1):114-118.
45. Wang S, Zhu J: ***Variable selection for model-based high-dimensional clustering and its application to microarray data***. *Biometrics* 2008, ***64***(2):440-448.
46. Istepanian RS, Sungoor A, Nebel JC: ***Linear predictive coding and wavelet decomposition for robust microarray data clustering***. *Conf Proc IEEE Eng Med Biol Soc* 2007, ***2007***:4629-4632.
47. Yuan A, He W: ***Semiparametric clustering method for microarray data analysis***. *J Bioinform Comput Biol* 2008, ***6***(2):261-282.
48. Han L, Zeng X, Yan H: ***Fuzzy clustering analysis of microarray data***. *Proc Inst Mech Eng H* 2008, ***222***(7):1143-1148.
49. Ovaska K, Laakso M, Hautaniemi S: ***Fast Gene Ontology based clustering for microarray experiments***. *BioData Min* 2008, ***1***(1):11.
50. Kim EY, Kim SY, Ashlock D, Nam D: ***MULTI-K: accurate classification of microarray subtypes using ensemble k-means clustering***. *BMC Bioinformatics* 2009, ***10***:260.

51. Savage RS, Heller K, Xu Y, Ghahramani Z, Truman WM, Grant M, Denby KJ, Wild DL: ***R/BHC: fast Bayesian hierarchical clustering for microarray data***. *BMC Bioinformatics* 2009, ***10***:242.
52. Chatterjee S, Bhattacharjee K, Konar A: ***A simple and robust algorithm for microarray data clustering based on gene population-variance ratio metric***. *Biotechnol J* 2009, ***4***(9):1357-1361.
53. Yi SG, Joo YJ, Park T: ***Rank-based clustering analysis for the time-course microarray data***. *J Bioinform Comput Biol* 2009, ***7***(1):75-91.
54. Benito M, Parker J, Du Q, Wu J, Xiang D, Perou CM, Marron JS: ***Adjustment of systematic microarray data biases***. *Bioinformatics* 2004, ***20***(1):105-114.
55. Choong MK, Levy D, Yan H: ***Study of microarray time series data based on Forward-Backward Linear Prediction and Singular Value Decomposition***. *Int J Data Min Bioinform* 2009, ***3***(2):145-159.
56. Ghosh D: ***Resampling methods for variance estimation of singular value decomposition analyses from microarray experiments***. *Funct Integr Genomics* 2002, ***2***(3):92-97.
57. Ghosh D: ***Singular value decomposition regression models for classification of tumors from microarray experiments***. *Pac Symp Biocomput* 2002:18-29.
58. Liou LS, Shi T, Duan ZH, Sadhukhan P, Der SD, Novick AA, Hissong J, Skacel M, Almasan A, DiDonato JA: ***Microarray gene expression profiling and analysis in renal cell carcinoma***. *BMC Urol* 2004, ***4***:9.
59. Liu L, Hawkins DM, Ghosh S, Young SS: ***Robust singular value decomposition analysis of microarray data***. *Proc Natl Acad Sci U S A* 2003, ***100***(23):13167-13172.
60. Wall ME, Dyck PA, Brettin TS: ***SVDMAN--singular value decomposition analysis of microarray data***. *Bioinformatics* 2001, ***17***(6):566-568.
61. Wu X, Dewey TG: ***From microarray to biological networks: Analysis of gene expression profiles***. *Methods Mol Biol* 2006, ***316***:35-48.
62. Hsu CW, Juan HF, Huang HC: ***Characterization of microRNA-regulated protein-protein interaction network***. *Proteomics* 2008, ***8***(10):1975-1979.
63. Hu X: ***Mining and analysing scale-free protein-protein interaction network***. *Int J Bioinform Res Appl* 2005, ***1***(1):81-101.
64. Kar G, Gursoy A, Keskin O: ***Human cancer protein-protein interaction network: a structural perspective***. *PLoS Comput Biol* 2009, ***5***(12):e1000601.
65. Liang H, Li WH: ***MicroRNA regulation of human protein protein interaction network***. *Rna* 2007, ***13***(9):1402-1408.
66. Lin CC, Juan HF, Hsiang JT, Hwang YC, Mori H, Huang HC: ***Essential core of protein-protein interaction network in Escherichia coli***. *J Proteome Res* 2009, ***8***(4):1925-1931.
67. Sen TZ, Kloczkowski A, Jernigan RL: ***Functional clustering of yeast proteins from the protein-protein interaction network***. *BMC Bioinformatics* 2006, ***7***:355.
68. Wu X, Zhu L, Guo J, Zhang DY, Lin K: ***Prediction of yeast protein-protein interaction network: insights from the Gene Ontology and annotations***. *Nucleic Acids Res* 2006, ***34***(7):2137-2150.
69. Wu Z, Zhao X, Chen L: ***Identifying responsive functional modules from protein-protein interaction network***. *Mol Cells* 2009, ***27***(3):271-277.
70. Xu J, Li Y: ***Discovering disease-genes by topological features in human protein-protein interaction network***. *Bioinformatics* 2006, ***22***(22):2800-2805.
71. Zhu M, Gao L, Li X, Liu Z, Xu C, Yan Y, Walker E, Jiang W, Su B, Chen X *et al*: ***The analysis of the drug-targets based on the topological properties in the human protein-protein interaction network***. *J Drug Target* 2009, ***17***(7):524-532.
72. Bourgain C, Abney M, Schneider D, Ober C, McPeek MS: ***Testing for Hardy-Weinberg equilibrium in samples with related individuals***. *Genetics* 2004, ***168***(4):2349-2361.
73. Cheng KF, Lin WJ: ***Retrospective analysis of case-control studies when the population is in Hardy-Weinberg equilibrium***. *Stat Med* 2005, ***24***(21):3289-3310.
74. Knapp M: ***On the asymptotic equivalence of allelic and trend statistic under Hardy-Weinberg equilibrium***. *Ann Hum Genet* 2008, ***72***(Pt 5):589.

75. Kuk AY, Zhang H, Yang Y: ***Computationally feasible estimation of haplotype frequencies from pooled DNA with and without Hardy-Weinberg equilibrium***. *Bioinformatics* 2009, ***25***(3):379-386.
76. Rogatko A, Slifker MJ, Babb JS: ***Hardy-Weinberg equilibrium diagnostics***. *Theor Popul Biol* 2002, ***62***(3):251-257.
77. Kuo CL, Feingold E: ***What's the best statistic for a simple test of genetic association in a case-control study?*** *Genet Epidemiol* 2009.
78. Infante-Rivard C, Mirea L, Bull SB: ***Combining case-control and case-trio data from the same population in genetic association analyses: overview of approaches and illustration with a candidate gene study***. *Am J Epidemiol* 2009, ***170***(5):657-664.
79. Tan Q, Zhao JH, Zhang D, Kruse TA, Christensen K: ***Power for genetic association study of human longevity using the case-control design***. *Am J Epidemiol* 2008, ***168***(8):890-896.
80. Won S, Elston RC: ***The power of independent types of genetic information to detect association in a case-control study design***. *Genet Epidemiol* 2008, ***32***(8):731-756.
81. Tsai HJ, Kho JY, Shaikh N, Choudhry S, Naqvi M, Navarro D, Matallana H, Castro R, Lilly CM, Watson HG *et al*: ***Admixture-matched case-control study: a practical approach for genetic association studies in admixed populations***. *Hum Genet* 2006, ***118***(5):626-639.
82. Hung RJ, Brennan P, Malaveille C, Porru S, Donato F, Boffetta P, Witte JS: ***Using hierarchical modeling in genetic association studies with multiple markers: application to a case-control study of bladder cancer***. *Cancer Epidemiol Biomarkers Prev* 2004, ***13***(6):1013-1021.