Contents Summer Bioinformatic school

Program:
Class 0) (BEFORE THE COURSE START)
Preparation of the work environment before coming to the course:

●
Installation of R, RStudio and packages that will be indicated.

Class 1)

- Introduction to R and RStudio.
- Variables and data types (Chapter 3 Statistics using R with biological examples), control structures, functions, subsets and vectorization (Chapter 4 Statistics using R with biological examples).
- Packages.
- Import, export and cleanup of data.
- Helps.

Class 2)

- Introduction to RNASeq
- Measurements of similarity and distance. Scale of the data.
- Clusters (clustering). Types of clusters.
- Clustering algorithms. Implementation in R.
- Biological examples.

Class 3)

- Complex systems. Biological examples and systems biology.
- Basic concepts of complex networks. Definition, topology and representation.
- Types of networks and their characteristics. Topological properties.
- Biological examples.
- The Igraph package, use of networks with R.

Class 4)

- Detection of communities in networks. Types of algorithms, modularity.
- Similarity in topological spaces. Similarity in networks, equivalences.
- Transcriptomics. Introduction to the analysis of transcriptomic data with weighted gene coexpression network analysis.

●
Introduction to the WGCNA package in R. Class 5)

- WGCNA.
- Heavy nets.
- Topological similarity (TOM).
- Grouping and detection of "important" genes.
- Functional enrichment of biological themes. What an enrichment.
- Semantic similarity.
- Complete the WGCNA work pipeline in R.

Recommended bibliography (manuals, tutorials and books):

- An introduction to R. [online]
- Peng R. (2015), R Programming for Data Science, Lean Publishing
- Everitt B. (2010), A Handbook of Statistical Analyses Using R, CRC Press.
- Verzani J. (2002), simpleR: Using R for Introductory Statistics. [pdf]
- Dalgaard P. (2008), Introductory Statistics with R, Springer
- Mangiafico S. (2015), AN R COMPANION FOR THE HANDBOOK OF BIOLOGICAL STATISTICS, New Brunswick, Rutgers University [pdf]
- McDonald J. (2014). HANDBOOK OF BIOLOGICAL STATISTICS, SPARKY HOUSE PUBLISHING. [pdf]
- Zeileis A. (2013), Extended Model Formulas in R: Multiple Parts and Multiple Responses. [pdf]

● Seefeld K. (2007), Statistics Using R with Biological Examples, Durham, NH, University of New Hampshire, Department of Mathematics & Statistics. [pdf]
●
Mangiafico, S. (2016). Summary and analysis of extension program evaluation in R. New Brunswick,
Rutgers University [pdf]
● Wickham h. (2009), ggplot2: Elegant graphics for data analysis, Springer
● Barabasi A. Network Science. [online]
● Tan, Steinbach & Kumar "Introduction to Data Mining". [online]
● Hastie, Tibshirani & Friedman, "The Elements of Statistical Learning", 2nd ed, Springer, 2009. [online]
● James, Witten, Hastie & Tibshirani, "An Introduction to Statistical Learning with Applications in R", 6th ed,
Springer, 2015 [online]

METHODS FOR THE CONFORMATIONAL STUDY OF PROTEINS AND THEIR INTERACTIONS (MECPI)

Objectives: This course addresses issues of modern structural biology of macromolecules,with special emphasis on the methodological aspects of experimental and computational techniques used to investigate the conformation of proteins and their interactions. Each technique is treated from its theoretical foundation, to the operational and instrumental aspects and finally in its concrete application to the structural and functional study of peptides and proteins.
Each method, its specific utility, as well as its scope and limitations. Permanent reference is made to the problem of protein folding and to specific molecular recognition phenomena in the study of the structure-relationship function in proteins. This mixed approach is intended to give the student a basis for examination of characteristics such as the structure, stability and dynamics of molecules of biological interest. As an achievement test, at the end the students prepare and individually present original work of their specific interest where applied in combined form two or more of the techniques covered throughout the course.
Techniques covered in theoretical-practical activities: X-ray crystallography, in vivo protein folding, circular dichroism (CD), fluorescence methods, folding mechanisms and theories, nuclear magnetic resonance (NMR), surface plasmon resonance (SPR), molecular dynamics (MD), molecular modeling by homology, computational drug design, electrospray mass spectrometry (ESI) and MALDI-TOF, fast kinetic methods, atomic force microscopy (AFM), photolabeling of soluble and membrane proteins, size exclusion chromatography (SEC- FPLC) and light scattering (LS)

Bibliography

- Bayley H (1983) Photogenerated Reagents in Biochemistry and Molecular Biology, Elsevier Science Publishers BV, Amsterdam
- Bodanszky M (1984) Principles of Peptide Synthesis, Springer-Verlag, Berlin
- Bodanszky M, Bodanszky A (1984) The Practice of Peptide Synthesis, Springer-Verlag, Berlin

- Branden C, Tooze J (1999) Introduction to Protein Structure. Second edition, Garland Publishing Inc., New York
- Campbell ID, Dwek RA (1984) Biological Spectroscopy, The Benjamin/Cummings Publishing Company Inc., Menlo Park, California
- Cavanagh J, Fairbrother WJ, Palmer III AG, Rance M, Skelton NJ (2007) Protein NMR Spectroscopy. Principles and Practice. Elsevier, China
- Chance M, Ed. (2008) Mass Spectrometry Analysis for Protein-Protein Interactions and Dynamics, Wiley, New Jersey
- Creighton TE, Ed. (1992) Protein Folding, WH Freeman & Co, New York
- Creighton TE (1993) Proteins. Structures and Molecular Properties. Second edition, WH Freeman & Co, New York
- Fasman GD, Ed. (1996) Circular Dichroism and the Conformational Analysis of Biomolecules, Plenum Press, New York
- Fersht A (1999) Structure and mechanism in protein science. A guide to enzyme catalysis and protein folding. Third edition, WH Freeman & Co, New York

School of Biomolecules Modeling

Program:

1) Concept of computational simulation in Science: relationship between experiment, theory and simulation. Molecular simulation in Chemistry, Biochemistry and Materials. Existing models for determining the surface potential energy. Proposal of simulation strategies for answer questions of chemical and biochemical interest.

2) Ab-initio methods. Hartree-Fock equations. Basic functions. Determination of molecular properties. Semi-empirical methods. General idea and CNDO implementations, MNDO, INDO. Semi-empirical models based on parameterization: AM1 and PM3 methods. Theory density functional. Fundamental theorems. Kohn and Sham implementation. Rank of applicability, advantages and disadvantages of the different electronic structure techniques.

3) Parameterized force field methods. Solvent models, explicit, implicit, water models (TIP3P, TIP4P, SPC, fluctuating load models). Force fields for biomolecules. AMBER and CHARMM potentials. Construction of a Force Field and Derivation of parameters (Union and non-union parameters and determination of partial charges by adjusting to electrostatic potential).

4) Statistical thermodynamics. Basic concepts. Application to simulation techniques. Assemblies. Partition function and thermodynamic properties. Ergodic hypothesis. Simulation scheme Monte Carlo. Molecular Dynamics Scheme. Technical details. Methods of integration of Newton's equations for molecular dynamics. Examples of Monte Carlo simulations and Molecular Dynamics. Determination of structural and dynamic properties. Thermostats (Berendsen, Nose). Langevin dynamics.

5) Free energy calculations: Thermodynamic functions Energy and Entropy. Sampling methods skewed (Umbrella Sampling). Methods based on thermodynamic transformations (integration thermodynamics, FEP perturbation theory). Guided Molecular Dynamics and approximations of no

equilibrium, relationship between work and reversibility: Jarzynski's equation (equality). Violations of second law. Implicit Ligand Sampling (ILS). Metadynamics.

6) Multi-scale methods. Modeling of reactive phenomena. Effects of the environment. Models of the continuous. Onsager schemes and PCM scheme. Hybrid quantum-classical methods (QM-MM). Additive schemes. Coupling between subsystems. Subtractive schemes (Oniom).

7) Computational implementations. Use of parallel implementations. Use of CPUs and boards graphics (GPUs).

8) Protein Simulation. Stability of protein dynamics and its characterization. Calculation of mean squared deviations (RMSD). Calculation of the mean fluctuation (RMSF). Clusterization, Normal Modes and Essential Modes. Correlation of Movements. Coefficients of involvement. Alosterism models, population change vs stereochemical Hemoglobin as an example of allosteric protein. Models of allosterism.

9) Simulation of Nucleic Acids. Evolution of force fields for DNA and RNA. Determination of the helical conformational space. Characterization of the torsional backbone and riboses. Calculation of the helical parameters. Existence of sub-states and structural sequence dependent polymorphisms. Calladine's Rules for DNA and its extension (ABC consortium). Calculation of the essential dynamics, flexibility, and bending of DNA. Calculation of the secondary structure and its classification according to Leontis & Westhof, and the pseudo-torsional η / θ for RNA. Particularities of the interaction of nucleic acids with proteins (direct- vs indirect-readout, conformational selection vs induced-fit, 2'OH as a switch conformational). Particularities of the interaction with cations (mono and di-valent) and the force ionic medium. Calculation of concentrations of ions, water, protein residues, or drugs using helical curvilinear coordinates. Applications of biological / biomedical interest.

10) Macromolecular complex prediction methods

Protein ligand interaction, prediction methods and calculation of affinities. Contributions to the binding free energy. Calculation of the energy term, prediction of the change in entropy of binding, prediction of change in free energy of solvation. Poisson Boltzman methods and Generalized from Born (mmpb (gb) sa). Algorithm-based complex prediction methods genetic (Autodock). Methods based on Fourier Transforms (FFT). Use of grids (FT-Dock). Scoring functions (Electrostatic partitioning, contact-vdw and solvation methods, use of contact atomic energies (ACE)). Protein-protein interaction. Prediction methods of protein-protein complexes, homo and heterodimers, formation of multimers. Methods of clustering (Clus-pro). Surface Complementarity Methods (Patch-Dock). Characterization of the complexes. Protein-protein interaction in transfer complexes electronics.

11) Multiscale and Coarse-Grain Models. Introduction to Coarse-Grain Simulations (GG). Derivation of models and parameters following bottom-up and top-down strategies. Models GG for membranes, proteins, DNA (from double helices to fragments of chromosomes),

polysaccharides and aqueous solvent for molecular dynamics. Multiscale methods for solutes
atomic with GG solvent or atomic solute / GG. Examples of GG force fields. The field of
SIRAH force and application examples. Limitations of the GG approximations.
12) Membrane simulations: Lipid bilayers as a model of biological membranes.
Phase transitions of lipid bilayers. Study of physical properties of bilayers by
Classical molecular dynamics simulations: correlation with experiments.
Asymmetric bilayers.
Pore formation Bio-medical applications: interaction of drugs with membranes and systems
by Drugdelivery. Other methods for the study of membranes through simulations.

Bibliography

1) Understanding molecular simulation; from algorithms to applications, Dan Frenkel, Berend Smit, AP, 2001.
2) Molecular Modeling, principles and applications, A. Leach, Pearson, 2001.
3) The art of molecular dynamics simulation, D. Rapaport, 2nd edition, Cambridge Press, 2004.
4) Quantum Chemistry, 7th edition, I.N. Levine, 2013.
5) Computer simulation of biomolecular systems, W.F. van Gunsteren, P.K. Weiner, Springer, 1997.
6) Computer simulation of protein structures and interactions, S. Fraga, J.M. Robert Parker, and J.M, Pocock, Springer, 2013.

Bioinformatics

- Origin of sequential variability. Introduction to the evolution of nucleic acids and proteins. Mutations and substitutions. Fixation. Population's genetics. Sequential and structural divergence. Homology Evolution convergent and divergent. Orthologous and paralogic proteins. Evolution of modules. Structural conditioning to sequential divergence. Neutral theory and selectionism.
- Sequential similarity. Paired and multiple sequential alignments. Identity and similarity between sequences. Statistical significance. Sequence comparison matrices. Similarity searches sequential. Local and global alignment algorithms. Heuristic searches. Databases of sequences, families and sequential motifs. Protein and acid databases nucleic. Use of individual sequences, arrays (profiles) and Hidden Markov Models (HMM) for the search for similarity. Processing of the results. Search for reasons sequential. Databases of sequential motifs and patterns.
- Structural similarity. Parameters to evaluate structural similarity. RMS alignment and calculation. Evaluation statistics of an alignment. Matrices of contacts. Structural databases. Search structural similarity. Non-redundant databases of structures.
- Secondary structure estimation. Methods for estimating the secondary structure. Segment estimation transmembrane. Polarity. Methods based on alignments and individual sequences. Consensus methods. Databases of secondary structures and their homologous sequences.
- Estimation of the tertiary structure.
Estimation of the tertiary structure by energy and sequential methods. Methods sequence-sequence, sequence-profiles, profile-profiles, hidden Markovian models. Threading. Estimation of tertiary structure at the genomic level. Region estimation coiled-coiled.
- Phylogenetic inference

Introduction to the probabilistic methods of molecular evolution. Evolution models molecular: sequential and codon, protein and nucleic acid. Maximum methods likelihood for calculating the evolutionary distance and topology. Statistical evaluation of topologies.
- Introduction to molecular modeling
Homology modeling. Estimation of the tertiary structure using ab initio calculations.
Structural and energy validation of the obtained models. Protein-protein interactions.
Prediction of oligomeric structures. Docking techniques.
- Structure-evolution integration: Prediction of biological function
Methods for functional prediction. Methods for Predicting Relevance Sites biological. Biological function assignment. Structural and functional divergence.

Bibliography

- Protein Evolution. Laszlo Patthy. Blackwell Science 1999.
- Computational Molecular Biology. Peter Clote and Rolf Backofen. Ed. Wiley. 2000
- Introduction to Bioinformatics. Anna Tramontano. Chapman & Hall/CRC. 2007
- Bioinformatics for Dummies. Jean-Michel Claverie and Cedric Notredame. Wiley. 2007.
- Molecular Evolution. A phylogenetic Approach. Roderic Page and Edward Holmes. Blackell Science. 1998.
- Molecular Evolution. Wen-Hsiung Li. Sinauer Associates. Inc. 1997.
- Bioinformatics. Genes, Proteins and Computers. CA Orengo, DT Jones and JM Thornton. Bios Scientific Publichers. 2003.
- Introduction to computational biology. An evolutionary approach. B Haubold and T Wiehe.Ed. Birhauser. 2006.

Methods in Machine Learning

1) General introduction.
Supervised vs unsupervised learning. Learning by reinforcement. Motivation and applications in bioinformatics.
2) Topics of inferential statistics.
Bases of inferential statistics. Statistics, why? Probab distribution, cumulative distrib. Examples Hypothesis test. Type I error and Type II error. Power.
Parametric vs non-parametric tests.
Test of interest: Shapiro-Wilk test, Fisher Exact Test, Mann-Witney-U test, Kolmogorov-Smirnov Test.
Correction by multiple testing.
Bonferroni. FDR.
Examples of application in bioinformatics. Differential expression. Gene Ontology.
3) Supervised learning. General concepts
Supervised learning models: regression and classification.
Binary classification vs multiple category classification. Decision threshold in classification models. Ideal predictor. Random predictor.
Generalization. Overfitting. Bias-Variance trade-off. Regularization.
Model parameters, fit via maximum likelihood.
Hyperparameters. Resampling methods. Cross validation.
Validation setup. Estimation of the error.
Class imbalance.
Better methods vs better variables.
4) Generalized Linear Models:

Univariate logistic regression. Formulation of the univariate and multivariate model.
Setting parameters. Examples
L1 and L2 regularization.
Regularization hyperparameter. Selection via cross validation.
Logistic regression for multiclass response vector.

5) Nonlinear models. Part 1
Decision trees. Intuitive idea. Stratification of the space of variables.
Criteria for cutting branches of the tree. Purity: Gini Impurity. Cross entropy.
Greedy approach for building the model. Overfitting. Pruning the tree to improve its generalizing power. Hyperparameters. Setting via cross validation.
Trees vs linear models. Pros-Cons of Tree Models.
6) Evaluation and comparison of models.
Evaluation of a classifier. Performance measures. Threshold dependency decision. Independent decision threshold measures. ROC and PR curves.
Aŕea under the AUC curves. Statistical distribution of AUC. Confidence interval.
Variance as a function of the sample size. Statistical comparison of classifiers. DeLong Test.
7) Nonlinear models. Part 2: Assembly models.
Bagging. Intuitive idea. Bagging trees. Formulation of the model. Sort by majority of votes. Hyperparameters.Estimation of the error via out of bag.
Importance of variables. Effect of co-linearities in the variables.
Random Forest. Idea and differentiation regarding bagging. Hyperparameters.
Extensive application cases.Boosting. Intuitive idea. Sequential nature.
Boosting trees. Hyperparameters. XGBoost. Qualities, advantages-disadvantages.
Hyperparameters.

Bibliography

1)Statistical Modeling and Machine Learning for Molecular Biology. Alan M. Moses University of Toronto, Canada. CHAPMAN & HALL/CRC Mathematical and Computational Biology Series. 2017.
2) An Introduction to Statistical Learning with Applications in R. Gareth James Daniela Witten Trevor Hastie Robert Tibshirani. Springer. 2015