MB&B 452b – Biological Data Science – Spring 2016

Based on lectures by Prof. Mark Gerstein + guests

James Diao

Table of Contents

Lecture 1: Introduction to the Class

Lecture 2: Genomics I (sequencing tech and genomes)

1. Methods of sequencing
    a. First-Gen
        i. Maxam-Gilbert sequencing: radiotag at 5' end, random breaks generate random lengths under varied conditions (G, A+G, C, C+T), run on gel and analyze.
        ii. Sanger sequencing: dideoxy method. Add chain-terminating nts (no 3'-OH). Random termination generates random lengths. Run on gel and analyze.
    b. Second-Gen
        i. Illumina: fluorescent reversibly-terminated nts
        ii. Ion Torrent: measure protons
        iii. PacBio: fluorescent nts (real-time).
    c. Third-Gen
        i. Nanopore-based (run DNA through a pore, measure changes in current)
        ii. Transistor-based (nanopore with field-effect transistor → electronic effect)
        iii. FRET-based (Forster resonance energy transfer- donor and acceptor chromophores).
2. Steps
    a. Sample/Library preparation: Isolation, shearing, blunting, A-tailing, ligation
    b. Sequencing: Flow cell loading, cluster generation, sequencing by synthesis, image analysis, de-multiplexing.
        i. Flow cell: lanes with lawns of oligos complementary to library adaptors.
        ii. Attachment, bridging, cluster generation.
        iii. Sequencing by synthesis with fluorescent, reversibly-blocked nts.
        iv. Multiplexing: barcodes identify samples, all run together.
    c. Data analysis: read QC filtering, alignment, etc.
3. Illumina output (fastq):
    a. [1] Read identifier, [2] sequence, [3] "+" (Q score id), [4] Q score
    b. 50-250 nt per read
    c. Short reads:
        i. May miss insertions/repeats, GC bias, scaffolding gaps.
        ii. Due to incomplete incorporation of bases.
    d. Paired end reads:
        i. Sequence both ends of a fragment (instead of 1).
        ii. Known distance (~length of fragment)
        iii. Easier to align (more sequence, anchor)
    e. Other bias: size selection, enzyme specificities, selective PCR amplification.
4. Alignment
    a. Overlapping reads → contigs → scaffolds → anchor on chromosomes.

Lecture 3: Genomics II (sequencing applications)

1. ChIP-Seq
    a. Cross-link, shear (+exonuclease?), pull-down with Ab, sequence
    b. Align, compare to input to look for enrichment, call peaks at significant sites.
    c. Problems: Ab non-specificity, bad cross-linking, etc.
2. Accessibility of chromatin: ATAC-seq, FAIRE-seq, MNase-Seq
3. Conformation of chromatin: 3-5 C, Hi-C
4. RNA-Seq
    a. Start with mRNA or total RNA, remove DNA, fragment RNA, RT to cDNA, ligate adaptors, amplify, select size, sequence cDNA ends.
    b. Technical issues: range of concentrations, strand-specificity, degradation, splicing, 2^o structure
    c. Normalization:
        i. Internal: reads per kilobase of feature length per million mappe reads.
        ii. External: reads relative to standard spike
5. Ribosome Footprinting: gives translation reading frame.

Lecture 4: Proteomics and P-P Interactions

1. Whole genome editing: UAG from STOP to new AA.
2. Mass Spec
    a. Measure m/z of ionized samples
    b. Ionizer, mass-filter, detector
    c. LC-MS: shotgun proteomics: proteins → trypsin, liquid chromatography to isolate peptide, MS/MS (gaps between m/z peaks correspond to AAs)
    d. Sequence coverage: misses chunks
    e. Proteome fractionation: separate with chromatography or Ab pull-down.
    f. Ionization techniques
        i. Electrospray: high voltage → ionized aerosol (evaporates to naked particles).
        ii. MALDI: matrix-assisted laser desorption ionization. Sample mixed into matrix material, laser ablates and desorbs sample. Hot plume of gases → ionization.
3. P-P Interactions:
    a. Yeast Two-Hybrid Assay:
        i. Gal4 AD + Prey
        ii. Gal4 BD + Bait
        iii. If active, colonies appear on -His plates.
    b. Tandem Affinity Purification (TAP) tagging
        i. Clone Protein + Cam-BD + TEV + IgG BD
        ii. Bait protein (of interest) binds interaction partners
        iii. Pull down with IgG and cleave with TEV protease
        iv. Pull down with Cam and separate with chelation.
        v. SDS-PAGE, tryptic digest, MS/MS
    c. Proximity Biotinylation: BirA adds biotin to interactors; pulled down by streptoavidin.

4. Quantitative Proteomics
    a. SILAC: stable isotope labeling with AA in cell culture
        i. Gives differential protein expression
        ii. 2 samples: light and heavy. Combine proteins and MS/MS. Look at ratio of abundances for light/heavy version.
        iii. Applications
            1. Specificity of interaction (expression level of pulldown)
            2. Phosphorylation (reduced expression of pulldown)
    b. TiO2 helps enrich for phosphorylated peptides

Lecture 5: X-Ray Crystallography and Cryo-EM

1. Crystallography
    a. Resolution limited by wavelength (diffraction limit). X-Rays give atomic detail.
    b. NMR: detects H-H distances in solution
    c. X-rays: detects direct positions in crystals (magnify signal with constructive interference).
        i. Needs 10^{12} copies of protein
        ii. Source: synchrotron.
        iii. Output: electron density map, structure model, unit cell type + dimensions
    d. Single-particle cryo-EM
        i. No crystals needed! 10^5 copies
        ii. Native: not stained or fixed
        iii. Single-particle: 3D reconstruction from images at different angles

Lecture 6: Databases in Biosciences

1. Data sources: drug research, social media, patient records, gene sequencing, medical test results, claims, home monitoring, mobile apps
2. 3Vs: volume, variety, velocity.
3. Data sources → database → API, interactive queries, download → User
4. DB: scalable, multiple users, queries and management.
5. Concepts
    a. Integrity, redundancy, dependency (linkages), security, quality (intended use)
6. Relational database management system (RDBMS): based on set theory
    a. Implementations: MS SQLServer, MySQL, Oracle, etc.
    b. Table (relation): represents some class (patients, hospitals, etc.)
        i. Consists of: attributes (columns) and object instances (rows)
        ii. Primary key: single/multiple columns with unique values.
        iii. Foreign key: key taken from a different table (i.e. ID of people with different phone numbers)
    c. Normalization
        i. Systematically organize tables to eliminate anomalies/redundancy
        ii. Additions, deletions, modifications made in just 1 table and propagated through foreign keys.

iii. Normalization reduces performance! Requires joins in report queries.
iv. Normal form: tables free of certain anomalies

d. Normal Forms
    i. First Normal Form (1NF):
        1. Columns: single entry, single data type, unique names, no order.
        2. Rows: unique, no order.
    ii. Second Normal Form (2NF):
        1. 1NF + all non-key columns are dependent on the key.
        2. Single-column keys are automatically 2NF.
    iii. Third Normal Form (3NF):
        1. 2NF + no transitive dependency (Zip → City)

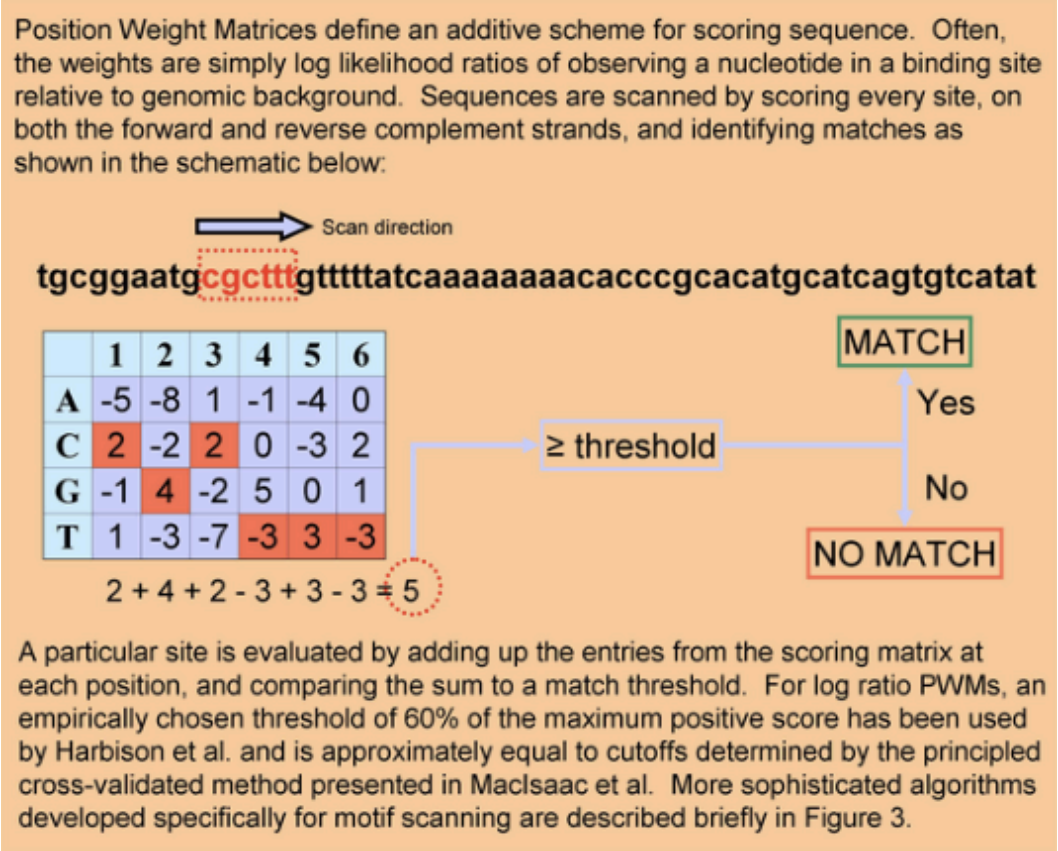
7. Entity Relationship Diagram (ERD)
    a. Description: data model diagram.
        i. Entity: collection of objects
        ii. Attribute, relationship
        iii. Cardinality: 1-1, 1-m, m-n
    b. Describes attributes + relationships of entities
8. OLTP: Online transaction processing
    a. Info systems (e.g. databases) that help data entry/retrieval
    b. Supports: insert, update, delete, select rows
9. Structured Query Language: help create, select, insert, delete, update data.
    a. CREATE DATABASE …
    b. CREATE TABLE … (item type, item type, item type)
    c. INSERT INTO … (item, item, …, item) VALUES (item, item, …, item)
    d. UPDATE … SET item=new.value WHERE attribute=a.value
    e. DELETE … WHERE attribute = a.value
    f. SELECT attributes FROM … WHERE attribute = a.value ORDER BY attribute
10. Star Schema
    a. Center holds IDs for different entities (patients, providers, clinics, procedures), each in their own table.

Lecture 7: Personal Genomics

1. Costs
    a. Sanger sequencing: $100 M
    b. Next-gen sequencing: <$1 M
    c. $3,100 for exam, blood draw, sequencing, risk report.
2. Genome variation
    a. SNPs, INDELs, SVs (>100 nt)
    b. Around 3M SNPs
3. iPOP (integrated personal omics profile) uses genome and other data over time.
    a. Incl. transcriptomic, proteomic, metabolomics, medical exams, etc.
    b. Longitudinal data tracks dynamic regulation during infection, etc.
    c. Diploid instead of hapoid (reference) finds allele-specific expression.
4. Process
    a. Raw reads (fastq) → human aligned reads (BAM)

Lecture 8: Sequence Analysis

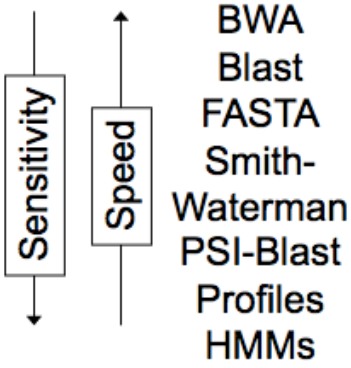
1. Sequence Comparison
    a. Alignment with Dynamic Programming
        i. Key idea: score (i, j) = previous best score + score here
        ii. Similarity (substitution) matrix: scores A-B match for all Ai Bj.
            1. PAM, Blossum, Gonnet.
            2. PAM 70 vs PAM 250 (distant)
            3. Freq of AA substitutions, log-odds: log2(freq-OBS / freq-EXP)
        iii. Gap Penalties: usually opening + extension * N
        iv. Needleman-Wunsch (wiki)
            1. Dot matrix at matches, sum matrix, highest number, traceback.
    b. Global alignment: matches whole thing
    c. Local alignment: matches part of it.
2. Multiple Sequence Alignment
    a. Progressive multiple alignment: progressively built from most closely related to less related. Local minimum problem. Parameter choice.
    b. Clustal uses average linkage clustering- mean of groups.
    c. Position weight matrix:



    d. Profile: position-specific scoring matrix of 21 (AA) columns and N rows. Chance of finding column (AA) at row (position)
    e. PSI-blast: position specific iterative BLAST.
        i. Input → profile, research, new query… build DB.

f. EM:
    i. Expectation step: uses temp parameters to compute likelihood
    ii. Maximization step: uses likelihood estimate to compute parameters
g. Gibbs sampling:
    i. Toss out an instance and use the rest to define a weight matrix.
    ii. Pick a new toss-out instance according to this matrix.
    iii. Return highest-scoring motif.

3. Alignment Speed and Complexity
    a. DP is $O(n^2)$, too slow.
    b. FASTA: hashes short words in query
    c. BLAST: more efficient query hashing. Takes overlapping words and calculates PAM matrix. Extends high-scoring pairs left and right maximally. O(n)
    d. BLAT: hashing the DB
        i. Huge hash table means faster scan but large memory usage.
    e. BWA/Bowtie: burrows-wheeler transform of DB: reversible cyclic permutation, sort, extract last column.



Lecture 9: Variant Identification

1. Detecting genomic variants
    a. Call SNPs, look for deletions (split reads, 0 coverage), duplications (2x coverage), insertions (overhangs), inversions (paired end mixups)
    b. AGE: Alignment with Gap Excision: SW (local) at both ends to find INDELs.
    c. Bayes Theorem to detect variants: P(Genotype | Data)
2. HMMs?
3. Mean-shift-based (MSB) segmentation
    a. Discontinuity-preserving smoothing (few assumptions)
    b. Each bin: attraction vector points toward bins with most similar signal
4. HR-PEM: High-resolution paired end mapping (???)
5. Pseudogenes and Duplications: not much
6. RDV (retroduplication variation) and Mobile Elements
    a. mRNA → inserts into genome
7. 1000 Genomes Project:
    a. Typical genome differs by 4-5M SNPs, 2k SVs.

Lecture 11: Unsupervised Data Mining

1. Saturation: fraction of genome coverage (y) by any 1:n rows (x)
2. Aggregation: find repeating signal and aggregate (usu. bell curve)
3. Clustering: by rows or columns
    a. Agglomerative: bottom up or top down.
    b. K-means: initialize, assign, recompute centers, reassign.
4. Networks: adjacency matrix (can be weighted)
5. SVD: $A = USV^T$
    a. $AV = US$: maps row space (V) to column space (U)
    b. A expressed as sum of rank-1 matrices ($s*u.v^T + \dots$)
    c. S is non-negative singular values from largest to smallest.
    d. Works on dependent datasets. Okay with non-normal, imprecise.
    e. REQUIRES LINEARITY and nonsparsity.
6. Weighted Gene Co-Expression Network
    a. Module detection: find with hierarchical clustering
7. Biplot: dimensionality reduction of points AND FEATURES
8. CCA: Canonical Correlation Analysis
    a. Takes two feature SETS that are linearly weighted and look at shared dimensions/variance.
    b. Correlation circle visualization: closer to outer circle = higher correlation. Variables in the same direction are correlated.

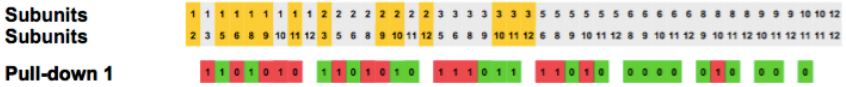
Lecture 12: Supervised Data Mining

1. Good rule: increases homogeneity—information theoretic entropy is popular: minimize S = -SUM(p log p). P = frequency of label.
2. Fisher discriminant analysis: find a linear combination that maximizes ratio of TOTAL(separation of class means) / SUM (class variances).
3. SVM: maximum margin hyperplane.
    a. Soft margin subject to cost * sum slack variables.
    b. Kernel adds non-linearity

Lecture 13: Skipped

Lecture 14: Predicting Networks

1. Examples
    a. Protein-protein (phosphorylation)
    b. Metabolic
    c. Regulatory (Chip-Seq)
        i. Call peaks and draw edge between TFs if they share a target
2. "Squared" scale from (N choose 2) possible interactions.
3. Predicting Networks via Bayesian Integration:
    a. R = weights * features + weight_0
       = weighted vote + bias
    b. Intersection: 0 dominant. Union: 1 dominant.
    c. w_i = TPR_i/FPR_i where prediction is P(feature | interaction)
    d. If n is small, you might have P = 0, which dominates the classifier.
       Replace with (TPR_i + m/k) / (FPR + m) where m is a weight (virtual instances) and k is the number of possible values of x.
4. Cross-validation
5. ROC curve (TPR by FPR)



$$L_1 = \frac{p(x_1 \mid GSTD+)}{p(x_1 \mid GSTD-)} = \frac{6/13}{11/32} = 1.34$$

$$L_0 = \frac{p(x_0 \mid GSTD+)}{p(x_0 \mid GSTD-)} = \frac{4/13}{14/32} = 0.70$$

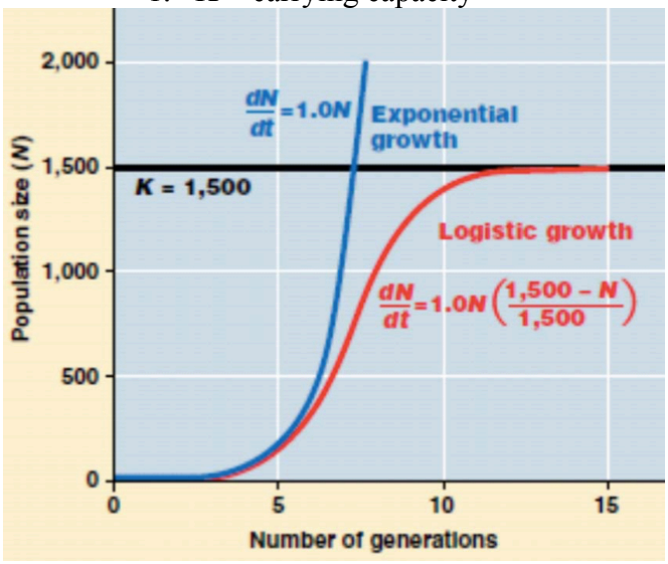
6. Many negatives => low prior => low PPV. Balanced examples is important.
7. Likelihood ratio: P(F | + I) / P(F | - I)

Lecture 15: Deep Learning

1. AI, Logic, Learning
    a. Building systems, reasoning, learning.
    b. Supervised: classification and regression. Requires input/output pairs.
    c. Unsupervised: find structure (clusters) or underlying distribution
    d. Semi-supervised: uses structure of data to inform supervised learning.
    e. Reinforcement learning: desired output given only after a sequence of actions.
2. Evaluating Performance
    a. Goal is generalization.
    b. Divide your data to training/validation/test, and/or use CV.
    c. Classification: Accuracy, Sensitivity, Specificity, TPR, FPR.
        i. ROC analysis for binary classifiers
    d. Regression: Sum of squares error and RMS error
3. Dimensionality and Overfitting
    a. Occam's razor: simplest explanation that fits the data.
    b. Curse of dimensionality: feature space grows quickly; data becomes sparse.
4. Artificial Neurons
    a. $y(x) = g(w^T x + w_0)$
    b. g can be sigmoid/logistic: $1/(1+e^{-z})$, tanh, +/- 1 step at z=0, rectified linear (ReLu)
    c. Differentiable activation function => gradient-based optimization.
        i. Move opposite the gradient repeatedly until Δerror < threshold.
    d. Limitation: monotonic activation function means linear decision boundary.
5. Multilayer NN
    a. Hidden units: makes data linearly separable
    b. Universal approximation theorem: ANNs can approximate any function to arbitrary accuracy with enough hidden units AND non-linear activation.
6. Error functions
    a. Regression: sum of square error: good for Gaussian noise
    b. Classification: cross entropy
7. Backpropagation
    a. Step 1: input is propagated forward and compared to the desired output.
    b. Step 2: error values are computed and propagated backwards until each neuron has an associated error value.
    c. Step 3: compute the gradient and update the weights.
8. ConvNets
    a. Four key ideas: Local connections, shared weights, pooling, many layers
    b. Feature map: each unit is connected to a local patch through a set of weights.
    c. Pooling units: compute maximum of patch
    d. Convolve with a small 3x3 filter matrix => convolved feature
        i. Edge detection, sharpen, box blur, Gaussian blur, etc.
9. Dropout (bagging)
    a. Sets output of ½ of the neurons to 0 in each pass.
    b. Forces neurons to learn more robust features.
    c. Equivalent to sampling a different architecture (same weights) with each input.

Lecture 16: Modeling and Simulation (Computational Immunology)

1. Statistical analysis
    a. Begin with large data set, find patterns, and generate predictions
    b. PCA, regression, network analysis
2. Mechanistic/Dynamic models
    a. Begin with hypothesis, write equations, run simulations, and generate predictions
    b. Dynamical systems, parameter estimation, ODE/PDE, stochastic models
3. Prediction
    a. Interpolation: within sample predictions
    b. Extrapolation: outside sample predictions
4. Modeling benefits: predictions, clarify/simplify, hypothesis generating
5. Modeling problems: requires assumptions (garbage in, garbage out), needs validation
6. Immunology
    a. Contains feedback loops and non-linear dynamics.
    b. Distributed system of 10^{12} cells and molecules.
7. Dynamic = over time
    a. Continuous/discrete, deterministic/stochastic
    b. ODEs: may not be solvable = simulation
        i. dN/dt = rN => can compute doubling time and half-life.
        ii. Steady-state: set derivatives = 0 (as time => Inf)
            1. Stable = small perturbations return to same state
        iii. Density dependence
        iv. Logistic model (S-shaped):
            1. K = carrying capacity



$$\frac{dN}{dt} = rN\left(1-\frac{N}{K}\right)$$

8. Modeling Interactions
    a. Law of mass action (mean-field assumption)
        i. Entities encounter each other according to relative abundance (Rate of rxn is proportional to [C])
    b. Phase plane analysis: nullclines plot where derivatives = 0 (cross at steady state)
        i. Plots typical trajectories into the state space.

9. Forward modeling
    a. Model generated from literature parameters; for simulating synthetic data
10. Inverse modeling
    a. Model designed to fit experimental data for quantifying parameters of interest.
    b. Minimizes difference between model and data.

Lecture 17: Modeling and Simulation II (Computational Immunology)

1. BrdU labels cells during S phase; rate of increase => proliferation
2. Models are identifiable if you can learn underlying parameters
    a. Parameters are identifiable if they affect the data and can be estimated from data.
3. Inverse Modeling – 6 steps
    a. Select an appropriate model
    b. Define cost function
    c. Adjust model parameters for best fit
    d. Examine goodness of fit
    e. Determine whether much better fit is possible
    f. Evaluate accuracy (confidence, uniqueness)
4. Euler's method: numerical solutions to ODEs.
    a. Move a short distance on the tangent; recompute slope/gradient, repeat.
5. Maximum likelihood estimation (inverse modeling)
    a. Maximize P(data | parameters)?
6. Goodness of Fit and Residuals Plot
    a. Look at distribution of residuals- equally and normally distributed; no trends.
7. F-statistic: deviations between groups / deviations within groups
    a. Forward selection: add variables as long as significant F-test
    b. Backwards selection: remove variables without significant F-test
8. Data points must exceed number of parameters.

Lecture 18: Modeling and Simulation III (Computational Immunology)

1. Accuracy of Estimated Model Parameters
    a. Monte Carlo Simulation
        i. Use estimate as true value to simulate distribution => synthetic data sets.
    b. Bootstrap method
        i. Resample from dataset with replacement => synthetic data sets
        ii. Only works if sample is representative. Does not rely on knowledge of measurement errors/noise.
    c. Synthetic data set => sampling distribution of parameter(s).
2. Examples:
    a. Model viral dynamics
    b. Model epidemics (SIR Model: Susceptible, Infectious, Removed).

$$\frac{dS}{dt} = -\beta SI$$

$$\frac{dI}{dt} = \beta SI - \mu I$$

$$\frac{dR}{dt} = \mu I$$

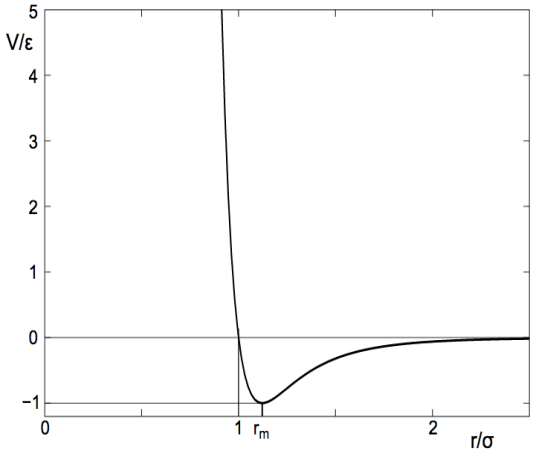
S is the population of susceptible individuals
I is the population of infectious individuals
R is the population of individuals who were infected, but have now recovered
β is the infection rate
μ is the recovery rate

3. Basic reproductive ratio (R_0) = average number of secondary cases caused by an infectious individual in a totally susceptible population.
    a. $R < 1$: disease dies out. $R > 1$: disease can invade.
4. Pseudo-Random Number Generators
    a. Seed is often system clock.
    b. Linear congruential generator: aI_i + c (mod m)
    c. Mersenne Twister: very long period
5. Simulating from other distributions:
    a. Needs: indefinite integral; f(x) must be invertible.
6. Boolean Network Models
    a. Qualitative: useful when kinetic parameters unknown
    b. Directed graph: nodes = elements, edges = regulatory relationships
    c. Nodes are either true or false: N nodes = 2^N states.
    d. Node state determined by transfer function (of neighbor states)
        i. Generally logical (NOT, AND, OR, XOR, etc.)
    e. Nodes are related functions (pro-inflammatory, cell types, bacteria, etc).
7. ODEs neglect spatial structure
    a. PDEs allow variation over time and space.
    b. Compartment modeling: elements in well-mixed compartments (+ movement between) tracked using ODEs.
8. Cellular Automata Models
    a. Conway's Game of Life: grid of cells, each in finite number of states.
9. Agent-based modeling (ABM):
    a. object-oriented, discrete-event, rule-based, stochastic.
    b. Views system as agents that follow rules.
10. Modeling frameworks
    a. Individual particle-based stochastic
    b. Particle number stochastic
    c. Concentration-based (non)spatial, (non)stochastic
11. XML encoding: markup language

Lecture 19: Protein Folding I

1. Protein-folding problem: find 3D structure from AA sequence
    a. Detailed: compute & minimize atomic-level free energy for all conformations
    b. Coarse: same, but residues + solvent
    c. Levinthal's paradox: number of conformations = #angles^(2N)
    You cannot just sample all of the states; too many of them.
    d. Smooth energy landscape = no intermediates.
2. Dihedral angles: from stereospecific bonds. Total angles = dihedral bonds + 2
3. Driving forces
    a. Folding: hydrophobicity, H-bonding, van der Waals, …, electrostatic
    b. Unfolding: entropy.
4. Ramachandran Plot: ok phi/psi angles. Beta is top-left, alpha is left, slightly bottom
5. # conformations AND # energy minima increase exponentially.

Lecture 20: Protein Folding II



1. Random close packing in protein cores
    a. Lennard-Jones potential:
    b. What is the packing fraction?
        i. 0.74 for hard spheres,
        0.64 for disordered,
        0.56 for all-atom in protein cores
        ii. Volume of residue / volume of container. Container found by summing volumes of Voronoi polyhedral enclosing each atom.
        iii. NOT as high as crystal close-packed, once you consider explicit hydrogens and calibrated radii.
    c. Force-fields necessary to model protein structure (not solid spheres)
2. Df = N – 3 (after bond length and angle constraints)
3. Hard-sphere model: repulsive interactions between non-bonded atoms.
4. Finding surface atoms: throw down random points. If the closest atom is more than 1.4 A away, it is a surface atom. (Imagine tracing sphere over the protein).
5. High df. Exponential growth of conformations/minima.
6. 3-letter BNL model (hydrophobic, neutral, hydrophilic)
7. Molecular dynamics
    a. Equations: $F = m\, d^2x/dt^2$
    b. $F = -dV/dr$

Lecture 22: Markov Chains I

1. Protein aggregation (bad!)
    a. Unfolded => nucleus => protofibrils => amyloid
    b. Exposed hydrophobic regions bind in (partially) unfolded protein.
    c. Caused by overproduction, stress, mutation
2. smFRET: single-molecule forster resonance energy transfer
    a. Tells you when two modified sites are close together
3. Coarse-grained model:
    a. Harmonic potential for angles and bond lengths
    b. Captures effective force and radius of gyration for intrinsically disordered proteins (IDPs)
    c. MAPt = microtubule associated protein tau -