

# Challenges

---

- How can we better support interdisciplinarity in the liberal arts?
- Should the first course prepare majors and/or serve mostly a non-major audience?

# Responses

---

- An interdisciplinary, problem-focused introductory course
- A (truly) interdisciplinary data analytics program

# Contents

## 1 Molecular Biology and Biological Chemistry 1

### The Genetic Material 2

- Nucleotides 2
- Orientation 3
- Base pairing 5
- The central dogma of molecular biology 6

### Gene Structure and Information Content 7

- Promoter sequences 7
- The genetic code 9
- Open reading frames 9
- Introns and exons 12

### Protein Structure and Function 13

- Primary structure 13
- Secondary, tertiary, and quaternary structure 14

### The Nature of Chemical Bonds 15

- Anatomy of an atom 17
- Valence 17
- Electronegativity 18
- Hydrophilicity and hydrophobicity 19

### Molecular Biology Tools 19

- Restriction enzyme digests 20
- Gel electrophoresis 21
- Blotting and hybridization 21
- Cloning 23
- Polymerase chain reaction 24
- DNA sequencing 25

iv

Contents

### Genomic Information Content 27

- C-value paradox 27
- Reassociation kinetics 28

### Chapter Summary 30

### Readings for Greater Depth 31

### Questions and Problems 31

## 2 Data Searches and Pairwise Alignments 33

### Dot Plots 34

### Simple Alignments 35

### Gaps 36

- Simple gap penalties 37
- Origination and length penalties 37

### Scoring Matrices 38

### Dynamic Programming: The Needleman and Wunsch Algorithm 41

### Global and Local Alignments 45

- Semiglobal alignments 45
- The Smith-Waterman Algorithm 46

### Database Searches 48

- BLAST and its relatives 48
- FASTA and related algorithms 50
- Alignment scores and statistical significance of database searches 51

### Multiple Sequence Alignments 52

### Chapter Summary 53

### Readings for Greater Depth 53

### Questions and Problems 54

## 3 Substitution Patterns 57

### Patterns of Substitutions within Genes 58

- Mutation rates 58
- Functional constraint 59
- Synonymous vs. nonsynonymous substitutions 61

- Indels and pseudogenes 62
- Substitutions vs. mutations 62
- Fixation 63

### Estimating Substitution Numbers 65

- Jukes-Cantor model 65
- Transitions and transversions 67
- Kimura's two-parameter model 67
- Models with even more parameters 68
- Substitutions between protein sequences 69

### Variations in Evolutionary Rates between Genes 70

### Molecular Clocks 71

- Relative rate test 71
- Causes of rate variation in lineages 73

### Evolution in Organelles 74

### Chapter Summary 74

### Readings for Greater Depth 74

### Questions and Problems 75

## 4 Distance-Based Methods of Phylogenetics 77

### History of Molecular Phylogenetics 78

### Advantages to Molecular Phylogenies 79

### Phylogenetic Trees 80

- Terminology of tree reconstruction 80
- Rooted and unrooted trees 81
- Gene vs. species trees 83
- Character and distance data 84

### Distance Matrix Methods 85

- UPGMA 86
- Estimation of branch lengths 88
- Transformed distance method 90
- Neighbor's relation method 91
- Neighbor-joining methods 92

### Maximum Likelihood Approaches 93

### Multiple Sequence Alignments 93

### Chapter Summary 94

vi

Contents

### Readings for Greater Depth 95

### Questions and Problems 95

## 5 Character-Based Methods of Phylogenetics 97

### Parsimony 98

- Informative and uninformative sites 98
- Unweighted parsimony 99
- Weighted parsimony 104

### Inferred Ancestral Sequences 104

### Strategies for Faster Searches 105

- Branch and bound 105
- Heuristic searches 107

### Consensus Trees 108

### Tree Confidence 109

- Bootstrapping 109
- Parametric tests 111

### Comparison of Phylogenetic Methods 112

### Molecular Phylogenies 112

- The tree of life 112
- Human origins 114

### Chapter Summary 114

### Readings for Greater Depth 115

### Questions and Problems 115

## 6 Genomics and Gene Recognition 117

### Prokaryotic Genomes 118

### Prokaryotic Gene Structure 120

- Promoter elements 121
- Open reading frames 124
- Conceptual translation 125
- Termination sequences 125

### GC Content in Prokaryotic Genomes 126

### Prokaryotic Gene Density 127

### Eukaryotic Genomes 127

### Eukaryotic Gene Structure 129

- Promoter elements 130
- Regulatory protein binding sites 131

### Open Reading Frames 133

- Introns and exons 134
- Alternative splicing 135

### GC Content in Eukaryotic Genomes 137

- CpG islands 137
- Isochores 141
- Codon usage bias 142

### Gene Expression 143

- cDNAs and ESTs 143
- Serial analysis of gene expression 145
- Microarrays 145

### Transposition 148

### Repetitive Elements 148

### Eukaryotic Gene Density 150

### Chapter Summary 151

### Readings for Greater Depth 151

### Questions and Problems 152

## 7 Protein and RNA Structure Prediction 155

### Amino Acids 156

### Polypeptide Composition 159

### Secondary Structure 160

- Backbone flexibility,  $\Phi$  and  $\Psi$  160
- Accuracy of predictions 161
- The Chou-Fasman and GOR methods 162

### Tertiary and Quaternary Structure 164

- Hydrophobicity 165
- Disulfide bonds 166
- Active structures vs. most stable structures 167

### Algorithms for Modeling Protein Folding 167

- Lattice models 168

viii

Contents

- Off-lattice models 170

- Energy functions and optimization 171

### Structure Prediction 172

- Comparative modeling 173
- Threading: Reverse protein folding 174

### Predicting RNA Secondary Structures 175

### Chapter Summary 176

### Readings for Greater Depth 177

### Questions and Problems 178

## 8 Proteomics 179

### From Genomes to Proteomes 180

### Protein Classification 181

- Enzyme nomenclature 181
- Families and superfamilies 182
- Folds 183

### Experimental Techniques 184

- 2D electrophoresis 184
- Mass spectrometry 185
- Protein microarrays 187

### Inhibitors and Drug Design 187

### Ligand Screening 188

- Ligand docking 189
- Database screening 190

### X-Ray Crystal Structures 191

### NMR Structures 197

### Empirical Methods and Prediction Techniques 197

### Post-Translational Modification Prediction 198

- Protein sorting 199
- Proteolytic cleavage 202
- Glycosylation 202
- Phosphorylation 203

### Chapter Summary 203

### Readings for Greater Depth 204

### Questions and Problems 205

# Contents

<b>Preface</b>	<b>xv</b>
<b>1 Introduction</b>	<b>1</b>
<b>2 Algorithms and Complexity</b>	<b>7</b>
2.1 What Is an Algorithm?	7
2.2 Biological Algorithms versus Computer Algorithms	14
2.3 The Change Problem	17
2.4 Correct versus Incorrect Algorithms	20
2.5 Recursive Algorithms	24
2.6 Iterative versus Recursive Algorithms	28
2.7 Fast versus Slow Algorithms	33
2.8 Big-O Notation	37
2.9 Algorithm Design Techniques	40
2.9.1 Exhaustive Search	41
2.9.2 Branch-and-Bound Algorithms	42
2.9.3 Greedy Algorithms	43
2.9.4 Dynamic Programming	43
2.9.5 Divide-and-Conquer Algorithms	48
2.9.6 Machine Learning	48
2.9.7 Randomized Algorithms	48
2.10 Tractable versus Intractable Problems	49
2.11 Notes	51
Biobox: Richard Karp	52
2.12 Problems	54

<b>3 Molecular Biology Primer</b>	<b>57</b>
3.1 What Is Life Made Of?	57
3.2 What Is the Genetic Material?	59
3.3 What Do Genes Do?	60
3.4 What Molecule Codes for Genes?	61
3.5 What Is the Structure of DNA?	61
3.6 What Carries Information between DNA and Proteins?	63
3.7 How Are Proteins Made?	65
3.8 How Can We Analyze DNA?	67
3.8.1 Copying DNA	67
3.8.2 Cutting and Pasting DNA	71
3.8.3 Measuring DNA Length	72
3.8.4 Probing DNA	72
3.9 How Do Individuals of a Species Differ?	73
3.10 How Do Different Species Differ?	74
3.11 Why Bioinformatics?	75
Biobox: Russell Doolittle	79
<b>4 Exhaustive Search</b>	<b>83</b>
4.1 Restriction Mapping	83
4.2 Impractical Restriction Mapping Algorithms	87
4.3 A Practical Restriction Mapping Algorithm	89
4.4 Regulatory Motifs in DNA Sequences	91
4.5 Profiles	93
4.6 The Motif Finding Problem	97
4.7 Search Trees	100
4.8 Finding Motifs	108
4.9 Finding a Median String	111
4.10 Notes	114
Biobox: Gary Stormo	116
4.11 Problems	119
<b>5 Greedy Algorithms</b>	<b>125</b>
5.1 Genome Rearrangements	125
5.2 Sorting by Reversals	127
5.3 Approximation Algorithms	131
5.4 Breakpoints: A Different Face of Greed	132
5.5 A Greedy Approach to Motif Finding	136
5.6 Notes	137

Biobox: David Sankoff	139
5.7 Problems	143
<b>6 Dynamic Programming Algorithms</b>	<b>147</b>
6.1 The Power of DNA Sequence Comparison	147
6.2 The Change Problem Revisited	148
6.3 The Manhattan Tourist Problem	153
6.4 Edit Distance and Alignments	167
6.5 Longest Common Subsequences	172
6.6 Global Sequence Alignment	177
6.7 Scoring Alignments	178
6.8 Local Sequence Alignment	180
6.9 Alignment with Gap Penalties	184
6.10 Multiple Alignment	185
6.11 Gene Prediction	193
6.12 Statistical Approaches to Gene Prediction	197
6.13 Similarity-Based Approaches to Gene Prediction	200
6.14 Spliced Alignment	203
6.15 Notes	207
Biobox: Michael Waterman	209
6.16 Problems	211
<b>7 Divide-and-Conquer Algorithms</b>	<b>227</b>
7.1 Divide-and-Conquer Approach to Sorting	227
7.2 Space-Efficient Sequence Alignment	230
7.3 Block Alignment and the Four-Russians Speedup	234
7.4 Constructing Alignments in Subquadratic Time	238
7.5 Notes	240
Biobox: Webb Miller	241
7.6 Problems	244
<b>8 Graph Algorithms</b>	<b>247</b>
8.1 Graphs	247
8.2 Graphs and Genetics	260
8.3 DNA Sequencing	262
8.4 Shortest Superstring Problem	264
8.5 DNA Arrays as an Alternative Sequencing Technique	265
8.6 Sequencing by Hybridization	268
8.7 SBH as a Hamiltonian Path Problem	271

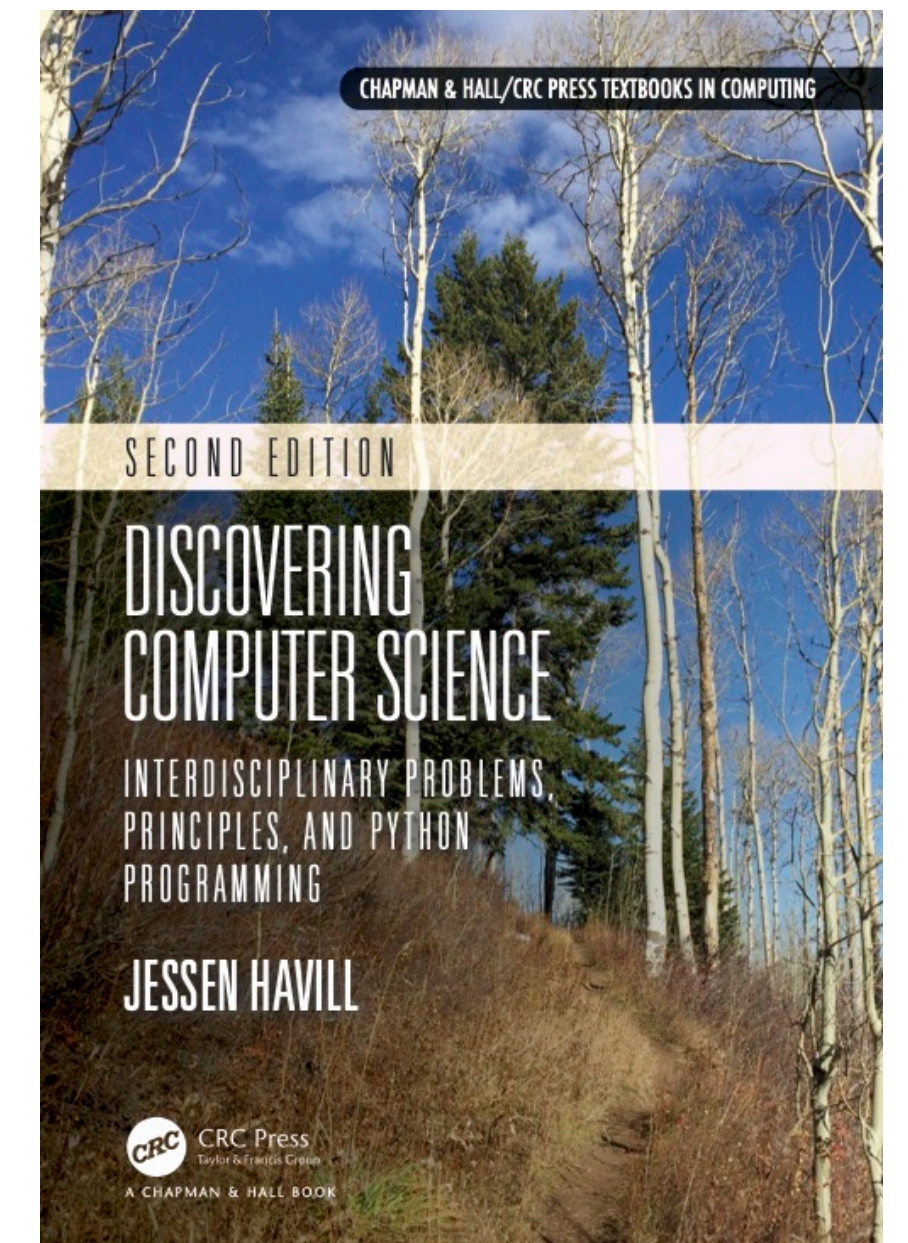
8.8 SBH as an Eulerian Path Problem	272
8.9 Fragment Assembly in DNA Sequencing	275
8.10 Protein Sequencing and Identification	280
8.11 The Peptide Sequencing Problem	284
8.12 Spectrum Graphs	287
8.13 Protein Identification via Database Search	290
8.14 Spectral Convolution	292
8.15 Spectral Alignment	293
8.16 Notes	299
8.17 Problems	302
<b>9 Combinatorial Pattern Matching</b>	<b>311</b>
9.1 Repeat Finding	311
9.2 Hash Tables	313
9.3 Exact Pattern Matching	316
9.4 Keyword Trees	318
9.5 Suffix Trees	320
9.6 Heuristic Similarity Search Algorithms	324
9.7 Approximate Pattern Matching	326
9.8 BLAST: Comparing a Sequence against a Database	330
9.9 Notes	331
Biobox: Gene Myers	333
9.10 Problems	337
<b>10 Clustering and Trees</b>	<b>339</b>
10.1 Gene Expression Analysis	339
10.2 Hierarchical Clustering	343
10.3 <i>k</i> -Means Clustering	346
10.4 Clustering and Corrupted Cliques	348
10.5 Evolutionary Trees	354
10.6 Distance-Based Tree Reconstruction	358
10.7 Reconstructing Trees from Additive Matrices	361
10.8 Evolutionary Trees and Hierarchical Clustering	366
10.9 Character-Based Tree Reconstruction	368
10.10 Small Parsimony Problem	370
10.11 Large Parsimony Problem	374
10.12 Notes	379
Biobox: Ron Shamir	380
10.13 Problems	384

<b>11 Hidden Markov Models</b>	<b>387</b>
11.1 <i>CG</i> -Islands and the “Fair Bet Casino”	387
11.2 The Fair Bet Casino and Hidden Markov Models	390
11.3 Decoding Algorithm	393
11.4 HMM Parameter Estimation	397
11.5 Profile HMM Alignment	398
11.6 Notes	400
Biobox: David Haussler	403
11.7 Problems	407
<b>12 Randomized Algorithms</b>	<b>409</b>
12.1 The Sorting Problem Revisited	409
12.2 Gibbs Sampling	412
12.3 Random Projections	414
12.4 Notes	416
12.5 Problems	417
<b>Using Bioinformatics Tools</b>	<b>419</b>
<b>Bibliography</b>	<b>421</b>
<b>Index</b>	<b>429</b>

# An interdisciplinary, problem-focused introductory course

---

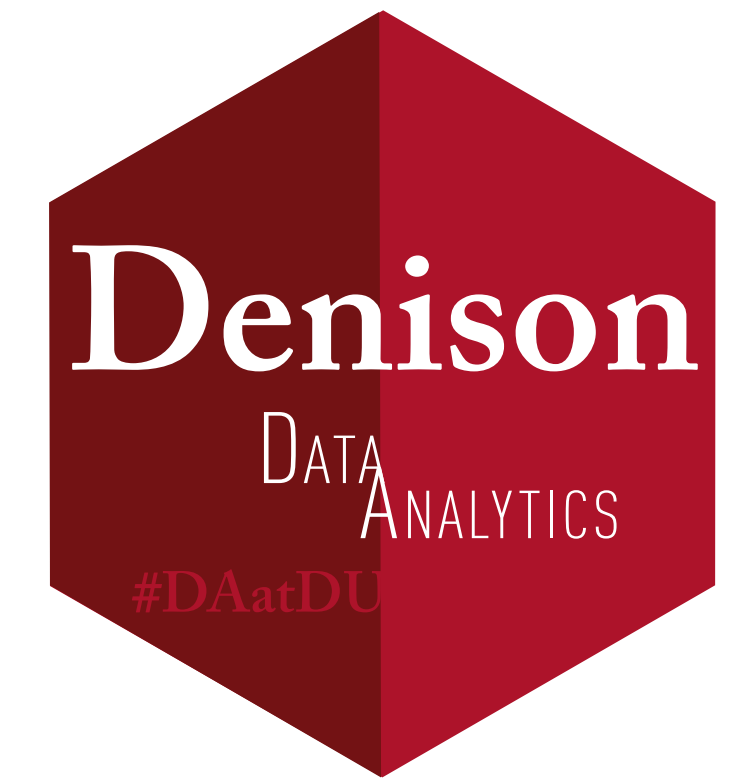
- Do students want to learn about if statements and while loops?
- Or do they want to learn how to solve problems?
- Multiple problem-focused “flavors”
- no prerequisites, for all students
- Python
- Polya’s four steps: understand, plan, code, look back
- Pair programming
- 2 subsequent courses—Intermediate CS and Data Structures—flesh things out for majors (in C++)



# A (Truly) Interdisciplinary Data Analytics program

---

- Data “Analytics” vs. Data “Science”
- academic program outside Math & CS with an interdisciplinary program committee
- DA faculty: ecology, political science, digital humanities, statistics, OR
- projects drawn from **diverse set of disciplines** with varied concerns
- **low barrier to entry**: no prerequisites, shorter Math & CS sequence
- size of major **allows broad exploration**, study abroad, etc.
- comfort with **uncertainty and ambiguity**
- **communication** with various audiences — visual, written, oral
- **ethical and social implications** of data collection and presentation



Biology
Computer Science
Economics
Environmental Science
Mathematics
Philosophy
Physics
Political Science
Psychology
Sociology