

# Index

- accession number
  - human, 2
- active sampling, 97
- aging, 300
- Alice in Wonderland*, 31
- alpha-helix, 29, 38, 118, 120, 121, 128, 141, 151, 186, 196, 242, 247
- alphabet, 1, 67, 72, 113, 128, 167, 236
  - merged, 118
  - reduced, 116
- alternative splicing, 4, 43
- Altschul, S.F., 35
- amino acids, 1, 26, 118, 167, 169
  - codons, 140
  - composition, 126, 129, 145
  - dihedral angle, 120
  - encoding, 115, 128, 139
  - genetic code, 26, 137
  - GES scale, 143
  - glycosylation, 41
  - hydrophobicity, 26, 133, 137, 141, 195
  - in beta-sheets, 115
  - in helix, 29, 38, 118
  - in HMM, 195
  - orthogonal encoding, 121
  - pathways, 137
  - substitution matrices, 35, 36, 209, 244, 267, 276
- Anastasia, 265
- ancestor, 95
- Anderson, A., 265
- antique DNA (aDNA), 265
- Arabidopsis thaliana*, 10, 20, 40, 42, 145, 149
- archaon, 7
- asymmetric window, 114
- asymmetric windows, 134
- background information, 49, 51, 53
- backpropagation, 83, 104, 111, 113, 121, 126, 128, 138, 139, 179, 246, 249
  - adaptive, 138, 140
  - learning order, 151
- bacteria, 9, 133
- bacteriophage, 7
- bacteriorhodopsin, 196
- Bayes theorem, 52, 249
- Bayesian framework, 48
- belief, 50, 95
- Bellman principle, 82
- bendability, 43, 44, 186, 219, 380, 381
- beta breakers, 38
- beta-sheet, 6, 18, 26, 38, 97, 115, 120, 121, 247
- blind prediction, 124, 131
- Blobel, G., 143
- Bochner's theorem, 395
- Boltzmann-Gibbs distribution, 85, 91, 92, 354, 362, 368
- Boltzmann-Gibbs distribution, 73, 74
- Boolean
  - algebra, 48
  - functions, 104
  - networks, 320
- brain
  - content-addressable retrieval, 2
  - memory, 2
- branch length, 266, 273
- branch point, 43, 212

- Burset, M., 153
- C-terminal, 115, 128
- C-value, 14  
paradox, 14
- cancer, 300
- capping, 38
- Carroll, L., 31
- CASP, 124, 131, 262
- cat, 7
- Cavalier-Smith, T., 143
- Chapman-Kolmogorov relation, 267
- Chargaff's parity rules, 230
- chimpanzee, 7
- Chomsky hierarchy, 277, 279, 280
- Chomsky normal form, 279
- chromatin, 44, 210
- chromosome, 7, 210, 230, 284  
components, 8  
unstable, 8
- classification, 97, 104
- classification error, 118
- Claverie, J-M., 13
- clustering, 44, 186, 191, 313
- Cocke-Kasami-Younger-algorithm, 290
- codon  
start, 235  
stop, 235  
usage, 44, 145, 147, 210
- codons, 26, 136, 137, 143, 210, 214  
start, 38  
stop, 26, 30, 141
- coin flip, 67, 71
- committee machine, 96
- communication, 9
- consensus sequences, 37, 165, 212, 236
- convolution, 107
- correlation coefficient, 122, 158, 209  
Matthews, 158  
Pearson, 158
- Cox-Jaynes axioms, 50, 266
- CpG islands, 147
- Creutzfeld-Jakob syndrome, 25
- Crick, F., 277
- cross-validation, 95, 124, 129, 134
- crystallography, 5, 120
- Cyber-T, 305, 307
- Darwin, C., 265
- data  
corpus, 51  
overrepresentation, 6  
redundancy, 4, 219  
storing, 2
- database  
annotation, 2  
bias, 129  
errors, 6  
noise, 2  
public, 2, 3
- database search  
iterative, 7
- decision theory, 347
- deduction, 48
- DEFINE program, 120
- development, 300
- dice, 67
- digital data, 1
- dinucleotides, 116
- Dirichlet distribution, 245
- discriminant function, 389
- distribution  
Boltzmann-Gibbs, 73
- DNA  
arrays, 299  
bending, 381  
binding sites, 320  
chip, 300  
helix types, 45  
library, 299  
melting, 14  
melting point, 45  
periodicity, 44, 212, 216  
reading frame, 210  
symmetries, 230
- DNA chips, 5
- DNA renaturation experiments, 27
- DNA sequencing, 1

- dog, 7
- DSSP program, 120, 131
- dynamic programming, 81, 172, 175, 240, 246, 249, 289, 290, 295
  - multidimensional, 184
- E. coli*, 38, 113, 135, 210
- email, 14
- encoding
  - adaptive, 128
- ensemble, 126, 128, 132
- ensembles, 96
- entropy, 74
  - maximum, 129
  - relative, 54, 69, 78, 109–111, 129
- ethics, 1
- evidence, 70
- evolution, 1, 8, 17, 56, 93, 137, 254
  - genetic code, 136
  - protein families, 116
- evolutionary information, 124
- evolutionary algorithms, 82, 93
- evolutionary events, 185, 209, 212
- evolutionary relationships, 196
- exon assembly, 147
- exon shuffling, 196
- exon-exon junction, 30
- exons, 103, 145, 147, 211
- extreme value distribution, 195, 219
  
- feature table, 3
- Fisher kernels, 391
- FORESST, 189
- forward-backward procedure, 83, 172, 174–176, 178, 180, 182, 291
- free energy, 73, 77, 85, 178
- functional features, 33
- fungi, 9
  
- Gamow, G., 17
- GenBank, 12, 15, 149, 152, 165, 219
- gene, 10
  - coregulated, 320
  - number in organism, 11
  - protein coding, 11
- gene pool, 8
  
- GeneMark, 210, 234
- GeneParser, 147
- genetic code, 136
- Genie, 234
- genome, 7, 16
  - circular, 7
  - diploid, 7
  - double stranded, 7
  - haploid, 7, 9
  - human, 9
  - mammalian, 15
  - single stranded, 7
  - size, 9
- GenomeScan, 234
- GenScan, 234
- Gibbs sampling, 89, 320, 373
- glycosylation, 3, 16, 34
- GRAIL, 147
- Grail, 234
- Guigo, R., 153
  
- halting problem, 280
- Hansen, J., 325
- hidden variables, 78
- Hinton, G.E., xviii
- histone, 44
- HMMs
  - used in word and language modeling, 240
- Hobohm algorithm, 219
- homology, 124, 126, 196, 275
- homology building, 33
- HSSP, 126, 131
- Hugo, V., 14
- human, 14
- human genome
  - chromosome size, 11
  - size, 11
- hybrid models, 239, 371, 383
- hybridization, 5
- hydrogen bond, 38, 120, 143
- hydrophobicity, 115, 118, 122, 186, 190
  - signal peptide, 133
- hydrophobicity scale, 141

- hyperparameters, 63, 95, 107, 170, 243, 389
- hyperplane, 114, 121, 390, 393
- hypothesis
  - complex, 49
- immune system, 24, 251, 321
- induction, 49, 104, 317
- infants, 4
- inference, 48, 70
- input representation, 114
- inside-outside algorithm, 291, 372
- inteins, 30
- intron, 235
  - splice sites, 3, 34, 40, 43, 103, 114, 145, 211, 212
- inverse models, 366
- Jacobs, R.E., 17
- Johannsen, W., 10
- Jones, D., 131
- k-means algorithm, 317
- Kabsch, W., 40
- Kernel methods, 389
- knowledge-based network, 123
- Krogh, A., 127, 208
- Lagrange multiplier, 74, 177, 318, 391, 394
- language
  - computer, 277
  - natural, 277
  - spelling, 2
- learning
  - supervised, 104
  - unsupervised, 104
- learning rate, 83
- likelihood, 67
- likelihood function, 75
- linguistics, 4, 26, 285
- lipid environment, 17
- lipid membrane, 143
- liposome-like vesicles, 143
- loss function, 347
- machine learning, 166
- mammoth, 265
- map, 31
- MAP estimate, 57, 58, 69, 85, 104, 177, 245
- MaxEnt, 54, 73, 75
- membrane proteins, 189, 195, 209
- MEME, 320
- Mercer's theorem, 396
- metabolic networks, 321
- Metropolis algorithm, 90, 91
  - generalizations, 91
- microarray expression data, 299, 320
- microarrays, 5
- mixture models, 63, 317
- model complexity, 48, 94
- models
  - graphical, 65, 73, 165
  - hierarchical, 63
  - hybrid, 63
- Monte Carlo, 59, 82, 87, 108, 250, 366, 389
  - hybrid methods, 93
- multiple alignment, 72, 124, 127, 129, 275, 292-294, 381
- mutual information, 160
- N-terminal, 115, 118, 128, 133, 136
- N-value paradox, 13
- Neal, R.M., xviii
- Needleman-Wunch algorithm, 34, 82
- NetGene, 146, 148
- NetPlantGene, 149
- NetTalk perceptron architecture, 113
- neural network, 126
- neural network, profiles, 126
- neural network
  - recurrent, 99, 122, 255, 320
  - weight logo, 141
- Nielsen, H., 36, 208
- nonstochastic
  - grammars, 289
- nucleosome, 210, 211, 221
- Ockham's Razor, 59
- orthogonal vector representation, 116

- overfitting, 126
- palindrome, 210, 278, 279, 281, 284, 285
- PAM matrix, 267, 276
- parameters
  - emission, 63, 170, 383
  - transition, 63, 75
- parse tree, 281, 292, 294, 297
- partition function, 57, 74, 76, 77, 90, 354, 362
- pathway, 320
- PDB, 22
- perceptron, 113
  - multilayer, 113
- Petersen, T.N., 132
- Pfam, 189
- phase transition, 76
- phonemes, 240
- phosphorylation, 16, 119
- phylogenetic information, 189, 293
- phylogenetic tree, 185, 265, 266, 273
- plants, 9
- polyadenylation, 147, 210
- polymorphism, 1
- position-specific scoring matrices, 131
- posttranslational modification, 16
- prior, 52, 53, 55, 57, 59, 74, 106, 107
  - conjugate, 56, 303
  - Dirichlet, 56, 69, 75, 170
  - gamma, 55
  - Gaussian, 55, 107
    - use in hybrid architectures, 243
  - uniform, 72
- profile, 6, 124, 126, 165, 219, 222
  - bending potential, 219, 381
  - emission, 214
- promoter, 115, 147, 221
- propositions, 50
- PROSITE, 190, 205
- protein
  - beta-sheet, 97
  - beta-sheet partners, 115
  - helix, 97
  - helix periodicity, 120, 128
  - length, 17
  - networks, 321
  - secondary structure, 6, 113, 121, 129, 189, 229
  - secretory, 16
  - tertiary structure, 121
- Protein Data Bank, 22
- protein folding, 73
- proteome, 16
- pruning, 97
- Prusiner, S.B., 25
- pseudo-genes, 12
- pseudoknots, 284, 288, 297
- PSI-BLAST, 7, 131
- PSI-PRED, 131
- Qian, N., 121
- quantum chemistry, 121
- reading frame, 29, 43, 145, 210, 214, 217, 218, 286
  - open, 31
- reductionism, 13
- redundancy reduction, 4, 219
- regression, 104, 308, 349, 387
- regularizer, 57, 76, 94, 171, 181, 252, 253
- regulatory circuits, 320
- relative entropy, 160
- renaturation kinetics, 14
- repeats, 26, 28, 279, 284
- representation
  - orthogonal, 115, 128
  - semiotic, 2
- ribosome, 38, 143, 145
- ribosome binding sites, 113
- Riis, S., 127
- ROC curve, 162, 204
- Rost, B., 25, 124
- rules, 113, 123, 151
  - Chou-Fasman, 123
- S. solfataricus, 38
- Sander, C., 25, 32, 40, 124
- Schneider, R., 32
- Schneider, T., 37

- secretory pathway, 16, 133
- Sejnowski, T.J., 121
- semiotic representation, 2
- sensitivity, 41, 162, 209
- sequence
  - data, 72
  - families, 5, 6
  - logo, 37, 134
- sequence space, 17
- Shine–Dalgarno sequence, 38, 210
- signal anchor, 208
- signal peptide, 114, 133, 207
- signalling networks, 321
- SignalP, 134, 207
- simulated annealing, 91, 116
- single nucleotide polymorphism, 33
- Smith–Waterman algorithm, 34, 82, 219, 295
- social security numbers, 2
- sparse encoding, 115
- specificity, 41, 162, 209
- speech recognition, 4, 113, 165, 167, 226
- splice site, 235
- splines, 104
- SSpro, 262
- statistical mechanics, 73, 88, 91, 96
- statistical model fitting, 47
- stochastic
  - grammars, 166, 254, 277, 282, 295
  - sampling, 82
  - units, 100
- Stormo, G., 113
- STRIDE program, 120
- string, 68
- Student distribution, 304
- support vector machines, 389
- SWISS-PROT, 19, 21, 136, 191, 193, 194, 198, 200, 203, 206
- systemic properties, 13
  
- TATA-box, 219, 235
- threshold gate, 104
- time series, 239
- TMHMM, 209
  
- training
  - balanced, 97, 126
- transcription initiation, 115, 221
- transfer free energy, 143
- transfer function, 100, 104
  - sigmoidal, 105
- translation initiation, 113, 136
- trinucleotides, 116, 137
- tsar, Nicholas II, 265
- t*-test, 300, 301, 304
- Turing machine, 280, 282
  - halting problem, 280
- twilight zone, 32, 209
  
- validation, 95, 103, 108, 252
- VC dimension, 94
- virus, 7, 285
- visual inspection, 3
- Viterbi algorithm, 82, 171, 175, 180–182, 184, 190, 191, 198, 206, 246, 251, 252, 271, 273, 274, 290, 292, 294
- von Heijne, G., 43
  
- Watson, J.D., 277
- Watson–Crick basepair, 286
- weight
  - decay, 107
  - logo, 141
  - matrix, 38, 136
  - sharing, 107, 128, 243
- weighting scheme, 96, 129
- White, S.H., 17
- winner-take-all, 139
  
- Ycas, M., 17
- yeast, 43, 230, 232