

Volume 5

Supplement 3

1978

ATLAS of PROTEIN SEQUENCE and STRUCTURE

Margaret O. Dayhoff



NATIONAL BIOMEDICAL RESEARCH FOUNDATION
GEORGETOWN UNIVERSITY MEDICAL CENTER
WASHINGTON, D. C. 20007



**ATLAS of PROTEIN
SEQUENCE and
STRUCTURE
VOLUME 5
SUPPLEMENT 3
1978**

**ATLAS of
PROTEIN SEQUENCE
and STRUCTURE**

Volume 5

**SUPPLEMENT 3
1978**

Margaret O. Dayhoff

National Biomedical Research Foundation

Georgetown University Medical Center

3900 Reservoir Road, N.W.

Washington, D.C. 20007

Tel. 202-625-2121

QD 431
. A1A85
v. 5, Suppl 3, 1978

ATLAS OF PROTEIN SEQUENCE AND STRUCTURE

Volume 5

SUPPLEMENT 3

1978

Library of Congress Card Catalogue Number 65-29342

ISBN 0-912466-07-3

© Copyright 1979

by

The National Biomedical Research Foundation

All Rights Reserved. Printed in the United States of America.
This book, or parts thereof, may not be reproduced
in any form without permission of the publishers.

Copies may be obtained from the publisher:

The National Biomedical Research Foundation
Post Office Box 629, Silver Spring, Maryland 20901

Also available from the publisher:

Protein Sequence Data Tape
Protein Segment Dictionary 78
Protein Data Search Services
Atlas of Protein Sequence and Structure
Volume 5
Supplement 1
Supplement 2

Table of Contents

List of Figures	viii
List of Alignments	x
List of Difference Matrices	xi
List of Tables	xii
List of Sequences	xiii
Preface	xx
 1 – SURVEY OF NEW DATA AND COMPUTER METHODS OF ANALYSIS	 1
<i>M. O. Dayhoff</i>	
 2 – PROTEIN SUPERFAMILIES	 9
<i>M. O. Dayhoff, W. C. Barker, L. T. Hunt, and R. M. Schwartz</i>	
 3 – PROTEIN DATA INTRODUCTION	 25
<i>L. T. Hunt, W. C. Barker, and M. O. Dayhoff</i>	
 4 – CYTOCHROMES	 29
<i>R. M. Schwartz and M. O. Dayhoff</i>	
 5 – OTHER ELECTRON TRANSPORT PROTEINS	 45
<i>R. M. Schwartz and M. O. Dayhoff</i>	
 6 – FLAVODOXIN AND OXIDOREDUCTASES	 59
<i>R. M. Schwartz, C. L. Young, and M. O. Dayhoff</i>	
 7 – SERINE PROTEASES	 73
<i>C. L. Young, W. C. Barker, C. M. Tomaselli, and M. O. Dayhoff</i>	
 8 – OTHER ENZYMES	 95
<i>W. C. Barker, C. L. Young, and M. O. Dayhoff</i>	
 9 – ENZYME INHIBITORS AND GROWTH FACTORS	 131
<i>L. K. Ketcham, W. C. Barker, and M. O. Dayhoff</i>	
 10 – HORMONES AND ACTIVE PEPTIDES	 145
<i>L. T. Hunt, F. D. Ledley, and M. O. Dayhoff</i>	
 11 – TOXINS	 165
<i>L. T. Hunt, S. Hurst-Calderone, and M. O. Dayhoff</i>	
 12 – IMMUNOGLOBULINS	 197
<i>W. C. Barker, L. K. Ketcham, and M. O. Dayhoff</i>	
 13 – GLOBINS	 229
<i>L. T. Hunt, S. Hurst-Calderone, and M. O. Dayhoff</i>	
 14 – NUCLEIC ACID-ASSOCIATED PROTEINS	 251
<i>L. T. Hunt, R. M. Schwartz, and M. O. Dayhoff</i>	
 15 – FIBROUS PROTEINS	 265
<i>L. T. Hunt and M. O. Dayhoff</i>	

16 – CONTRACTILE SYSTEM PROTEINS	273
<i>W. C. Barker, L. K. Ketcham, and M. O. Dayhoff</i>	
17 – MISCELLANEOUS PROTEINS	285
<i>L. T. Hunt, W. C. Barker, C. L. Young, and M. O. Dayhoff</i>	
18 – NUCLEIC ACID DATA INTRODUCTION	311
<i>R. M. Schwartz and M. O. Dayhoff</i>	
19 – TRANSFER RNA	313
<i>R. M. Schwartz and M. O. Dayhoff</i>	
20 – RIBOSOMAL AND OTHER RNAs	327
<i>R. M. Schwartz and M. O. Dayhoff</i>	
21 – GENOME NUCLEIC ACIDS	338
<i>R. M. Schwartz and M. O. Dayhoff</i>	
22 – A MODEL OF EVOLUTIONARY CHANGE IN PROTEINS	345
<i>M. O. Dayhoff, R. M. Schwartz, and B. C. Orcutt</i>	
23 – MATRICES FOR DETECTING DISTANT RELATIONSHIPS	353
<i>R. M. Schwartz and M. O. Dayhoff</i>	
24 – DUPLICATIONS IN PROTEIN SEQUENCES	359
<i>W. C. Barker, L. K. Ketcham, and M. O. Dayhoff</i>	
25 – COMPOSITION OF PROTEINS	363
<i>M. O. Dayhoff, L. T. Hunt, and S. Hurst-Calderone</i>	
APPENDIX	374
CUMULATIVE INDEX OF PROTEIN SEQUENCES	377
CUMULATIVE INDEX OF STEREO-PAIR DRAWINGS	389
AUTHOR INDEX	390
SUBJECT INDEX	395

LIST OF FIGURES

1. Phylogenetic tree of prokaryotes and eukaryotes	28-29
2. Evolutionary tree derived from c-type cytochrome sequences	30
3. The cytochrome c evolutionary tree	31
4. Evolutionary tree of cytochromes c ₂ and c ₅₅₁	32
5. Ferredoxin evolutionary tree	45
6. Azurin-plastocyanin evolutionary tree	49
7. Stereo-pair drawing of the α -carbon backbone of thioredoxin from <i>Escherichia coli</i>	52
8. Schematic drawing of the main-chain conformation of thioredoxin-S ₂ from <i>E. coli</i>	52
9. Amino acid sequence of thioredoxin-S ₂ from <i>E. coli</i>	52
10. Stereo-pair drawing of the α -carbon backbone of one chain of bovine superoxide dismutase (Cu-Zn)	59

11. Diagram showing the eight strands of β -sheet and their topological connectivity for one chain of bovine superoxide dismutase	59
12. Evolutionary tree of the bacterial serine proteases with a representative eukaryote branch	73
13. Evolutionary tree of the eukaryote serine proteases and bacterial trypsin	74
14. Evolutionary tree of the kringle regions of plasminogen and prothrombin	75
15. Stereo-pair drawing of the α -carbon backbone of trypsin	82
16. Stereo-pair drawing of <i>Streptomyces griseus</i> protease B	83
17. Molecular architecture of alpha-lytic protease from myxobacter 495	86
18. Molecular architecture of protease B from <i>Streptomyces griseus</i>	86
19. Molecular architecture of trypsin from <i>Streptomyces griseus</i>	86
20. Molecular architecture of bovine trypsinogen	86
21. Molecular architecture of pig elastase	86
22. Molecular architecture of bovine chymotrypsinogen A	86-87
23. Molecular architecture of bovine factor X	86-87
24. Molecular architecture of bovine prothrombin	86-87
25. Molecular architecture of human plasminogen	87
26. Pathways of blood coagulation, fibrinolysis, and kinin formation	88-89
27. Evolutionary tree of phospholipases	95
28. Stereo-pair drawing of bovine pancreatic phospholipase A ₂	98
29. Stereo-pair drawing of penicillopepsin	104
30. Schematic drawing of the secondary structure of penicillopepsin	104
31. Evolutionary tree of carbonic anhydrases B and C	110
32. Stereo-pair drawing of one chain of triosephosphate isomerase	111
33. Stereo-pair drawing of the dimer of triosephosphate isomerase	111
34. Evolutionary tree of basic protease inhibitors	131
35. Evolutionary tree of PSTI-type protease inhibitors	135
36. Evolutionary tree of Bowman-Birk protease inhibitors	137
37. Evolutionary tree of calcitonins	148
38. Evolutionary tree of insulin-related proteins	150
39. Evolutionary tree of snake long neurotoxins and related venom proteins	169
40. Evolutionary tree of snake short neurotoxins and related proteins	173
41. Evolutionary tree of snake cytotoxins	177
42. Evolutionary tree of the superfamily of proteroglyph snake toxins and related proteins	181
43. Stereo-pair drawing of short neurotoxin 1, variant b (erabutoxin b), from the broad-banded blue sea snake	182
44. Schematic representation of antiparallel β -pleated sheet structure in short neurotoxin 1, variant b (erabutoxin b)	182
45. Evolutionary tree of immunoglobulin C regions	197
46. Topology of human immunoglobulin C-region domains	199
47. Evolutionary tree of immunoglobulin kappa chain V regions	206
48. Evolutionary tree of immunoglobulin lambda chain V regions	206
49. Evolutionary tree of immunoglobulin heavy chain V regions	207
50. Stereo-pair drawing of the human immunoglobulin lambda chain Mcg	212
51. Schematic drawing of the human immunoglobulin lambda chain Mcg	212
52. Evolutionary tree of the globin superfamily	229
53. Evolutionary tree of myoglobin	239
54. Schematic drawing of the structure of a chromatin fiber	251
55. Schematic representation of a possible model for the chromatin subunit structure	251
56. A possible way in which a branched protein A24 molecule could fit into an octameric nucleosome core	251
57. Evolutionary tree of the troponin C superfamily	273
58. Structure of striated muscle fiber	274
59. Thick and thin filaments of muscle	274
60. Proteins of the contractile element of skeletal muscle	274

61.	Stereo-pair drawing of the dimer of the human plasma protein prealbumin	290
62.	Stereo-pair drawing of the channel in the human prealbumin tetramer	290
63.	Schematic drawing of the main-chain conformation in the prealbumin dimer	291
64.	Schematic drawing of the opposed β -sheets at the center of the prealbumin tetramer	291
65.	Stereo-pair diagram of the atomic arrangement in a single hormone-binding site of prealbumin	291
66.	Stereo-pair drawing of myohemerythrin	294
67.	Stereo-pair drawing of the hemerythrin monomer	294
68.	Quaternary structure of hemerythrin	295
69.	Schematic drawing of iron coordination in hemerythrin	295
70.	Ribonucleotide chemical structure	312
71.	Evolutionary tree of initiator methionine transfer RNA	314
72.	Evolutionary tree of phenylalanine transfer RNA	316
73.	Transfer RNA cloverleaf structure	318
74.	Evolutionary tree of 5S ribosomal RNA	327
75.	Generalized secondary structure of the 5S ribosomal RNA molecule	328
76.	Schematic drawing of the genes that code for proteins in the ϕ X174 genome	338
77.	Schematic drawing of the genes that code for proteins in the MS2 genome	342
78.	Simplified phylogenetic tree	345
79.	Matrix of accepted point mutations derived from the tree of Figure 78	345
80.	Numbers of accepted point mutations accumulated from closely related sequences	346
81.	Sample computation of relative mutability	347
82.	Mutation probability matrix for the evolutionary distance of 1 PAM	348
83.	Mutation probability matrix for the evolutionary distance of 250 PAMs	350
84.	Log odds matrix for 250 PAMs	352
85.	Scoring matrix for the evolutionary distance of 250 PAMs	354
86.	Dependence of alignment scores on the total number of mutations	354
87.	Alignment scores as a function of the evolutionary distance of the mutation data matrices	355
88.	Segment comparison scores as a function of the evolutionary distance of the mutation data matrix	355
89.	Chemical structures of the twenty amino acids	364

LIST OF ALIGNMENTS

1.	Cytochrome c superfamily	33-35
2.	Ferredoxins	46-47
3.	Azurin and plastocyanin	50-51
4.	Lactate dehydrogenase	60-61
5.	Glyceraldehyde 3-phosphate dehydrogenase	62-63
6.	Dihydrofolate reductase	65
7.	Alpha-lytic protease and proteases A and B	76-77
8.	Eukaryote serine proteases and bacterial trypsin	79-81
9.	Kringle regions from prothrombin and plasminogen	84
10.	Calcium-binding regions of prothrombin and factor X	85
11.	Related regions of epidermal growth factor and factor X	85
12.	Phospholipases	96-97
13.	Pancreatic and related ribonucleases	100-101
14.	Carboxypeptidases A and B	102-103
15.	Pepsin superfamily	105-107
16.	Carbonic anhydrases B and C	108-109
17.	Triosephosphate isomerase	112-113
18.	Basic protease inhibitors	133
19.	Pancreatic secretory trypsin inhibitor and related sequences	134
20.	Bowman-Birk and related protease inhibitors	136
21.	Corticotropin-lipotropin	147

22.	Calcitonins	149
23.	Superfamily of proinsulin-related proteins	151
24.	Snake long neurotoxins and related venom proteins	166-167
25.	Snake short neurotoxins and related proteins	170-171
26.	Snake cytotoxins	174-175
27.	Superfamily of snake toxins and related proteins	178-179
28.	Scorpion neurotoxins	183
29.	Plant toxins	184
30.	Immunoglobulin C-region domains	200-201
31.	Immunoglobulin heavy chain C regions	202-205
32.	Globin superfamily	230-233
33.	Myoglobin	235-237
34.	Alpha crystallins	265-266
35.	Troponin C superfamily	276-277
36.	Human fibrinogen superfamily	286-288
37.	Kappa casein	289
38.	Hemerythrin and myohemerythrin	292-293
39.	Neurophysins	296
40.	Initiator methionine transfer RNA	315
41.	Phenylalanine transfer RNA	317
42.	5S ribosomal RNA	329-331

LIST OF DIFFERENCE MATRICES

1.	Cytochrome c superfamily	36
2.	Bacterial ferredoxins	48
3.	Plant-type ferredoxins	48
4.	Azurin	51
5.	Plastocyanin	51
6.	Lactate dehydrogenase	61
7.	Glyceraldehyde 3-phosphate dehydrogenase	64
8.	Dihydrofolate reductase	64
9.	Alpha-lytic protease and proteases A and B	77
10.	Eukaryote serine proteases and bacterial trypsin	78
11.	Kringle regions from prothrombin and plasminogen	78
12.	Phospholipases	98
13.	Pancreatic and related ribonucleases	99
14.	Pepsin superfamily	107
15.	Carbonic anhydrases B and C	110
16.	Triosephosphate isomerase	113
17.	Basic protease inhibitors	132
18.	Pancreatic secretory trypsin inhibitor and related sequences	132
19.	Bowman-Birk protease inhibitors	137
20.	Corticotropin-lipotropin	148
21.	Calcitonins	149
22.	Superfamily of proinsulin-related proteins	150
23.	Snake long neurotoxins and related venom proteins	168
24.	Snake short neurotoxins and related proteins	172
25.	Snake cytotoxins	176
26.	Superfamily of snake toxins and related proteins	180
27.	Scorpion neurotoxins	183
28.	Plant toxins	184
29.	Immunoglobulin C-region domains	198
30.	Immunoglobulin heavy chain C regions	205

31. Immunoglobulin kappa chain V regions	208-209
32. Immunoglobulin lambda chain V regions	210
33. Immunoglobulin heavy chain V regions	211
34. Globin superfamily	234
35. Myoglobin	238
36. Alpha crystallins	266
37. Troponin C superfamily, longer sequences	275
38. Parvalbumins and calcium-binding protein	275
39. Kappa casein	289
40. Hemerythrin and myohemerythrin	293
41. Initiator methionine transfer RNA	314
42. Phenylalanine transfer RNA	316
43. 5S ribosomal RNA	332

LIST OF TABLES

1. Rates of mutation acceptance	3
2. New protein superfamilies and families	4
3. One- and three-letter amino acid abbreviations	26
4. Special punctuation in sequences	26
5. Molecular weights	26
6. Nomenclature of blood-clotting factors	89
7. Invertebrate peptide similar to melanotropin	158
8. Invertebrate peptides similar to glucagon	159
9. Gastrin-related and similar peptides	160
10. Eledoisin-related and similar peptides	161
11. Hypothalamic and similar peptides	162
12. Bradykinin-related peptides	163
13. Miscellaneous active peptides	164
14. Nomenclature of snake toxins	165
15. Fibrinopeptides	285
16. One-character nucleoside abbreviations	311
17. Three-character nucleoside abbreviations	312
18. Transfer RNA sequences reported in the <i>Atlas</i>	313
19. The genetic code	318
20. 5S ribosomal RNA sequences reported in the <i>Atlas</i>	328
21. Relative mutabilities of the amino acids	347
22. Normalized frequencies of the amino acids in the accepted point mutation data	347
23. Correspondence between observed differences and the evolutionary distance	351
24. Comparison of matrices for calculating alignment scores	356
25. Comparison of matrices for calculating segment comparison scores	357
26. Comparison of mutation data matrices for calculating alignment scores	357
27. Comparison of mutation data matrices for calculating segment comparison scores	358
28. Proteins with previously detected duplications	360
29. New proteins with detected duplications	361
30. Superfamilies containing duplicated proteins	362
31. Average percent of amino acids in proteins	363
32. Average percent of amino acid groups in proteins	363
33. Composition of protein families in amino acid groups	365
34. Protein families in order of length	372
35. Probabilities of standardized scores for the normal distribution	374
36. Amino acid PAM scale	375
37. Nucleotide PAM scale	375

LIST OF SEQUENCES

An asterisk indicates a reference entry.

CYTOCHROMES		Cytochrome b ₂ (lactate dehydrogenase)	
Cytochrome c — Guanaco*	37	core — Baker's yeast*	44
Cytochrome c — Donkey*	37		
Cytochrome c — Mouse*	37	OTHER ELECTRON TRANSPORT PROTEINS	
Cytochrome c, testis-specific — Mouse	37	Ferredoxin — <i>Clostridium</i> M-E	53
Cytochrome c — Bonito*	37	Ferredoxin — <i>Peptococcus aerogenes</i>	
Cytochrome c — Tuna*	37	(= <i>Micrococcus aerogenes</i>)*	53
Cytochrome c — Starfish	38	Ferredoxin II — <i>Chlorobium limicola</i>	53
Cytochrome c — Common brandling worm	38	Ferredoxin — <i>Chromatium vinosum</i>	53
Cytochrome c — Mediterranean fruit fly*	38	Ferredoxin — <i>Bacillus stearothermophilus</i>	53
Cytochrome c — Locust	38	Ferredoxin — <i>Halobacterium halobium</i>	54
Cytochrome c — Freshwater prawn	38	Ferredoxin — <i>Nostoc muscorum</i>	54
Cytochrome c — <i>Schizosaccharomyces pombe</i>	39	Ferredoxin — <i>Spirulina maxima</i>	54
Cytochrome c — Potato*	39	Ferredoxin — <i>Spirulina platensis</i>	54
Cytochrome c — Arum	39	Ferredoxin — <i>Aphanothece sacrum</i>	55
Cytochrome c — Niger	39	High potential iron-sulfur protein —	
Cytochrome c — <i>Enteromorpha intestinalis</i>	39	<i>Rhodopseudomonas gelatinosa</i>	55
Cytochrome c — <i>Crithidia oncopelti</i> *	40	Thioredoxin — <i>Escherichia coli</i> *	55
Cytochrome c — <i>Crithidia fasciculata</i>	40	Rubredoxin — <i>Desulfovibrio gigas</i>	55
Cytochrome c — <i>Euglena gracilis</i> *	40	Rubredoxin — <i>Desulfovibrio vulgaris</i>	55
Cytochrome c — <i>Tetrahymena pyriformis</i>	40	Plastocyanin — Broad bean	56
Cytochrome c ₂ — <i>Rhodospirillum photometricum</i>	40	Plastocyanin — Dock	56
Cytochrome c ₂ — <i>Rhodopseudomonas acidophila</i>	41	Plastocyanin — Lettuce	56
Cytochrome c ₂ , iso-1 — <i>Rhodospirillum molischianum</i>	41	Plastocyanin — Vegetable marrow	56
Cytochrome c ₂ , iso-2 — <i>Rhodospirillum molischianum</i>	41	Plastocyanin — Spinach	57
Cytochrome c ₂ , iso-1 — <i>Rhodospirillum fulvum</i>	41	Plastocyanin — Shepherd's purse	57
Cytochrome c ₂ — <i>Rhodospirillum tenue</i>	42	Plastocyanin — Dog's mercury	57
Cytochrome c ₂ — <i>Rhodopseudomonas gelatinosa</i>	42	Plastocyanin — <i>Anabaena variabilis</i>	57
Cytochrome c ₂ — <i>Rhodopseudomonas viridis</i>	42	FLAVODOXIN AND OXIDOREDUCTASES	
Cytochrome c ₂ — <i>Rhodomicrobium vannielii</i>	42	Flavodoxin — <i>Desulfovibrio vulgaris</i> *	66
Cytochrome c ₅₅₀ — <i>Paracoccus denitrificans</i> (= <i>Micrococcus denitrificans</i>)*	43	Flavodoxin — <i>Azotobacter vinelandii</i>	66
Cytochrome c ₆ (f) — <i>Alaria esculenta</i> *	43	Alcohol dehydrogenase — Horse*	66
Cytochrome c ₆ (f) — <i>Plectonema boryanum</i>	43	Alcohol dehydrogenase — Yeast	66
Cytochrome c ₅₅₅ — <i>Prosthecochloris aestuarii</i> *	43	Lactate dehydrogenase M chain — Pig	67
Cytochrome c ₅₅₅ — <i>Chlorobium limicola</i> *	43	Lactate dehydrogenase H chain — Pig	67
Cytochrome c ₇ (c _{551.5}) — <i>Desulfuromonas acetoxidans</i> *	43	Lactate dehydrogenase M chain — Spiny dogfish	67
Cytochrome c' — <i>Rhodospirillum rubrum</i>	44	Ribitol dehydrogenase — <i>Enterobacter aerogenes</i> (= <i>Klebsiella aerogenes</i>)	68
Cytochrome b ₅ — Bovine	44	Peroxidase — Horseradish	68
Cytochrome b ₅ — Horse	44	Superoxide dismutase (Cu-Zn) — Bovine	68
		Glyceraldehyde 3-phosphate dehydrogenase — Lobster*	69
		Glyceraldehyde 3-phosphate dehydrogenase — <i>Thermus aquaticus</i>	69
		Glyceraldehyde 3-phosphate dehydrogenase — <i>Bacillus stearothermophilus</i>	69
		NADP-specific glutamate dehydrogenase — <i>Neurospora crassa</i>	70
		Dihydrofolate reductase — Mouse	70

Dihydrofolate reductase — <i>Escherichia coli</i> B (RT500)	71	<i>N, O</i> -Diacetylmuramidase — <i>Chalaropsis</i> sp.	122
Dihydrofolate reductase — <i>Streptococcus faecium</i>	71	Beta galactosidase — <i>Escherichia coli</i>	123
SERINE PROTEASES		Carboxypeptidase B — Bovine	123
Prothrombin — Bovine*	90	Streptococcal proteinase — <i>Streptococcus pyogenes</i>	124
Factor X (Stuart factor) — Bovine	90	Pepsinogen — Bovine (fragment)	124
Plasminogen — Human	91	Prochymosin — Bovine	125
Trypsinogen — Bovine*	92	Penicillopepsin — <i>Penicillium janthinellum</i>	125
Trypsin — Spiny dogfish	92	Asparaginase — <i>Escherichia coli</i>	126
Kallikrein — Pig (fragments)	92	Penicillinase — <i>Staphylococcus aureus</i> PC-1*	126
Haptoglobin beta chain — Human (fragment)*	92	Penicillinase, membrane-bound — <i>Bacillus licheniformis</i> 749/C	126
Trypsin — <i>Streptomyces griseus</i>	93	Aldolase A — Rabbit*	127
Protease A — <i>Streptomyces griseus</i>	93	Carbonic anhydrase B — Human*	127
Protease B — <i>Streptomyces griseus</i>	93	Carbonic anhydrase C — Human*	127
OTHER ENZYMES		Carbonic anhydrase C — Rabbit	127
Aspartate aminotransferase, cytoplasmic — Pig*	114	Carbonic anhydrase C — Bovine	128
Adenylate kinase, isoenzyme 1 — Human	114	Carbonic anhydrase C — Sheep	128
DNA-dependent RNA polymerase alpha chain — <i>Escherichia coli</i>	114	Triosephosphate isomerase — Chicken*	129
RNA replicase beta chain — Bacteriophage MS2	115	Triosephosphate isomerase — Coelacanth*	129
Thiosulfate sulfurtransferase (rhodanese) — Bovine	115	Triosephosphate isomerase — <i>Bacillus stearothermophilus</i>	129
Diphtheria toxin, fragment A — Corynebacterium beta	116	ENZYME INHIBITORS AND GROWTH FACTORS	
Prophospholipase A ₂ — Pig	116	Venom basic protease inhibitor II — Ringhals	138
Prophospholipase A ₂ — Horse	116	Venom basic protease inhibitor II — Cape cobra	138
Phospholipase A ₂ , notechis 5 — Australian tiger snake	117	Basic protease inhibitor — Red sea turtle (1 homology region)	138
Phospholipase A ₂ , notexin — Australian tiger snake	117	Isoinhibitor K (BPI type) — Snail (<i>Helix pomatia</i>)*	138
Phospholipase A ₂ I — Forest cobra	117	Nerve growth factor — Indian cobra	139
Phospholipases A ₂ III and II — Forest cobra	118	Pancreatic secretory trypsin inhibitor — Human	139
Phospholipase A ₂ — Ringhals	118	Acrosin inhibitor A ₁ (PSTI type) — Pig	139
Phospholipase A ₂ — Honey bee*	118	Protease inhibitor (PSTI type), submandibular gland — Dog	140
Ribonuclease St — <i>Streptomyces erythreus</i>	119	Ovomucoid (PSTI-type protease inhibitor) — Japanese quail	140
Ribonuclease U ₂ — <i>Ustilago sphaerogena</i>	119	Subtilisin inhibitor (PSTI type) — <i>Streptomyces albogriseolus</i>	140
Ribonuclease — Bison, eland, topi, and brindled gnu*	119	Antithrombin-III (heparin cofactor) — Human	141
Ribonuclease — Red deer and roe deer*	120	Thrombin inhibitor (hirudin) — Leech (<i>Hirudo medicinalis</i>)	141
Ribonuclease — Arabian and Bactrian camels	120	Protease inhibitor (Bowman-Birk) — Lima bean	142
Ribonuclease — Chinchilla	121	Protease inhibitor C-II (Bowman-Birk) — Soybean	142
Ribonuclease — Coypu	121	Protease inhibitor D-II (Bowman-Birk) — Soybean	142
Ribonuclease A — Guinea pig	121	Protease inhibitor II (Bowman-Birk) — Garden bean	142
Ribonuclease B — Guinea pig	121		
Ribonuclease — Muskrat	121		
Lysozyme — Human*	122		
Lysozyme — Chachalaca	122		
Lysozyme — Bacteriophage T4*	122		

Bromelain inhibitor — Pineapple	143	Enkephalins — Pig and bovine	161
Carboxypeptidase inhibitor — Potato	143	Ranatensin — Frog	161
Epidermal growth factor — Human	143	Ranatensin C — Bullfrog	161
HORMONES AND ACTIVE PEPTIDES		Ranatensin R — Frog	161
Corticotropin — Finback whale*	152	Litorin — Australian frog	161
Melanotropin alpha — Arabian camel*	152	Somatostatin — Pig	162
Corticotropin (and melanotropin alpha) —		Hypothalamic tetradecapeptide — Pig	162
Spiny dogfish	152	Neurohypophyseal hexapeptide — Pig	162
Melanotropins beta 1 and 2 — Arabian		Thyroliberin — Toad and newt	162
camel	152	Gonadoliberin — Pig and sheep	162
Melanotropin beta — Spiny dogfish	152, 158	Mating hormone alpha-1 — Baker's yeast	162
Melanotropin beta — Spotted dogfish	153	Mating hormone alpha-1B — Baker's yeast	162
Lipotropin beta (and beta-endorphin) —		Bradykinin — Bovine and frogs	163
Human	153	Kinin — Wasps	163
Beta-endorphin — Arabian camel and sheep*	153	Vespalakinin 1 — Yellow jacket	163
Follitropin alpha chain — Human	153	Vespalakinin-M — Hornet	163
Thyrotropin beta chain — Pig	154	Vespalakinin-X — Hornet	163
Follitropin beta chain — Human	154	6-Thr-bradykinin — Frog and wasp	163
Lutropin beta chain — Human	154	Polistes kinin R — Wasp	163
Prolactin — Pig	154	Ranakinin R — Frog	163
Prolactin — Sheep*	155	Thymic factor — Pig	164
Somatotropin — Horse*	155	Eosinophilotactic peptides — Human	164
Thymopoietin II — Bovine	155	Inhibitory peptide — Cat	164
Thymosin alpha ₁ — Bovine	155	Growth-modulating peptide — Human	164
Calcitonin — Rat	155	Delta sleep-inducing peptide — Rabbit	164
Calcitonin — Eel	156	Posterior pituitary peptide — Pig	164
Parathyroid hormone — Pig (fragment)*	156	Contraceptive peptide — Hamster zygote	164
Glucagon — Arabian camel*	156	Proctolin — Cockroach	164
Glucagon — Chicken*	156	TOXINS	
Vasoactive intestinal peptide — Chicken	156	Long neurotoxin 1 — Broad-banded blue	
Pancreatic hormone — Chicken*	156	sea snake	185
Insulin — Rattlesnake*	156	Long neurotoxin 1 — Middle Asian cobra	185
Insulin-like growth factor I — Human	157	Long neurotoxin 1 — Ethiopian cobra	185
Insulin-like growth factor II — Human	157	Long neurotoxin 1 — Indian cobra*	185
Relaxin — Pig	157	Long neurotoxin 2 — West African green	
Melanotropin alpha — Bovine	158	mamba	185
Melanotropin beta — Bovine	158	Venom protein CM-13b — Banded Egyptian	
Melanotropin beta — Horse	158	cobra	186
Distal-retinal-pigment hormone — Prawn	158	Venom protein S ₄ C ₁₁ — Forest cobra	186
Vasotocin — Sturgeons*	158	Short toxin 2 — West African green mamba	186
Adipokinetic hormone — Locusts	159	Short neurotoxins 3 and 4 — Banded	
Red pigment-concentrating hormone —		Egyptian cobra	186
Prawns	159	Short neurotoxin 2 — Banded Egyptian	
Cholecystokinin-pancreozymin — Pig	160	cobra	187
Big gastrin — Human	160	Short neurotoxin 1 — <i>Naja naja philippinensis</i>	
Big gastrin — Pig	160	and <i>N. n. samarensis</i> *	187
Caerulein — Australian frog	160	Short neurotoxin 1 — Broad-banded blue	
Caerulein — African frog	160	sea snake*	187
Paragonial peptide PS-1 — Fruit fly	160	Short neurotoxin 1 — Reef sea snake	
Eledoisin — Octopus	161	(<i>Aipysurus laevis</i>)	188
Kassinin — African frog	161	Short neurotoxin 1 — Yellow-bellied sea snake	
Uperolein — Australian frogs	161	(<i>Pelamis platurus</i>)*	188
Ameletin — Rat	161	Cytotoxin 3 — Formosan cobra*	188
Substance P — Bovine	161	Cytotoxin 2 — Middle Asian cobra	188

Cytotoxins 2 and 4 — Formosan cobra	188	Ig kappa-B9 chain C region — Rabbit	217
Cytotoxin 1 — Formosan cobra*	189	Ig lambda chain V-I region — Human Vor*	217
Cytotoxin 1 — Middle Asian cobra	189	Ig lambda chain V-II region — Human Boh	217
Cytotoxin 10 — Banded Egyptian cobra	189	Ig lambda chain V-II and C regions — Human	
Cytotoxin 9 — Banded Egyptian cobra	189	Bur	218
Cytotoxin 1 — Cape cobra	189	Ig lambda chain V-II region — Human Mcg	218
Cytotoxin 3 — Cape cobra	190	Ig abnormal lambda chain V-II region —	
Cytotoxin 3 — Banded Egyptian cobra	190	Human Sm	218
Cytotoxin 8 — Banded Egyptian cobra	190	Ig lambda chain V-II region — Human Tro	219
Cytotoxin 4 — Banded Egyptian cobra	190	Ig lambda chain V-V region — Human Del	219
Cytotoxin 2 — Banded Egyptian cobra	191	Ig lambda-1 chain precursor V and C	
Cytotoxin 2 — Cape cobra	191	regions — Mouse MOPC 104E	219
Cytotoxins 5, 6, and 7 — Banded Egyptian		Ig lambda chain V and C regions — Pig	220
cobra	191	Ig heavy chain V-II region — Human Newm	220
Cytotoxin 11 — Banded Egyptian cobra	191	Ig heavy chain V-III region — Human Bro	220
Cytotoxin 3 — Forest cobra	192	Ig heavy chain V-III region — Human Lay	221
Crotamine — South American rattlesnake	192	Ig heavy chain V-III region — Human Pom	221
Neurotoxin I — North American scorpion	192	Ig heavy chain V region — Mouse M167	221
Neurotoxin 2 — North American scorpion	192	Ig heavy chain V region — Mouse M603	221
Neurotoxin 3 — North American scorpion	193	Ig heavy chain V regions — Mouse TEPC 15,	
Neurotoxin 1 — North American scorpion	193	S107, and H8	222
Melittin — Ceylon bee	193	Ig heavy chain V region — Mouse M511	222
Neurotoxin B-IV — Heteronemertine worm	193	Ig heavy chain V region — Mouse ABE-47N	222
Toxin II — Sea anemone (<i>Anemonia</i>		Ig heavy chain V-A1 region — Rabbit BS-5	223
<i>sulcata</i>)	194	Ig heavy chain V-A2 region — Rabbit BS-1	223
Anthopleurin A — Sea anemone		Ig heavy chain V-A2 region — Rabbit K-25	223
(<i>Anthopleura xanthogrammica</i>)	194	Ig alpha-1 chain V-III and C regions —	
Purothionin A-I — Wheat	194	Human Bur	224
Purothionin A-II — Wheat	195	Ig alpha-1 chain V-III and C regions — Human	
Cholera enterotoxin, beta chain —		Tro	224
<i>Vibrio cholerae</i>	195	Ig epsilon chain V-I and C regions — Human	
		Nd	225
IMMUNOGLOBULINS		Ig gamma-1 chain V-III and C regions —	
Ig kappa chain V-I region — Human Au*	213	Human Nie*	225
Ig kappa chain V-I region — Human Lay	213	Ig gamma-3 heavy chain disease protein —	
Ig kappa chain V-I region — Human Rei*	213	Human Zuc	225
Ig kappa chain V-II region — Human Fr	213	Ig gamma chain C region — Rabbit	226
Ig kappa chain V-III region — Human Pom	213	Ig gamma-1 chain V and C regions — Mouse	
Ig kappa chain V-IV region — Human Len	214	MOPC 21	226
Ig kappa chain V region — Mouse MOPC 173	214	Ig gamma-2A chain V and C regions — Mouse	
Ig kappa chain precursor V region — Mouse		MOPC 173	227
MOPC 41	214		
Ig kappa chain V and C regions — Rat S211	214	GLOBINS	
Ig kappa chain V region — Rabbit 120		Hemoglobin alpha chain — Marmoset*	240
(fragment)	215	Hemoglobin alpha chain — Tarsier	240
Ig kappa chain V region — Rabbit 3315	215	Hemoglobin alpha chains — Dog*	240
Ig kappa chain V region — Rabbit 3368	215	Hemoglobin alpha chain — Horse*	240
Ig kappa chain V region — Rabbit 3374	215	Hemoglobin alpha chain — Llama	240
Ig kappa chain V region — Rabbit 3381		Hemoglobin alpha chain — Pig	241
(fragment)	216	Hemoglobin alpha chain — Opossum	241
Ig kappa chain V region — Rabbit 3547	216	Hemoglobin alpha chain — Platypus	241
Ig kappa chain V region — Rabbit K-25	216	Hemoglobin alpha chain — Goose	241
Ig kappa chain V region — Rabbit K16-167	216	Hemoglobin alpha-1 chain — Chicken	242
Ig kappa chain V region — Rabbit XP-1		Hemoglobin alpha chain — Port Jackson	
(fragment)	217	shark	242

Hemoglobin beta chain — Marmoset*	242	30S ribosomal protein S13 — <i>Escherichia coli</i>	255
Hemoglobin beta chain — Tarsier	242	30S ribosomal protein S15 — <i>Escherichia coli</i>	255
Hemoglobin beta chain — Rat	243	30S ribosomal protein S16 — <i>Escherichia coli</i>	255
Hemoglobin beta chain — Yellow-cheeked vole	243	30S ribosomal protein S18 — <i>Escherichia coli</i>	255
Hemoglobin beta chain — Dog*	243	30S ribosomal protein S20 — <i>Escherichia coli</i>	256
Hemoglobin beta chain — Badger	243	30S ribosomal protein S21 — <i>Escherichia coli</i>	256
Hemoglobin beta fetal chain — Sheep	244	50S ribosomal protein L5 — <i>Escherichia coli</i>	256
Hemoglobin beta chain — Platypus	244	50S ribosomal protein L8 (aggregate) — <i>Escherichia coli</i> *	256
Hemoglobin beta chain — Chicken*	244	50S ribosomal protein L10 — <i>Escherichia coli</i>	257
Hemoglobin beta chain — Bullfrog C	244	50S ribosomal protein L16 — <i>Escherichia coli</i>	257
Myoglobin — Eastern highland gorilla and orangutan*	245	50S ribosomal protein L18 — <i>Escherichia coli</i>	257
Myoglobin — European hedgehog	245	50S ribosomal protein L25 — <i>Escherichia coli</i>	257
Myoglobin — Dog, bat-eared fox, and Cape hunting dog	245	50S ribosomal protein L27 — <i>Escherichia coli</i>	258
Myoglobin — Common treeshrew	245	50S ribosomal protein L29 — <i>Escherichia coli</i>	258
Myoglobin — Potto and slow loris	246	50S ribosomal protein L30 — <i>Escherichia coli</i>	258
Myoglobin — Rabbit	246	50S ribosomal protein L32 — <i>Escherichia coli</i>	258
Myoglobin — California gray whale, pike whale, and Amazon River dolphin	246	50S ribosomal protein L33 — <i>Escherichia coli</i>	259
Myoglobin — Black Sea dolphin, Atlantic bottle-nosed dolphin, and killer whale*	247	50S ribosomal protein L34 — <i>Escherichia coli</i>	259
Myoglobin — Dwarf sperm whale*	247	Coat protein — Tobacco mosaic virus cowpea	259
Myoglobin — Zebra*	247	Coat protein — Alfalfa mosaic virus 425	260
Myoglobin — Pig	247	Coat protein — Alfalfa mosaic virus S	260
Myoglobin — Sheep and red deer*	248	Coat protein — Bacteriophage Q beta*	261
Myoglobin — Opossum	248	Coat protein — Bacteriophage fd*	261
Myoglobin — Platypus	248	Coat protein — Bacteriophage Pf1	261
Myoglobin — Chicken	248	Assembly protein — Bacteriophage MS2	261
Globin CTT-II beta — Midge larva	249	Internal protein I — Bacteriophage T4	262
Globin — Gastropod mollusc (<i>Busycon canaliculatum</i>)	249	Internal protein II — Bacteriophage T4	262
Leghemoglobin I — Broad bean	249	Gene B protein — Bacteriophage phi-X174	262
Leghemoglobin I — Yellow lupin	249	Gene D protein — Bacteriophage phi-X174	263
		Gene E protein — Bacteriophage phi-X174	263
		Gene G protein — Bacteriophage phi-X174	264
		Gene J protein — Bacteriophage phi-X174	264
NUCLEIC ACID-ASSOCIATED PROTEINS		FIBROUS PROTEINS	
Ubiquitin — Human and bovine	252	Alpha crystallin A chain — Rhesus monkey and human	267
Histone H2A — Bovine and rat*	252		
Histone H2B — Brown trout	252		
Histone H4 — Sea urchin*	252		
Histone H5 — Chicken (fragments)	253		
Galline — Chicken	253		
Protamine — Caspian sturgeon	253		
Lac repressor — <i>Escherichia coli</i> *	253		
30S ribosomal protein S4 — <i>Escherichia coli</i> *	254		
30S ribosomal protein S6 — <i>Escherichia coli</i>	254		
30S ribosomal protein S8 — <i>Escherichia coli</i>	254		
30S ribosomal protein S9 — <i>Escherichia coli</i>	254		
30S ribosomal protein S12 — <i>Escherichia coli</i>	254		

Alpha crystallin A chain — Dog, cat, pig, horse, rabbit, and rat*	267	Fibrinogen beta (B) chain — Human	299
Alpha crystallin A chain — Red kangaroo and opossum	267	Fibrinogen gamma chain — Human	300
Alpha crystallin B chain — Human*	268	Kappa casein — Goat	301
Keratin, high-sulfur fraction, IIIA3 — South African angora goat*	268	Myelin basic protein — Chimpanzee*	301
Keratin, high-sulfur fraction, IIIB2 — South African angora goat	268	Neurophysin 1 — Bovine	301
Keratin, feather — Silver gull	268	Neurophysin 1 — Human (fragment)	302
Collagen alpha 1(I) chain — Chicken skin (fragments)	269	Neurophysin 2 — Sheep and bovine	302
Collagen alpha 1(II) chain — Bovine cartilage (fragments)	270	Neurophysin 2 — Pig	303
Collagen alpha 1(III) chain — Human liver and skin (fragments)	271	Glycophorin A — Human	303
Collagen alpha 2 chain — Bovine skin (fragment)	272	Haptoglobin alpha 1 and 2 chains — Human*	303
		Haptoglobin alpha chain — Pig	304
CONTRACTILE SYSTEM PROTEINS		Amyloid protein AA — Human JL*	304
Flagellin — <i>Bacillus subtilis</i> 168*	278	Amyloid protein AA — Pekin duck	304
Tropomyosin alpha chain, skeletal muscle — Rabbit	278	C-reactive protein — Human	305
Troponin C, skeletal muscle — Human*	278	Complement C3a anaphylatoxin — Human	305
Troponin C, skeletal muscle — Rabbit*	278	Complement C3a anaphylatoxin — Pig	305
Troponin C, skeletal muscle — Chicken	278	Serum albumin — Human	306
Troponin C, skeletal muscle — Frog	279	Serum albumin precursor — Bovine*	306
Troponin C, cardiac muscle — Bovine	279	Gamma-carboxyglutamic acid-containing protein — Bovine	307
Calcium-dependent regulator protein — Bovine	279	Metallothionein-1B — Horse	307
Myosin DTNB light chain — Rabbit	280	Statherin — Human	307
Parvalbumin — Frog*	280	Myohemerythrin — <i>Themiste</i> (= <i>Dendrostomum</i>) <i>zostericola</i>	308
Parvalbumin — Whiting	280	Hevein — Rubber tree	308
Parvalbumin II — Pike	280	Allergen RA5 — Ragweed	308
Parvalbumin — Cod	281	Monellin, chain A — Serendipity berry	309
Parvalbumin — Rabbit	281	Monellin, chain B — Serendipity berry	309
Parvalbumin — Thornback ray	281		
Calcium-binding protein, intestinal — Pig	282	TRANSFER RNA	
Troponin T, skeletal muscle — Rabbit	282	Arginine (I C G) transfer RNA 2 — <i>Escherichia coli</i> *	319
Troponin I, skeletal muscle — Rabbit	283	Aspartic acid (Q G U C) transfer RNA 1 — <i>Escherichia coli</i> *	319
Troponin I, cardiac muscle — Rabbit	283	Asparagine (Q G U U) transfer RNA — <i>Escherichia coli</i>	319
		Cysteine (G C A) transfer RNA — Baker's yeast*	319
MISCELLANEOUS PROTEINS		Glutamic acid (S N U U C) transfer RNA 1 — <i>Escherichia coli</i> *	319
Fibrinopeptide A — Italian water buffalo	285	Glutamine (+U U G) transfer RNA 1 — <i>Escherichia coli</i> *	319
Fibrinopeptide A — Lamprey	285	Glutamine (C U G) transfer RNA 2 — <i>Escherichia coli</i> *	319
Fibrinopeptide B — Italian water buffalo	285	Glutamine (+U U G) and leucine (+U A A) transfer RNA precursor — Bacteriophage T4*	320
Fibrinopeptide B — Lamprey	285	Glycine (+U C C) transfer RNA 2 — <i>Escherichia coli</i>	320
Lipid-binding protein A-II — Rhesus monkey (Gln-II)	297	Glycine (G C C) transfer RNA 1 — Wheat	320
Apovitellenin I — Chicken	297	Histidine (Q G U G) transfer RNA 1 — <i>Escherichia coli</i> *	320
Apovitellenin I — Pekin duck	297	Leucine (U A G) transfer RNA — Baker's yeast	320
Murein-lipoprotein precursor — <i>Escherichia</i> <i>coli</i> B	297		
Fibrinogen alpha (A) chain — Human (fragments)	298		

Lysine (+U U U) transfer RNA — <i>Escherichia coli</i>	321	RIBOSOMAL AND OTHER RNAs	
Methionine (C A U) transfer RNA 3 —		5S ribosomal RNA — <i>Clostridium pasteurianum</i>	333
Brewer's yeast and baker's yeast	321	5S ribosomal RNA — <i>Bacillus brevis</i>	333
<i>N</i> -Formyl-methionine (C A U) transfer RNA —		5S ribosomal RNA — <i>Bacillus firmus</i>	333
<i>Bacillus subtilis</i>	321	5S ribosomal RNA — <i>Bacillus pasteurii</i>	333
<i>N</i> -Formyl-methionine (C A U) transfer RNA —		5S ribosomal RNA 1 — <i>Bacillus licheniformis</i>	333
<i>Thermus thermophilus</i>	322	5S ribosomal RNA 2 — <i>Bacillus licheniformis</i>	333
<i>N</i> -Formyl-methionine (C A U) transfer RNA —		5S ribosomal RNA — <i>Bacillus megaterium</i> *	334
<i>Anacystis nidularis</i>	322	5S ribosomal RNA — <i>Bacillus stearothermophilus</i> *	334
Formylatable methionine (C A U) transfer		5S ribosomal RNA precursor — <i>Bacillus subtilis</i>	334
RNA — Brewer's yeast*	322	5S ribosomal RNA — <i>Streptococcus faecalis</i>	334
Formylatable methionine (C A U) transfer		5S ribosomal RNA — <i>Lactobacillus brevis</i>	334
RNA — Salmon, human, and South		5S ribosomal RNA — <i>Photobacterium phosphoreum</i>	334
African toad*	322	5S ribosomal RNA — <i>Proteus vulgaris</i>	335
Phenylalanine (OMG A A) transfer		5S ribosomal RNA — Rye, dwarf bean,	
RNA — <i>Bacillus stearothermophilus</i>	323	sunflower, and tomato	335
Phenylalanine (G A A) transfer RNA —		5S ribosomal RNA — Broad bean	335
Chloroplast, <i>Euglena gracilis</i>	323	5S ribosomal RNA precursor — <i>Drosophila melanogaster</i>	335
Phenylalanine (OMG A A) transfer RNA —		5S ribosomal RNA — Chicken*	335
Baker's yeast*	323	5S ribosomal RNA — South African toads*	336
Phenylalanine (OMG A A) transfer RNA —		5S ribosomal RNA — Iguana*	336
Pea*	323	5S ribosomal RNA — Bovine and dolphin*	336
Phenylalanine (OMG A A) transfer RNA —		2S RNA — <i>Drosophila melanogaster</i>	336
Human*	323	4.5S RNA precursor — <i>Escherichia coli</i>	336
Serine (5AU G A) transfer RNA 1 —		5.8S ribosomal RNA — Rat, human, mouse,	
<i>Escherichia coli</i> *	324	and turtle	337
Threonine (G G U) transfer RNA —		6S messenger RNA — Bacteriophage	
<i>Escherichia coli</i> *	324	lambda*	337
Threonine (I G U) transfer RNA 1A and			
1B — Brewer's yeast	324	GENOME NUCLEIC ACIDS	
Tyrosine (QG U A) transfer RNA 1 and 2 —		Genome (DNA) — Bacteriophage phi-X174	339-341
<i>Escherichia coli</i> *	324	Genome (RNA) — Bacteriophage MS2	343-344
Valine (G A C) transfer RNA 2 — <i>Bacillus</i>			
<i>stearothermophilus</i>	324		
Valine (I A C) transfer RNA 1 — Baker's			
yeast*	325		
Valine (+U A C) transfer RNA 2A — Baker's			
yeast*	325		

22 A Model of Evolutionary Change in Proteins

M.O. Dayhoff, R.M. Schwartz, and B.C. Orcutt

In the eight years since we last examined the amino acid exchanges seen in closely related proteins,¹ the information has doubled in quantity and comes from a much wider variety of protein types. The matrices derived from these data that describe the amino acid replacement probabilities between two sequences at various evolutionary distances are more accurate and the scoring matrix that is derived is more sensitive in detecting distant relationships than the one that we previously derived.^{2,3} The method used in this chapter is essentially the same as that described in the *Atlas*, Volume 3⁴ and Volume 5.¹

Accepted Point Mutations

An accepted point mutation in a protein is a replacement of one amino acid by another, accepted by natural selection. It is the result of two distinct processes: the first is the occurrence of a mutation in the portion of the gene template producing one amino acid of a protein; the second is the acceptance of the mutation by the species as the new predominant form. To be accepted, the new amino acid usually must function in a way similar to the old one: chemical and physical similarities are found between the amino acids that are observed to interchange frequently.

Any complete discussion of the observed behavior of amino acids in the evolutionary process must consider the frequency of change of each amino acid to each other one and the propensity of each to remain unchanged. There are $20 \times 20 = 400$ possible comparisons. To collect a useful amount of information on these, a great many observations are necessary. The body of data used in this study includes 1,572 changes in 71 groups of closely related proteins appearing in the *Atlas* volumes through Supplement 2.

The mutation data were accumulated from the phylogenetic trees and from a few pairs of related sequences. The sequences of all the nodal common ancestors in each tree are routinely generated. Consider, for example, the much simplified artificial phylogenetic tree of Figure 78.

The matrix of accepted point mutations calculated from this tree is shown in Figure 79. We have assumed that the likelihood of amino acid X replacing Y is the same as that of Y replacing X, and hence 1 is entered in box YX as well as in box XY. This assumption is reasonable, because this likelihood should depend on the product of the frequencies of occurrence of the two amino acids and on their chemical and physical similarity. As a consequence of this assumption, no change in amino acid frequencies over evolutionary distance will be detected.

By comparing observed sequences with inferred ancestral sequences, rather than with each other, a sharper

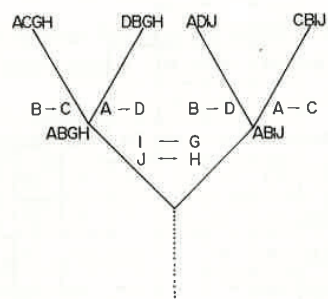


Figure 78. Simplified phylogenetic tree. Four "observed" proteins are shown at the top. Inferred ancestors are shown at the nodes. Amino acid exchanges are indicated along the branches.

	A	B	C	D	G	H	I	J
A			1	1				
B			1	1				
C	1	1						
D	1	1						
G							1	
H								1
I					1			
J						1		

Figure 79. Matrix of accepted point mutations derived from the tree of Figure 78.

change, as well as those that did. For this we need to know the probability that each amino acid will change in a given small evolutionary interval. We call this number the "relative mutability" of the amino acid.

In order to compute the relative mutabilities of the amino acids, we simply count the number of times that each amino acid has changed in an interval and the number of times that it has occurred in the sequences and thus has been subject to mutation. The relative mutability of each amino acid is proportional to the ratio of changes to occurrences. Figure 81 illustrates the computation for a simple case in which B changes relatively often, A less often, and D never.

Aligned sequences	A	D	A
Amino acids	A	D	B
Changes	1	1	0
Frequency of occurrence (total composition)	3	1	2
Relative mutability	.33	1	0

Figure 81. Sample computation of relative mutability. The two aligned sequences may be two experimentally observed sequences or an observed sequence and its inferred ancestor.

In calculating relative mutabilities from many trees, the information from sequences of different lengths and evolutionary distances is combined. Each relative mutability is still a ratio. The numerator is the total number of changes of this amino acid on all branches of all protein trees considered. The denominator is the total exposure of the amino acid to mutation, that is, the sum for all branches of its local frequency of occurrence multiplied by the total number of mutations per 100 links for that branch.

The relative mutabilities of the amino acids are shown in Table 21. On the average, Asn, Ser, Asp, and Glu are most mutable and Trp and Cys are least mutable.

The immutability of cysteine is understandable. Cysteine is known to have several unique, indispensable functions. It is the attachment site of heme groups in cytochrome and of FeS clusters in ferredoxin. It forms cross-links in other proteins such as chymotrypsin or ribonuclease. It seldom occurs without having an important function.

The substitution of one of the larger amino acids of distinctive shape and chemistry for any other is rather uncommon. At the other extreme, the low mutability of glycine must be due to its unique smallness that is advantageous in many places. Even though serine sometimes functions in the active center, it much more often per-

Table 21
Relative Mutabilities of the Amino Acids^a

Asn	134	His	66
Ser	120	Arg	65
Asp	106	Lys	56
Glu	102	Pro	56
Ala	100	Gly	49
Thr	97	Tyr	41
Ile	96	Phe	41
Met	94	Leu	40
Gln	93	Cys	20
Val	74	Trp	18

^a The value for Ala has been arbitrarily set at 100.

forms a function of lesser importance, easily mimicked by several other amino acids of similar physical and chemical properties. On the average it is highly mutable.

Amino Acid Frequencies in the Mutation Data

The relative frequencies of exposure to mutation of the amino acids are shown in Table 22. These frequencies, f_i , are approximately proportional to the average composition of each group multiplied by the number of mutations in the tree. The sum of the frequencies is 1.

Mutation Probability Matrix for the Evolutionary Distance of One PAM

We can combine information about the individual kinds of mutations and about the relative mutability of the amino acids into one distance-dependent "mutation probability matrix" (see Figure 82). An element of this matrix, M_{ij} , gives the probability that the amino acid in column j will be replaced by the amino acid in row i after a given evolutionary interval, in this case 1 PAM.

Table 22
Normalized Frequencies of the Amino Acids in the Accepted Point Mutation Data

Gly	0.089	Arg	0.041
Ala	0.087	Asn	0.040
Leu	0.085	Phe	0.040
Lys	0.081	Gln	0.038
Ser	0.070	Ile	0.037
Val	0.065	His	0.034
Thr	0.058	Cys	0.033
Pro	0.051	Tyr	0.030
Glu	0.050	Met	0.015
Asp	0.047	Trp	0.010

The nondiagonal elements have the values:

$$M_{ij} = \frac{\lambda m_j A_{ij}}{\sum_i A_{ij}}$$

where

A_{ij} is an element of the accepted point mutation matrix of Figure 80,

λ is a proportionality constant, and

m_j is the mutability of the j th amino acid, Table 21.

The diagonal elements have the values:

$$M_{jj} = 1 - \lambda m_j$$

Consider a typical column, that for alanine. The total probability, the sum of all the elements, must be 1. The

probability of observing a change in a site containing alanine (the sum of all the elements except M_{AA}) is proportional to the mutability of alanine. The same proportionality constant, λ , holds for all columns. The individual nondiagonal terms within each column bear the same ratio to each other as do the observed mutations in the matrix of Figure 80.

The quantity $100 \times \sum f_i M_{ij}$ gives the number of amino acids that will remain unchanged when a protein 100 links long, of average composition, is exposed to the evolutionary change represented by this matrix. This apparent evolutionary change depends upon the choice of λ , in this case chosen so that this change is 1 mutation. Since there are almost no superimposed changes, this also represents 1 PAM of change. If λ had been four times as large, the initial matrix would have represented 4 PAMs; the discussion which follows would not be changed noticeably.

	ORIGINAL AMINO ACID																			
	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
REPLACEMENT AMINO ACID	Ala	Arg	Asn	Asp	Cys	Gln	Glu	Gly	His	Ile	Leu	Lys	Met	Phe	Pro	Ser	Thr	Trp	Tyr	Val
A Ala	9867	2	9	10	3	8	17	21	2	6	4	2	6	2	22	35	32	0	2	18
R Arg	1	9913	1	0	1	10	0	0	10	3	1	19	4	1	4	6	1	8	0	1
N Asn	4	1	9822	36	0	4	6	6	21	3	1	13	0	1	2	20	9	1	4	1
D Asp	6	0	42	9859	0	6	53	6	4	1	0	3	0	0	1	5	3	0	0	1
C Cys	1	1	0	0	9973	0	0	0	1	1	0	0	0	0	1	5	1	0	3	2
Q Gln	3	9	4	5	0	9876	27	1	23	1	3	6	4	0	6	2	2	0	0	1
E Glu	10	0	7	56	0	35	9865	4	2	3	1	4	1	0	3	4	2	0	1	2
G Gly	21	1	12	11	1	3	7	9935	1	0	1	2	1	1	3	21	3	0	0	5
H His	1	8	18	3	1	20	1	0	9912	0	1	1	0	2	3	1	1	1	4	1
I Ile	2	2	3	1	2	1	2	0	0	9872	9	2	12	7	0	1	7	0	1	33
L Leu	3	1	3	0	0	6	1	1	4	22	9947	2	45	13	3	1	3	4	2	15
K Lys	2	37	25	6	0	12	7	2	2	4	1	9926	20	0	3	8	11	0	1	1
M Met	1	1	0	0	0	2	0	0	0	5	8	4	9874	1	0	1	2	0	0	4
F Phe	1	1	1	0	0	0	0	1	2	8	6	0	4	9946	0	2	1	3	28	0
P Pro	13	5	2	1	1	8	3	2	5	1	2	2	1	1	9926	12	4	0	0	2
S Ser	28	11	34	7	11	4	6	16	2	2	1	7	4	3	17	9840	38	5	2	2
T Thr	22	2	13	4	1	3	2	2	1	11	2	8	6	1	5	32	9871	0	2	9
W Trp	0	2	0	0	0	0	0	0	0	0	0	0	0	1	0	1	0	9976	1	0
Y Tyr	1	0	3	0	3	0	1	0	4	1	1	0	0	21	0	1	1	2	9945	1
V Val	13	2	1	1	3	2	2	3	3	57	11	1	17	1	3	2	10	0	2	9901

Figure 82. Mutation probability matrix for the evolutionary distance of 1 PAM. An element of this matrix, M_{ij} , gives the probability that the amino acid in column j will be replaced by the amino acid in row i after a given evolutionary interval, in this case

1 accepted point mutation per 100 amino acids. Thus, there is a 0.56% probability that Asp will be replaced by Glu. To simplify the appearance, the elements are shown multiplied by 10,000.

Simulation of the Mutational Process

For evaluating statistical methods of detecting relationships, for developing methods of measuring evolutionary distances between proteins, and for determining the accuracy of programs to construct evolutionary trees, we need to have examples of proteins at known evolutionary distances. The mutation probability matrix provides the information with which to simulate any amount of evolutionary change in an unlimited number of proteins. Further, we can start with one protein and simulate its separate evolution in duplicated genes or in divergent organisms. By considering many groups of sequences related by the same evolutionary history, a measure is readily obtained of the expected deviations due to random fluctuations in the evolutionary process.

If we only require that, on the average, one mutation takes place in the evolutionary interval of 1 PAM, we can use a simulation requiring one random number for each amino acid in the sequence, as follows: To determine the fate of the first amino acid, say Ala, a uniformly distributed random number between 0 and 1 is obtained. The first column of the mutation probability matrix (Figure 82) gives the relative probability of each possible event that may befall Ala (neglecting deletion for simplicity). If the random number falls between 0 and .9867, Ala is left unchanged. If the number is between .9867 and .9868, it is replaced with Arg, if it is between .9868 and .9872, it is replaced with Asp, and so forth. Similarly, a random number is produced for each amino acid in the sequence, and action is taken as dictated by the corresponding column of the matrix. The result is a simulated mutant sequence. Any number of these can be generated; their average distance from the original is 1 PAM although some may have no mutations and some may have two or more. The effects on the sequence of a longer period of evolution may be simulated by successive applications of the matrix to the sequence resulting from the last application.

For simulations in which a predetermined number of changes are required, a two-step process involving two random numbers for each mutation can be used. Starting with a given sequence, the first amino acid that will mutate is selected: the probability that any one will be selected is proportional to its mutability (Table 21). Then the amino acid that replaces it is chosen. The probability for each replacement is proportional to the elements in the appropriate column of Figure 82. Starting with the resultant sequence, a second mutation can be simulated, and so on, until a predetermined number of changes have been made. In this process, superimposed and back mutations may occur.

The 1 PAM matrix can be multiplied by itself N times to yield a matrix that predicts the amino acid replacements to be found after N PAMs of evolutionary change in a sequence of average composition. On the average, the results of the simulations above match the predictions of the corresponding matrices.

Mutation Probability Matrices for Other Distances

The mutation probability matrix M_1 , corresponding to 1 PAM, has a number of interesting properties (see Figure 82). If, in a simulation, it is applied to a protein with the average amino acid composition given in Table 22, on the average, the composition of the resulting mutated proteins will be unchanged. Repeated applications of the matrix to proteins of any other composition will give mutants that change toward average composition; any such matrix has implicit in it some particular asymptotic composition.

There is a different mutation probability matrix for each evolutionary interval. These can be derived from the one for 1 PAM by matrix multiplication. If the 1-PAM matrix is multiplied by itself an infinite number of times, each column of the resulting matrix approaches the asymptotic amino acid composition:

$$M_{\infty} = \begin{pmatrix} f_A & f_A & f_A & f_A & \dots \\ f_R & f_R & f_R & f_R & \dots \\ f_N & f_N & f_N & f_N & \dots \\ \vdots & \vdots & \vdots & \vdots & \ddots \end{pmatrix}$$

At a great distance, there is very little relationship information left in the matrix. For example, at a distance of 2,034 PAMs, all of the matrix values are within 5% of their limiting values except for the Trp-Trp element, which is 75% higher than the limit, and the Cys-Cys element, which is 11% higher.

The matrix for 0 PAMs is simply a unit diagonal; no amino acid would have changed:

$$M_0 = \begin{pmatrix} 1 & 0 & 0 & \dots \\ 0 & 1 & 0 & \dots \\ 0 & 0 & 1 & \dots \\ \vdots & \vdots & \vdots & \ddots \end{pmatrix}$$

The mutation probability matrix for 250 PAMs is shown in Figure 83. At this evolutionary distance, only one amino acid in five remains unchanged. However, the

amino acids vary greatly in their mutability; 55% of the tryptophans, 52% of the cysteines and 27% of the glycines would still be unchanged, but only 6% of the highly mutable asparagines would remain. Several other amino acids, particularly alanine, aspartic acid, glutamic acid, glycine, lysine, and serine are more likely to occur in place of an original asparagine than asparagine itself at this evolutionary distance! This is understandable from the data giving the preferred mutations and the relative mutabilities. Asparagine is highly mutable, therefore it changes to other amino acids. These are less mutable and may not change again. This effect is much more conspicuous in the case of methionine. Surprisingly, a methionine originally present would have changed to leucine in 20% of the cases, but would remain methionine in only 6%. Over one-third of the mutations in methionine are specifically to leucine (Figure 80). Leucine is less than one-half as mutable as methionine (Table 21).

From the series of distance-dependent mutation probability matrices, we can compute detailed answers to the question "How does the evolutionary process affect the similarity of related protein sequences?"

Estimation of Evolutionary Distance

There is a different mutation probability matrix for each evolutionary interval measured in PAMs. For each such matrix, we can calculate the percentage of amino acids that will be observed to change on the average in the interval by the formula:

$$100(1 - \sum_i f_i M_{ii})$$

Table 23 shows the correspondence between the observed percent difference between two sequences and the evolutionary distance in PAMs. We use this scale to estimate

		ORIGINAL AMINO ACID																			
		A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
		Ala	Arg	Asn	Asp	Cys	Gln	Glu	Gly	His	Ile	Leu	Lys	Met	Phe	Pro	Ser	Thr	Trp	Tyr	Val
REPLACEMENT AMINO ACID	A Ala	13	6	9	9	5	8	9	12	6	8	6	7	7	4	11	11	11	2	4	9
	R Arg	3	17	4	3	2	5	3	2	6	3	2	9	4	1	4	4	3	7	2	2
	N Asn	4	4	6	7	2	5	6	4	6	3	2	5	3	2	4	5	4	2	3	3
	D Asp	5	4	8	11	1	7	10	5	6	3	2	5	3	1	4	5	5	1	2	3
	C Cys	2	1	1	1	52	1	1	2	2	2	1	1	1	1	2	3	2	1	4	2
	Q Gln	3	5	5	6	1	10	7	3	7	2	3	5	3	1	4	3	3	1	2	3
	E Glu	5	4	7	11	1	9	12	5	6	3	2	5	3	1	4	5	5	1	2	3
	G Gly	12	5	10	10	4	7	9	27	5	5	4	6	5	3	8	11	9	2	3	7
	H His	2	5	5	4	2	7	4	2	15	2	2	3	2	2	3	3	2	2	3	2
	I Ile	3	2	2	2	2	2	2	2	2	10	6	2	6	5	2	3	4	1	3	9
	L Leu	6	4	4	3	2	6	4	3	5	15	34	4	20	13	5	4	6	6	7	13
	K Lys	6	18	10	8	2	10	8	5	8	5	4	24	9	2	6	8	8	4	3	5
	M Met	1	1	1	1	0	1	1	1	1	2	3	2	6	2	1	1	1	1	1	2
	F Phe	2	1	2	1	1	1	1	1	3	5	6	1	4	32	1	2	2	4	20	3
	P Pro	7	5	5	4	3	5	4	5	5	3	3	4	3	2	20	6	5	1	2	4
	S Ser	9	6	8	7	7	6	7	9	6	5	4	7	5	3	9	10	9	4	4	6
	T Thr	8	5	6	6	4	5	5	6	4	6	4	6	5	3	6	8	11	2	3	6
	W Trp	0	2	0	0	0	0	0	0	1	0	1	0	0	1	0	1	0	55	1	0
	Y Tyr	1	1	2	1	3	1	1	1	3	2	2	1	2	15	1	2	2	3	31	2
	V Val	7	4	4	4	4	4	4	4	5	4	15	10	4	10	5	5	5	7	2	4

Figure 83. Mutation probability matrix for the evolutionary distance of 250 PAMs. To simplify the appearance, the elements are shown multiplied by 100. In comparing two sequences of average amino acid frequency at this evolutionary distance, there is a 13% probability that a position containing Ala in the first

sequence will contain Ala in the second. There is a 3% chance that it will contain Arg, and so forth. The relationship of two sequences at a distance of 250 PAMs can be demonstrated by statistical methods.

Table 23
Correspondence between Observed Differences
and the Evolutionary Distance

Observed Percent Difference	Evolutionary Distance in PAMs
1	1
5	5
10	11
15	17
20	23
25	30
30	38
35	47
40	56
45	67
50	80
55	94
60	112
65	133
70	159
75	195
80	246
85	328

evolutionary distances from matrices of percent difference between sequences. These estimated distances were used in the computations of evolutionary trees in this book. The differences predicted for a given PAM distance differ by up to 2.5% from those that we reported in Volume 5. A more complete scale is given in Table 36 of the Appendix.

Relatedness Odds Matrix

The elements, M_{ij} , of the mutation probability matrix for each distance give the probability that amino acid j will change to i in a related sequence in that interval. The normalized frequency f_i gives the probability that i will occur in the second sequence by chance.

The terms of the relatedness odds matrix are then:

$$R_{ij} = \frac{M_{ij}}{f_i}$$

The odds matrix is symmetrical. Each term gives the probability of replacement per occurrence of i per occurrence of j .

Amino acid pairs with scores above 1 replace each other more often as alternatives in related sequences than in random sequences of the same composition whereas those with scores below 1 replace each other less often.

The information in the 250-PAM odds matrix has proven very useful in detecting distant relationships between sequences. When one protein is compared with another, position by position, one should multiply the odds for each position to calculate an odds for the whole protein. However, it is more convenient to add the logarithms of the matrix elements. The log of the 250-PAM odds matrix is shown in Figure 84.

The Chemical Meaning of Amino Acid Mutations

Patterns have been visible in the accepted point mutations since the beginning of protein sequence work. Isoleucine-valine and serine-threonine were frequently observed alternatives. It was obvious that this interchangeability had something to do with their chemical similarities. In the large amount of information that now exists, far more detailed correlations are visible, and many more functional inferences can be made.

In the log odds matrix of Figure 84, the order of the amino acids has been rearranged to show clearly the groups of chemically similar amino acids that tend to replace one another: the hydrophobic group; the aromatic group; the basic group; the acid, acid-amide group; cysteine; and the other hydrophilic residues. Some groups overlap: the basic and acid, acid-amide groups tend to replace one another to some extent, and phenylalanine interchanges with the hydrophobic group more often than chance expectation would predict. These patterns are imposed principally by natural selection and only secondarily by the constraints of the genetic code: they reflect the similarity of the functions of the amino acid residues in their weak interactions with one another in the three-dimensional conformation of proteins. Some of the properties of an amino acid residue that determine these interactions are: size, shape, and local concentrations of electric charge; the conformation of its van der Waals surface; and its ability to form salt bonds, hydrophobic bonds, and hydrogen bonds.

Computing Relationships between Sequences

We use log odds matrices as scoring matrices for detecting very distant relationships between proteins. Such scoring matrices, based ultimately on accepted point mutations, can discriminate significant relationships from

Figure 84. Log odds matrix for 250 PAMs. Elements are shown multiplied by 10. The neutral score is zero. A score of -10 means that the pair would be expected to occur only one-tenth as frequently in related sequences as random chance would predict, and

random coincidences better than simpler scoring systems. Mere counts of identities and matrices based only on the changes predicted by the genetic code are not sufficiently complex. It is obvious that there is a good deal of information in the detailed nature of both the nonidentities and the identities. Certain combinations of different amino acids are positive evidence of relatedness, and others are contraindications. The log odds matrix for 250 PAMs, which we have found to be a very effective scoring matrix for detecting distant relationships, is compared with other matrices in chapter 23.

1. Dayhoff, M.O., Eck, R.V., and Park, C.M., *in* Atlas of Protein Sequence and Structure 1972, Vol.5, ed. Dayhoff, M.O., pp.89-99, Nat. Biomed. Res. Found., Washington, D.C., 1972
2. Schwartz, R.M., and Dayhoff, M.O., *in* Evolution of Protein Molecules, ed. Matsubara, H., and Yamanaka, T., pp.1-16, Japan Sci. Soc. Press, Tokyo, 1978
3. Schwartz, R.M., and Dayhoff, M.O., *in* Origin of Life, ed. Noda, H., pp.457-469, Center for Academic Pub. Japan/Japan Sci. Soc. Press, Tokyo, 1978
4. Dayhoff, M.O., and Eck, R.V., Atlas of Protein Sequence and Structure 1967-68, pp.33-45, Nat. Biomed. Res. Found., Silver Spring, Md., 1968