Bioinformatics for Geneticists. Edited by Michael R. Barnes and Ian C. Gray

ISBNs: 0-470-84393-4 (HB); 0-470-84394-2 (PB)

APPENDIX I

IA IUPAC Nucleotide Ambiguity Codes

IUPAC Code	Meaning	Complement
A	A	T
C	C	G
G	G	C
T/U	T	A
M	A or C	K
R	A or G	Y
W	A or T	W
S	C or G	S
Y	C or T	R
K	G or T	M
V	A or C or G	B
H	A or C or T	D
D	A or G or T	H
B	C or G or T	V
N	G or A or T or C	N

IB IUPAC Amino Acid Codes

IUPAC Amino Acid Code	Three Letter Code	Amino Acid
A	Ala	Alanine
C	Cys	Cysteine
D	Asp	Aspartate
E	Glu	Glutamate
F	Phe	Phenylalanine
G	Gly	Glycine
H	His	Histidine
I	Ile	Isoleucine
K	Lys	Lysine
L	Leu	Leucine
M	Met	Methionine
N	Asn	Asparagine
P	Pro	Proline
Q	Gln	Glutamine
R	Arg	Arginine

IUPAC Amino Acid Code	Three Letter Code	Amino Acid
S	Ser	Serine
T	Thr	Threonine
V	Val	Valine
W	Trp	Tryptophan
Y	Tyr	Tyrosine

IC Human Codon Usage Table

Second Codon					
First Codon	U	C	A	G	Last Codon
U	Phe	Ser	Tyr	Cys	**U**
	Phe	Ser	Tyr	Cys	**C**
	Leu	Ser	**Stop**	**Stop**	**A**
	Leu	Ser	**Stop**	Trp	**G**
C	Leu	Pro	His	Arg	**U**
	Leu	Pro	His	Arg	**C**
	Leu	Pro	Gln	Arg	**A**
	Leu	Pro	Gln	Arg	**G**
A	Ile	Thr	Asn	Ser	**U**
	Ile	Thr	Asn	Ser	**C**
	Ile	Thr	Lys	Arg	**A**
	Met	Thr	Lys	Arg	**G**
G	Val	Ala	Asp	Gly	**U**
	Val	Ala	Asp	Gly	**C**
	Val	Ala	Glu	Gly	**A**
	Val	Ala	Glu	Gly	**G**

Bioinformatics for Geneticists. Edited by Michael R. Barnes and Ian C. Gray

ISBNs: 0-470-84393-4 (HB); 0-470-84394-2 (PB)

APPENDIX II

Amino Acid Substitution Matrices

More information on these matrices is available on the following www site (www.russell.embl-heidelberg.de/aas).

IIA — All Protein Types

	ALA	ARG	ASN	ASP	CYS	GLN	GLU	GLY	HIS	ILE	LEU	LYS	MET	PHE	PRO	SER	THR	TRP	TYR	VAL
ALA		Arg (-2)	Asn (0)	Asp (0)	Cys (-2)	Gln (0)	Glu (0)	Gly (1)	His (-1)	Ile (-1)	Leu (-2)	Lys (-1)	Met (-1)	Phe (-3)	Pro (1)	Ser (1)	Thr (1)	Trp (-6)	Tyr (-3)	Val (0)
ARG	Ala (-2)		Asn (0)	Asp (-1)	Cys (-4)	Gln (1)	Glu (-1)	Gly (-3)	His (2)	Ile (-2)	Leu (-3)	Lys (3)	Met (0)	Phe (-4)	Pro (0)	Ser (0)	Thr (-1)	Trp (2)	Tyr (-4)	Val (-2)
ASN	Ala (0)	Arg (0)		Asp (2)	Cys (-4)	Gln (1)	Glu (1)	Gly (0)	His (2)	Ile (-2)	Leu (-3)	Lys (1)	Met (-2)	Phe (-3)	Pro (0)	Ser (1)	Thr (0)	Trp (-4)	Tyr (-2)	Val (-2)
ASP	Ala (0)	Arg (-1)	Asn (2)		Cys (-5)	Gln (2)	Glu (3)	Gly (1)	His (1)	Ile (-2)	Leu (-4)	Lys (0)	Met (-3)	Phe (-6)	Pro (-1)	Ser (0)	Thr (0)	Trp (-7)	Tyr (-4)	Val (-2)
CYS	Ala (-2)	Arg (-4)	Asn (-4)	Asp (-5)		Gln (-5)	Glu (-5)	Gly (-3)	His (-3)	Ile (-2)	Leu (-6)	Lys (-5)	Met (-5)	Phe (-4)	Pro (-3)	Ser (0)	Thr (-2)	Trp (-8)	Tyr (0)	Val (-2)
GLN	Ala (0)	Arg (1)	Asn (1)	Asp (2)	Cys (-5)		Glu (2)	Gly (-1)	His (3)	Ile (-2)	Leu (-2)	Lys (1)	Met (-1)	Phe (-5)	Pro (0)	Ser (-1)	Thr (-1)	Trp (-5)	Tyr (-4)	Val (-2)
GLU	Ala (0)	Arg (-1)	Asn (1)	Asp (3)	Cys (-5)	Gln (2)		Gly (0)	His (1)	Ile (-2)	Leu (-3)	Lys (0)	Met (-2)	Phe (-5)	Pro (-1)	Ser (0)	Thr (0)	Trp (-7)	Tyr (-4)	Val (-2)
GLY	Ala (1)	Arg (-3)	Asn (0)	Asp (1)	Cys (-3)	Gln (-1)	Glu (0)		His (-2)	Ile (-3)	Leu (-4)	Lys (-2)	Met (-3)	Phe (-5)	Pro (0)	Ser (1)	Thr (0)	Trp (-7)	Tyr (-5)	Val (-1)
HIS	Ala (-1)	Arg (2)	Asn (2)	Asp (1)	Cys (-3)	Gln (3)	Glu (1)	Gly (-2)		Ile (-2)	Leu (-2)	Lys (0)	Met (-2)	Phe (-2)	Pro (0)	Ser (-1)	Thr (-1)	Trp (-3)	Tyr (0)	Val (-2)
ILE	Ala (-1)	Arg (-2)	Asn (-2)	Asp (-2)	Cys (-2)	Gln (-2)	Glu (-2)	Gly (-3)	His (-2)		Leu (2)	Lys (-2)	Met (2)	Phe (1)	Pro (-2)	Ser (-1)	Thr (0)	Trp (-5)	Tyr (-1)	Val (4)
LEU	Ala (-2)	Arg (-3)	Asn (-3)	Asp (-4)	Cys (-6)	Gln (-2)	Glu (-3)	Gly (-4)	His (-2)	Ile (2)		Lys (-3)	Met (4)	Phe (2)	Pro (-3)	Ser (-3)	Thr (-2)	Trp (-2)	Tyr (-1)	Val (2)
LYS	Ala (-1)	Arg (3)	Asn (1)	Asp (0)	Cys (-5)	Gln (1)	Glu (0)	Gly (-2)	His (0)	Ile (-2)	Leu (-3)		Met (0)	Phe (-5)	Pro (-1)	Ser (0)	Thr (0)	Trp (-3)	Tyr (-4)	Val (-2)
MET	Ala (-1)	Arg (0)	Asn (-2)	Asp (-3)	Cys (-5)	Gln (-1)	Glu (-2)	Gly (-3)	His (-2)	Ile (2)	Leu (4)	Lys (0)		Phe (0)	Pro (-2)	Ser (-2)	Thr (-1)	Trp (-4)	Tyr (-2)	Val (2)
PHE	Ala (-3)	Arg (-4)	Asn (-3)	Asp (-6)	Cys (-4)	Gln (-5)	Glu (-5)	Gly (-5)	His (-2)	Ile (1)	Leu (2)	Lys (-5)	Met (0)		Pro (-5)	Ser (-3)	Thr (-3)	Trp (0)	Tyr (7)	Val (-1)
PRO	Ala (1)	Arg (0)	Asn (0)	Asp (-1)	Cys (-3)	Gln (0)	Glu (-1)	Gly (0)	His (0)	Ile (-2)	Leu (-3)	Lys (-1)	Met (-2)	Phe (-5)		Ser (1)	Thr (0)	Trp (-6)	Tyr (-5)	Val (-1)
SER	Ala (1)	Arg (0)	Asn (1)	Asp (0)	Cys (0)	Gln (-1)	Glu (0)	Gly (1)	His (-1)	Ile (-1)	Leu (-3)	Lys (0)	Met (-2)	Phe (-3)	Pro (1)		Thr (1)	Trp (-2)	Tyr (-3)	Val (-1)
THR	Ala (1)	Arg (-1)	Asn (0)	Asp (0)	Cys (-2)	Gln (-1)	Glu (0)	Gly (0)	His (-1)	Ile (0)	Leu (-2)	Lys (0)	Met (-1)	Phe (-3)	Pro (0)	Ser (1)		Trp (-5)	Tyr (-3)	Val (0)
TRP	Ala (-6)	Arg (2)	Asn (-4)	Asp (-7)	Cys (-8)	Gln (-5)	Glu (-7)	Gly (-7)	His (-3)	Ile (-5)	Leu (-2)	Lys (-3)	Met (-4)	Phe (0)	Pro (-6)	Ser (-2)	Thr (-5)		Tyr (0)	Val (-6)
TYR	Ala (-3)	Arg (-4)	Asn (-2)	Asp (-4)	Cys (0)	Gln (-4)	Glu (-4)	Gly (-5)	His (0)	Ile (-1)	Leu (-1)	Lys (-4)	Met (-2)	Phe (7)	Pro (-5)	Ser (-3)	Thr (-3)	Trp (0)		Val (-2)
VAL	Ala (0)	Arg (-2)	Asn (-2)	Asp (-2)	Cys (-2)	Gln (-2)	Glu (-2)	Gly (-1)	His (-2)	Ile (4)	Leu (2)	Lys (-2)	Met (2)	Phe (-1)	Pro (-1)	Ser (-1)	Thr (0)	Trp (-6)	Tyr (-2)	

IIB Extracellular Proteins

	ALA	ARG	ASN	ASP	CYS	GLN	GLU	GLY	HIS	ILE	LEU	LYS	MET	PHE	PRO	SER	THR	TRP	TYR	VAL
ALA		Arg (0)	Asn (0)	Asp (-1)	Cys (-4)	Gln (0)	Glu (0)	Gly (0)	His (0)	Ile (0)	Leu (0)	Lys (0)	Met (0)	Phe (-1)	Pro (0)	Ser (0)	Thr (0)	Trp (-2)	Tyr (-1)	Val (0)
ARG	Ala (0)		Asn (0)	Asp (0)	Cys (-5)	Gln (0)	Glu (0)	Gly (0)	His (0)	Ile (0)	Leu (-1)	Lys (1)	Met (0)	Phe (-1)	Pro (0)	Ser (0)	Thr (0)	Trp (-1)	Tyr (0)	Val (0)
ASN	Ala (0)	Arg (0)		Asp (1)	Cys (-6)	Gln (0)	Glu (0)	Gly (0)	His (0)	Ile (-1)	Leu (-2)	Lys (0)	Met (-1)	Phe (-2)	Pro (0)	Ser (0)	Thr (0)	Trp (-3)	Tyr (-1)	Val (-1)
ASP	Ala (-1)	Arg (0)	Asn (1)		Cys (-7)	Gln (0)	Glu (0)	Gly (0)	His (0)	Ile (-2)	Leu (-2)	Lys (0)	Met (-2)	Phe (-2)	Pro (0)	Ser (0)	Thr (0)	Trp (-3)	Tyr (-2)	Val (-1)
CYS	Ala (-4)	Arg (-5)	Asn (-6)	Asp (-7)		Gln (-5)	Glu (-6)	Gly (-6)	His (-5)	Ile (-5)	Leu (-5)	Lys (-6)	Met (-5)	Phe (-5)	Pro (-6)	Ser (-5)	Thr (-5)	Trp (-5)	Tyr (-4)	Val (-4)
GLN	Ala (0)	Arg (0)	Asn (0)	Asp (0)	Cys (-5)		Glu (0)	Gly (0)	His (0)	Ile (-1)	Leu (-1)	Lys (0)	Met (0)	Phe (-2)	Pro (0)	Ser (0)	Thr (0)	Trp (-1)	Tyr (-1)	Val (0)
GLU	Ala (0)	Arg (0)	Asn (0)	Asp (0)	Cys (-6)	Gln (0)		Gly (-1)	His (0)	Ile (-1)	Leu (-1)	Lys (0)	Met (0)	Phe (-2)	Pro (0)	Ser (0)	Thr (0)	Trp (-1)	Tyr (-1)	Val (0)
GLY	Ala (0)	Arg (0)	Asn (0)	Asp (0)	Cys (-6)	Gln (0)	Glu (-1)		His (0)	Ile (-2)	Leu (-2)	Lys (-1)	Met (-2)	Phe (-3)	Pro (0)	Ser (0)	Thr (0)	Trp (-2)	Tyr (-2)	Val (-2)
HIS	Ala (0)	Arg (0)	Asn (0)	Asp (0)	Cys (-5)	Gln (0)	Glu (0)	Gly (0)		Ile (-1)	Leu (-1)	Lys (0)	Met (-1)	Phe (-1)	Pro (0)	Ser (0)	Thr (0)	Trp (-1)	Tyr (0)	Val (-1)
ILE	Ala (0)	Arg (0)	Asn (-1)	Asp (-2)	Cys (-5)	Gln (-1)	Glu (-1)	Gly (-2)	His (-1)		Leu (1)	Lys (0)	Met (0)	Phe (0)	Pro (-1)	Ser (-1)	Thr (0)	Trp (-1)	Tyr (0)	Val (2)
LEU	Ala (0)	Arg (-1)	Asn (-2)	Asp (-2)	Cys (-5)	Gln (-1)	Glu (-1)	Gly (-2)	His (-1)	Ile (1)		Lys (-1)	Met (1)	Phe (0)	Pro (0)	Ser (-1)	Thr (0)	Trp (-2)	Tyr (-1)	Val (1)
LYS	Ala (0)	Arg (1)	Asn (0)	Asp (0)	Cys (-6)	Gln (0)	Glu (0)	Gly (-1)	His (0)	Ile (0)	Leu (-1)		Met (-1)	Phe (-2)	Pro (0)	Ser (0)	Thr (0)	Trp (-2)	Tyr (-1)	Val (0)
MET	Ala (0)	Arg (0)	Asn (-1)	Asp (-2)	Cys (-5)	Gln (0)	Glu (0)	Gly (-2)	His (-1)	Ile (0)	Leu (1)	Lys (-1)		Phe (0)	Pro (-1)	Ser (-1)	Thr (0)	Trp (-1)	Tyr (-1)	Val (0)
PHE	Ala (-1)	Arg (-1)	Asn (-2)	Asp (-2)	Cys (-5)	Gln (-2)	Glu (-2)	Gly (-3)	His (-1)	Ile (0)	Leu (0)	Lys (-2)	Met (0)		Pro (-2)	Ser (-2)	Thr (-1)	Trp (1)	Tyr (2)	Val (0)
PRO	Ala (0)	Arg (0)	Asn (0)	Asp (0)	Cys (-6)	Gln (0)	Glu (0)	Gly (0)	His (0)	Ile (-1)	Leu (0)	Lys (0)	Met (-1)	Phe (-2)		Ser (0)	Thr (0)	Trp (-3)	Tyr (-1)	Val (0)
SER	Ala (0)	Arg (0)	Asn (0)	Asp (0)	Cys (-5)	Gln (0)	Glu (0)	Gly (0)	His (0)	Ile (-1)	Leu (-1)	Lys (0)	Met (-1)	Phe (-2)	Pro (0)		Thr (1)	Trp (-1)	Tyr (-1)	Val (-1)
THR	Ala (0)	Arg (0)	Asn (0)	Asp (0)	Cys (-5)	Gln (0)	Glu (0)	Gly (0)	His (0)	Ile (0)	Leu (0)	Lys (0)	Met (0)	Phe (-1)	Pro (0)	Ser (1)		Trp (-1)	Tyr (-1)	Val (0)
TRP	Ala (-2)	Arg (-1)	Asn (-3)	Asp (-3)	Cys (-5)	Gln (-1)	Glu (-1)	Gly (-2)	His (-1)	Ile (-1)	Leu (-2)	Lys (-2)	Met (-1)	Phe (1)	Pro (-3)	Ser (-1)	Thr (-1)		Tyr (1)	Val (-1)
TYR	Ala (-1)	Arg (0)	Asn (-1)	Asp (-2)	Cys (-4)	Gln (-1)	Glu (-1)	Gly (-2)	His (0)	Ile (0)	Leu (-1)	Lys (-1)	Met (-1)	Phe (2)	Pro (-1)	Ser (-1)	Thr (-1)	Trp (1)		Val (0)
VAL	Ala (0)	Arg (0)	Asn (-1)	Asp (-1)	Cys (-4)	Gln (0)	Glu (0)	Gly (-2)	His (-1)	Ile (2)	Leu (1)	Lys (0)	Met (0)	Phe (0)	Pro (0)	Ser (-1)	Thr (0)	Trp (-1)	Tyr (0)	

IIC Intracellular Proteins

	ALA	ARG	ASN	ASP	CYS	GLN	GLU	GLY	HIS	ILE	LEU	LYS	MET	PHE	PRO	SER	THR	TRP	TYR	VAL
ALA		Arg (0)	Asn (-1)	Asp (-1)	Cys (0)	Gln (0)	Glu (0)	Gly (0)	His (-1)	Ile (0)	Leu (0)	Lys (0)	Met (0)	Phe (-1)	Pro (0)	Ser (0)	Thr (0)	Trp (-2)	Tyr (-1)	Val (0)
ARG	Ala (0)		Asn (0)	Asp (0)	Cys (-1)	Gln (0)	Glu (0)	Gly (0)	His (0)	Ile (-1)	Leu (-1)	Lys (1)	Met (0)	Phe (-2)	Pro (0)	Ser (0)	Thr (0)	Trp (-1)	Tyr (-1)	Val (-1)
ASN	Ala (-1)	Arg (0)		Asp (1)	Cys (-1)	Gln (0)	Glu (0)	Gly (0)	His (0)	Ile (-2)	Leu (-2)	Lys (0)	Met (-1)	Phe (-2)	Pro (-1)	Ser (0)	Thr (0)	Trp (-2)	Tyr (-1)	Val (-2)
ASP	Ala (-1)	Arg (0)	Asn (1)		Cys (-2)	Gln (0)	Glu (1)	Gly (0)	His (0)	Ile (-3)	Leu (-3)	Lys (0)	Met (-2)	Phe (-3)	Pro (0)	Ser (0)	Thr (0)	Trp (-2)	Tyr (-2)	Val (-2)
CYS	Ala (0)	Arg (-1)	Asn (-1)	Asp (-2)		Gln (-2)	Glu (-2)	Gly (-1)	His (0)	Ile (0)	Leu (0)	Lys (-1)	Met (0)	Phe (0)	Pro (-2)	Ser (0)	Thr (0)	Trp (-1)	Tyr (0)	Val (0)
GLN	Ala (0)	Arg (0)	Asn (0)	Asp (0)	Cys (-2)		Glu (1)	Gly (0)	His (0)	Ile (-2)	Leu (-1)	Lys (0)	Met (0)	Phe (-2)	Pro (0)	Ser (0)	Thr (0)	Trp (-2)	Tyr (-1)	Val (-1)
GLU	Ala (0)	Arg (0)	Asn (0)	Asp (1)	Cys (-2)	Gln (1)		Gly (-1)	His (0)	Ile (-2)	Leu (-2)	Lys (0)	Met (-1)	Phe (-2)	Pro (0)	Ser (0)	Thr (0)	Trp (-2)	Tyr (-1)	Val (-1)
GLY	Ala (0)	Arg (0)	Asn (0)	Asp (0)	Cys (-1)	Gln (0)	Glu (-1)		His (-1)	Ile (-3)	Leu (-3)	Lys (0)	Met (-2)	Phe (-3)	Pro (0)	Ser (0)	Thr (-1)	Trp (-2)	Tyr (-2)	Val (-2)
HIS	Ala (-1)	Arg (0)	Asn (0)	Asp (0)	Cys (0)	Gln (0)	Glu (0)	Gly (-1)		Ile (-2)	Leu (-1)	Lys (0)	Met (-1)	Phe (-1)	Pro (-1)	Ser (0)	Thr (0)	Trp (0)	Tyr (1)	Val (-1)
ILE	Ala (0)	Arg (-1)	Asn (-2)	Asp (-3)	Cys (0)	Gln (-2)	Glu (-2)	Gly (-3)	His (-2)		Leu (2)	Lys (-1)	Met (1)	Phe (0)	Pro (-2)	Ser (-2)	Thr (0)	Trp (-1)	Tyr (0)	Val (2)
LEU	Ala (0)	Arg (-1)	Asn (-2)	Asp (-3)	Cys (0)	Gln (-1)	Glu (-2)	Gly (-3)	His (-1)	Ile (2)		Lys (-1)	Met (2)	Phe (1)	Pro (-2)	Ser (-2)	Thr (-1)	Trp (0)	Tyr (0)	Val (1)
LYS	Ala (0)	Arg (1)	Asn (0)	Asp (0)	Cys (-1)	Gln (0)	Glu (0)	Gly (0)	His (0)	Ile (-1)	Leu (-1)		Met (0)	Phe (-2)	Pro (0)	Ser (0)	Thr (0)	Trp (-1)	Tyr (-1)	Val (-1)
MET	Ala (0)	Arg (0)	Asn (-1)	Asp (-2)	Cys (0)	Gln (0)	Glu (-1)	Gly (-2)	His (-1)	Ile (1)	Leu (2)	Lys (0)		Phe (1)	Pro (-1)	Ser (-1)	Thr (0)	Trp (0)	Tyr (0)	Val (0)
PHE	Ala (-1)	Arg (-2)	Asn (-2)	Asp (-3)	Cys (0)	Gln (-2)	Glu (-2)	Gly (-3)	His (-1)	Ile (0)	Leu (1)	Lys (-2)	Met (1)		Pro (-2)	Ser (-2)	Thr (-1)	Trp (1)	Tyr (2)	Val (0)
PRO	Ala (0)	Arg (0)	Asn (-1)	Asp (0)	Cys (-2)	Gln (0)	Glu (0)	Gly (0)	His (-1)	Ile (-2)	Leu (-2)	Lys (0)	Met (-1)	Phe (-2)		Ser (0)	Thr (0)	Trp (-2)	Tyr (-1)	Val (-1)
SER	Ala (0)	Arg (0)	Asn (0)	Asp (0)	Cys (0)	Gln (0)	Glu (0)	Gly (0)	His (0)	Ile (-2)	Leu (-2)	Lys (0)	Met (-1)	Phe (-2)	Pro (0)		Thr (0)	Trp (-2)	Tyr (-1)	Val (-1)
THR	Ala (0)	Arg (0)	Asn (0)	Asp (0)	Cys (0)	Gln (0)	Glu (0)	Gly (-1)	His (0)	Ile (0)	Leu (-1)	Lys (0)	Met (0)	Phe (-1)	Pro (0)	Ser (0)		Trp (-2)	Tyr (-1)	Val (0)
TRP	Ala (-2)	Arg (-1)	Asn (-2)	Asp (-2)	Cys (-1)	Gln (-2)	Glu (-2)	Gly (-2)	His (0)	Ile (-1)	Leu (0)	Lys (-1)	Met (0)	Phe (1)	Pro (-2)	Ser (-2)	Thr (-2)		Tyr (2)	Val (-1)
TYR	Ala (-1)	Arg (-1)	Asn (-1)	Asp (-2)	Cys (0)	Gln (-1)	Glu (-1)	Gly (-2)	His (1)	Ile (0)	Leu (0)	Lys (-1)	Met (0)	Phe (2)	Pro (-1)	Ser (-1)	Thr (-1)	Trp (2)		Val (0)
VAL	Ala (0)	Arg (-1)	Asn (-2)	Asp (-2)	Cys (0)	Gln (-1)	Glu (-1)	Gly (-2)	His (-1)	Ile (2)	Leu (1)	Lys (-1)	Met (0)	Phe (0)	Pro (-1)	Ser (-1)	Thr (0)	Trp (-1)	Tyr (0)	

IID Membrane Proteins

	ALA	ARG	ASN	ASP	CYS	GLN	GLU	GLY	HIS	ILE	LEU	LYS	MET	PHE	PRO	SER	THR	TRP	TYR	VAL
ALA		Arg (-1)	Asn (-1)	Asp (0)	Cys (0)	Gln (-2)	Glu (0)	Gly (1)	His (-3)	Ile (0)	Leu (-2)	Lys (-2)	Met (-1)	Phe (-2)	Pro (0)	Ser (2)	Thr (1)	Trp (-4)	Tyr (-3)	Val (0)
ARG	Ala (-1)		Asn (2)	Asp (1)	Cys (-1)	Gln (6)	Glu (2)	Gly (0)	His (5)	Ile (-3)	Leu (-3)	Lys (9)	Met (0)	Phe (-4)	Pro (-3)	Ser (-1)	Thr (-1)	Trp (5)	Tyr (-1)	Val (-2)
ASN	Ala (-1)	Arg (2)		Asp (6)	Cys (-1)	Gln (3)	Glu (1)	Gly (-2)	His (3)	Ile (-3)	Leu (-4)	Lys (5)	Met (-2)	Phe (-4)	Pro (-2)	Ser (2)	Thr (1)	Trp (-3)	Tyr (-1)	Val (-3)
ASP	Ala (0)	Arg (1)	Asn (6)		Cys (-3)	Gln (2)	Glu (8)	Gly (2)	His (3)	Ile (-3)	Leu (-5)	Lys (3)	Met (-3)	Phe (-6)	Pro (-2)	Ser (0)	Thr (0)	Trp (-4)	Tyr (-2)	Val (-3)
CYS	Ala (0)	Arg (-1)	Asn (-1)	Asp (-3)		Gln (-3)	Glu (-3)	Gly (-1)	His (-1)	Ile (-1)	Leu (-1)	Lys (-3)	Met (-1)	Phe (1)	Pro (-3)	Ser (2)	Thr (0)	Trp (1)	Tyr (3)	Val (0)
GLN	Ala (-2)	Arg (5)	Asn (3)	Asp (2)	Cys (-3)		Glu (7)	Gly (-2)	His (7)	Ile (-4)	Leu (-2)	Lys (6)	Met (-2)	Phe (-4)	Pro (0)	Ser (-1)	Thr (-2)	Trp (0)	Tyr (0)	Val (-4)
GLU	Ala (0)	Arg (2)	Asn (1)	Asp (8)	Cys (-3)	Gln (7)		Gly (3)	His (2)	Ile (-4)	Leu (-5)	Lys (1)	Met (-3)	Phe (-6)	Pro (-3)	Ser (0)	Thr (-1)	Trp (-3)	Tyr (-5)	Val (-2)
GLY	Ala (1)	Arg (0)	Asn (-2)	Asp (3)	Cys (-1)	Gln (-1)	Glu (3)		His (-3)	Ile (-2)	Leu (-4)	Lys (-1)	Met (-3)	Phe (-5)	Pro (-1)	Ser (1)	Thr (0)	Trp (-2)	Tyr (-5)	Val (-1)
HIS	Ala (-3)	Arg (5)	Asn (3)	Asp (3)	Cys (-1)	Gln (7)	Glu (2)	Gly (-3)		Ile (-4)	Leu (-4)	Lys (4)	Met (-3)	Phe (-3)	Pro (-4)	Ser (-2)	Thr (-2)	Trp (-1)	Tyr (6)	Val (-4)
ILE	Ala (0)	Arg (-3)	Asn (-3)	Asp (-3)	Cys (-1)	Gln (-4)	Glu (-4)	Gly (-2)	His (-4)		Leu (1)	Lys (-4)	Met (1)	Phe (-1)	Pro (-3)	Ser (-1)	Thr (0)	Trp (-3)	Tyr (-4)	Val (2)
LEU	Ala (-2)	Arg (-3)	Asn (-4)	Asp (-5)	Cys (-1)	Gln (-2)	Glu (-5)	Gly (-4)	His (-4)	Ile (1)		Lys (-4)	Met (1)	Phe (1)	Pro (-1)	Ser (-2)	Thr (-1)	Trp (-2)	Tyr (-3)	Val (0)
LYS	Ala (-2)	Arg (9)	Asn (5)	Asp (3)	Cys (-3)	Gln (6)	Glu (1)	Gly (-1)	His (4)	Ile (-4)	Leu (-4)		Met (-1)	Phe (-5)	Pro (-4)	Ser (-1)	Thr (-2)	Trp (3)	Tyr (1)	Val (-4)
MET	Ala (-1)	Arg (0)	Asn (-2)	Asp (-3)	Cys (-1)	Gln (-2)	Glu (-3)	Gly (-3)	His (-3)	Ile (1)	Leu (1)	Lys (-1)		Phe (0)	Pro (-3)	Ser (-2)	Thr (0)	Trp (-2)	Tyr (-3)	Val (1)
PHE	Ala (-2)	Arg (-4)	Asn (-4)	Asp (-6)	Cys (1)	Gln (-4)	Glu (-6)	Gly (-5)	His (-3)	Ile (-1)	Leu (1)	Lys (-5)	Met (0)		Pro (-4)	Ser (-1)	Thr (-2)	Trp (-3)	Tyr (2)	Val (-1)
PRO	Ala (0)	Arg (-3)	Asn (-2)	Asp (-2)	Cys (-4)	Gln (0)	Glu (-3)	Gly (-1)	His (-4)	Ile (-3)	Leu (-1)	Lys (-4)	Met (-3)	Phe (-4)		Ser (-1)	Thr (-1)	Trp (-6)	Tyr (-5)	Val (-3)
SER	Ala (2)	Arg (-1)	Asn (2)	Asp (0)	Cys (1)	Gln (-1)	Glu (0)	Gly (1)	His (-2)	Ile (-1)	Leu (-2)	Lys (-1)	Met (-2)	Phe (-1)	Pro (-1)		Thr (2)	Trp (-3)	Tyr (0)	Val (-1)
THR	Ala (1)	Arg (-1)	Asn (1)	Asp (0)	Cys (0)	Gln (-2)	Glu (-1)	Gly (0)	His (-2)	Ile (0)	Leu (-1)	Lys (-2)	Met (0)	Phe (-2)	Pro (-1)	Ser (2)		Trp (-4)	Tyr (-3)	Val (0)
TRP	Ala (-3)	Arg (5)	Asn (-3)	Asp (-4)	Cys (1)	Gln (0)	Glu (-3)	Gly (-2)	His (-1)	Ile (-3)	Leu (-2)	Lys (3)	Met (-2)	Phe (-3)	Pro (-6)	Ser (-3)	Thr (-4)		Tyr (-3)	Val (-2)
TYR	Ala (-3)	Arg (-1)	Asn (-1)	Asp (-2)	Cys (3)	Gln (0)	Glu (-5)	Gly (-5)	His (6)	Ile (-4)	Leu (-2)	Lys (1)	Met (-3)	Phe (2)	Pro (-5)	Ser (0)	Thr (-3)	Trp (-3)		Val (-4)
VAL	Ala (0)	Arg (-2)	Asn (-3)	Asp (-3)	Cys (0)	Gln (-4)	Glu (-2)	Gly (-1)	His (-4)	Ile (2)	Leu (0)	Lys (-4)	Met (1)	Phe (-1)	Pro (-3)	Ser (-1)	Thr (0)	Trp (-2)	Tyr (-4)	

Bioinformatics for Geneticists. Edited by Michael R. Barnes and Ian C. Gray

ISBNs: 0-470-84393-4 (HB); 0-470-84394-2 (PB)

GLOSSARY OF TERMS AND ABBREVIATIONS

BLAST Basic Local Alignment Search Tool — a tool for identifying sequences in a database that match a given query sequence. Statistical analysis is applied to judge the significance of each match. Matching sequences may be homologous to, or related to, the query sequence. There are several versions of BLAST:

- **BLASTP** compares an amino acid query sequence against a protein sequence database
- **BLASTN** compares a nucleotide query sequence against a nucleotide sequence database
- **BLASTX** compares a nucleotide query sequence translated in all reading frames against a protein sequence database
- **TBLASTN** compares a protein query sequence against a nucleotide sequence database dynamically translated in all reading frames
- **TBLASTX** compares the six-frame translations of a nucleotide query sequence against the six-frame translations of a nucleotide sequence database.

BLAT BLAST-Like Alignment Tool. BLAT might superficially appear to be like BLAST, also being a tool for detecting subsequences that match a given query sequence, however BLAT and BLAST have a number of differences. BLAT was developed at the UCSC; it searches the human genome by keeping an index of the entire genome in memory. The index consists of all non-overlapping 11-mers except for repeat sequences. A BLAT search of the human genome will quickly find sequences of 95% and greater similarity of length 40 bases or more. It may miss more divergent or shorter sequence alignments (see the UCSC FAQ for more details on this tool — http://genome.ucsc.edu/FAQ.html).

CDS Coding sequence.

Contig Map A map depicting the relative order of overlapping (contiguous) clones representing a complete genomic or chromosomal segment.

DAS (Distributed Annotation System) DAS is a protocol for browsing and sharing genome sequence annotations across the Internet, allowing users to search and compare annotations from several sources. Ensembl provides a DAS reference server giving access to a wide range of specialist annotations of the human genome (see http://www.ensembl.org/das/ for more detail).

Data Mining The ability to query very large databases in order to satisfy a hypothesis ("top-down" data mining); or to interrogate a database in order to generate new hypotheses based on rigorous statistical correlations ("bottom-up" data mining).

Domain (protein) A region of special biological interest within a single protein sequence. However, a domain may also be defined as a region within the three-dimensional structure of a protein that may encompass regions of several distinct protein sequences that accomplishes a specific function. A domain class is a group of domains that share a common set of well-defined properties or characteristics.

Electronic PCR (ePCR) An electronic process analogous to lab based PCR. Two primers are used to map a sequence feature (e.g. a SNP). To validate the position both primers must map in the same vicinity spanning a defined distance, effectively producing an electronic PCR product.

Expressed Sequence Tag (EST) A short sequence read from an expressed gene derived from a cDNA library. Databases storing large numbers of ESTs can be used to gauge the relative abundance of different transcripts in cDNA libraries and the tissues from which they are derived. An EST can also act as a physical tag for the identification, cloning and full length sequencing of the corresponding cDNA or gene.

FASTA format FASTA format, originally devised for Lipman & Pearson's FASTA (Fast-All) sequence alignment algorithm, is one of the simplest and most widely accepted formats for sequences, taking the form of a simple header preceded by a "$>$" sign and sequence on the following line, e.g.

$>$sequence_id
gataggctgagcgatgcgatgctagctagctagc

Golden Path The golden path is a term applied to the first and subsequent assemblies of the human genome.

Hidden Markov model (HMM) A joint statistical model for an ordered sequence of variables. The result of stochastically perturbing the variables in a Markov chain (the original variables are thus "hidden"), where the Markov chain has discrete variables which select the "state" of the HMM at each step. The perturbed values can be continuous and are the "outputs" of the HMM. A Hidden Markov Model is equivalently a coupled mixture model where the joint distribution over states is a Markov chain. Hidden Markov models are valuable in bioinformatics because they allow a search or alignment algorithm to be trained using unaligned or unweighted input sequences; and because they allow position-dependent scoring parameters such as gap penalties, thus more accurately modelling the consequences of evolutionary events on sequence families.

Homology (strict) Two or more biological species, systems or molecules that share a common evolutionary ancestor. (general) Two or more gene or protein sequences that share a significant degree of similarity, typically measured by the amount of identity (in the case of DNA), or conservative replacements (in the case of protein), that they register along their lengths. Sequence "homology" searches are typically performed with a query DNA or protein sequence to identify known genes or gene products that share significant similarity and hence might inform on the ancestry, heritage and possible function of the query gene.

***in silico* (biology)** (Lit. computer mediated). The use of computers to simulate, process, or analyse a biological experiment.

NCBI National Center for Biotechnology Information, Washington, D.C., USA.

Open reading frame (ORF) Any stretch of DNA that potentially encodes a protein. Open reading frames start with a start codon, and end with a termination codon. No termination codons may be present internally. The identification of an ORF is the first indication that a segment of DNA may be part of a functional gene.

Ortholog/Paralog Paralogs are genes related by duplication within a genome. Orthologs retain the same function in the course of evolution, whereas paralogs evolve new functions, even if these are related to the original one.

PERL PERL is the short form acronym for Practical Extraction and Report Language. Perl is relatively straightforward up to a certain level — this has encouraged its development as the primary language of biological computing.

Relational Database A database that follows E. F. Coddís 11 rules, a series of mathematical and logical steps for the organization and systemization of data into a software system that allows easy retrieval, updating, and expansion. A relational database management system (RDBMS) stores data in a database consisting of one or more tables of rows and columns. The rows correspond to a record (tuple); the columns correspond to attributes (fields) in the record. RDBMSs use Structured Query Language (SQL) for data definition, data management, and data access and retrieval. Relational and object-relational databases are used extensively in bioinformatics to store sequence and other biological data.

Secondary structure (protein) The organization of the peptide backbone of a protein that occurs as a result of hydrogen bonds e.g. alpha helix, Beta pleated sheet.

Sequence Tagged Site (STS) A unique sequence from a known chromosomal location that can be amplified by PCR. STSs act as physical markers for genomic mapping and cloning.

Single Nucleotide polymorphism (SNP) A DNA sequence variation resulting from substitution of one nucleotide for another.

SQL Structured Query Language. A type of programming language used to construct database queries and perform updates and other maintenance of relational databases, SQL is not a fully-fledged language that can create standalone applications, but it is powerful enough to create interactive routines in other database programs.

Substitution matrix A model of protein evolution at the sequence level resulting in the development of a set of widely used substitution matrices. These are frequently called Dayhoff, MDM (Mutation Data Matrix), BLOSUM or PAM (Percent Accepted Mutation) matrices. They are derived from global alignments of closely related sequences. Matrices for greater evolutionary distances are extrapolated from those for lesser ones.

Tertiary structure (protein) Folding of a protein chain via interactions of its side-chain molecules including formation of disulphide bonds between cysteine residues.

UCSC University of California Santa Cruz

UTR Untranslated region. The non coding region of an mRNA transcript flanking either side of the open reading frame.

Bioinformatics for Geneticists. Edited by Michael R. Barnes and Ian C. Gray

ISBNs: 0-470-84393-4 (HB); 0-470-84394-2 (PB)

INDEX

Note: page numbers in *italics* refer to figures and tables