

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 |

■ 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 (2) | | 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) | Leu (3) | 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) | Pro (1) | 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) | Thr (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 (-2) | Leu (-2) | Lys (-1) | Met (-2) | Phe (-3) | Pro (0) | Ser (0) | Thr (0) | Trp (-2) | Tyr (-2) | Val (-2) |
| 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 (-1) | 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) | Phe (-2) | 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) | Pro (0) | 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) | Ser (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) | Thr (-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 (-1) | Pro (-1) | Ser (0) | Thr (0) | Trp (-2) | Tyr (-1) | Val (-2) |
| ASP | Ala (-1) | Arg (0) | Asn (1) | | Cys (-2) | Gln (-2) | 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) | Pro (0) | 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) | Ser (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) | Thr (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) | Trp (-2) | Tyr (-1) | 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) | Tyr (2) | Val (-1) |
| 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) | Val (0) |

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 sequences.

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
gataggctgagcgatgcatgctagctagctagc
```

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. Codd's 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.

INDEX

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

- AAUAAA polyadenylation signal mutations 260
- AC113611 80
- accession numbers 35
 - primary 31, 32
- ACE/ID polymorphism 48
- actin promoter model 283
- affected sib-pairs (ASPs) 219
- Alagille syndrome 267–8
- alanine 298, 299, 300
- ALLASS program 231
- alleles
 - cytosine 262
 - differences 274
 - frequency 45–6, 222, 311
 - genetic disease 42, 250, 251
 - identical by descent (IBD) 219, 222–3, 225
 - identical by state (IBS) 219
 - minor 311
 - risk 167
 - sharing 219
 - transmitted 225
- ALU repeat sequences 275
- Alzheimer's disease
 - ApoE* 166
 - late-onset 9
- AMC (Academic Medical Centre) tag-to-gene mapping 327, 334, 336
- AMIGO browser 136, 137
- amino acids 266, 291–314
 - aliphatic side chains 298
 - amphipathic 298
 - aromatic 298–9
 - behaviour 292–6
 - chemical property classification 296–8
 - classification 296–8
 - environment defining 266, 267
 - function 299–311
 - hydrophobic 297, 298–9
 - mutations 311–13
 - key in evolution 312–13
 - matrices 296, 297
 - physical property classification 296–8
 - polar 299
 - polymorphisms 266–7, 268
 - properties 298–9
 - protein structure 294
 - side-chain size 297
 - single nucleotide polymorphisms 311–12
 - site-directed mutagenesis 312
 - small 299
 - stacking interactions 298–9
 - structural property classification 296–8
 - structure 299–311
 - subsets 297
 - substitution 294, 299–311
 - matrices 296, 379, 380–3
 - tools 316
- analysis of variance (ANOVA) 236
- ANALYZE program 225
- angiotensin converting enzyme (ACE) 48
- animal models 15, 16
- annealing temperature (TM) 207–8, 211
- anticipation 47
 - triplet repeat 47–8
- APOE* gene 9
- ApoE* gene 166
- Arachne assembly 129
- Arg148Cys missense mutation 267
- arginine 299, 303–4, 305
- ARLEQUIN program 170, 175, 231–3, 234, 235
- asparagine 299, 306
- aspartate 299, 305–6
- aspartylprotease genes 78
- ASPEX program 220
- Assay by Design™ Genomic Assay Service 208
- association analysis 223–4
- association studies 10–11
 - markers 11–12
- asthma 10
- AU-rich elements (AREs) 263

- BACE* gene 76–8
 accession numbers 77–8
 gene portal inspection 79–80
 nomenclature 78
BACE mRNA 77, 78
BACE2 gene
 detailed view 195, 196
 linkage disequilibrium 201
 promoter region repeat 196–7
Bacillus stearothermophilus 312
 bacterial artificial chromosome (BAC) 12
 clone-based genome map 96
 clones 152
 FISH-mapped 325
 mouse clone sequencing 129
 physical maps 151–2
 mouse genome 126–7
 rat library 128
 rat physical map 128
 sequence overlaps 97–8
 BioCarta graphical biochemical pathway maps 325
 biochemical pathway dissection 15–16
 bioinformatics
 applications 5
 role 6
 biological information on internet 23
 biological sciences, web resources 24
 biological sequence databases 31–6
 primary 31–3
 secondary 33–4
 nucleic acids 34–5
 BioMedNet platform 25
 bipolar disorder, candidate gene identification 190–5
 BLAST database 7, 23, 31, 180
BACE gene searches 77
 comprehensive search 88
 against Ensembl 88–9
 Ensembl use 104, 106
 LocusLink/RefSeq 35
 matching 104
 mouse assemblies 129
 nucleotide searches 76
 protein interaction networks 354
 proteome identification 349
 searching 57, 60, 264
 similarity search 205
 BLASTN 99
 BLASTP 101
 BLASTZ 266
 BLAT 100, 266
 data retrieval 107
 mouse 74, 83, 84, 86
 nucleotide searches 76
 sequence search tool 184
 BLOCK program 220
 BLOSUM matrices 296
 Bonferroni correction 175
 boolean searching 26, 27–8
BRCA1 gene 9, 44
 breast cancer 9
 susceptibility gene 44
Caenorhabditis elegans 15
 cancer
 DNA sequence changes 320
 gene overexpression 336, 337
 gene silencing 336
 genomic aberrations 50
 Mitelman Map 50, 62–3, 325
 Cancer Genome Anatomy Project (CGAP) 50, 51, 63, 324–5
 Digital Gene Expression Displayer 324, 325
 Gene Library Summarizer 324, 325
 Genetic Annotation Initiative (GAI) 65–6
 Library Finder Tool 324
 SAGEmap tag-to-gene mapping 327, 328
 cardiovascular disease 9
 mouse mutagenesis projects 135
 case–control cohorts 166, 167
 case–control studies 10
 CATG sequence 337
 errors 328–9
 causal variants 167
Cd36 17
 cDNA 266, 323
 3'-end of clones 326, 327–8
 5'-end of clones 327
 CGAP resource 324, 325
 libraries 374
 Celera Genomics (CG) 96
 draft genome assembly 96–7
 human genome assembly 96–7, 98
 marker inconsistencies 154
 CEPH families 146, 147
CFTR gene 9, 44
 CGAP *see* Cancer Genome Anatomy Project
 Charcot–Marie–Tooth disease type 1A (CMT1A) 49
 chloroplasts 292–3
 chromatin loop 275
 chromosomal loops 275
 chromosome(s)
 aberrations
 in cancer 50, 62–3, 320, 325
 gross 49–50

- abnormalities 48–9
 - gene expression 336, 338
 - recurrent duplications/deletions 49
 - sequencing in mouse 129
- chromosome 1p loss 320
- chromosome 2p 337
- chromosome 3 340
- chromosome 11 338
 - radiation hybrid map 338, 339
- chromosome 17p11.2 49
- chromosome 21
 - hard-to-clone DNA 153
 - human 170, 172
 - physical separation of two copies 170
- chromosome 22q11.2 49, 50
- cis*-regulatory elements 260
- CLUMP program 173, 174–5
- ClustalW 31
- Clusters of Orthologous Groups (COGs) of proteins 101
- coding sequence (CDS) 34
- codons, initiator 261, 262
- common disease/common variant (cd/cv) hypothesis 251
- compartmentalized shotgun assembler 97
- COMPEL database 278
- complex disease 9, 10
 - environmental factors 373
 - genes 251, 373
 - linkage peak 180–1, 182
 - magnitude of effect 251
 - processes 15
 - rodents 137
 - splicing abnormalities 259
- complex trait holistic analysis 375–6
- Composite Interval Mapping (CIM) 236
- composite likelihood methods 231
- CONSED program 205, 206
- CONSENSUS 277, 285
- Contig Explorer (iCE) 127
- contigs
 - physical maps 151–2
 - sequences 108
 - initial 96
- CoreSearch 277, 285
- CpG islands 108, 109
- Crohn's disease 9
 - NOD2* gene 166, 172
- crosses, experimental 236–9
- cysteine 292, 299, 303, 308–9
- cystic fibrosis 9, 44, 280
- cytogenetic maps 148–9
- cytogenetic studies, Mitelman map 62, 63
- cytosine allele 262
- cytosol 292–3
- Danio rerio* 15
- data
 - accessing 7
 - curation 111
 - functional 15
 - indexing 23
 - integration for rat/mouse bioinformatics 121–2
 - management 6–8
 - mining 6–8, 22
 - resources 6
 - storage, retrieval and handling systems 4
 - sub-division of biological on internet 23
- database(s) 4, 6
 - biological sequence 31–6
 - comprehensive searching 88–90
 - controlled vocabulary 28
 - gene nomenclature 28
 - genetic 7
 - genetic marker 60–1
 - genetic variation 50, 51
 - insertion/deletion polymorphisms 49
 - locus-specific 30
 - microsatellite 60–1
 - model organism for rat/mouse 122–4
 - mouse mutagenesis 135
 - mutations 46
 - non-nuclear/somatic 61–3
 - proteins 36, 73
 - public 7
 - primers 208
 - sequence variation 58
 - unindexed 25
 - VNTR 48
 - see also individual named databases*
- Database of Interacting Proteins (DIP) 366
 - visualization tool 356
- Database of Transcribed Sequences (DoTS) 109
- dbEST 84
 - BACE* gene searches 77
 - CGAP subset 324
 - POLYBAYES SNP discovery tool 206
- dbSNP 7, 8, 12, 35–6, 46, 51
 - BACE* gene searches 77
 - candidate SNPs 55–6, 197
 - data submission 52–3, 55
 - export of list of SNPs 199
 - flanking sequences 209
 - HGVbase relationship 57
 - mouse 133

- dbSNP (*continued*)
 primer designs 208
 searching 52, 53, 54
 sequence variations 204
 short insertion/deletion polymorphisms 49
 tools 64
- dbSTS 35, 60
- development defects, mouse mutagenesis
 projects 135
- diabetes 10
 insulin gene 176
 PPAR γ association 14
- disease
 gene association with phenotype 176
 monogenic 181
 post-translational modification abnormalities
 295–6
 rodent models 137
 single nucleotide polymorphisms causing
 311–12
see also complex disease; genetic disease;
 Mendelian disorders
- disease-susceptibility gene 14–15
- Distributed Annotation System 113–14
- disulphide bonds 308
- DMAP program 231
- DNA 4
 amplification 209
 chips 207
 chromatin-associated 275
 chromosomal 275
 deletion 40
 elements 276
 genomic 275
 amplification 204
 hard-to-clone 153
 insertion 40
 internet searching 30–1
 mitochondrial 61, 62
 nucleosomal 275
 physical separation of two copies 170
 polymorphisms 252
 regulatory 276
 repetitive 275
 sequence changes 320
 sequencing methods 204
 variants 15
see also cDNA
- DNA Database of Japan (DDBJ) 31
- DNA mapping panels, mouse 124
- DNA markers, rat 123
- DNA microarrays 323–324
 technology 321–323
- domain–domain interactions 367
- Down syndrome cell adhesion molecule
 (DSCAM) gene 258, 374
- Drosophila melanogaster* 15, 258, 374
- drug target identification 10
- DUP25 interstitial duplication 49
- e-PCR 43, 153, 187
- EcoCyc database 359
- EHPLUS program 170, 173, 175, 226, 227
 haplotype-based association testing 229,
 230
 haplotype frequency estimation 227, 228
 linkage disequilibrium 230
- ELAV family 263
- EMBL database 31
 mouse sequences 121
 proteome identification 349
 rat sequences 121
 USAGE application 334
- enhancers 275
- Ensembl 7, 8, 22
 BLAST use 88–9, 104, 106
 candidate gene selection 190–5
 characterization of genetic/physical locus
 199–201
 data retrieval 106
 definition of known/novel genes across
 genomic region 188–90
 duplication detection in genomic assemblies
 186
 genome sequence annotation 104–6
 genome viewing applications 61, 63, 79,
 157
 genomic features 104–5
 identification of known and novel markers
 195–9
 marker mapping to human genome 182
 marker panel design 199–201
 motif recognition 86–7
 mouse genome assembly 74, 126, 127
 novel gene analysis 81, 86–8
 promoter analysis 257
 pros and cons 183
 proteins 73
 regulatory element analysis 257
 sequence characterization 173
 unspliced ESTs 374
- Entrez Map View 155–6, 157
- Entrez-PubMed 23, 30, 31
- ENU mutagenesis projects 134–5
- environmental factors 9, 373
- enzymes 294
- ePCR program 99
- eSAGE program 334

- Escherichia coli*
 gene fusion events 359
 IDDP prediction method 364
 inference of protein interactions 365
 ETDT program 225
 European Bioinformatics Institute (EBI) 102
 Radiation Hybrid Database (RHdb) 125
 European Mouse Mutant Resource (EMMA)
 134
 Exofish program 107
 exonic splicing enhancer (ESE) regions 176,
 256, 258, 259, 260
 exonic splicing silencers (ESS) 256, 258, 259,
 260
 exons
 recognition 276
 silent variants 252
 expectation–maximization (EM)
 algorithm 226, 236, 266
 maximum likelihood estimate (MLE) 170
 Expert Protein Analysis System 36
 expressed sequence tags (ESTs) 10
 accession numbers 35
 analysis of clusters 204
 BACE gene 77
 CATG-sequencing errors 328–9
 cDNA sequence resources 7
 CGAP–GAI database 65–6
 clones 73
 correctly spliced mRNA transcript 189
 G-cap selected 266
 gene portal inspection 79–80
 libraries 14, 328
 matches 83
 mouse 73
 novel gene analysis 82–3, 84
 orthologous 73
 overlapping 72
 POLYBAYES SNP discovery tool 206
 rat 124
 secondary databases 35
 sequence tagged sites 153
 sequencing errors 328–9
 spliced 82, 190
 TBLASTN 84, 86
 UniGene
 clusters 75
 record 191–2
 unspliced 190, 374
 virtual mRNA 85
see also dbEST
- family-based cohorts 167
 FANTOM Consortium 130
 FAST 25
 FASTLINK program 218–19
 feature identifiers 32
 Fgenesh program 106
 fibroblast growth factor 294, 295
 Fisher's Exact test 235
 Flicker program 350–1
 fluorescent *in situ* hybridization (FISH)
 BAC mapping 325
 data 128
 mapping 149
 FlyBase 103
 FOXP2 gene 102
 coding sequence 107
 Genome Channel browser 109, 110, 111
 genomic region 105–7, 108, 109
 FPC program 127
 fragile X syndrome 47
 Frataxin 9
 Friedreich's ataxia 9, 47
Fugu rubripes 15
- gap sizes 154–5
 GASSOC program 225
 gbPAT 88
 records 33
 GDB database 46, 61, 157–9
 chromosome aberrations 49
 genetic maps 146
 insertion/deletion polymorphisms 49
 maps 154
 radiation hybrid maps 151
 text-based data mining 158
 gel image analysis 348–9, 350–1
 GEMS Launcher program 283, 284
 GenBank database 31, 36
 accession numbers 35
 author responsibility 32–3
 BAC clone 96
 clones 326, 327
 comprehensive search 88
Homo sapiens CAGH44 mRNA 102
 human mRNA 72–3
 mouse sequences 121
 PAC sequences 339
 patent subdivision 33
 rat sequences 121
- gene(s)
 aliases 78
 analysis of novel 81–4, 85, 86–8
 anatomy 256, 258–64
 annotation 71–2, 123
 tools 374
 associated 251
 candidate 14–15

- gene(s) (*continued*)
- genetic analysis 12, 13
 - identification 173–6
 - selection for analysis 14
 - characterization 15
 - clustering 284
 - co-expression 284
 - co-regulation 284
 - complex disease 251
 - complex trait holistic analysis 375–6
 - computational prediction 98
 - controlled nomenclature 28
 - definition 374
 - known/novel across genomic region 188–90
 - density in RIDGES 340
 - detection 74
 - disease contribution 373
 - function 136
 - fusion events 359, 360
 - modifier 251
 - name dictionaries 366
 - nomenclature 78, 123
 - standards 122
 - normal function 256
 - number in human genome 273–4, 374
 - ontology 136–7
 - overlapping 75
 - prediction 100–1
 - tools 374
 - promoter region anatomy 257
 - rat 123
 - redundant 251
 - regulation
 - networks 358–9
 - tools for functional analysis 255
 - regulatory elements 257
 - selection 175
 - candidate 190–5
 - sequence similarity 100–1
 - splicing 255, 258–9
 - structure prediction 100
 - transcripts 374
 - see also* mutations; open reading frames (ORFs)
- Gene and Position Predictor (GAPP) 123, 131
- gene expression
- analysis 320–1
 - control 274
 - genome-wide surveys 132, 136
 - genotype variation 132
 - informatics 320–1
 - patterns 15
 - regulation 258
 - SAGE measure 192
 - in silico* resources 14
 - single nucleotide polymorphisms 281
 - technologies for measurement 321–3
 - tumour tissues 320
 - see also* RIDGES
- Gene Expression Database (GXD) 123
- gene identifier (GI) numbers 32
- gene knock-out animals 15
- gene model 75
 - approximate 189
- gene neighbourhood method 359–61
- Gene Ontology Consortium 293
- Gene Ontology (GO) database 123, 136–7
 - annotation 103
- gene products
 - evidence cascade 72–5
 - normal function 256
- GeneChip™ technology 323
- gene–gene interactions 15
- GENEHUNTER program 219, 220, 226, 236
- GeneMap99 334–6
- GeneReviews 30
- Genethon human linkage map 146
- Genetic Annotation Initiative (GAI) 65–6
- genetic disease 4
 - alleles 42, 250, 251
 - associations 42
 - gene mapping 100
 - gene splicing mechanisms 258–9
 - mutations 250
 - phenotype 4
 - phenotypic variability 250
 - VNTR-mediated 47–8
- genetic distance 154
- genetic drift, random 169
- genetic load 41
- genetic maps 144, 145–8
 - accuracy 154
 - construction 145
 - draft sequence curation 152–3
 - haplotype 148
 - human 144, 145–8
 - integration 155, 156, 158
 - linkage 145
 - linkage disequilibrium 148
 - mouse 124–5
 - rat 127–8
 - SNP-based haplotype 148
 - UDB database 159
 - value 154
- genetic marker databases 60–1
- genetic study designs 8–12, 13
- genetic traits, complex 373–4

- genetic variation 250
 - data integration 42–3
 - see also* human genetic variation
- genetical genomics 17
- genetics 4
 - bioinformatics
 - applications 5
 - role 6
 - comparative 15–17
 - inverse 265, 266
- GENEWISE
 - exon structure prediction 104
 - sequence similarity 100–1
- GenMapDB 152
- Genomatix Promoter Resource (GPR) 283
- Genomatix sequence analysis tools 282
- genome
 - annotation 123
 - coordinates 78
 - cross-species comparisons 8
 - data 4
 - genetic variation data integration 42–3
 - manipulation techniques in mouse 133
 - maps 145
 - protein encoding 274
 - regulatory regions 101
 - viewer 79–80
 - see also* human genome
- Genome Biology* 22
- Genome Channel browser 108–9, 110, 111
 - FOXP2 gene 109, 110, 111
 - gene prediction 108–9
 - ratio analysis 200–1
- genome database *see* GDB database
- Genome Monitoring Table 95
- genome sequence
 - annotation 99–109, 110, 111, 112, 113–14
 - data curation 111
 - nucleotide level 99–101
 - process level 102–3
 - protein level 101–2, 111, 113
 - specific 103–9, 110, 111
 - assembly 96–9
 - resources for rat/mouse 128–31
 - splice site location prediction 259
 - websites 112–13
- genome-wide genetic/physical distance ratio 200
- genome-wide microarray projects 192
- genomic assemblies, duplication detection 185–8
- genomic control 167
- genomic DNA 154–5
- genomic fragments, random 153
- genomic instability 49
- genomic prediction 73–4
- genomic region, definition of known/novel
 - genes 188–90
- genomics 4, 15–17
 - comparative 131–2, 359
 - functional 135–7
 - genetical 17
- genotoxic stress 259
- genotype
 - expected frequency 175
 - linkage to phenotype 24
- genotyping, high-throughput methods 42
- Genscan 74
 - gene prediction 82, 100–1, 104
- Giemsa bands 200
- GigAssembler 97
- glutamate 299, 306
- glutamine 299, 306–7
- glycine 299, 309, 310
- glycosylation 295
- GNF Gene Expression Atlas Ratio 192, 193–5
- GOLD program 230–1
- Golden Path genome browser 7, 71, 72, 173
 - draft dataset 180
 - gene location 76–8
 - marker inconsistencies 154
 - missing genes 80–1
 - raw sequence data 76–7
 - STS marker positions 155
 - template for genetics 181
 - unified assembly 78
- golden triangle track 266
- Golgi apparatus 293
- Google 24, 25, 27
- guilt-by-association rule 366–7, 368
- haemoglobin 291
 - mutations 313
- haplotype map 197–8, 199–200
 - human chromosome 21 172
- haplotype tags 148, 197, 375
 - design tool 57
- haplotypes 12, 170–2
 - association testing 228–9
 - construction 170, 172, 175
 - frequency
 - determination 175, 375
 - differences 173
 - estimation 227–8
 - genetic maps 148, 172
 - length 170
 - linkage disequilibrium analysis 234

- haplotypes (*continued*)
 - marker sets 171
 - patterns 198
 - reconstruction 226–9
 - statistical analysis 170, 173
- HAPPY program 237
- Hardy–Weinberg equilibrium 175, 226
- header records 32
- Helicobacter pylori*
 - gene fusion events 360
 - inference of protein interactions 365
 - interaction network 355
- heterogenous stocks 17
- HEXB* gene 265
- HGMD database 46, 57–8
- HGVbase 36, 56–7
 - short insertion/deletion polymorphisms 49
 - SRS database 58
- high-performance liquid chromatography (HPLC) 350
- High Throughput Genomic Sequences (HTGS) 80–1
- HighWire 29–30
- histidine 298, 299, 302–3
- HIV-1 infection 280
- HMMER program 102, 296
- Homo sapiens* CAGH44 mRNA 102
- homology support 74
- HTRA3 80, 81
- HTRA4 80, 81
- htSNP program 171
- Human Gene Mutation Database *see* HGMD database
- human genetic variation 40–2
 - databases 50, 51
 - forms 43–50
 - mechanisms 43–50
 - quantity 41
- human genome 3
 - annotation 98–109, 110, 111, 112, 113–14
 - browsers 7, 12, 52
 - completion 98–9
 - of sequencing 144
 - data
 - interrogation 71–2
 - overload 374–5
 - draft sequence 95
 - genetic data integration 7
 - genetic maps 154
 - high-throughput technologies 103
 - locus 168
 - mapping 375
 - marker localization 184
 - misassembly rate 98
 - number of genes 273–4, 374
 - other vertebrate comparisons 74
 - QC 154
 - sequence
 - characterization 173
 - physical maps 153–4
 - sequencing 373
 - web-based tools 168
 - websites 112–13
- Human Genome Browser (HGB) *see* UCSC human genome browser
- Human Genome Variation database *see* HGVbase
- Human Proteomics Initiative (HPI), SwissProt sequences 73
- Human Transcriptome Map (HTM) 320–1, 336–7, 338, 339, 340–1
 - annotation 337, 339
 - construction 334–6
 - relational database 334–6
 - RIDGES 340
 - tags
 - antisense 337, 339
 - unreliable 337
- human variation 375
- Huntingtin 9
- Huntington's disease 9, 47
- hypermutable 42
- hypothesis construction 22
- Improbizer tool 266
- indexing of data 23
- Induced Mutant Resource, mouse 133–4
- Information Retrieval 365–6
- insertion/deletion (INDEL) polymorphisms 40, 48–9
 - dbSNP database 52
- insulin gene, diabetes type 1 176
- Interacting Domain Profile Pair (IDPP) 362–3, 364
- interactome 351
- Internal Ribosome Entry Site (IRES) 262
- International Human Genome Sequencing Consortium (IHGSC) 96–8
- International Protein Index (IPI) 73
 - analysis 88
 - ORFs 75
- internet 4, 6
 - biological data sub-division 23
 - heart 30
 - resources 22–37
 - search methods 22
 - see also* search engines
- InterPro database 102, 367

- InterProScan 31
- intronic sequence
 identification 266
 removal 279
- intronic splicing enhancers (ISE) 256, 258, 259, 260
- intronic splicing silencers (ISS) 256, 258, 259, 260
- introns 276
 variants 252
- inverse genetics 265, 266
- inverted repeat detection 265
- isochore boundary identification 200–1
- isoleucine 298, 300
- Jackson Lab Radiation Hybrid Map 124, 125
- Jagged 1 267
- journals, full-text 29–30
- KEGG graphical biochemical pathway maps 325
- keyword retrieval technique 364–5
- knowledge management 6
- Kozak sequence 261–2, 264
- KWOK 54
- Kyoto Encyclopedia of Genes and Genomes* 14
- laboratory information management systems (LIMS) 7
- lactate dehydrogenase 312, 313
- least squares regression 236
- leucine 298, 300–1
- likelihood methods 231
- likelihood ratio 219–20
- linkage analysis 9–10, 165–76, 218–23
 non-parametric 219–20, 222–3
 parametric 218–19
 recombination 9
 study population 166–7
- linkage disequilibrium 10, 11, 12, 41–2, 168–9, 229–35
 absolute 169
 definition 229
 genetic nature of region 180
 genotypes with unknown phase 233–4
 haplotype analysis 235
 human variation 375
 maps 148, 199–200, 375
 marker maps 197
 measures 168
 recombination frequency 200
 two polymorphisms 168–9
- linkage maps 145–8, 146–8
 genome-wide 145–6
- LINKAGE program 221, 225
- Linux 22
- literature
 digests 30
 mining 365–6
 search 6, 14, 28–9
 full-text journals 29–30
- locus
 characterization of genetic/physical 199–201
 defining 180–4
 genetic characterization 182
 genome sequence extraction 184–5
 human genome 168
 known/novel genes 189
 refinement 173–6
 sequence characterization 167–8
- locus-specific databases 57
- LocusLink 52, 64
- LocusLink/RefSeq (LLRS) 34–5, 36
- LOD (log of the odds) score 10
 GDB database 61
 linkage measurement 218
 locus definition 182
 maps 154, 156
 quantitative trait locus mapping 236
 quantitative trait NPL analysis 222
- LOKI program 220
- long interspersed elements (LINEs) 275
- lysine 299, 304–5
- lysosomes 293
- malate dehydrogenase 312, 313
- MALDI/TOF peptide mass fingerprinting 349
- Malecot isolation 231
- Mammalian Gene Collection 130–1, 324
- Map Manager QTX program 237–9, 240
- MapMaker 145
- MAPMAKER/EXP program 236
- MAPMAKER/QTL program 236, 237
- MAPMAKER/SIBS program 219–23, 224, 236
- mapping, simple interval 236
- MapQTL program 237
- maps
 construction of marker 197–8
 cytogenetic 148–9
 genome 145
 haplotype 148, 172
 integration 155, 156, 158
 linkage disequilibrium 148, 197, 199–200, 375
 master 156
see also genetic maps; physical maps; radiation hybrid maps

- MapViewer 146, 154
 genome viewer 157–8
 radiation hybrid maps 151
- MARFinder 275
- markers
 density 9, 11, 42
 single nucleotide polymorphisms 147
 disease association 175
 genomic sequence
 identification/extraction 184–5
 integrity checking 185–8
 identification of known and novel 195–9
 localization to human genome 184
 map
 construction 197–8
 order 188
 mapping to human genome 182
 multi-allelic 225
 order resolution 154–5, 159
 panel design 199–201
 sequence 184
 sets 171
 statistical analysis 173, 175
- Markov Chain method 220, 235, 236
- Marshfield genetic linkage map 146, 147
- Marshfield website 49
- MaskerAid 207, 209
- mass spectrometry 349–50
- MatInspector program 284
- Matrix-Assisted Laser Desorption/Ionisation (MALDI)/Time-of-Flight (TOF)-based peptide mass fingerprinting 349
- Matrix/Scaffold Attachment Regions (S/MARs) 275
- maximum likelihood estimate (MLE), expectation–maximization (EM) 170
- maximum likelihood methods 231
- mDNA 207
- MEDLINE 25, 28
 record analysis 366
- MEGABLAST 84
- Mendelian disorders 8, 9
 genes 10
 mutations affecting mRNA splicing 259
 regulatory regions 257
- MERLIN program 220, 225, 226
- metabolic pathways 356, 357, 358
- MetaCrawler 25
- metasearch engines 25
- methionine 298, 301
- MFOLD program 263
- MHC locus 128
- microarrays 15
 data 14
 tracks 193–5
 expression 15
 technology costs 194
- microsatellites 7, 40, 47
 data 197
 databases 60–1
 genome map 145
 identifying in sequence data 196–7
 mouse 133
 polymorphic sequences 279
 Sequence Tagged Sites 60
see also single tandem repeat (STR) markers
- minisatellites 41, 47
- MIPS database 358, 365, 366–7
- Mitelman Map of Chromosome Aberrations in Cancer 50, 62–3, 325
- mitochondria 292–3
- MITOMAP database 61–2
- Molecular Biology Database Collection 33
- Molecular Modelling Database of 3D structures (MMDB) 31
- Moment Method 231
- monogenic disorders *see* Mendelian disorders
- monogenic traits
 linkage analysis 172–3
 mapping 16
- Monte Carlo simulation test 174, 220, 236
 SAGE tags 331, 333
- motif recognition 86–7
- mouse
 bioinformatics 120–1
 data integration 121–2
 BLAT 74
 cDNA clone resources 130–1
 chemical mutagenesis 133
 chromosome sequencing 129
 consensus linkage map 124
 disease models 137
 DNA mapping panels 124
 functional genomics 135–7
 gene targeting 133
 genetic maps 124–5
 genetic variants 133
 genome 15, 16–17, 106
 assembly 74
 manipulation techniques 133
 sequence 126, 127
 genome sequencing
 comparative genomics 131–2
 completion 144
 initiative 128–9
 resources 129–30
 systematic genome-wide approaches 132

- Induced Mutant Resource 133–4
- microsatellites 133
- model organism databases 122–4
- monogenic trait mapping 16
- phenotypic variants 134
- physical maps 125–7
- radiation hybrid maps 124–5
- single nucleotide polymorphisms 133
- strains
 - characterization resources 134
 - nomenclature standards 122
 - transgenic 133
 - whole genome shotgun data 128
- Mouse Atlas and Gene Expression Database Project 103
- Mouse Genome Database 103, 122, 123, 124
 - DNA mapping panels 124
 - maps of curated orthologues 131
- Mouse Genome Informatics (MGI) 123
- Mouse Genome Resources, NCBI 130
- Mouse Genome Sequencing Project (MGS) 123
- Mouse Phenome Database (MPD) 134
- Mouse Tumor Biology Database (MTB) 123
- MouseBLAST 129–30
- mRNA
 - 3' UTR 75, 79, 261
 - regulatory elements 263
 - RNA instability signals 279
 - 5' UTR 75, 79, 261
 - IRES 262
 - regulatory elements 263
 - RNA instability signals 279
 - accession numbers 77–8
 - analysis tools/databases 263–4
 - correctly spliced transcript 189–90
 - extended 72–3, 84
 - folding 263
 - Homo sapiens* CAGH44 102
 - human 72–3
 - LocusLink/RefSeq (LLRS) 34–5
 - mature 258
 - mutations affecting splicing 259
 - regulatory 264
 - regulatory control of processing/translation 263
 - regulatory elements 264
 - secondary structure stability 262–3
 - splice variants 190
 - transcripts 75
 - polymorphisms 260–1
 - UTRs 252
 - virtual 72, 84, 85, 86
- mtDNA 61, 62
- multigenic disease *see* complex disease
- Multimapper program 236
- Multiple QTL Mapping (MQM) 236
- mutagenesis, site-directed 312
- Mutant Mouse Regional Resource Centres 134
- mutations 8, 40, 41, 291–2
 - AAUAAA polyadenylation signal 260
 - Alagille syndrome 267
 - amino acids 311–13
 - correlated 367
 - databases 46, 57–8, 59, 60
 - non-nuclear/somatic 61–3
 - disease 250
 - haemoglobin 291
 - human data 7
 - key in amino acid evolution 312–13
 - loss 44–5
 - mapping 4
 - matrices 296, 297
 - Mendelian 44
 - multiple founder 231
 - natural history 44–6
 - non-nuclear 61–3
 - phenotypic 204
 - point 50, 311
 - potentially deleterious 269
 - single nucleotide polymorphism relationship 43–6
 - somatic 50
 - databases 61–3
 - survival 274
 - tools for visualization 63–6
- MYCN oncogene 321
- myotonic dystrophy 47
- National Centre for Biotechnology Information
 - see* NCBI
- NC160 Cell Line Project 192, 193–5
- NCBI Assembly tool 106, 108
- NCBI browser 173
- NCBI database 60
 - genome viewer 79–80
 - Human BAC resource page 152
 - human genome sequencing 97–8
 - Mouse Genome Resources 130
 - novel gene analysis 81, 82–3, 88
 - radiation map 186
 - see also* LocusLink
- NCBI MapViewer tool 50, 60, 63, 107–8
 - characterization of genetic/physical locus 199–201
 - duplication detection in genomic assemblies 186, 187

- NCBI MapViewer tool (*continued*)
 FOXP2 gene genomic region 108, 109
 identification of known and novel markers
 195–9
 marker localization to human genome 184
 marker mapping to human genome 182
 marker panel design 199–201
 pros and cons 183
- NCBI RefSeq, human transcripts 73
 neighbourhood quality standard (NQS) 205
 neurological disorders, mouse models 135
 neutral theory of evolution 41
NF1 gene 201
NOD2 gene 9
 Crohn's disease 166, 172
 Notch receptor family 267
 nuclear matrix 275
 nucleic acid sequences, secondary database
 34–5
Nucleic Acids Research 33
 nucleosomes 275
 nucleotide sequences
 analysis of novel regulatory elements and
 motifs 264–6
 databases 30–1
- Oak Ridge National Laboratory (ORNL)
 Genome Channel 108–9, 110, 111
- On-line Mendelian Inheritance in Man (OMIM)
 6, 14, 30, 58, 60
 chromosome aberrations 49
 insertion/deletion polymorphisms 49
 literature search 31
 LocusLink 35
 MITOMAP linkage 62
 predicted protein–protein links 365
- Open Reading Frame EST sequencing 324
 open reading frames (ORFs) 72, 73, 256, 258
 identification 261
 International Protein Index 75
 novel protein 84, 85
 yeast two-hybrid system 351
- ORESTES 324
 organelles 292–3
 ovarian cancer 9
- palindromic motif detection 265
 pentanucleotide sequences 263
 peptide mass fingerprinting 349
 PFAM 36
 phage display 354
 phenotype
 evolutionary selection 311
 genetic interactions 103
 mouse variants 134
 similarity 137
 single nucleotide polymorphism effects 311
 variance 220
- phenotype–genotype correlation 8, 24
 phenylalanine 298, 301–2
 phosphorylation 295
 Phusion 129
 physical distance 154
 physical locus analysis 12, 13
 physical maps 144, 148–51
 accuracy 154
 bacterial artificial chromosome 151–2
 contig 151–2
 cytogenetic 148–9
 draft sequence curation 152–3
 FISH 149
 human 148–51
 human genome sequence 153–4
 integration 155, 156, 158
 mouse 125–7
 radiation hybrid 149–51
 rat 128
 UDB database 159
 value 154
 yeast artificial chromosome 151–2
- PIMRider program 355, 356
 PIPMAKER 131
 PIScout program 356
 PLABQTL program 236
 PMPLUS program 226, 227, 229
 Point Accepted Mutation (PAM) matrices 296
 polyA signal 326, 328
 polyA sites 108, 109
 polyA tail 326, 328, 337
 polyacrylamide gel electrophoresis (PAGE)
 348–9
 polyadenylation signals 260
 POLYBAYES 205–6, 207
 polymerase chain reaction (PCR)
 allele-specific 170
 assay design for SNPs 204, 208–10
 DNA amplification 209
 electronic 43, 153, 187
 primer design 207–8
 primers 211
 annealing temperature 207–8, 211
- polymerase II promoters 277, 278
 polymorphic markers 145
 polymorphic microsatellite sequences 279
 polymorphisms 41, 176
 amino acids 266–7, 268
 approximate localization 253
 AU-rich motif disruption in 3' UTR
 sequence 263

- candidate 251–2
- computational discovery 205
- DNA 252
- functional in genes/gene regulatory sequences 254–5
- identification of potentially functional 195, 196
- mapping 4
- mRNA transcript 260–1
- multiple 9
- non-synonymous coding 266–8
- predictive functional analysis 250–69
 - decision-making 252, 253, 256–7, 266
 - gene promoter characterization 257
 - novel regulatory element/motif analysis 264–6
 - putative splicing elements 259–60
 - regulatory region characterization 257
- regulatory regions 266
- in silico* predictions 268–9
- splice region 259
- see also* single nucleotide polymorphisms (SNP)
- Polyphred 205, 206, 207
- population stratification 166–7, 175
- PPAR γ , diabetes association 14
- Primer3 208, 209, 210, 211
 - pooled sequencing 211
- Probe-Set data 31
- proline 298, 299, 309–11
- promoter(s) 278, 281, 282, 283
 - modules 278
 - prediction tools 26–7
 - TF-sites 277–8
- promoter region anatomy 257
- promoter sequences 282–3
 - analysis 265
 - deletions/insertions 278
- PromoterInspector program 283
- PROSITE 36
- protein(s)
 - amino acids 266, 267
 - behaviour 292–6
 - structural 294
 - substitution matrices 379, 380–3
 - annotation 268
 - auto-activator bait 354–5
 - cellular location 266, 292–3
 - Clusters of Orthologous Groups (COGs) 101
 - complexes
 - co-localization 15
 - purification 348
 - correlated mutations 367
 - cytosolic 292
 - databases 36, 73
 - docking propositions 367
 - domain–domain interactions 367
 - duplication 293
 - environments 292–3
 - evolution 293–4
 - expression data analysis 350–1
 - expression networks 350
 - extracellular 292
 - amino acid substitution matrices 381
 - function 294
 - functional analysis 58
 - gene fusion events 359, 360
 - gene neighbourhood 359–61
 - gene regulation networks 358–9
 - guilt-by-association rule 366–7, 368
 - homologues 101, 293, 294
 - ID numbers 32
 - Interacting Domain Profile Pair (IDPP) 362–3, 364
 - interaction inferences 362–3
 - interaction networks 351–4
 - analysis 355–6
 - automated validation 364–5
 - building 354
 - cell pathways 356, 357, 358–9
 - centrality 368
 - combined methods 361–2
 - comparative genomics 359
 - exploitation 366–7
 - false-negative/positive interactions 354–5, 368
 - inference mechanism across organisms 362
 - lethality 368
 - literature mining 365–6
 - manual validation 365
 - prediction 359–63
 - prediction assessment/validation 363–6
 - prediction rule deduction 367–8
 - shape analysis 367–8
 - intracellular 382
 - matches 88
 - membrane 383
 - metabolic pathways 356, 357, 358
 - microarrays 354
 - orthologues 101, 102, 293, 294
 - pairs 101
 - paralogues 101, 102, 293, 294
 - phylogenetic profiles 361
 - post-translational modification 294–6
 - prey 355
 - prey gene sequencing 354

- protein(s) (*continued*)
 regulation 347
 sequence information 346
 signal transduction networks 358
 speciation 293
 structure 293
 websites 36
 yeast two-hybrid system 351–2, 353
- protein-binding sites 276
- protein kinase 294
- Protein Mutation Database (PMD) 58
- protein sequence databases, internet searching 30–1
- protein–protein interactions 15, 351–2, 353, 354, 356
 domain–domain interactions 367
 predicted links 365
- proteome
 cell pathways 356, 357, 358–9
 false-negative/-positive interactions 354–5
 identification 349–50
 interactions 351
 purification 348
 separation 348–9
see also protein(s), interaction networks
- proteome-wide characterization 350
- proteomic informatics 347
- proteomics
 classical 347–51
 definition 346
see also protein(s), interaction networks;
 proteome
- pseudogenes 264
- PSI-BLAST program 296
- PubMed 23, 24, 28–9
 full-text journals 30
 GenBank accession numbers 35
 literature search 31
- pufferfish genome 106
- Purkinje cell protein 4 193, 194
- pyruvate metabolic pathway 356, 357, 358
- QTDT program 223
- QTL Cartographer 236, 237
- quantitative trait loci 17
 multiple 236
 rat 123, 124
 rodents 137
- quantitative trait locus mapping 236–9
 heterogeneous stocks 237
 simple interval mapping 236, 239–40
- quantitative traits 220, 221
 non-parametric linkage analysis 222–3
- quotation mark use 26
- radiation hybrid code (RH-code) 334
- Radiation Hybrid Database (RHdb) 125
- radiation hybrid maps
 duplication detection in genomic assemblies 186–7
 human 149–51, 154
 Human Transcriptome Map 336
 mouse 124–5
 rat 128
 resolution 188
- RANTES gene 280
- rat
 bioinformatics 120–2
 disease models 137
 functional genomics 135–7
 genes 123
 genetic maps 127–8
 genome 15, 16–17
 genome sequencing
 comparative genomics 131–2
 initiative 131
 systematic genome-wide approaches 132
 genomic resources 127
 model organism databases 122–4
 monogenic trait mapping 16
 physical maps 128
 radiation hybrid maps 128
 strains 121, 122
- Rat Genome Database 122, 123–4
- RatMap 123, 127, 128, 131
- RealSNP 208
- recombinant inbred line 236
- recombination 180
 events 145
- reduced representation shotgun (RRS)
 sequencing 54
- reference sequence (RefSeq) 34, 35
- Reference SNPs (RefSNPs) 52
- regions of increased gene expression *see* RIDGES
- regulatory elements 257, 276
 bioinformatic characterization 15
 identification 283–4
- regulatory regions 276
- relational database management system (RDBMS) 335, 336
- repeat sequences 108, 109
- RepeatMasker 205, 207, 209, 211
- restriction fragment length polymorphisms (RFLPs) 11
- RIDGES 321, 338, 339–40, 341
 statistical evaluation 340
- RIKEN Exploration Research Group 130
- RING-finger 86–7

- risk allele 167
- RNA
 instability signals 279
 non-coding 101, 264
 processing 279
 regulatory 374
 see also mRNA
- RNA-binding proteins 263
- Rosetta-stone method 359
- RPCI BAC libraries 126–7
- RS number 35, 36
- RT-PCR data 14
- Saccharomyces cerevisiae* 15
 proteome 103
- Saccharomyces* Genome Database 103
- SAGE 14, 192, 219, 220, 320, 323
 analysis 321
 computational resources 333–4
 concatemer sequences 322, 325
 critical values 331, 332
 data processing 325–34
 DNA microarray technology comparison 323
 expression profiles 192, 321
 Human Transcriptome Map 336, 337, 340–1
 libraries 323, 324, 325, 330–1
 Human Transcriptome Map construction 334, 335
 RIDGES 339
 statistical tests 331–3
 principles 322
 statistical analysis resources 330
 tag-to-gene mapping 325–6, 327, 330, 333
 tags 322, 325–30, 331–3
 anti-sense 330
 Human Transcriptome Map construction 334, 337
 sense 330
 sequence errors 334
 Z-test 333, 334
- SAGE300 331, 333
- SAGEmap
 tag-to-gene mapping 327, 328, 333, 334
 Virtual Northern tool 325
- SAHA nucleotide searches 76
- salt bridges 304, 305
- SANGER 54
- SCA10 gene 47
- schizophrenia 10
- Schizosaccharomyces pombe* 15
- Scirus 23, 25–6, 27
 point mutations 50
- SDS-PAGE 348, 349
- search engines 23, 24–6
 domain restriction 26–7
 filtering results 26
 quotation mark use 26
 search syntax 26–7
- sequence data 30–1
 biological sequence databases 31–6
- Sequence Retrieval Server (SRS) 30, 31, 32, 58
- sequence tagged sites (STS) 60, 145
 genetic maps 152–3
 markers 151
 physical maps 152–3
 rat 124
 see also dbSTS
- sequence variation database 58
- Sequencher 205, 206–7
- Serial Analysis of Gene Expression *see* SAGE
- serine 299, 307
- serine/arginine-rich (SR) protein family 258
- serine proteases 80, 81
- short insertion/deletion polymorphisms (SIDPs) 49
- short interspersed elements (SINES) 275
- short tandem repeats *see* microsatellites
- sib transmission disequilibrium (S-TDT) 167, 225
- SIBPAIR program 219
- sickle cell disease 291
- signal transduction networks 358
- silencers 275
- simple interval mapping 236, 239–40
- SimWalk2 223, 225, 226
- single base extension (SBE) reactions 211
- single gene disorders *see* Mendelian disorders
- single marker association 239–40
- single nucleotide polymorphisms (SNP) 8, 11, 35, 168
 allelic discrimination methods 208
 amino acids 311–12
 annotation 195
 assay methods 208–9
 assay validation 212–13
 BACE gene 77
 base-wise multiple alignment 205
 biallelic 168
 candidate 55–6, 57, 66, 197
 identification 198–9
 CATG sequence 329, 330
 common 41
 consortium 12

- single nucleotide polymorphisms (SNP)
(continued)
 cystic fibrosis new binding site creation 280
 data 7, 54–5
 integration 43
 databases 51–7
 mouse 133
 density 250
 design 211, 212
 parameters 209–10
 detection
 algorithm 205
 methods 205–7, 208
 discovery 175, 204
 de novo 205
 global 204
 method 55
 non-sequencing methods 207
 targeted 204
 disease
 gene mapping 100
 susceptibility 204
 disease-causing 311–12
 effects in promoters 278
 exonic region 250
 experimental parameters 209–10
 exporting 197, 199
 flanking sequences 209
 gene coding sequence 280
 gene expression 281
 genetic maps 325
 genome map 145
 genome-wide 207
 genome-wide linkage map 145–6
 genomic 204
 overlap 54–5
 genotyping 175
 guide sequence 213
 haplotype
 and linkage disequilibrium maps 148
 patterns 198
 tags 197
 human variation 375
 identification 205
 for marker set selection 171, 172
 of potentially functional 195, 196
 intragenic region 250
 intronic region 250
 large-scale discovery 42
 location 276, 279, 280
 map locations 212–13
 mapping to promoter region 283
 marker density 147
 missense 311–12
 mouse 133
 mRNA secondary structure 263
 mutation
 events 40
 relationship 43–6
 natural history 44–6
 non-coding 273–85
 evaluation 280
 non-synonymous (nsSNP) 250, 311
 nucleotide difference scanning 205
 phenotypic effects 311
 physical maps 325
 pooled sequencing 211, 212
 potential effects 274–5, 276, 279
 primer
 design 209–10
 selection 210–11, 212
 private 55, 56
 reaction formats 208
 reference 52
 regulatory 277–8, 279–80
 functional consequence estimation 284–5
 in silico detection/evaluation 281–2
 regulatory analysis 277, 281–2
 regulatory DNA 276
 regulatory networks 281
 regulatory regions 279–80, 280
 relationship evolution 169
 repeat masking 209, 211
 score 205
 sequence trace data 206
 single base extension reactions 211
 statistical analysis 175
 structural elements 276–7
 TF binding site influence 279–80, 282
 tools for visualization 63–6
 TSC linkage map 147–8
 validation 285
 variation 43–6
 see also dbSNP
 single tandem repeat (STR) markers 9, 11, 12,
 172–3
 allele frequency distribution 172–3, 174
 analysis 173–5
 genotyping 174
 identifying in sequence data 196–7
 multiallelic 168
 mutation rate 172, 173
 polymorphic 9
 polymorphism testing 174
 see also microsatellites
 slipped-strand mispairing 48
 small nuclear ribonucleoprotein (snRNP)
 complexes 258

- SMARTest 275
 The SNP Consortium (TSC) Allele Frequency Project 208
 SNP-HAP program 226
 SNPper 52, 64–5, 199, 208
 flanking sequences 209
 identification of known and novel markers 195–9
 SOLAR program 220
 solenoids 275
 SPAD database 358
 spinocerebellar ataxia 10 (SCA10) 47
 splice acceptor/donor sites 259
 splice signals 279
 splice site prediction
 location 259
 tools 259–60
 spliceosome 258
 splicing 279
 SPLINK program 220
 SQL (Structured Query Language) 335–6
 SSAHA 99
 statistical analysis tools 218–40
 association analysis 223–5
 haplotype reconstruction 226–9
 quantitative trait locus mapping 236–9
 single marker association 239–40
 see also linkage analysis; linkage disequilibrium
 STRAT program 167
 structural elements 276–7
 STRUCTURE program 167
 Swiss-3Image 36
 SWISS-PROT keyword overlap 364
 SwissProt/TrEMBL (SPTR) 36, 73
 synteny groups 360–1

 Tag n Tell 57, 198
 tandem affinity purification (TAP) 348
 Tandem Mass Spectrometry 349
 Tandem Repeat Finder 196–7
 tandem repeat polymorphisms 46–8
 see also variable number of tandem repeats (VNTR)
 Taqman™ assays 208
 TBLASTN expressed sequence tags 84, 86
 TDTHAP program 226
Tetraodon nigroviridis (pufferfish) 15, 106
 TG deletion 265
 The SNP Consortium (TSC) Allele Frequency Project 208
 threonine 299, 307–8
 TIGR human gene index 35, 73, 84
 EST assemblies 88

 TNG radiation hybrid map 98, 149
 Trans-NIH BAC Sequencing Program 129
 transcription control 277–8
 transcription factor binding sites (TF-sites) 276, 277–8, 282
 single nucleotide polymorphism influence 279–80
 synergistic/antagonistic pairs 278
 transcription factors 265
 transcription start sites 111, 257, 266, 282
 transcriptional control regions, *in silico* recognition 74
 transcriptome translation products 346
 TRANSFAC database 264, 265, 358
 transgenic animals 15
 translation initiation 261–2
 transmission disequilibrium test (TDT) 167, 173, 225
 TRANSMIT program 227
 Transterm database 264
 TrEMBL 36
 TRES tool 265–6
 triplet repeat expansion diseases 47
 trypsin 294
 tryptophan 298, 302
 TSC map 146, 147–8
 tumour classification 15
 Twinscan program 111
 tyrosine 298, 299, 302
 tyrosine kinase receptor signal transduction network 358

 UCSC human genome browser 63
 candidate gene selection 190–5
 characterization of genetic/physical locus 199–201
 chromosome aberrations 49–50
 definition of known/novel genes across genomic region 188–90
 duplication detection in genomic assemblies 186
 genome viewer 157
 genomic sequence
 annotation 106–7
 assembly 98
 identification/extraction between two markers 184–5
 integrity checking between markers 185
 Genscan prediction 82
 GNF gene expression atlas ratios 193, 194
 human genome misassembly rate 98
 identification of known and novel markers 195–9
 maps 155
 marker mapping to human genome 182

- UCSC human genome browser (*continued*)
marker panel design 199–201
microarray data tracks 193, 194
mouse whole genome shotgun data 129
novel gene analysis 81–4, 86, 88
promoter analysis 257
pros and cons 183
regulatory element analysis 257
repeat saving 209
sequence characterization 173
SNP visualization 195, 196
unspliced ESTs 374–5
- UCSC site 7, 8
genome viewer 79
mouse genome assembly 74
viewing applications 61
- UK Mouse Genome Centre 124, 125
- Unified Database for Human Genome Mapping (UDB) 154, 159
- UniGene clusters 35, 75, 82, 84, 191–2
algorithm 330
chromosome 3 340
chromosome 11 338
errors 339
EST clones 328
homology assignments 87
Human Transcriptome Map construction 334
identity matches 88
SAGE analysis 326
tags 337, 339
- UniSTS 35
- University of California, Santa Cruz *see* UCSC
- untranslated regions (UTRs) 75, 79, 252, 256–7, 261
RNA instability signals 279
sequence identification 266
- UTRdb 263–4
- valine 298, 301
- variable number of tandem repeats (VNTR) 40, 41, 46–8
databases 48
- velocardiofacial syndrome (VCFS) 49
- version numbers 32
- Virtual Comparative Map (VCMaP) 124, 131, 132
- VISTA 131
- VITESSE program 218–19
- vocabulary, controlled 28
- Whitehead Institute for Biomedical Research 124, 125
- wildcards 28
- WNT genes 24
- yeast artificial chromosome (YAC) 12
clones 152
physical maps 151–2
- yeast protein interactions 366–7
map 367–8
- Yeast Proteome Database 366
- yeast two-hybrid system 351–2, 353, 354
- zoo-FISH 131
- Zucker, Michael 276–7