


Elementary Sequence Analysis

Brian Golding, Dick Morton and Wilfried Haerty

Department of Biology
McMaster University
Hamilton, Ontario
L8S 4K1

These notes are in Adobe Acrobat format (they are available upon request in other formats) and they can be obtained from the website <http://helix.biology.mcmaster.ca/courses.html>. Some of the programs that you will be using in this course and which will be run locally can be found at <http://evol.mcmaster.ca/p3S03.html>.

The “blue text” should designate links within this document while the “red text” designate links outside of this document. Clicking on the latter should activate your web browser and load the appropriate page into your browser. If these do not work please check your Acrobat reader setup. The web links are accurate to the best of our knowledge but the web changes quickly and we cannot guarantee that they are still accurate. The links designated next to the JAVA logo, , require that JAVA be installed on your computer.

These notes are used in Biology 3S03. The purpose of this course is to introduce students to the basics of bioinformatics and to give them the opportunity to learn to manipulate and analyze DNA/protein sequences. Of necessity only some of the more simple algorithms will be examined.

The course will hopefully cover ...

- databases of relevance to molecular biology.
- some common network servers/sites that provide access to these databases.
- methods to obtain sequence analysis software and data.
- methods of sequence alignment.
- methods of calculating genetic distance.
- methods of phylogenetic reconstruction.
- methods for detecting patterns and codon usage.
- methods for detecting gene coding regions.

The formal part of the course will consist of two approximately one hour lectures each week. Weekly assignments will be provided to practice and explore the lecture material. In addition there will be an optional tutorial to help students with these assignments or other problems. These assignments will be 40% of your grade and three, in class quizzes will make up the remainder.

We would appreciate any comments, corrections or updates regarding these notes.

Golding@McMaster.CA

Morton@McMaster.CA

HaertyW@McMaster.CA

Table of Contents in Brief

In order to speed download, I place here links to the individual chapters in pdf format. The contents of these are shown on the following 'Contents' pages but note that the links will function only for the individual chapter included here.

[Preliminaries](#)
[Basic Unix](#)
[Genomics](#)
[Databases](#)
[Sequence File Formats](#)
[Sequence Alignment](#)
[Distance Measures](#)
[Database Searching](#)
[Reconstructing Phylogenies](#)
[Pattern analysis](#)
[Exon analysis](#)

Contents

1	Preliminaries	1
1.1	Resources	1
1.1.1	Electronic Resources	1
1.1.2	Textbooks	2
1.1.3	Journal sources	7
1.2	Biological preliminaries	10
1.2.1	Some notes on terminology	10
1.2.2	Letter Codes for Sequences	11
2	Computer skills preliminaries	13
2.1	UNIX Operating Systems	13
2.1.1	Logging on/off	14
2.1.2	UNIX File System	14
2.1.3	Commands	17
2.1.4	Help	19
2.1.5	Redirection	20
2.1.6	Shells	20
2.1.7	Special 'hidden' files	21
2.1.8	Background Processes	21
2.1.9	Utilities	22
2.1.10	Editors	22
2.2	Exchange among computers	24
2.2.1	ssh	24
2.2.2	Mail	24
2.3	Scripts-Languages	25
2.4	Obtaining LINUX	25
3	Genomics	27
3.1	Where the data comes from	27
3.2	How DNA is sequenced	27

3.3	First Generation Methods	28
3.4	The reality of sequencing includes errors	32
3.5	From sequence to genome	33
3.6	Second (Next) Generation Sequencing	37
3.7	Paired sequences	43
3.8	Third Generation Sequencing	44
3.9	Upcoming Sequencing Technologies	45
3.10	Types of sequencing	46
3.10.1	Exome sequencing	46
3.10.2	RAD-tag seq	47
3.10.3	BAsE-seq	47
3.10.4	RNA-seq	47
3.10.5	BS-seq	48
3.10.5.1	TAB-seq	48
3.10.5.2	NOMe-seq	49
3.10.6	Regulatory sequencing: DNase-seq/FAIRE-seq	49
3.10.7	ChIP-seq	49
3.10.7.1	CLIP-seq	49
3.10.8	Hi-C	50
3.11	Other kinds of biological data	50
3.11.1	Microarrays	51
3.11.2	Mass spectrometry methods	56
3.11.3	Textual information	56
4	Databases	59
4.1	Introduction	59
4.2	N.C.B.I.	62
4.3	E.M.B.L.	67
4.4	D.D.B.J.	68
4.5	SwissProt	69
4.6	Organization of the entries	71
4.7	Other Major Databases	73
4.8	Remote Database Entry retrieval	76
4.8.1	Entrez	76
4.8.2	NCBI retrieve	77
4.8.3	EMBL get	79
4.8.4	Others	80
4.9	Reliability	80

5	Sequence File Formats	83
5.1	Genbank/EMBL	83
5.2	FASTA	85
5.3	FASTQ	86
5.4	SAM/BAM format	87
5.5	Stockholm format	88
5.6	GDE	90
5.7	NEXUS	92
5.8	PHYLIP	93
5.9	ASN	94
5.10	BSML format	97
5.11	PDB file format	97
6	Sequence Alignment	103
6.1	Dot Plots	103
6.1.1	The Exact Way	103
6.1.2	Identity Blocks	105
6.2	Alignments	113
6.2.1	The Needleman and Wunsch Algorithm	113
6.2.2	The Smith-Waterman Algorithm	116
6.3	Testing Significance	117
6.4	Gaps and Indels	120
6.4.1	“Natural” Gap Weights - Thorne, Kishino & Felsenstein	120
6.5	Multiple Sequence Alignments	121
7	Distance Measures	125
7.1	Nucleotide Distance Measures	125
7.1.1	Simple counts as a distance measure	125
7.1.2	Jukes - Cantor Correction	126
7.1.3	Kimura 2-parameter Correction	128
7.1.4	Tamura - Nei Correction	128
7.1.5	Uneven spatial distribution of substitutions	129
7.1.6	Synonymous - nonsynonymous substitutions	130
7.2	Amino acid distance measures	130
7.2.1	PAM Matrices	131
7.2.2	BLOSUM Matrices	133
7.2.3	GONNET Matrix	134
7.3	Gap Weighting	135

8	Database Searching	137
8.1	Are there homologues in the database?	137
8.1.1	FASTA	137
8.1.1.1	Instructions	137
8.1.1.2	FASTA output	139
8.1.1.3	FASTA format	142
8.1.1.4	Statistical Significance	144
8.1.2	BLAST	145
8.1.2.1	BLAST output	146
8.1.2.2	BLAST format	150
8.1.3	MPsrch	152
8.1.3.1	MPsrch output	153
8.1.3.2	MPsrch format	155
8.2	BLOCKS	156
8.2.1	BLOCKS output	157
8.2.2	Getting the Block	158
8.3	SSearch	164
8.4	Why you should routinely check your sequence	164
9	Reconstructing Phylogenies	165
9.1	Introduction	165
9.1.1	Purpose	165
9.1.2	Trees of what	165
9.1.3	Terminology	167
9.1.4	Controversy	169
9.2	Distance Methods	169
9.3	Parsimony Methods	171
9.4	Other Methods	174
9.4.1	Compatibility methods	174
9.4.2	Maximum Likelihood methods	174
9.4.3	Method of Invariants	175
9.4.4	Quartet Methods	176
9.5	Consensus Trees	178
9.6	Bootstrap trees	178
9.7	Warnings	181
9.8	Available Packages	182
9.9	PHYLIP	186
9.9.1	PHYLIP Contents	186

10 Pattern Analysis	199
10.1 Base Composition: first order patchiness	199
10.1.1 Genome Patchiness	199
10.2 Dinucleotide Composition: second order patchiness	200
10.3 Strand Asymmetry	201
10.3.1 Chargaff's Rules	201
10.3.2 Replication Asymmetry	202
10.3.3 Transcriptional Asymmetry	203
10.3.4 Codon Selection	204
10.4 Simple Sequence Repeats	204
10.5 Sequence Complexity	204
10.5.1 Information Theory	204
10.5.2 Sequence Window Complexity	206
10.6 Finding Pattern in DNA Sequences	207
10.6.1 Consensus Sequences	207
10.6.2 Matrix Analysis of Sequence Motifs	208
10.6.3 Sequence Conservation and Sequence Logos	209
11 Exon Analysis	213
11.1 Open Reading Frames	213
11.2 Gene Recognition	213
11.2.1 Splice Sites	214
11.2.2 Codon Usage	215
11.2.3 Gene Prediction Software	218
11.2.4 Hidden Markov Models (HMM)	219
11.2.5 Comparison of Programs	219

Chapter 5

Sequence File Formats

There are many formats that sequence data can be presented in. Each has advantages over the others (e.g. some are small and compact; others contain lots of information) and different programs require different formats as their input. The major databases permit sequences to be stored on your local computer in more than one format and there are programs that will convert one format to another. The most popular of these is a program called `readseq` by D.G. Gilbert (available for **UNIX**, **DOS** and **APPLE** machines).

The GenBank and EMBL formats have been discussed above. Both the GenBank and EMBL formats are highly stylized and strictly controlled to conform to consistent standards. Other popular formats are ASN.1, DNASTrider, Fitch, GCG, GDE, HENNIG86, IG/Stanford, MSF, NBRF, NEXUS, PIR/CODATA, Pearson/Fasta, Phylip - Interleaved, Phylip - Sequential, and Plain/Raw, I will not present all here but rather just a smattering.

Most formats will ignore case and this can therefore often be used to add information about the sequences. While the GenBank and EMBL formats can contain the character '-', they generally do not contain these characters and these formats were not intended to convey the kind of information that includes homologous sites between multiple sequences (the dashes indicate conceptual gaps in the sequences that have been inserted so that homologous parts of the sequence from each species are in the same location).

5.1 Genbank/EMBL

As a quick review, these two formats would be ...

```
LOCUS      MPU28721      607 bp      DNA                ROD      28-JUN-1995
DEFINITION Mus pahari adenine phosphoribosyltransferase (APRT) gene, complete
            cds.
ACCESSION  U28721
NID        g881573
KEYWORDS   .
SOURCE     shrew mouse.
            ORGANISM  Mus pahari
            Eukaryotae; mitochondrial eukaryotes; Metazoa; Chordata;
            Vertebrata; Eutheria; Rodentia; Sciurognathi; Myomorpha; Muridae;
            Murinae; Mus.
REFERENCE  1 (bases 1 to 2283)
            AUTHORS   Fieldhouse,D. and Golding,G.B.
            TITLE      Rates of substitution in closely related rodent species
            JOURNAL    Unpublished
REFERENCE  2 (bases 1 to 2283)
            AUTHORS   Fieldhouse,D.
            TITLE      Direct Submission
            JOURNAL    Submitted (07-JUN-1995) Dan Fieldhouse, Biology, McMaster
            University, 1280 Main Street West, Hamilton, ON, L8S 4K1, Canada
FEATURES   Location/Qualifiers
            source     1..2283
                    /organism="Mus pahari"
                    /db_xref="taxon:10093"
            gene       join(46..125,256..362,1509..1642,1847..1925,2044..2186)
```

```

        /gene="APRT"
CDS      join(46..125,256..362,1509..1642,1847..1925,2044..2186)
        /gene="APRT"
        /EC_number="2.4.2.7"
        /note="purine salvage enzyme"
        /codon_start=1
        /product="adenine phosphoribosyltransferase"
        /db_xref="PID:g881574"
        /translation="MSESELKLVARRIRSFDPDFPIPGVLFDRDISPLLKDPDSFRASIR
LLASHLKSTHSGKIDYIAGLDSRGFLFGPSLAQELGVGCVLIRKQGKLPGPPTISASYA
LEYGKAELEIQKDALEPGQORVVIVDDLLATGGTMFAACDLLHQLRAEVVECVSLVELT
SLKGRERLGPPIFFSLLQYD"
BASE COUNT      87 a      228 c      145 g      147 t
ORIGIN
1  cctgcggata  ctcacctcct  ccttgtctcc  tacaagcacg  cggccatgtc  cgagtctgag
61 ttgaaactgg  tggcgcgggc  catccgcagc  ttccccgact  tccccatccc  gggcgtgctg
121 ttcaggtgcg  gtcacgagcc  ggcgaggcgt  tggcgccgta  ctctcatccc  ccggcgccag
181 cgcggtggga  gccttgggga  tcttgcgggg  cctctgcccc  gccacacgcg  gtcactctcc
241 tgtccttggt  cccagggata  tctcgcccct  cttgaaagat  cgggactcct  tccgagcttc
301 catccgcctc  ctggccagtc  acctgaagtc  cacgcacagc  ggcaagatcg  actatatogc
361 agggcaaggt  ggccttgcta  ggccttactc  atccccacg  gtcctatccc  ctatcccctt
421 tcccctcgtg  tcacccacag  tctaccccac  acccatccat  tctttcttta  acctctgact
481 ctctcctcct  ggtttctcac  tgccttggtg  gcttggtcac  cccggatgaa  ctccgtaggg
541 gtctccttcc  cctgcttggt  accctaaggt  gccttcggtg  cttgttcgta  gagacgaact
601 ctgctctc
//

```

and

```

ID      MP28721      standard; DNA; ROD; 607 BP.
XX
AC      U28721;
XX
NI      g881573
XX
DT      04-JUL-1995 (Rel. 44, Created)
DT      04-JUL-1995 (Rel. 44, Last updated, Version 1)
XX
DE      Mus pahari adenine phosphoribosyltransferase (APRT) gene, complete
DE      cds.
XX
KW      .
XX
OS      Mus pahari (shrew mouse)
OC      Eukaryotae; mitochondrial eukaryotes; Metazoa; Chordata; Vertebrata;
OC      Mammalia; Eutheria; Rodentia; Sciurognathi; Muridae; Murinae; Mus.
XX
RN      [1]
RP      1-2283
RA      Fieldhouse D., Golding G.B.;
RT      "Rates of substitution in closely related rodent species";
RL      Unpublished.
XX
RN      [2]
RP      1-2283
RA      Fieldhouse D.;
RT      ;
RL      Submitted (07-JUN-1995) to the EMBL/GenBank/DDBJ databases.
RL      Dan Fieldhouse, Biology, McMaster University, 1280 Main Street West,
RL      Hamilton, ON, L8S 4K1, Canada
XX
DR      SWISS-PROT; P47956; APT_MUSPA.
XX
CC      NCBI gi: 881573
XX
FH      Key          Location/Qualifiers
FH
FT      source          1. .2283
FT                      /organism="Mus pahari"
FT      CDS            join(46..125,256..362,1509..1642,1847..1925,2044..2186)
FT                      /codon_start=1
FT                      /db_xref="PID:g881574"
FT                      /db_xref="SWISS-PROT:P47956"
FT                      /note="purine salvage enzyme; Method: conceptual
FT                      translation supplied by author. NCBI gi: 881574"
FT                      /gene="APRT"
FT                      /EC_number="2.4.2.7"
FT                      /product="adenine phosphoribosyltransferase"
FT                      /translation="MSESELKLVARRIRSFDPDFPIPGVLFDRDISPLLKDPDSFRASIRL

```

```

FT          LASHLKSTHSGKIDYIAGLDSRGLFPGPSLAQELGVGCVLIRKQKLPGPPTISASYALE
FT          YGKAELEIQKDALEPGQRVVIVDDLLATGGTMFAACDLHLQLRAEVVECVSLVELTSLK
FT          GRERLGPPIFFSLLQYD"
XX
SQ Sequence 607 BP; 87 A; 228 C; 145 G; 147 T; 0 other;
      CCTGCGGATA CTCACCTCCT CTTGTCTCC TACAAGCAGC CGGCCATGTC CGAGTCTGAG      60
      TTGAAACTGG TGGCGCGGCG CATCCGCAGC TTCCCCGACT TCCCCATCCC GGGCGTGCTG      120
      TTCAGGTGCG GTACAGAGCC GGCAGGGCGT TGGCGCCGTA CTCTCATCCC CCGCGCAGG      180
      CGCGTGGGCA GCCTTGGGA TCTTGCGGGG CCTCTGCCCG GCCACACGCG GTCACTCTCC      240
      TGTCTTGTGT CCCAGGGATA TCTCGCCCCT CTTGAAAGAT CCGGACTCCT TCCGAGCTTC      300
      CATCCGCTCC CTGGCCAGTC ACCTGAAGTC CACGCACAGC GGCAAGATCG ACTATATCGC      360
      AGGCAAGGT  GGCCTTGCTA GGCCGTACTC ATCCCCCAGC GTCTATCCC CTATCCCCTT      420
      TCCCCTCGTG TCACCCACAG TCTACCCAC ACCCATCCAT TCTTTCTTTA ACCTCTGACT      480
      CTTCTCCTTT GGTTCCTCAC TGCCTTGGAC GCTTGTTCAC CCCGGATGAA CTCCGTAGGC      540
      GTCCTCCCTC CCTGCTTGGT ACCCTAAGGT GCCCTCGGTG CTTGTTCTGA GAGACGAACT      600
      CTGCTCT
//

```

In the Genbank format, sequence information is set aside with key words. The entire entry begins with the keyword LOCUS at the beginning of a line and ends with //. Different features are set off with different keywords; the sequence information itself with the keyword ORIGIN.

The EMBL format is similar but with two-letter codes at the beginning of each line to designate different features of the entry (much easier to program). The entire entry begins with the key ID at the beginning of a line and ends with //.

5.2 FASTA

By far the simplest format is termed the *fasta* (also known as the Pearson format). This sequence format contains the minimal amount of information. A *fasta* file will contain just a ‘>’ sign (at the beginning of a line) to indicate the beginning of a new sequence and a word (phrase) to serve as the sequence title. The sequence information itself follows immediately. No other information is stored within a *fasta* file. As an example, I will use a proportion of the *Mus pahari*, *Mus spicilegus* and *Gerbillus campestris* APRT gene sequences. These sequences would appear as ...

```

>MPU28721      650 bp      1/31/98 14:18:24, 650 bases, F8A0A666 checksum.
-----CCTGCGGATACTCA
CCTCCTCCTTGTCTCCTACAAGCACGCGGCCATGTCGAGTCTGAGTTGA
AAGTGGTGGCGCGGCGCATCCCGCAGCTTCCCCGACTTCCCCATCCCGGGC
GTGCTGTTTCAAGTGCAGTACAGAGCCGCGGAGGCGTTGGCGCCGTA
CATCCC-CCGCGCAGGCGCGTGGGAGCCTTGGGATCTTGGCGGCGCT
CTGCCCCGCCACAGCGG-TCACTCTCCTGTCTTGTTCAGGGATATC
TGCCCCCTTTGAAAGATCCGGACTCCTTCCGAGCTTCCATCCGCTCCT
GGCCAGTCACTGAAGTCCACGACAGCGGCAAGATCGACTATATCGCAG
GGCAAGGTGGCCTTGTAGGCGGACTCATCCCCACGGTCTATCCCT
ATCCCCCTTTCCC-TCGTGTACCCACAGTCTACCCACACCCATCCATT
CTTTCTTAACTCTGACTTCTCCTCTTGGTTTCTACTGCCTTGGAGC
CTTGTTCACCCGGATGAAGTCCGTTAGGCGTCTCCCTTCCCTGCTTGGTA
CCCTAAGG----TGCCCTCGGTGCTTGTTCGTAGAGACGAACTCTGCTCT
>MSU28720      650 bp      1/31/98 14:18:24, 650 bases, 450AB895 checksum.
-----TCGGGATTGACGTGAATTTAGCGTGTGATACCTA
CCTCCTCCTTGCCTCCTACAGCACGCGGCCATGTCCGAACCTGAGTTGA
AAGTGGTGGCGCGGCGCATCCCGCAGCTTCCCCGACTTCCCAATCCCGGGC
GTGCTGTTTCAAGTGCAGTACAGAGCCGCGGAGGCGTTGGCGCCGTA
CATCCC-CCGCGCAGGCGCGTAGGACAGCTCGGGATCTTGGCGGCGCT
CTGCCCCGCCACAGCGGGTCACTCTCCTGTCTTGTTCAGGGATATC
TGCCCCCTTTGAAAGACCCGGACTCCTTCCGAGCTTCCATCCGCTCCT
GGCCAGTCACTGAAGTCCACGACAGCGGCAAGATCGACTACATCGCAG
GGCA--GTGGCCTTGTAGGCGGTGCTCGTCCCCACGGTCTAGCCCCCT
ATCCCCCTTTCCCCTCGTGTACCCACAGTCTGCCCCACACCCATCCATT
CTTTCTTCAACTCTGACACTTCTCCTTGGTTTCTCACTGCCTTGGAGC
CTTGTTCACCCGGATGAAGTATGTAGGAGTCTCCCTTCCCTGCTAGGTA
CCCTAAGGCATCTGCCCTCGGTGCTTGTTCGTAGAGACGAACTCTGCTCT
>GCU28961      650 bp      1/31/98 14:18:24, 650 bases, 606DF2D9 checksum.
CCTCCGCCCTTGTTCCTGGGACAGGCTTGACCTAGCCAGTTGACACCTC
ACCTCCGCCCTTCTCT-CACGCACGCGGCCATGGCGGAACCCGAGTTGC
AGCTGGTGGCGCGGCGCATCCCGCAGCTTCCCCGACTTCCCCATCCCGGGC
GTGCTGTTTCAAGTGCAGTCCACAGCCGCGGAGGCGTTGGCGCTGCTCCT
CAGCCCTCCGCGCAGGCGCGTGTCTTCCGGATCTTGGCGGCGCT
CCGCCAGCCATACCCAAAGTACCATCCTG----TGTTCAGGGATATC
TGCCCCCTTCTGAAAGACCCGGACTCCTTCCGAGCTTCCATCCGCTCCT
GGCCAACCATCTGAAGTCAAGCATGGCGGCAAAATCGACTACATCGCAG
GGCA--GTGTTCTTGTAGGCGGTGCCCTTCCC-ACTGTAGGGCGGCC

```

```
ATCCCGTGTTCCTCC-----TTTTTCGTGTCACCCACACCCACCCCTC
CTTCTCTGACACTCCCAAGTTCCCT----GTCCCTCTCGCCTTGGTCC
CATATTCACCCCGGATGA-CTGCGGAGTCTCCACCCCTCTGACCTCTGCT
CTCAAAGC-----CTGTCCCTAC---TAGAGAGGAACCTCTGCTCT
```

Note that although it is a simple format, sequence alignment information (more on this later) can be indicated by the dashes.

5.3 FASTQ

This FASTQ format specification is modified from <http://maq.sourceforge.net/fastq.shtml>.

FASTQ format stores sequences and Phred qualities in a single file (Phred quality scores are so named after a popular software package and have become the standard method to quantify the reliability of the base call). It is concise and compact. FASTQ was first widely used in the Sanger Institute. Although Solexa/Illumina files look like FASTQ files, they scale the quality scores differently. In the quality string, if you can see a character with its ASCII code higher than 90, your file is probably in the Solexa/Illumina format. Just to make things more confusing Illumina created a third version “Illumina 1.3+ FASTQ” format.

An example from work done at McMaster,

```
@HWI-EAS038:8:1:8:697#0/1
AGACTGGCTGGAGCATGTCTATGACGGACTATGATG
+HWI-EAS038:8:1:8:697#0/1
aaa`[[`a`^[^U^_YPU[[`ZU^VSTZVX_TB BBB
@HWI-EAS038:8:1:8:1326#0/1
AGACTACCGTGTCTCGTCACGACACGGTCGACGACCAC
+HWI-EAS038:8:1:8:1326#0/1
a^a^`aa`\ZUZVPV`\SP`]aSPQSRNXWBBBBBB
@HWI-EAS038:8:1:8:1305#0/1
AGACTCGAAACGCCTTTCTGGAACACGAAAGGTCTC
+HWI-EAS038:8:1:8:1305#0/1
aXa`_`^`aaa_W[\`^`^`^VT]a`_[T`^`\`WSW^W[
```

The FASTQ format specification comes in four lines. The first line begins with an ‘@’ symbol and is followed by the sequence name. The second line contains the base call (in this case for each of 36 nucleotides). The third line begins with a ‘+’ symbol and may (or may not) repeat the sequence name. The fourth line contains a symbol that measures the quality score for the corresponding base call as listed on the second line. There should be one symbol for each base call. Another read will follow with another four lines.

The symbol on the fourth line uses an ASCII character (American Standard Code for Information Interchange) to encode the quality score. Part of an ASCII table is reproduced here.

	30	40	50	60	70	80	90	100	110	120
0	(2	<	F	P	Z	d	n	x	
1)	3	=	G	Q	[e	o	y	
2	*	4	>	H	R	\	f	p	z	
3	!	5	?	I	S]	g	q	{	
4	”	,	@	J	T	^	h	r		
5	#	-	7	A	K	U	_	i	s	}
6	\$.	8	B	L	V		j	t	~
7	%	/	9	C	M	W	a	k	u	DEL
8	&	0	:	D	N	X	b	l	v	
9	'	1	;	E	O	Y	c	m	w	

Given a character q , the corresponding Phred quality score can be calculated with:

1. QNAME Query template/pair NAME
In the case above, the name read “name” is the name of the machine that generated the sequence (HWUSI-EAS1786), the run id (60), the flowcell id (FC62MTAAXX), the flowcell lane (1), the tile number within the flowcell lane (1), and the (x,y) coordinates of the cluster on the tile.
2. FLAG bitwise FLAG
A number that describes various features of the read (e.g. if the sequence is reverse complemented). See samtools.sourceforge.net/SAM1.pdf for a listing of the flags.
3. RNAME Reference sequence NAME
The name of the sequence to which the read was mapped.
4. POS The left most coordinate of the read
using the number of the sequence in the reference genome.
5. MAPQ MAPping Quality
The map quality is Phred-scaled. A value of 255 is used for an unknown map quality.
6. CIAGR extended CIGAR string
This string describes features of the match between the read and the reference sequence. In the cases above it is ‘[0-9 M’ indicating a perfect match for the length of the read. The format is a number followed by a letter. The number indicate the number of bases and the letter designates a category; M for match, I for an insert in the read, D for a deletion in the read, N for a region skipped, etc.
7. MRNM Mate Reference sequence NaMe
In the cases above it is ‘*’ meaning that there is no mate; these were unpaired reads.
8. MPOS Mate POSition
The bp location in the reference genome where the leftmost bp of the mate read maps.
9. TLEN inferred Template LENgth
The length of the insert between mate pairs.
10. SEQ query SEQUENCE
The sequence of the read.
11. QUAL query QUALity
The quality is given as ord(ASCII)-33 (Sanger Phred score).
12. OPT variable OPTional fields
These fields are in the format TAG:VTYPE:VALUE. Many TAGs (e.g. X*:*) can be defined by the user-program. Two common, predefined TAGs are MD:Z (a tag for mismatching positions) and NM:i (a tag for the edit distance from the read to the reference genome)

The BAM format is a binary version of the SAM format. It is designed to improve data handling performance, to speed the analysis and to reduce file sizes.

5.5 Stockholm format

This section is modified from a posting originally at EBI (see also http://en.wikipedia.org/wiki/Stockholm_format).

The Stockholm Format is used by the Pfam database, by the popular program HMMER that uses hidden Markov chain models to compare protein sequences and by the Belvu software. The major feature difference from some of the other formats noted here is that it has a system for marking up features in a multiple alignment. These mark-up annotations are preceded by a ‘magic’ label.

The file format must begin with a line that declares the format and the version being used. Currently it should be

```
# STOCKHOLM 1.0
```

This is then followed by either markup annotations or sequence alignments. The sequence alignments follow the format of

```
< seqname> <aligned sequence>
< seqname> <aligned sequence>
< seqname> <aligned sequence>
.
.
//
```

where <seqname> is the “sequence name” and the “//” indicates the end of the alignment. Sequence letters may include any characters except whitespace. Gaps may be indicated by “.” or “-”. Wrap-around alignments are allowed in principle, mainly for historical reasons, but are not used in e.g. Pfam. Wrapped alignments are discouraged since they are much harder to parse. Hence this format is best adapted to protein sequences.

There are four types of alignment mark-up, indicated in the following manner.

```
#=GF <feature> <Generic per-File annotation, free text>
#=GC <feature> <Generic per-Column annotation, exactly 1 char per column>
#=GS <seqname> <feature> <Generic per-Sequence annotation, free text>
#=GR <seqname> <feature> <Generic per-Sequence AND per-Column markup,
      exactly 1 char per column>
```

Mark-up lines may include any characters except whitespace. Use underscore (“_”) instead of space. Many different “features” can be recognized or simply free text can be used. Some of the more interesting per-column (GR) annotations are

```
#=GR
```

Feature	Description	Markup letters
SS	Secondary Structure	[HGIEBTSCX]
SA	Surface Accessibility (0=0%-10%; ...; 9=90%-100%)	[0-9X]
TM	TransMembrane	[Mio]
PP	Posterior Probability (0=0.00-0.05; 1=0.05-0.15; *=0.95-1.00)	[0-9*]
LI	LIgand binding	[*]
AS	Active Site	[*]
pAS	AS - Pfam predicted	[*]
sAS	AS - from SwissProt	[*]
IN	INtron (in or after)	[0-2]

An example is,

```
# STOCKHOLM 1.0
#=GF ID CBS
#=GF AC PF00571
#=GF DE CBS domain
#=GF AU Bateman A
#=GF CC CBS domains are small intracellular modules mostly found
#=GF CC in 2 or four copies within a protein.
```

```

#=GF SQ 67
#=GS O31698/18-71 AC O31698
#=GS O83071/192-246 AC O83071
#=GS O83071/259-312 AC O83071
#=GS O31698/88-139 AC O31698
#=GS O31698/88-139 OS Bacillus subtilis
O83071/192-246      MTCRAQLIAVPRASSLAE..AIACAQKM....RVS RVPVYERS
#=GR O83071/192-246 SA 999887756453524252..55152525....36463774777
O83071/259-312      MQHVSAPVVFVFECTRLAY..VQHKLRAH....SRAVAIVLDEY
#=GR O83071/259-312 SS CCCCCHHHHHHHHHHHHHH..EEEEEEEE...EEEEEEEEEEEE
O31698/18-71        MIEADKVAHVQVGNLEH..ALLVLTKT....GYTAIPVLDPS
#=GR O31698/18-71 SS CCCHHHHHHHHHHHHHHHH..EEEEEEEE...EEEEEEEEHHH
O31698/88-139       EVMLTDIPRLHINDPIMK..GFGMVINN.....GFVCVENDE
#=GR O31698/88-139 SS CCCCCCHHHHHHHHHHHHH..HEEEEEEE...EEEEEEEEEEEH
#=GC SS_cons        CCCCCHHHHHHHHHHHHHH..EEEEEEEE...EEEEEEEEEEEH
O31699/88-139       EVMLTDIPRLHINDPIMK..GFGMVINN.....GFVCVENDE
#=GR O31699/88-139 AS _____*_____
#=GR_O31699/88-139_IN _____1_____2_____0_____
//

```

5.6 GDE

The GDE format can also contain alignment information but note that it may have an 'offset' value. This (often annoying) feature permits a compact storage of sequence information at the tails of the sequence. An 'offset' of 36 means to insert 36 '-' in front of the sequence in order to properly line it up with the other sequences. This format can also contain all the information that is present in a GenBank format but does so simply as a 'comment' enclosed in quotation marks and any information may appear within the comment field. The example *Mus pahari*, *Mus spicilegus* and *Gerbillus campestris* APRT gene sequences in a GDE format would appear as ...

```

{
name "MPU28721"
type "DNA"
longname Mus pahari
sequence-ID "U28721"
descrip "Mus pahari adenine phosphoribosyltransferase (APRT) gene, complete cds"
creator "Fieldhouse,D. and Golding,G.B."
offset 36
creation-date 1/31/98 14:18:24
direction 1
strandedness 1
comments "
NID g881573
KEYWORDS .
SOURCE shrew mouse.
Eukaryotae; mitochondrial eukaryotes; Metazoa; Chordata;
Vertebrata; Eutheria; Rodentia; Sciurognathi; Myomorpha; Muridae;
Murinae; Mus.
REFERENCE 1 (bases 1 to 2283)
TITLE Rates of substitution in closely related rodent species
JOURNAL Unpublished
REFERENCE 2 (bases 1 to 2283)
TITLE Direct Submission
JOURNAL Submitted (07-JUN-1995) Dan Fieldhouse, Biology, McMaster
University, 1280 Main Street West, Hamilton, ON, L8S 4K1, Canada
FEATURES
Location/Qualifiers
source 1..2283
/organism='Mus pahari'
/db_xref='taxon:10093'
gene join(46..125,256..362,1509..1642,1847..1925,2044..2186)
/gene='APRT'
CDS join(46..125,256..362,1509..1642,1847..1925,2044..2186)
/gene='APRT'
/EC_number='2.4.2.7'
/note='purine salvage enzyme'

```

```

/codon_start=1
/product='adenine phosphoribosyltransferase`
/db_xref='PID:g881574`
/translation='MSESELKLVARRIRSFDPFPIPGLVFRDISPLLKDPDSFRASIR
LLASHLKSTHSGKIDYIAGLDSRGFLFGPSLAQELGVCVLRKQGLPGPTISASYA
LEYGKAELEIQKDALEPGQVVIVDDLLATGGTMFAACDLLHQLRAEVVECVSLVELT
SLKGRERLGPPIFFFSLQYD`
BASE COUNT      485 a      696 c      590 g      512 t
"
sequence "CCTGCGGATACTCACCTCCTCCTT
GTCTCCTACAAGCACGCGGCCATGTCCGAGTCTGAGTTGAAACTGGTGGCGCGGCATC
CGCAGCTTCCCCGACTTCCCAATCCCGGGCGTGTTCAGGTGCGGTACAGACCGGGC
AGGCGTTGGCGCGTACTCTCATCCC-CCGGCGCAGGCGGTGGGAGCCTTGGGGATCT
TGCGGGCCTCTGCCCGGCCACACGCGG-TCACTCTCCTGTCTTGTCCAGGGATATC
TCGCCCTCTTGAAAGATCCGGACTCCTCCGAGCTTCCATCCGCCTCTGGCCAGTCAC
CTGAAGTCCACGCACAGCGGCAAGATCGACTATATCGCAGGGCAAGGTGGCCTTGTAGG
CCGTACTCATCCCCACGGTCTATCCCCATCCCCCTTCCCC-TCGTGTACCCACAGT
CTACCCACACCCATCCATTCTTTCTTAACCTCTGACTCTTCCCTCTGGTTCTCACT
GCCTTGGACGCTTGTTCACCCCGGATGAACTCCGTAGGCGTCTCCCTTCCCTGCTGGTA
CCCTAAGG----TGCCTCGGTGCTTGTTCGTAGAGACGAACCTCTGCTCT"
}
{
name "MSU28720"
type "DNA"
longname Mus spicilegus
sequence-ID "U28720"
descrip "Mus spicilegus adenine phosphoribosyltransferase (APRT) gene,"
creator "Fieldhouse,D. and Golding,G.B."
offset 15
creation-date 1/31/98 14:18:24
direction 1
strandedness 1
comments "
NID      g881575
KEYWORDS .
SOURCE   Steppe mouse.
         Eukaryotae; mitochondrial eukaryotes; Metazoa; Chordata;
         Vertebrata; Eutheria; Rodentia; Sciurognathi; Myomorpha; Muridae;
         Murinae; Mus.
REFERENCE 1 (bases 1 to 2117)
         TITLE Rates of substitution in closely related rodent species
         JOURNAL Unpublished
REFERENCE 2 (bases 1 to 2117)
         TITLE Direct Submission
         JOURNAL Submitted (07-JUN-1995) Dan Fieldhouse, Biology, McMaster
         University, 1280 Main Street West, Hamilton, ON, L8S 4K1, Canada
FEATURES   Location/Qualifiers
         source      1..2117
                 /organism='Mus spicilegus`
                 /db_xref='taxon:10103`
         gene        join(67..146,278..384,1355..1488,1675..1753,1860..2002)
                 /gene='APRT`
         CDS         join(67..146,278..384,1355..1488,1675..1753,1860..2002)
                 /gene='APRT`
                 /EC_number='2.4.2.7`
                 /note='purine salvage enzyme`
                 /codon_start=1
                 /product='adenine phosphoribosyltransferase`
                 /db_xref='PID:g881576`
                 /translation='MSEPELKLVARRIRSFDPFPIPGLVFRDISPLLKDPDSFRASIR
                 LLASHLKSTHSGKIDYIAGLDSRGFLFGPSLAQELGVCVLRKQGLPGPTVSASYS
                 LEYGKAELEIQKDALEPGQVVIVDDLLATGGTMFAACDLLHQLRAEVVECVSLVELT
                 SLKGRERLGPPIFFFSLQYD`
BASE COUNT      413 a      652 c      564 g      488 t"
sequence "TCGGGATTGACGTGAATTTAGCGTGTGATACCTACCTCCTCCTT
GCCTCCTACACGCACGCGGCCATGTCCGAACCTGAGTTGAAACTGGTGGCGCGGCATC
CGCAGCTTCCCCGACTTCCCAATCCCGGGCGTGTTCAGGTGCGGTACAGACCGGGC
AGGCGTTGGCGCGTACGCTCATCCC-CCGGCGCAGGCGGTAGGCAGCCTCGGGATCT
TGCGGGCCTCTGCCCGGCCACACGCGGGTCACTCTCCTGTCTTGTCCAGGGATATC
TCGCCCTCTTGAAAGACCCGGACTCCTCCGAGCTTCCATCCGCCTTGGCCAGTCAC
CTGAAGTCCACGCACAGCGGCAAGATCGACTACATCGCAGGGCA--GTGGCCTTGTAGG
CCGTGCTCGTCCCCACGGTCTTAGCCCTATCCCCTTCCCCCTCGTGTACCCACAGT
CTGCCCCACACCCATCCATTCTTTCTCAACCTCTGACTTCTCCTTGGTTCCTCACT
GCCTTGGACGCTTGTTCACCCCGGATGAACTATGTAGGAGTCTCCCTTCCCTGCTAGGTA
CCCTAAGGCATCTGCCCTCGGTGCTTGTTCCTAGAGACGAACCTCTGCTCT"
}
{
name "GCU28961"
type "DNA"
longname Gerbillus campestris

```

```

sequence-ID "U28961"
descrip "Gerbillus campestris adenine phosphoribosyltransferase (APRT) gene,"
creator "Yazdani,F. and Golding,G.B."
creation-date 1/31/98 14:18:24
direction 1
strandedness 1
comments "
NID g899456
KEYWORDS .
SOURCE Gerbillus campestris.
Eukaryotae; mitochondrial eukaryotes; Metazoa; Chordata;
Vertebrata; Eutheria; Rodentia; Sciurognathi; Myomorpha; Muridae;
Gerbillinae; Gerbillus.
REFERENCE 1 (bases 1 to 2076)
TITLE Rates of substitution in closely related rodent species
JOURNAL Unpublished
REFERENCE 2 (bases 1 to 2076)
TITLE Direct Submission
JOURNAL Submitted (12-JUN-1995) Fariborz Yazdani, Biology, McMaster
University, 1280 Main Street West, Hamilton, Ont L8S 4K1, Canada
FEATURES
source 1..2076
/organism='Gerbillus campestris'
/db_xref='taxon:41199'
gene join(81..160,289..395,1313..1446,1649..1727,1828..1970)
/gene='APRT'
exon >81..160
/gene='APRT'
CDS join(81..160,289..395,1313..1446,1649..1727,1828..1970)
/gene='APRT'
/EC_number='2.4.2.7'
/note='purine salvage enzyme'
/codon_start=1
/product='adenine phosphoribosyltransferase'
/db_xref='PID:g899457'
/translation='MAEPELQLVARRIRSFDFPPIPGVLFRRDISPLLKDPDSFRASIR
LLANHLKSKHGKIDYIAGLDSRGFLFGPSLAQELGLGCVLIRKRGKLPGPTVSASYA
LEYGKAELEIQKDALEPGQKVVIVDDLLATGGTMCACQLLQQLRAEVVECVSLVELT
SLKGREKLGVPVFFSLLQYE'
intron 161..288
/gene='APRT'
exon 289..395
/gene='APRT'
intron 396..1312
/gene='APRT'
exon 1313..1446
/gene='APRT'
intron 1447..1648
/gene='APRT'
exon 1649..1727
/gene='APRT'
intron 1728..1827
/gene='APRT'
exon 1828..>1970
/gene='APRT'
BASE COUNT 385 a 666 c 577 g 448 t"
sequence "
CCTCCGCCCTTGTTCTCGGGACAGGCTTGACCCTAGCCAGTTGACACCTCACCTCCGCC
TTCTCT-CACGCACGCGGCCATGGCGGAACCCGAGTTGCAGCTGGTGGCGCGGCATC
CGCAGCTTCCCGACTTCCCATCCCGGGCGTGTGTTTCAGGTGCGTCCACGAGCCGCC
AGCGTGGCGTGCCTCAGCCCTCCGGCGCAGGCGGTGAGCTGTCTCCGGGATCT
TGCGGGGCTCCGCCAGCCATACCAAGTACCATCCTG----TGTTCCAGGGATATC
TCGCCCTCTGAAAGACCCGACTCCTCCGAGCTTCCATCCGTCTCCTGGCCAACCAT
CTGAAGTCCAAGCATGGCGGCAAAATCGACTACATCGCAGGCGA--GTGTTCTGTAGG
CCGTGCCGTTCCC-ACTGTAGGGCCGCCATCCCGTGTTC-----TTTTTCGT
GTCACCCACACCCCTCCTTCTCTGACACTCCCAAGTTCCT----GTCTCTCT
GCCTTGGTCCATATTCACCCCGGATGA-CTGCGGAGTCTCCACCTCTGACCTCTGCT
CTCAAAGC-----CTGTCCTAC---TAGAGAGGAAGTCTGCTCT"
}

```

5.7 NEXUS

The popular PAUP, MacClade and Mr. Bayes programs (and others) use a NEXUS format (Maddison, Swofford and Maddison 1997. *Syst. Biol.*, 46, 590-621). The primary feature of this format is its modularity. Files identify themselves with the key phrase ``#NEXUS'' at the beginning of the file. Each block of information begins with ``BEGIN -

- ' ' ; and ends with ` `END; ' '. Comments can be enclosed within square brackets. For these sequences a simple translation would be

```
#NEXUS

[Name: MPU28721      Len:   650  Check:  643A358]
[Name: MSU28720     Len:   650  Check:  FDC8BCDB]
[Name: GCU28961     Len:   650  Check:  D8AFF697]

BEGIN TAXA;
  DIMENSIONS NTAX=3;
  TAXLABELS MPU28721 MSU28720 GCU28961;
END;

BEGIN CHARACTERS;
  DIMENSIONS NCHAR=650;
  FORMAT MISSING=? DATATYPE=DNA INTERLEAVE GAP=-;
  MATRIX
MPU28721  -----CCTG  CGGATACTACCTCCTCCTT  GTCTCCTACAAGCACGCGGC  CATGTCCGAGTCTGAGTTGA
MSU28720  -----TCGGG  ATTGACGTGAATTTAGCGTG  CTGATACCTACCTCCTCCTT  GCCTCCTACACGCACGCGGC  CATGTCCGAACCTGAGTTGA
GCU28961  CCTCCGCCCTTGTTCCTGGG  ACAGGCTTGACCTAGCCAG  TTGACACCTACCTCCGCC  TTCTCT-CACGCACGCGGC  CATGGCGGAACCCGAGTTGC

MPU28721  AACTGGTGGCGCGGCATC  CGCAGCTTCCCGACTTCCC  CATCCCGGGCGTGTGTTCA  GGTGCGGTCACGAGCCGCG  AGGCGTTGGCGCCGACTCT
MSU28720  AACTGGTGGCGCGGCATC  CGCAGCTTCCCGACTTCCC  AATCCCGGGCGTGTGTTCA  GGTGCGGTCACGAGCCGCG  AGGCGTTGGCGCCGACTACG
GCU28961  AGCTGGTGGCGCGGCATC  CGCAGCTTCCCGACTTCCC  CATCCCGGGCGTGTGTTCA  GGTGCGTCCACGAGCCGCC  AGGCGTTGGCGCTGCGTCCT

MPU28721  CATCCC-CCGGCGCAGGCGC  GTGGGCAGCCTTGGGGATCT  TCGGGGCTCTGCCCGGCC  ACACGCGG-TCACTCTCCTG  TCCTTGTTCACAGGGATATC
MSU28720  CATCCC-CCGGCGCAGGCGC  GTAGGCAGCCTCGGGGATCT  TCGGGGCTCTGCCCGGCC  ACACGCGGGTCACTCTCCTG  TCCTTGTTCACAGGGATATC
GCU28961  CAGCCCTCCGGCGCAGGCGC  GTGAGCTGTCTCCGGGATCT  TCGGGGCTCTGCCCGGCC  ATACCCAAGTCACCATCTG  ----TGTTCCACAGGATATC

MPU28721  TCGCCCCTTTGAAAGATCC  GGACTCCTTCCGAGCTTCCA  TCCGCCTCTGGCCAGTCAC  CTGAAGTCCACGCACAGCGG  CAAGATCGACTATATCGCAG
MSU28720  TCGCCCCTTTGAAAGATCC  GGACTCCTTCCGAGCTTCCA  TCCGCCTCTGGCCAGTCAC  CTGAAGTCCACGCACAGCGG  CAAGATCGACTACATCGCAG
GCU28961  TCGCCCCTCTGAAAGATCC  GGACTCCTTCCGAGCTTCCA  TCCGCTCTTGCCCAACCAT  CTGAAGTCCAAGCATGGCGG  CAAATCGACTACATCGCAG

MPU28721  GGCAAGGTGGCCTTGCTAGG  CCGTACTCATCCCCACGGT  CCTATCCCCTATCCCCTTTC  CCC-TCGTGTACCCACAGT  CTACCCACACCCATCCATT
MSU28720  GCGA--GTGGCTTGCTAGG  CCGTGTCTGTCACCCACGGT  CCTAGCCCCTATCCCCTTTC  CCCCTCGTGTACCCACAGT  CTGCCACACCCATCCATT
GCU28961  GCGA--GTGTTCTTGCTAGG  CCGTGCCCGTTCCC-ACTGT  CAGGGCCGCCATCCCGTGT  CCC-----TTTTTCGT  GTCACCCACACCCACCCCTC

MPU28721  CTTTCTTAACTCTGACTC  TTCCTCCTTGGTTTCTCACT  GCCTTGGACGCTTGTTCACC  CCGGATGAACCTCGTAGGCG  TCTCCCTTCCCTGCTTGGTA
MSU28720  CTTTCTTCAACTCTGACAC  TTCCTCCTTGGTTTCTCACT  GCCTTGGACGCTTGTTCACC  CCGGATGAACCTATGTAGGAG  TCTCCCTTCCCTGCTAGGTA
GCU28961  CTTTCTTGACTCTCCCAAG  TTCCT----GTTCTCTCT  GCCTTGGTCCCATATTCACC  CCGGATGA-CTGCGGAGTCT  CCCACCCTCTGACTCTGCT

MPU28721  CCCTAAGG----TGCCCTCG  GTGCTTGTTCGTAGAGACGA  ACTCTGCTCT
MSU28720  CCCTAAGGCATCTGCCCTCG  GTGCTTGTTCGTAGAGACGA  ACTCTGCTCT
GCU28961  CTCAAAGC-----CT  GTCCTAC---TAGAGAGGA  ACTCTGCTCT

;
END;
BEGIN TREES;
  TREE tree1 = (MPU28721, (MSU28720,GCU28961));
  TREE tree2 = (MSU28720, (MPU28721,GCU28961));
END;
BEGIN NOTES;
  PICTURE TAXON=3 FORMAT=GIF SOURCE=FILE
  PICTURE=a_rodent.gif
END;
```

The major blocks of data that the file format permits are TAXA, CHARACTERS, UNALIGNED, DISTANCES, SETS, ASSUMPTIONS, CODONS, TREES and NOTES. Only a few of these are shown above and each permits many other options. Note that the file format permits things such as the phylogeny (or tree) of a group of species to be stored, pictures of the organisms to be stored or referenced, along with many other capabilities.

5.8 PHYLIP

The PHYLIP programs are also very popular and other programs have incorporated the sequence format used by these programs. There are two formats that can be used, an interleaved and a sequential format. The `phylip-interleaved` format begins with two numbers on the first line. The first number gives the number of taxa or different sequences in the file. The second number gives the overall length of the sequences. On the next line the sequence information begins preceded by a sequence title of no more than 10 characters. The APRT sequences in this format (interleaved) would be

```

MPU28721 -----CCTG CGGATACTCA
MSU28720 -----TCGGG ATTGACGTGA ATTTAGCGTG CTGATACCTA
GCU28961 CCTCCGCCCT TGTTCCTGGG ACAGGCTTGA CCCTAGCCAG TTGACACCTC

CCTCCTCCTT GTCTCCTACA AGCACGCGGC CATGTCCGAG TCTGAGTTGA
CCTCCTCCTT GCCTCCTACA CGCACGCGGC CATGTCCGAA CCTGAGTTGA
ACCTCCGCCCT TTCTCT-CA CGCACGCGGC CATGGCGGAA CCCGAGTTGC

AACTGGTGGC GCGGCGCATC CGCAGCTTCC CCGACTTCCC CATCCCGGGC
AACTGGTGGC GCGGCGCATC CGCAGCTTCC CCGACTTCCC AATCCCGGGC
AGCTGGTGGC GCGGCGCATC CGCAGCTTCC CCGACTTCCC CATCCCGGGC

GTGCTGTTCA GGTGCGGTCA CGAGCCGGCG AGGCGTTGGC GCCGTA CTCT
GTGCTGTTCA GGTGCGGTCA CGAGCCGGCG AGGCGTTGGC GCCGTA CTCT
GTGCTGTTCA GGTGCGGTCA CGAGCCGGCG AGGCGTTGGC GCTGCGTCTC

CATCCC-CCG GCGCAGGCGC GTGGGCAGCC TTGGGGATCT TCGGGGGCCT
CATCCC-CCG GCGCAGGCGC GTAGGCAGCC TCGGGGATCT TCGGGGGCCT
CAGCCCTCCG GCGCAGGCGC GTGAGCTGTC TCCGGGATCT TCGGGGGCCT

CTGCCCGGCC ACACGCGG-T CACTCTCCTG TCCTTGTTC CAGGGATATC
CTGCCCGGCC ACACGCGGGT CACTCTCCTG TCCTTGTTC CAGGGATATC
CGGCCAGCC ATACCCAAGT CACCATCCTG ----TGTTCC CAGGGATATC

TCGCCCTCTT TGAAAGATCC GGACTCCTTC CGAGCTTCCA TCCGCTCTCT
TCGCCCTCTT TGAAAGATCC GGACTCCTTC CGAGCTTCCA TCCGCTCTCT
TCGCCCTCTT TGAAAGATCC GGACTCCTTC CGAGCTTCCA TCCGCTCTCT

GGCCAGTCAC CTGAAGTCCA CGCACAGCGG CAAGATCGAC TATATCGCAG
GGCCAGTCAC CTGAAGTCCA CGCACAGCGG CAAGATCGAC TATATCGCAG
GGCCAACCAT CTGAAGTCCA AGCATGGCGG CAAAATCGAC TATATCGCAG

GGCAAGTGGC CTTTGCTAGG CCGTACTCAT CCCCCACGGT CCTATCCCCT
GCGA--GTGG CTTTGCTAGG CCGTACTCAT CCCCCACGGT CCTATCCCCT
GCGA--GTGT TCTTGCTAGG CCGTACTCAT CCCCCACGGT CCTATCCCCT

ATCCCCTTTC CCC-TCGTGT CACCCACAGT CTACCCACACA CCCATCCATT
ATCCCCTTTC CCCCTCGTGT CACCCACAGT CTACCCACACA CCCATCCATT
ATCCCCTTTC CCC----- --TTTTTCGT GTCACCCACA CCCACCCCTC

CTTTCTTAA CCTCTGACTC TTCCTCCTTG GTTTCTCACT GCCTTGGAGC
CTTTCTTAA CCTCTGACTC TTCCTCCTTG GTTTCTCACT GCCTTGGAGC
CTTTCTTAA CCTCTGACTC TTCCTCCTTG GTTTCTCACT GCCTTGGAGC

CTTGTTACC CCGGATGAAC TCCGTAGGCG TCTCCCTTCC CTGCTGGTAA
CTTGTTACC CCGGATGAAC TATGTAGGAG TCTCCCTTCC CTGCTGGTAA
CATATTACC CCGGATGA-C TCGGAGTCT CCCACCTTCT GACCTCTGCT

CCCTAAGG-- --TGCCCTCG GTGCTTGTTC GTAGAGACGA ACTCTGCTCT
CCCTAAGGCA TCTGCCCTCG GTGCTTGTTC GTAGAGACGA ACTCTGCTCT
CTCAAAGC-- -----CT GTCCTAC-- -TAGAGAGGA ACTCTGCTCT

```

5.9 ASN

The Abstract Syntax Notation (`asn`) format is intended to be read by computer rather than humans. It was developed at NCBI. It is included here to demonstrate the broad variety of sequence formats in use. For just the *Mus spicilegus* sequence (complete entry) it would be

```

Seq-entry ::= set {
  level 1 ,
  class nuc-prot ,
  descr {
    title "Mus spicilegus adenine phosphoribosyltransferase (APRT) gene, and
translated products" ,
    update-date
      std {
        year 1995 ,
        month 6 ,
        day 16 } ,
    source {
      org {
        taxname "Mus spicilegus" ,
        common "steppe mouse" ,
        db {
          {

```

```

        db "taxon" ,
        tag
        id 10103 } } ,
    orgname {
        name
        binomial {
            genus "Mus" ,
            species "spicilegus" } ,
        lineage "Eukaryotae; mitochondrial eukaryotes; Metazoa; Chordata;
Vertebrata; Eutheria; Rodentia; Sciurognathi; Myomorpha; Muridae; Murinae;
Mus" ,
        gcode 1 ,
        mgcode 2 } } } ,
    pub {
        pub {
            gen {
                serial-number 1 } ,
            gen {
                cit "Unpublished" ,
                authors {
                    names {
                        std {
                            {
                                name
                                name {
                                    last "Fieldhouse" ,
                                    initials "D." } } } ,
                            {
                                name
                                name {
                                    last "Golding" ,
                                    initials "G.B." } } } } } ,
                title "Rates of substitution in closely related rodent species" } } } ,
        pub {
            pub {
                gen {
                    serial-number 2 } ,
                sub {
                    authors {
                        names {
                            std {
                                {
                                    name
                                    name {
                                        last "Fieldhouse" ,
                                        initials "D." } } } } } ,
                    imp {
                        date
                        std {
                            year 1995 ,
                            month 6 ,
                            day 7 } ,
                        pub
                        str "Dan Fieldhouse, Biology, McMaster University, 1280 Main
Street West, Hamilton, ON, L8S 4K1, Canada" } ,
                        medium other } } } } ,
        seq-set {
            seq {
                id {
                    genbank {
                        name "MSU28720" ,
                        accession "U28720" } ,
                        gi 881575 } ,
                    descr {
                        title "Mus spicilegus adenine phosphoribosyltransferase (APRT) gene,
complete cds." ,
                        genbank {
                            source "Steppe mouse." ,
                            div "ROD" } ,
                        create-date
                        std {
                            year 1995 ,
                            month 6 ,
                            day 28 } ,
                        molinfo {
                            biomol genomic } } ,
                    inst {
                        repr raw ,
                        mol dna ,
                        length 2117 ,

```



```

seq-data
  ncbi2na 'DA8F86E0FC9B9E317175D7E5D711919A53B58178BE01EBA669935927D56
1F54356A6E7BD2B9AD189698A6FA65B19D355699299B2925DAA37E6A97795A5119AB4775ED7EF5
4A8CDD9577E0215A1D7D627D4D65DFA52D1782D46449A42361C4D9298BA5F9CA5B9DB5546B5C95
7355FD55DBB4544B7954454D4F7F7D05DE11F5D7EBD74797E867EF455A381CECA2DD5F579CAC57
0A4DE576B9FBD722181DE77B5FBB5205297577FCA91027A524D784929EA215E8175238684D7E7C
AAC977A8E07231C00F2B05FAFAA6E9B97A9217425EB27D2A9EFDD54A1C45AA4DFD5FBD5D1109FB
BC041E7B717A7539789F4804572A49E0ED4528BB522A2AEA45522048BA57AC2E74A85121FF971F
47D73EB157A539D480F2A4ECECD7D51849C8E793F80AE90894532BA5789EF48292B28D5429E23A
5152C94D06F7DC9EB2D242172EF5C90BBE17654C7E97F23D5395749D4D5105F575F15C12B721D4
AA7D7BFA57D5C9D289EA6EA7BB9D35A012A82796A551EED25D73DDE8B3A82B0989EEEC8A0A92AD
F3469C52EDCA2C0EEAE7488AF884FAB4AFC45152019D8A72A2BA51FBD93721DDDF11C7D7B7929E
27A035202397C815A9222EB4FBA385D7A5128AC0814150840487D02A5295ED7AB9E1C24089F831
77F77B57D554A053BF9A5EE379E45275A9E0BAE8BBB897AE89E176782A4A88A7285CC5BDF7775D
4B3878A27A723AD11579D5249D4A079FAE9D254A65C2E17FB89C52595FFB8BBC0'H } ,
  annot {
  {
    data
      ftable {
      {
        data
          gene {
            locus "APRT" } ,
          location
          mix {
            int {
              from 66 ,
              to 145 ,
              id
                gi 881575 } ,
            int {
              from 277 ,
              to 383 ,
              id
                gi 881575 } ,
            int {
              from 1354 ,
              to 1487 ,
              id
                gi 881575 } ,
            int {
              from 1674 ,
              to 1752 ,
              id
                gi 881575 } ,
            int {
              from 1859 ,
              to 2001 ,
              id
                gi 881575 } } } } } } } ,
    seq {
      id {
        gi 881576 } ,
      descr {
        title "adenine phosphoribosyltransferase" ,
        molinfo {
          tech concept-trans } } ,
      inst {
        repr raw ,
        mol aa ,
        length 180 ,
        seq-data
          iupacaa "MSEPELKLIVARRIRSFDFPIPGVLFDRDISPLLKDPDSFRASIRLLASHLKSTHSGKID
YIAGLDSRGLFPGPSLAQELGVGCVLIRKQKLPPTVSASYSLEYGKAELEIQKDALEPGQRRVVIVDDLLATGGTMF
AACDLLHLQLRAEVVECVSLVELTSLRGRERLGPPIFFSLLQYD" } ,
        annot {
          {
            data
              ftable {
              {
                data
                  prot {
                    name {
                      "adenine phosphoribosyltransferase" } ,
                    ec {
                      "2.4.2.7" } } ,
                  location
                    whole
                      gi 881576 } } } } } } ,
          annot {

```

```
{
  data
  ftable {
    {
      data
      cdregion {
        frame one ,
        code {
          id 1 } } ,
      comment "purine salvage enzyme" ,
      product
      whole
      gi 881576 ,
      location
      mix {
        int {
          from 66 ,
          to 145 ,
          id
          gi 881575 } ,
        int {
          from 277 ,
          to 383 ,
          id
          gi 881575 } ,
        int {
          from 1354 ,
          to 1487 ,
          id
          gi 881575 } ,
        int {
          from 1674 ,
          to 1752 ,
          id
          gi 881575 } ,
        int {
          from 1859 ,
          to 2001 ,
          id
          gi 881575 } } ,
      xref {
        {
          data
          gene {
            locus "APRT" } } } } } } }
```

5.10 BSML format

The Bioinformatic Sequence Markup Language is another format that is rapidly gaining popularity and is designed to be read mostly by computers (viewers are developed to present human readable forms). The BSML format is based on the XML - extended markup language. XML is heralded as the replacement for HTML (hypertext markup language — basically the format used on the internet and read by your favorite internet browser). The primary feature that makes XML an improvement on HTML is that XML is an extendable language. New features and new objects can be defined within XML itself and does not require an entire rewriting of the language (as would HTML to add new features). BSML is a language set up with objects (data structures) predefined that are useful for bioinformatic research. The BSML data specification (DTD) was created to solve the data management problems and yet to include support for complicated structures such as sequence annotations, sequence restriction enzyme digestions, phylogenies and so on.

5.11 PDB file format

Of quite a different nature than the files listed above, this file is meant to store the three dimensional location of atoms within a molecule. This data structure file is perhaps the most difficult to maintain or to alter because programs must parse these files very precisely in order to produce three dimensional structures of the encoded molecules. There are two formats for the three dimensional structure data stored in the PDB. The first is the rather old flat file format described in detail at <http://www.wwpdb.org/docs.html> and the second is the mmCIF format (the macromolecular Crystallographic Information File) described in detail at <http://www.sdsc.edu/pb/cif/papers/methenz.html>.

The flat file format (with many rows deleted — indicated by the dots in the center of a row) looks as follows ...

```

HEADER      OXIDOREDUCTASE                      26-MAY-98   ZOCC
TITLE       BOVINE HEART CYTOCHROME C OXIDASE AT THE FULLY OXIDIZED
TITLE       2 STATE
COMPND      MOL_ID: 1;
COMPND      2 MOLECULE: CYTOCHROME C OXIDASE;
COMPND      3 CHAIN: A, B, C, D, E, F, G, H, I, J, K, L, M, N, O, P, Q,
COMPND      4 R, S, T, U, V, W, X, Y, Z;
COMPND      5 SYNONYM: FERROCYTOCHROME C\OXYGEN OXIDOREDUCTASE;
COMPND      6 EC: 1.9.3.1;
COMPND      7 OTHER_DETAILS: THIS ENZYME IS A HYBRID PROTEIN COMPLEX AND
COMPND      8 IS A HOMODIMER. ONE MONOMER IS COMPOSED OF 13 DIFFERENT
COMPND      9 SUBUNITS AND SEVEN METAL CENTERS, HEME A, HEME A3, CUA,
COMPND     10 CUB, MG, NA AND ZN.
SOURCE      MOL_ID: 1;
SOURCE      2 ORGANISM_SCIENTIFIC: BOS TAURUS;
SOURCE      3 ORGANISM_COMMON: BOVINE;
SOURCE      4 ORGAN: HEART;
SOURCE      5 TISSUE: HEART MUSCLE;
SOURCE      6 ORGANELLE: MITOCHONDRION
KEYWDS      OXIDOREDUCTASE, CYTOCHROME (C)-OXYGEN, CYTOCHROME C
KEYWDS      2 OXIDASE
EXPDTA      X-RAY DIFFRACTION
AUTHOR      T.TSUKIHARA,M.YAO
REVDAT      1   13-JAN-99 ZOCC   0
REMARK      1
REMARK      1 REFERENCE 1
REMARK      1 AUTH   S.YOSHIKAWA,K.SHINZAWA-ITOH,R.NAKASHIMA,R.YAONO,
REMARK      1 AUTH 2 E.YAMASHITA,N.INOUE,M.YAO,M.J.FEI,C.P.LIBEU,
REMARK      1 AUTH 3 T.MIZUSHIMA,H.YAMAGUCHI,T.TOMIZAKI,T.TSUKIHARA
REMARK      1 TITL  REDOX-COUPLED CRYSTAL STRUCTURAL CHANGES IN BOVINE
REMARK      1 TITL 2 HEART CYTOCHROME C OXIDASE
REMARK      1 REF   SCIENCE                      V. 280  1723 1998
REMARK      1 REFN  ASTM SCIEAS  US ISSN 0036-8075          0038
REMARK      1 REFERENCE 2

```

.....

```

REMARK      2
REMARK      2 RESOLUTION. 2.3  ANGSTROMS.
REMARK      3
REMARK      3 REFINEMENT.
REMARK      3 PROGRAM       : X-PLOR 3.84
REMARK      3 AUTHORS        : BRUNGER
REMARK      3
REMARK      3 DATA USED IN REFINEMENT.
REMARK      3 RESOLUTION RANGE HIGH (ANGSTROMS) : 2.3
REMARK      3 RESOLUTION RANGE LOW  (ANGSTROMS) : 15.
REMARK      3 DATA CUTOFF             (SIGMA(F)) : 2.0
REMARK      3 DATA CUTOFF HIGH        (ABS(F)) : 100000.0
REMARK      3 DATA CUTOFF LOW         (ABS(F)) : 0.1
REMARK      3 COMPLETENESS (WORKING+TEST) (%) : 88.88
REMARK      3 NUMBER OF REFLECTIONS       : 278049
REMARK      3
REMARK      3 FIT TO DATA USED IN REFINEMENT.
REMARK      3 CROSS-VALIDATION METHOD           : THROUGHOUT
REMARK      3 FREE R VALUE TEST SET SELECTION : RANDOM
REMARK      3 R VALUE                          (WORKING SET) : 0.209

```

.....

```

REMARK      4
REMARK      4 ZOCC COMPLIES WITH FORMAT V. 2.2, 16-DEC-1996
REMARK      6
REMARK      6 THIS ENZYME IS A MULTI-COMPONENT PROTEIN COMPLEX AND IS A
REMARK      6 HOMODIMER. EACH MONOMER IS COMPOSED OF 13 DIFFERENT
REMARK      6 SUBUNITS AND SIX METAL CENTERS: HEME A, HEME A3, CUA, CUB,
REMARK      6 MG, NA, AND ZN. THE SIDE CHAINS OF H 240 AND Y244 OF
REMARK      6 MOLECULES A AND N ARE LINKED TOGETHER BY A COVALENT BOND.
REMARK      6 THE ELECTRON DENSITY OF REGION FROM D(Q) 1 TO D(Q) 3,
REMARK      6 E(R) 1 TO E(R) 4, H(U) 1 TO H(U) 6, J(W) 59, K(X) 1 TO
REMARK      6 K(X) 5, K(X) 53 TO K(X) 54 AND M(Z) 41 TO M(Z) 43 IS
REMARK      6 NOISY AND THE MODEL OF THIS REGION HAS AMBIGUITY.
REMARK     200
REMARK     200 EXPERIMENTAL DETAILS
REMARK     200 EXPERIMENT TYPE                : X-RAY DIFFRACTION
REMARK     200 DATE OF DATA COLLECTION        : MAY-1996
REMARK     200 TEMPERATURE                    (KELVIN) : 283
REMARK     200 PH                            : 6.8

```

Elementary Sequence Analysis

edited by Brian Golding, Dick Morton and Wilfried Haerty August 2015

99

REMARK 200 NUMBER OF CRYSTALS USED : 32
REMARK 200
REMARK 200 SYNCHROTRON (Y/N) : Y
REMARK 200 RADIATION SOURCE : PHOTON FACTORY
REMARK 200 BEAMLINE : 6A, 6B
REMARK 200 X-RAY GENERATOR MODEL : NULL
REMARK 200 MONOCHROMATIC OR LAUE (M/L) : M

.....

DBREF	20CC	A	1	514	SWS	P00396	COX1_BOVIN	1	514
DBREF	20CC	B	1	227	SWS	P00404	COX2_BOVIN	1	227
DBREF	20CC	C	1	261	SWS	P00415	COX3_BOVIN	1	261
DBREF	20CC	D	4	147	SWS	P00423	COX4_BOVIN	26	169
DBREF	20CC	E	5	109	SWS	P00426	COXA_BOVIN	5	109
DBREF	20CC	F	1	98	SWS	P11949	COXB_BOVIN	1	98

.....

SEQRES	1	A	514	MET	PHE	ILE	ASN	ARG	TRP	LEU	PHE	SER	THR	ASN	HIS	LYS
SEQRES	2	A	514	ASP	ILE	GLY	THR	LEU	TYR	LEU	LEU	PHE	GLY	ALA	TRP	ALA
SEQRES	3	A	514	GLY	MET	VAL	GLY	THR	ALA	LEU	SER	LEU	LEU	ILE	ARG	ALA
SEQRES	4	A	514	GLU	LEU	GLY	GLN	PRO	GLY	THR	LEU	LEU	GLY	ASP	ASP	GLN
SEQRES	5	A	514	ILE	TYR	ASN	VAL	VAL	VAL	THR	ALA	HIS	ALA	PHE	VAL	MET
SEQRES	6	A	514	ILE	PHE	PHE	MET	VAL	MET	PRO	ILE	MET	ILE	GLY	GLY	PHE
SEQRES	7	A	514	GLY	ASN	TRP	LEU	VAL	PRO	LEU	MET	ILE	GLY	ALA	PRO	ASP
SEQRES	8	A	514	MET	ALA	PHE	PRO	ARG	MET	ASN	ASN	MET	SER	PHE	TRP	LEU

.....

SEQRES	1	B	227	MET	ALA	TYR	PRO	MET	GLN	LEU	GLY	PHE	GLN	ASP	ALA	THR
SEQRES	2	B	227	SER	PRO	ILE	MET	GLU	GLU	LEU	LEU	HIS	PHE	HIS	ASP	HIS
SEQRES	3	B	227	THR	LEU	MET	ILE	VAL	PHE	LEU	ILE	SER	SER	LEU	VAL	LEU
SEQRES	4	B	227	TYR	ILE	ILE	SER	LEU	MET	LEU	THR	THR	LYS	LEU	THR	HIS
SEQRES	5	B	227	THR	SER	THR	MET	ASP	ALA	GLN	GLU	VAL	GLU	THR	ILE	TRP
SEQRES	6	B	227	THR	ILE	LEU	PRO	ALA	ILE	ILE	LEU	ILE	LEU	ILE	ALA	LEU
SEQRES	7	B	227	PRO	SER	LEU	ARG	ILE	LEU	TYR	MET	MET	ASP	GLU	ILE	ASN

.....

HET	HEA	A	515	60	PROTOPORPHYRIN IX CONTAINS FE(II)
HET	HEA	A	516	60	PROTOPORPHYRIN IX CONTAINS FE(II)
HET	CU	A	517	1	
HET	MG	A	518	1	
HET	NA	A	519	1	
HET	PER	A	520	2	
HET	CU	B	228	1	
HET	CU	B	229	1	
HET	ZN	F	99	1	
HET	HEA	N	515	60	PROTOPORPHYRIN IX CONTAINS FE(II)
HET	HEA	N	516	60	PROTOPORPHYRIN IX CONTAINS FE(II)
HET	CU	N	517	1	
HET	MG	N	518	1	
HET	NA	N	519	1	
HET	PER	N	520	2	
HET	CU	O	228	1	
HET	CU	O	229	1	
HET	ZN	S	99	1	

HETNAM HEA HEME-A
HETNAM CU COPPER (II) ION
HETNAM MG MAGNESIUM ION
HETNAM NA SODIUM ION
HETNAM PER PEROXIDE ION
HETNAM ZN ZINC ION

FORMUL 27 HEA 4 (C49 H62 N4 O6 FE1)
FORMUL 28 CU 6 (CU1 2+)
FORMUL 29 MG 2 (MG1 2+)
FORMUL 30 NA 2 (NA1 1+)
FORMUL 31 PER 2 (O2 2-)
FORMUL 32 ZN 2 (ZN1 2+)

HELIX	1	1	PHE	A	2	TRP	A	6	1	5
HELIX	2	2	HIS	A	12	LEU	A	41	1	30
HELIX	3	3	ASP	A	51	ILE	A	87	1	37
HELIX	4	4	PRO	A	95	MET	A	117	1	23
HELIX	5	5	ALA	A	141	ASN	A	170	1	30
HELIX	6	6	LEU	A	183	ASN	A	214	1	32

.....

SHEET 1 A 5 LEU B 116 SER B 120 0

```

SHEET 2 A 5 TYR B 105 TYR B 110 -1 N TYR B 110 O LEU B 116
SHEET 3 A 5 LEU B 95 HIS B 102 -1 N HIS B 102 O TYR B 105
SHEET 4 A 5 ILE B 150 SER B 156 1 N ARG B 151 O LEU B 95
SHEET 5 A 5 ASN B 180 LEU B 184 -1 N LEU B 184 O ILE B 150
SHEET 1 B 3 VAL B 142 PRO B 145 0
SHEET 2 B 3 ILE B 209 VAL B 214 1 N GLU B 212 O VAL B 142
SHEET 3 B 3 GLY B 190 GLY B 194 -1 N GLY B 194 O ILE B 209
SHEET 1 C 2 HIS B 161 VAL B 165 0
SHEET 2 C 2 LEU B 170 ALA B 174 -1 N ALA B 174 O HIS B 161
SHEET 1 D 3 ASN F 47 SER F 51 0
SHEET 2 D 3 GLY F 86 PRO F 93 1 N LYS F 90 O ASN F 47
SHEET 3 D 3 GLN F 80 CYS F 82 -1 N CYS F 82 O GLY F 86
SHEET 1 E 2 LYS F 55 CYS F 60 0
SHEET 2 E 2 ILE F 70 HIS F 75 -1 N LEU F 74 O ARG F 56
SHEET 1 F 5 LEU O 116 SER O 120 0
SHEET 2 F 5 TYR O 105 TYR O 110 -1 N TYR O 110 O LEU O 116
SHEET 3 F 5 LEU O 95 HIS O 102 -1 N HIS O 102 O TYR O 105
SHEET 4 F 5 ILE O 150 SER O 156 1 N ARG O 151 O LEU O 95
SHEET 5 F 5 ASN O 180 LEU O 184 -1 N LEU O 184 O ILE O 150
SHEET 1 G 3 VAL O 142 PRO O 145 0
SHEET 2 G 3 ILE O 209 VAL O 214 1 N GLU O 212 O VAL O 142
SHEET 3 G 3 GLY O 190 GLY O 194 -1 N GLY O 194 O ILE O 209
SHEET 1 H 2 HIS O 161 VAL O 165 0
SHEET 2 H 2 LEU O 170 ALA O 174 -1 N ALA O 174 O HIS O 161
SHEET 1 I 3 ASN S 47 SER S 51 0
SHEET 2 I 3 GLY S 86 PRO S 93 1 N LYS S 90 O ASN S 47
SHEET 3 I 3 GLN S 80 CYS S 82 -1 N CYS S 82 O GLY S 86
SHEET 1 J 2 LYS S 55 CYS S 60 0
SHEET 2 J 2 ILE S 70 HIS S 75 -1 N LEU S 74 O ARG S 56
SSBOND 1 CYS H 29 CYS H 64
SSBOND 2 CYS H 39 CYS H 53
SSBOND 3 CYS U 29 CYS U 64
SSBOND 4 CYS U 39 CYS U 53
LINK FE HEA A 515 NE2 HIS A 61
LINK FE HEA A 515 NE2 HIS A 378
LINK FE HEA A 516 NE2 HIS A 376
LINK FE HEA A 516 O1 PER A 520
LINK CU CU A 517 ND1 HIS A 240
LINK CU CU A 517 NE2 HIS A 290
    
```

.....

```

ATOM 1 N MET A 1 55.242 340.693 224.088 1.00 68.90 N
ATOM 2 CA MET A 1 54.908 339.282 224.487 1.00 71.09 C
ATOM 3 C MET A 1 54.673 338.307 223.329 1.00 66.66 C
ATOM 4 O MET A 1 55.350 337.285 223.238 1.00 67.66 O
ATOM 5 CB MET A 1 53.723 339.248 225.450 1.00 79.30 C
ATOM 6 CG MET A 1 54.110 339.452 226.915 1.00 87.90 C
ATOM 7 SD MET A 1 55.300 338.229 227.515 1.00 97.07 S
ATOM 8 CE MET A 1 54.166 336.799 228.014 1.00 96.59 C
ATOM 9 N PHE A 2 53.673 338.579 222.494 1.00 61.89 N
ATOM 10 CA PHE A 2 53.412 337.739 221.322 1.00 56.50 C
ATOM 11 C PHE A 2 54.569 337.917 220.303 1.00 53.31 C
ATOM 12 O PHE A 2 55.076 336.947 219.739 1.00 53.84 O
ATOM 13 CB PHE A 2 52.077 338.127 220.683 1.00 55.21 C
ATOM 14 CG PHE A 2 51.737 337.334 219.459 1.00 54.54 C
ATOM 15 CD1 PHE A 2 51.050 336.138 219.565 1.00 55.24 C
ATOM 16 CD2 PHE A 2 52.126 337.775 218.200 1.00 55.62 C
ATOM 17 CE1 PHE A 2 50.756 335.388 218.432 1.00 58.99 C
ATOM 18 CE2 PHE A 2 51.839 337.035 217.059 1.00 57.84 C
ATOM 19 CZ PHE A 2 51.155 335.840 217.171 1.00 58.36 C
ATOM 20 N ILE A 3 55.010 339.158 220.116 1.00 47.37 N
    
```

.....

```

HETATM 4147 CU CU A 517 67.173 310.978 190.358 1.00 16.27 CU
HETATM 4148 MG MG A 518 62.605 315.176 179.115 1.00 19.26 MG
HETATM 4149 NA NA A 519 42.250 318.661 179.405 1.00 26.18 NA
HETATM 4150 O1 PER A 520 64.953 309.772 191.618 1.00 10.28 O
HETATM 4151 O2 PER A 520 65.314 311.367 191.209 1.00 15.28 O
ATOM 4152 N MET B 1 50.114 302.768 167.666 1.00 37.69 N
ATOM 4153 CA MET B 1 49.455 303.851 168.484 1.00 36.15 C
ATOM 4154 C MET B 1 48.691 303.239 169.660 1.00 34.30 C
ATOM 4155 O MET B 1 48.549 302.024 169.753 1.00 34.54 O
ATOM 4156 CB MET B 1 48.490 304.694 167.641 1.00 35.38 C
    
```

.....

```

ATOM 28892 O SER Z 43 155.003 299.215 171.486 1.00 99.03 O
    
```

```
ATOM 28893 CB SER Z 43 152.512 300.170 170.193 1.00 99.03 C
ATOM 28894 OG SER Z 43 151.462 300.982 169.639 1.00 99.02 O
ATOM 28895 OXT SER Z 43 154.021 299.630 173.431 1.00 99.03 O
TER 28896 SER Z 43
CONNECT 351 350 4149
CONNECT 474 472 473 4027
CONNECT 1836 1835 1838 4147
CONNECT 2239 2237 2238 4147
CONNECT 2249 2247 2248 4147
```

.....

```
CONNECT264472635326446
CONNECT265522625626551
MASTER 370 0 18 98 30 0 2 928870 26 308 292
END
```

The file begins with the keywords, HEADER, TITLE, COMPND, SOURCE, KEYWDS, EXPDTA, that describe the nature of the molecule to which this file pertains. The keywords AUTHOR and REVDAT give the authors responsible for this file and its revision history. The REMARK keyword indicates descriptive entries about the molecular structure (they are numbered according to their category). These remarks provide enormous detail regarding the structure. DBREF supplies cross references to entries of this molecule in other databases. SEQRES is the beginning of the actual sequence information of the molecule. Note the molecule can consist of multiple chains; in this case labelled chain A – chain Z. The HET, HETNAM and FORMUL fields contain information about atoms/molecules that are associated with the molecule in question (for HET, the fields here are a letter code for each “HET” atom(s), the letter identifying the chain, insertion code, number of records with a HET entry, and some descriptive text), their chemical name (in this case a HEME group, copper ion, ...) and their chemical formula. The HELIX, SHEET, and TURN (not shown above) give information about the secondary structure of the molecule. Information about connections in the molecule are shown by SSBOND, LINK, HYDBND, SLTBRG, and CISPEP (the last three not shown in the above structure).

And much more information is provided by other fields too numerous to list here. The business end is in the ATOM field. This contains a numbered list of atoms (in this case 28,895 of them), the atom name, the (amino acid) residue name, the chain identifier number, the residue sequence number, then three numbers that describe the x, y, z coordinates of this atom in Angstrom units, an occupancy number, a temperature factor and finally an element symbol.

The TER field indicates the end a section. The CONECT section provides further information on chemical connectivity. The MASTER and END fields are used to describe the number of records of different types and to signal the end of the file.