THE SMALL BIOINFORMATICIAN

THE GENETIC CODE (in RNA, T=U)

	T	C	A	G	
T	TTT Phe (F)	TCT Ser (S)	TAT Tyr (Y)	TGT Cys (C)	T
	TTC "	TCC "	TAC "	TGC "	C
	TTA Leu (L)	TCA "	TAA STOP	TGA STOP	A
	TTG "	TCG "	TAG STOP	TGG Trp (W)	G
C	CTT Leu (L)	CCT Pro (P)	CAT His (H)	CGT Arg (R)	T
	CTC "	CCC "	CAC "	CGC "	C
	CTA "	CCA "	CAA Gln (Q)	CGA "	A
	CTG "	CCG "	CAG "	CGG "	G
A	ATT Ile (I)	ACT Thr (T)	AAT Asn (N)	AGT Ser (S)	T
	ATC "	ACC "	AAC "	AGC "	C
	ATA "	ACA "	AAA Lys (K)	AGA Arg (R)	A
	ATG Met (M)	ACG "	AAG "	AGG "	G
G	GTT Val (V)	GCT Ala (A)	GAT Asp (D)	GGT Gly (G)	T
	GTC "	GCC "	GAC "	GGC "	C
	GTA "	GCA "	GAA Glu (E)	GGA "	A
	GTG "	GCG "	GAG "	GGG "	G

AMINO ACIDS: NOT POLAR - POLAR NOT CHARGED - ACID - BASIC

N=aNy base	S=G/C (Strong-3H bonds)		W=A/T (Weak-2H bonds)
R=A/G (puRine)	Y=C/T (pYrimidine)	K=G/T (Keto)	M=A/C (aMino)
B=C/G/T	D=A/G/T	H=A/C/T	V=A/C/G

HUMAN GENOME (2001-2006)
Aploid size=**3.2 Gb**
CHROMOSOMES 22 (×2) Autosomes
Largest: 1 (246 Mb) Smallest: 21 (45 Mb)
Sex chromosomes: X (155 Mb) — Y (50 Mb)
mtDNA (16,569 bp)
32% of DNA: PROTEIN-CODING GENES
About 20-25,000 Models: 24,764 Known: **18,245**
Mean: 760/chromosome (33/band 550, 1/175 kb)
Gene Mean size=**57 kb**
From 0.1 kb (*KRTAP22-1*) to 2,305 kb (*CNTNAP2*)
Mean intergenic distance: 119 kb
mRNA Mean size **2.9 kb**
0.2 kb 5´UTR + 1.6 kb CDS (540 AA) +
1.0 kb 3´UTR
Exons Number: 1 to 363 (*TTN*) Mean=11
Size: Mean=**280 bp** (2 to 22,753 bp)
Total: 5.3% of gene sequences
Introns Number: 0 to 362 (*TTN*) Mean=10
Size: Mean=**5.9 kb** (31 to 740,920 bp)
rRNA 28S=4,718 bp 18S=1,874 bp

Splice Junctions
Exon 1 (C/A)AG**GT**(A/G)AGT...(T/C)$_{11}$NC**AG**G Exon 2 [Polyadenylation signal: AATAAA]
↑-------Intron--------↑

Eucaryote promoters

-100	-75	-50	-25	
GTGG(AAA/TTT)G	GG(C/T)CAATCT	CNCCGCCC	TATAA	gccGCCRCCAUGG
Enhancer core seq.	"CAT" box	Tjian GC box	Hogness box	Kozak seq.

Procaryote promoters

-35	-10	
TTGACA	TATAAT	AAGGAGGU(N)$_{3-6}$ AUG
	Pribnow box	Shine-Dalgarno seq.

Translation start:

BLASTN (query nt/database nt) Program Advanced Options (Default) www.ncbi.nlm.nih.gov
-gapopen Cost to open a gap = 5 **-gapextend** Cost to extend a gap = 2 **-penalty** Penalty for a mismatch = -3
-reward Reward for a match = 1 **-evalue** Expectation value (E) [Real] = 10.0 **-word_size** Word size = 11
-num_descriptions Number of one-line descriptions [Int] = 100 **-num_alignments** Number of alignments to show [Int] = 100
BLASTP (aa/aa), BLASTX (nt/aa), TBLASTN (aa/nt) Programs Advanced Options (Default)
-gapopen Cost to open a gap = 11 **-gapextend** Cost to extend a gap = 1 **-word_size** Word size = 3

ASCII n.	Sigla WORD	TexEdit	UNIX	FileMaker	Simbolo	Tastiera
009	TAB ^9	^t	\t	End field	Tabulatore →|	ctrl-I
010	LF ^10	^n (cancellino)	\n	End record	Fine riga	ctrl-J
013	CR ^13	^c	\r	End record	A capo	ctrl-M

ScreenShot |Win: Alt-PrintScreen (window), PrintScreen (screen), paste in Paint |Mac, ⌘-Shift-3, ⌘-Shift-4 (select)

THE SMALL BIOINFORMATICIAN

THE GENETIC CODE - REVERSE!

	A	G	T	C	
A	AAA Phe (F)	AGA Ser (S)	ATA Tyr (Y)	ACA Cys (C)	A
	GAA "	GGA "	GTA "	GCA "	G
	TAA Leu (L)	TGA "	TTA STOP	TCA STOP	T
	CAA "	CGA "	CTA STOP	CCA Trp (W)	C
G	AAG Leu (L)	AGG Pro (P)	ATG His (H)	ACG Arg (R)	A
	GAG "	GGG "	GTG "	GCG "	G
	TAG "	TGG "	TTG Gln (Q)	TCG "	T
	CAG "	CGG "	CTG "	CCG "	C
T	AAT Ile (I)	AGT Thr (T)	ATT Asn (N)	ACT Ser (S)	A
	GAT "	GGT "	GTT "	GCT "	G
	TAT "	TGT "	TTT Lys (K)	TCT Arg (R)	T
	CAT Met (M)	CGT "	CTT "	CCT "	C
C	AAC Val (V)	AGC Ala (A)	ATC Asp (D)	ACC Gly (G)	A
	GAC "	GGC "	GTC "	GCC "	G
	TAC "	TGC "	TTC Glu (E)	TCC "	T
	CAC "	CGC "	CTC "	CCC "	C

AMINO ACIDS: NOT POLAR - POLAR NOT CHARGED - ACID - BASIC

Splice Junctions - REVERSE! -

[Polyadenylation signal: TTTATT] Exon 2 CCTGN(G/A)$_{11}$...ACT(C/T)ACCT(T/G) Exon 1
↑------- Intron 1 -------↑

Eucaryote promoters - REVERSE! -

-25	-50	-75	-100
CCATGGYGGCggc TTATA	GGGCGGNG	AGATTG(A/G)CC	C(AAA/TTT)CCAC
Kozak seq.	Hogness box	Tjian GC box	"CAT" box	Enhancer core seq.

Procaryote promoters - REVERSE! -

-10	-35
CAT---CC(C/T)CC(C/T)T	ATTATA TGTCAA
	Pribnow box

Gene categories (from Hattori et al., Nature, 2000)

Category	**1**	**Known human genes** (from the literature or public databases):
	1.1	defined functional association (e.g., transcription factor)
	1.2	unknown function (e.g., KIAA series of large cDNAs)
Category	**2**	**Novel genes with similarities over ~total length**:
	2.1	homology to a characterized cDNA (25–100% aa identity)
	2.2	similarity to a putative ORF predicted in silico
Category	**3**	**Novel genes with similarity to confined protein regions:**
	3.1	functional domain (e.g., zinc fingers)
	3.2	regions of a known protein without known functional association
Category	**4**	**Novel genes defined solely by gene prediction:**
	4.1	predicted and supported by spliced EST match(es)
	4.2	spliced EST(s)
	4.3	predicted genes composed only of a pattern of consistent exons
Category	**5**	**Pseudogenes** (gene-derived DNA sequences no longer expressed as protein products)

73% of DNA:EXTRAGENIC [Unique or low copy number; Repeats: Tandem or interspersed]

TANDEM REP.	**bp**	**~Copies**	**Chromosome location(s)**
Satellite (blocks of 0.1-3 Mb)	**5-171**		**Especially at centromeres**
Alpha (alphoid DNA)	171	8 × 10e5	- All Centromeres; 3-5% of the DNA of each chr.
Beta (*Sau*3A family)	68	5 × 10e4	- Centromeres of 1, 9, 13-14-15, 21-22, Y
Simple repeats:			
Satellite 1 (AT-rich)	25-48		- Most centromeres; other heterochromatic regions
Satellite 2 and 3	5		- All chromosomes
Minisatellite (blocks of 0.1-20 kb) **6-64**			**Telomeres of all chromosomes**
Telomeric family (TTAGGG)	6	2-35 × 10e4	All Telomeres
Hypervariable (GGGCAGGAXG)	9-64	3 × 10e4	All, often near Telomeres; expressed: MUC1 (1q)
Microsatellite (blocks of <150 bp) **1-4**			**Dispersed throughout all chromosomes**
$(A)_n/(T)_n$	1	3 × 10e7	All (0.3% of nuclear genome)
$(CA)_n/(TG)_n$	2	7 × 10e6	All (0.5% of nuclear genome; higly polymorphic)
$(CT)_n/(AG)_n$	2	3 × 10e6	All (0.2% of nuclear genome)
	3		Pathogenic expans. is possible (in coding DNA)

INTERSPERSED REP. (45%)		**bp**	**~Copies**	
SINE	(Short Intersp. Nuclear Elements)		13.1% of genome	
	Alu [Primate; from 7SL RNA)	**250-300**	1.1 × 10e6	Dimer 130+160 (~10.6% of genome)
	MIR [Mammalian Intersp. Rep.)	**130**	4.7 × 10e5	(~2.5% of genome)
LINE	(Long Intersp. Nuclear Elements)		20.4% of genome	
	L1 (Line-1) (Kpn) family	**800-6 100**	5.1 × 10e5	5′UTR+ORF1(p40)+ORF2(RT)+3′UTR
	L2 (Line-2) family	**250**	3.1 × 10e5	
LTR	(Long Terminal Repeats)		8.3% of genome	
	HERV (Hum. Endog. Retrovir.)	**6-11000**	2.0 × 10e5	