

SECTION 4

**BIOLOGICAL SEQUENCE ANALYSIS
AND CHARACTERIZATION**

CHAPTER 12

Predictive Functional Analysis of Polymorphisms: An Overview

MICHAEL R. BARNES

Genetic Bioinformatics
GlaxoSmithKline Pharmaceuticals
Harlow, Essex, UK

- 12.1 Introduction
 - 12.1.1 Moving from associated genes to disease genes
 - 12.1.2 Candidate polymorphisms
- 12.2 Principles of predictive functional analysis of polymorphisms
 - 12.2.1 Defining the boundaries of normal function in genes and gene products
 - 12.2.2 A decision tree for polymorphism analysis
- 12.3 The anatomy of promoter regions and regulatory elements
- 12.4 The anatomy of genes
 - 12.4.1 Gene splicing
 - 12.4.2 Splicing mechanisms, human disease and functional analysis
 - 12.4.3 Functional analysis of polymorphisms in putative splicing elements
 - 12.4.4 Polyadenylation signals
 - 12.4.5 Analysis of mRNA transcript polymorphism
 - 12.4.6 Initiation of translation
 - 12.4.7 mRNA secondary structure stability
 - 12.4.8 Regulatory control of mRNA processing and translation
 - 12.4.9 Tools and databases to assist mRNA analysis
- 12.5 Pseudogenes and regulatory mRNA
- 12.6 Analysis of novel regulatory elements and motifs in nucleotide sequences
 - 12.6.1 TRES (<http://bioportal.bic.nus.edu.sg/tres/>)
 - 12.6.2 Improbizer
- 12.7 Functional analysis on non-synonymous coding polymorphisms
- 12.8 A note of caution on the prioritization of *in silico* predictions for further laboratory investigation
- 12.9 Conclusions
- References

12.1 INTRODUCTION

Human genetic disease is generally characterized by a profound range of phenotypic variability manifested in variable age of onset, severity, organ specific pathology and response to drug therapy. The causes underlying this variability are likely to be equally diverse, influenced by differing levels of genetic and environmental modifiers. The vast majority of human genetic variants are likely to be neutral in effect, but some may cause or modify disease phenotypes. The challenge for bioinformatics is to identify the genetic variants which are most likely to show a non-neutral allelic effect. Geneticists studying complex disease are already seeking to identify these genetic determinants by genetic association of phenotypes with markers. The literature is now replete with reported associations, but moving from associated marker to disease allele is proving to be very difficult. So why are we so unsuccessful in making this transition? Disregarding false positive associations (which may make up the bulk of reported associations to date!) it may be that the diverse effects of genetic variation are helping disease alleles to elude us. Genetic variation can cause disease at any number of stages between promotion of gene transcription to post-translational modification of protein products. Many geneticists have chosen to focus their efforts on the most obvious form of variation—non-synonymous coding variation in genes. While this category of variation is undoubtedly likely to contribute considerably to human disease, this may overlook many equally important categories of variation in the genome, namely the effects of variation on gene transcription, temporal and spatial expression, transcript stability and splicing.

Clearly all polymorphisms are not equal. Analysis of polymorphism distribution across the human genome shows significant variations in polymorphism density and allele frequency distribution. Chakravarti (1999) showed an immediate difference between the density of SNPs in exonic regions and intragenic and intronic regions. SNPs occurred at 1.2-kb average intervals in coding regions and 0.9-kb intervals in intragenic and intronic regions. These differences point to different selection intensities in the genome, particularly in protein coding regions, where SNPs may result in alteration of amino acid sequences (non-synonymous SNPs (nsSNPs)) or the alteration of gene regulatory sequences. These observations are intuitive—natural selection is obviously likely to be strongest across gene regions, essentially encapsulating the objective of genetics—to identify non-neutral alleles with a role in disease.

So how should we go about identifying disease alleles? One approach used to identify disease mutations is to directly screen strong candidate genes for mutations present in affected but not unaffected family members. This approach is very useful in the study of monogenic diseases and cancers, where transmission of the disease allele can generally be demonstrated to be restricted to affected individuals/tissues. But in the case of complex disease the odds of identifying disease alleles by population screening of candidate genes would seem to be very high and proving their role is problematic as disease alleles are likely to be present in cases and controls. Instead we detect common marker alleles in LD with rarer disease alleles. This methodical approach to disease gene hunting localizes disease alleles rather than actually identifying them directly, the next step is to identify the disease allele from a range of alleles in LD with the associated marker. To conclusively identify this allele a functional mechanism for the allele in the disease needs to be identified.

12.1.1 Moving from Associated Genes to Disease Genes

Many potential associations have been reported between markers and disease phenotypes. Aside from the potential for false positive association, magnitude of effect in complex disease is also a problem. There may be a few gene variants with major effects, but generally complex disease is very heterogeneous and polygenic, it therefore follows that studies of single gene variants will be inconclusive and inconsistent—this is just something we have to work with. We may also find a bewildering array of complex disease genes with somewhat indirect roles in disease, such as modifier genes and redundant genes, that have many effects on phenotype. Understanding the mode of action of these associated alleles will help in determining how susceptibility genes may give rise to a multifactorial phenotype. Bioinformatics may be critical in this process. Follow-up studies need to be designed to ask the right questions, to ensure that the right candidates are tested and to confirm the biological role of positive associations. It may also be necessary to attempt to characterize polymorphisms with a potential functional impact, to help to identify the molecular mechanisms by a combination of bioinformatics and laboratory follow-up. Many of these informatics approaches are similar to the approaches originally used to identify candidates, but by necessity these analyses benefit from a far more detailed approach as in-depth analyses transfer to in-depth laboratory investigation.

Moving from an ‘associated gene’ to a ‘disease gene’ is not a purely academic objective. Genetics may sometimes be our only insight into the nature of a disease, such insights may help us to restore the normal function of disease genes in patients, develop drugs and better still it may help prevent disease in the first place. Better diagnosis and treatments are also prospects afforded by better understanding of the pathology of disease. A validated ‘disease gene’ is one of the most tangible progressions towards this end.

12.1.2 Candidate Polymorphisms

To turn the arguments for association analysis on their head, there is also theory that suggests that the direct identification of disease alleles may not be entirely futile. The common disease/common variant (cd/cv) hypothesis predicts that the genetic risk for common diseases will often be due to disease-predisposing alleles with relatively high frequencies (Reich and Lander, 2001). There is not enough evidence to prove or disprove this hypothesis, however several examples of common disease variants have been identified, some of which are listed in Table 12.1, the allele frequency of these variants in the public databases is also listed.

The possibility that many disease alleles may be common, presents an intriguing challenge for genetics (and bioinformatics), if the cd/cv hypothesis holds true, then a substantial number of disease alleles may already be present in polymorphism databases or the human genome sequence. These might be termed ‘candidate polymorphisms’. To extend this idea, just as genes with a putative biological role in disease are often prioritized for genetic association analysis, ‘candidate polymorphisms’ can be prioritized based on a predicted effect on the structure and function of regulatory regions, genes, transcripts or proteins. Thus selection of candidate polymorphisms is an extension of the candidate gene selection process—but in this case a link needs to be established between a predicted functional allelic effect and a target phenotype. As discussed earlier, DNA polymorphism can impact almost any biological process. Much of the literature in this area

TABLE 12.1 Disease Alleles Supporting the Common Disease/Common Variant Hypothesis

Gene (Allele)	Minor Allele Freq. (In dbSNP)	Disease/Trait Association	OMIM Review
APOE ϵ 4	16% (14%)	Alzheimer's and cardiovascular disease	107741
Factor V ^{Leiden} R506Q	2–7% (ND)	Deep vein thrombosis	227400
KCNJ11 E23K	14% (25%)	Type II diabetes	600937
COMT V158M	0.1–62% (45%)	Catechol drug pharmacogenetics	116790

has focused on the most obvious form of variation — non-synonymous changes in coding regions of genes. Alterations in amino acid sequences have accounted for a great number of diseases. Coding variants may impact protein folding, active sites, protein–protein interactions, protein solubility or stability. But the effects of DNA polymorphism are by no means restricted to coding regions, variants in regulatory regions may alter the consensus of transcription factor binding sites or promoter elements; variants in the untranslated regions (UTR) of mRNA may alter mRNA stability; variants in the introns and silent variants in exons may alter splicing efficiency.

Approaches for evaluating the potential functional effects of DNA polymorphisms are almost limitless, but there are very few tools designed specifically for this task. Instead almost any bioinformatics tool which makes a prediction based on a DNA or protein sequence can be commandeered to analyse polymorphisms — simply by analysing wild-type and mutant sequences and looking for an alteration in predicted outcome by the tool. Polymorphisms can also be evaluated at a simple level by looking at physical considerations of the properties of genes and proteins or they can be evaluated in the context of a variant within a family of homologous or orthologous genes or proteins.

12.2 PRINCIPLES OF PREDICTIVE FUNCTIONAL ANALYSIS OF POLYMORPHISMS

Faced with the extreme diversity of disease, analysis of polymorphism data calls for equally diverse methods to assess functional effects that might lead to these phenotypes. The complex arrangements that regulate gene transcription, translation and function are all potential mechanisms through which disease could act and so analysis of potential disease alleles needs to evaluate almost every eventuality. Figure 12.1 illustrates the logical decision-making process that needs to be applied to the analysis of polymorphisms and mutations. The tools and approaches for the analysis of variation are completely dependent on the location of the variant within a gene or regulatory region. Many of these questions can be answered very quickly using genomic viewers such as Ensembl or the UCSC human genome browser (see Chapter 5 for a tutorial on these tools). Placing a polymorphism in full genomic context is useful to evaluate variants in terms of location

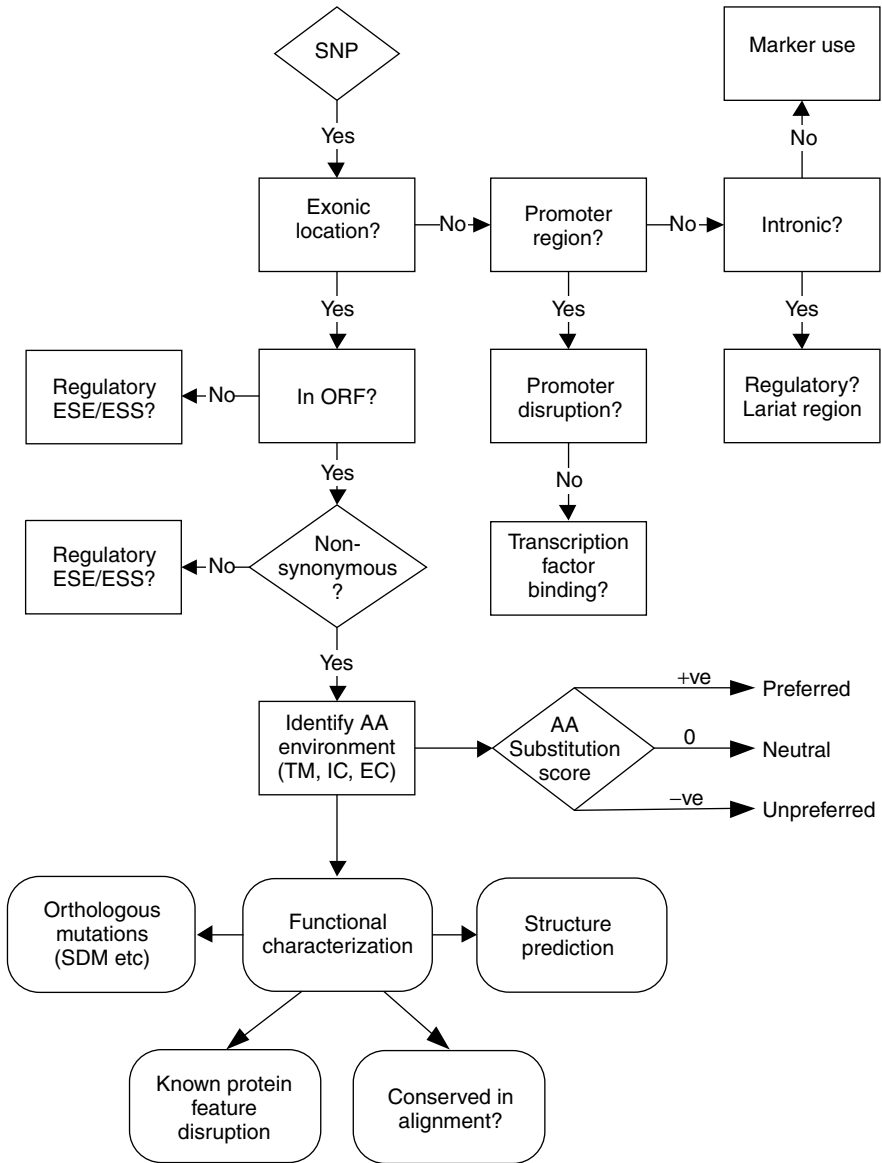


Figure 12.1 A decision tree for polymorphism analysis.

within or near genes (exonic, coding, UTR, intronic, promoter region) and other functionally significant features, such as CPG islands, repeat regions or recombination hotspots. Once approximate localization is achieved, specific questions need to be asked to place the polymorphism in a specific genic or intergenic region. This will help to narrow down the potential range of functional effects attributable to a variant, which will in turn help to identify the appropriate laboratory follow-up approach to evaluate function. Tables 12.2

TABLE 12.2 Functional Polymorphisms in Genes and Gene Regulatory Sequences

Location	Gene/Disease	Mechanism
Transcription factor binding	TNF in cerebral malaria	−376A SNP introduces OCT1 binding site—altering TNF expression, associated with four-fold increased susceptibility to cerebral malaria. (Knight <i>et al.</i> , 1999)
Promoter	CYP2D6	Common — 48T > G substitution disrupts the TATA box of the CYP2D6 promoter, causing 50% reduction in expression. (Pitarque <i>et al.</i> , 2001)
Promoter	RANTES in HIV progression	−28G mutation increases transcription of the RANTES gene slowing HIV-1 disease progression (Liu <i>et al.</i> , 1999)
<i>cis</i> -regulatory element	Bruton's tyrosine kinase in X-linked agammaglobulinemia	+5G/A (intron 1) shows reduced BTK transcriptional activity, suggesting a novel <i>cis</i> -acting element, involved in BTK downregulation but not splicing (Jo <i>et al.</i> , 2001)
Lariat region	HNF-4α	NIDDM-associated C/T substitution in polypyrimidine tract in intron 1b in an important <i>cis</i> -acting element directing intron removal (lariat region) (Sakurai <i>et al.</i> , 2000)
Splice donor/acceptor sites	ATP7A in Menke disease	Mutation in donor splice site of exon 6 of ATP7A causes a lethal disorder of copper metabolism (Moller <i>et al.</i> , 2000)
Cryptic donor/acceptor sites	β-glucuronidase gene (GUSB) in MPS VII	A 2-bp intronic deletion creates a new donor splice site activating a cryptic exon in intron 8 (Vervoort <i>et al.</i> , 1998)
Exonic splicing enhancers (ESE)	BRCA1 in breast cancer	Both silent and nonsense exonic point mutations were demonstrated to disrupt splicing in BRCA1 with differing phenotypic penetrance (Liu <i>et al.</i> , 2001)
Intronic splicing enhancers (ISE)	Alpha galactosidase in Fabry disease	G > A transversion within 4 bp of splice acceptor results in greatly increased alternative splicing (Ishii <i>et al.</i> , 2002)
Exonic splicing silencers (ESS)	CD45 in multiple sclerosis	Silent C77G disrupts ESS that inhibits the use of the 5' exon four splice sites (Lynch and Weiss, 2001)
Intronic splicing silencers (ISS)	TAU in dementia with parkinsonism	Mutations in TAU intron 11 ISS cause disease by altering exon 10 splicing (D'Souza and Schellenberg, 2000)

TABLE 12.2 (continued)

Location	Gene/Disease	Mechanism
Polyadenylation signal	FOXP3 in IPEX syndrome	A→G transition within the polyadenylation signal leads to unstable mRNA with 5.1 kb extra UTR (Bennett <i>et al.</i> , 2001)

TABLE 12.3 Tools for Functional Analysis of Gene Regulation and Splicing

Tool	URL
Promoter prediction	
NNPP	http://www.fruitfly.org/seq_tools/promoter.html
CorePromoter	http://sciclio.cshl.org/genefinder/CPROMOTER/
Promoter Scan II	http://www.molbiol.ox.ac.uk/promoterscan.htm
Orange	http://www.witi.cs.uni-magdeburg.de/~grabe/orange/
Transcription factor binding site prediction	
TRANSFAC	http://transfac.gbf.de/TRANSFAC/
FastM/ModelInspector	http://genomatix.gsf.de/cgi-bin/fastm2/fastm.pl
TESS	http://www.cbil.upenn.edu/tess/
TFSEARCH	http://www.cbrc.jp/research/db/TFSEARCH.html
Splice site prediction	
NETGENE	http://genome.cbs.dtu.dk/services/NetGene2/
Splice Site Prediction	http://www.fruitfly.org/seq_tools/splice.html
SpliceProximalCheck	http://industry.ebi.ac.uk/~thanaraj/SpliceProximalCheck.html
Gene prediction and ORF finding	
Genscan	http://genes.mit.edu/GENSCAN.html
Genie	http://www.fruitfly.org/seq_tools/genie.html
ORF Finder	http://www.ncbi.nlm.nih.gov/gorf/gorf.html
Detection of novel regulatory elements and comparative genome analysis	
PipMaker	http://bio.cse.psu.edu/pipmaker/
TRES	http://bioportal.bic.nus.edu.sg/tres/
Improbizer	http://www.soe.ucsc.edu/~kent/improbizer/
Regulatory Vista	http://www-gsd.lbl.gov/vista/rVistaInput.html
Integrated platforms for gene, promoter and splice site prediction	
Webgene	http://www.itba.mi.cnr.it/webgene/
BCM Gene Finder	http://dot.imgen.bcm.tmc.edu:9331/gene-finder/gf.html

and 12.3 illustrate some carefully selected examples of non-coding polymorphisms in genes and transcripts, these publications were specifically selected as each also includes a detailed laboratory based follow-up to evaluate each form of polymorphism. We refer the reader to these publications as a potential guide to assist in laboratory investigation.

12.2.1 Defining the Boundaries of Normal Function in Genes and Gene Products

Beyond the general localization of variants that general bioinformatics tools, such as Ensembl, can afford, there is a further more detailed context to many known regulatory elements in genes and gene regulatory regions. Our knowledge of these elements is still very sparse, but certain elements are relatively well defined. Many of these elements have been defined by mutations in severe Mendelian phenotypes. By definition this suggests that many elements which may have moderate effects on gene function are less likely to have been identified as they are less likely to have come to the attention of physicians. In the case of complex disease it may be very difficult to distinguish genuine disease susceptibility alleles from the normal spectrum of variability in human individuals.

12.2.2 A Decision Tree for Polymorphism Analysis

The first step in our decision tree for polymorphism analysis (Figure 12.1) is a simple question—is the polymorphism located in an exon? Answering this accurately may not always be simple or even possible with only *in silico* resources. As we have already seen in the previous section, delineation of genes is really the key step in all subsequent analyses, once we know the location of a gene all other functional elements fall into place based on their location in and around genes. In Chapter 4 we presented a detailed examination of the art of delineating genes, including methods for extending sequences to identify the true boundaries of a gene, not just its coding region. This activity may seem superfluous in the ‘post genome’ era, but the fact is that we still know very little about the full diversity of genes and the vast majority of genes are still incompletely characterized. Gene prediction and gene cloning has generally focused on the open reading frame—the protein coding sequence (ORF/CDS) of genes. For the most part UTR sequences have

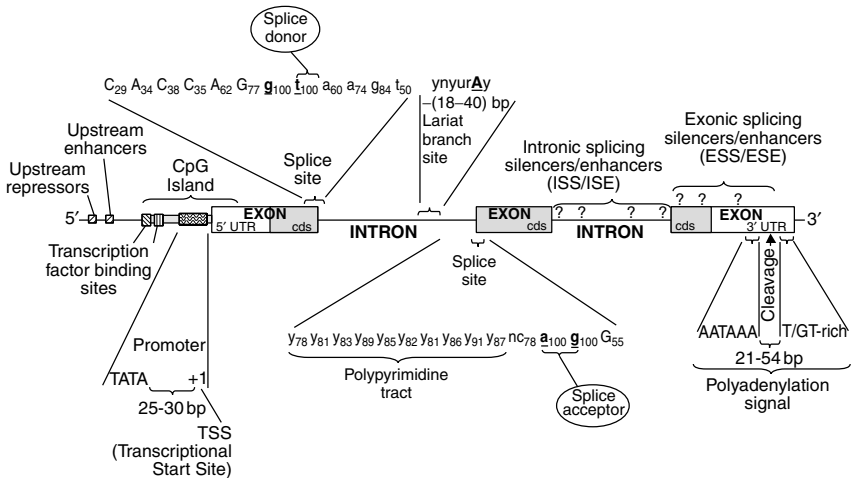


Figure 12.2 The anatomy of a gene. This figure illustrates some of the key regulatory regions which control the transcription, splicing and post-transcriptional processing of genes and transcripts. Polymorphisms in these regions should be investigated for functional effects.

been neglected in the rush to find an ORF and a protein. In the case of polymorphism analysis, these sequences should not be overlooked as the extreme 5' and 3' limits of UTR sequence delineate the true boundaries of genes. This delineation of gene boundaries is illustrated in a canonical gene model in Figure 12.2. As the model shows, most of the known regulatory elements in genes are localized to specific regions based on the location of the exons. So for example, the promoter region is generally located in a 1–2-kb region immediately upstream of the 5' UTR and splice regulatory elements flank intron/exon boundaries. Many of these regulatory regions were first identified in Mendelian disorders and now some are also being identified in complex phenotypes. Table 12.2 lists some of the disease mutations and polymorphisms that have helped to shape our knowledge of this complex area.

12.3 THE ANATOMY OF PROMOTER REGIONS AND REGULATORY ELEMENTS

Prediction of eukaryotic promoters from genomic sequence remains one of the most challenging tasks for bioinformatics. The biggest problem is over-prediction; current methods will on average predict promoter elements at 1-kb intervals across a given genomic sequence. This is in stark contrast to the estimated average 40–50-kb distance of functional promoters in the human genome (Reese *et al.*, 2000). Although it is possible that some of these predicted promoters may be expressed cryptically, the vast majority of predictions are likely to be false positives. To avoid these false predictions it is essential to provide promoter prediction tools with the appropriate sequence region, that is, the region immediately upstream of the gene transcriptional start site (TSS). It is important to define the TSS accurately; it is certainly insufficient to simply take the sequence upstream from the start codon as 5' UTR can often span additional 5' exons in higher eukaryotes (Reese *et al.*, 2000). As Uwe Ohler of the *Drosophila* genome project so eloquently stated, 'without a clear idea of the TSS location we may well be looking for a needle in the wrong haystack' (Ohler, 2000). If we can identify the TSS, the majority of RNA polymerase promoter elements are likely to be located within 150 bp, although some may be more distant so it may be important to analyse 2 kb or more upstream, particularly when the full extent of the 5' UTR or TSS is not well defined.

Once a potential TSS has been identified there are many tools which can be applied to identify promoter elements and transcription factor binding sites. The human genome browsers (UCSC and Ensembl) are the single most valuable resources for the analysis of promoters and regulatory elements. Specifically, Ensembl annotates putative promoter regions using the Eponine tool. The UCSC browser annotates known transcription factor binding sites from the Transfac database and novel predicted regulatory elements in the 'golden triangle' track (see Section 12.6.2 below). These are very useful for rapid evaluation of the location of variants in relation to these features, although this data needs to be used with caution as whole genome analyses may over-predict or overlook evidence for alternative gene models. The analysis approaches for promoter and transcription binding site analysis are reviewed thoroughly in Chapter 13.

Characterization of gene promoters and regulatory regions is not only valuable for functional analysis of polymorphisms, but it can also provide important information about the regulatory cues that govern the expression of a gene, which may be valuable for pathway expansion to assist in the elucidation of the function of candidate genes and disease-associated genes.

12.4 THE ANATOMY OF GENES

12.4.1 Gene Splicing

Alternative splicing is an important mechanism for regulation of gene expression which can also expand the coding capacity of a single gene to allow production of different protein isoforms, which can have very different functions. The recent completion of the human genome draft has given an interesting new insight into this form of gene regulation. Despite initial estimates of a human gene complement of > 100 K genes, direct analysis of the sequence suggests that humans may only have 30–40 K genes, which is only a two- to three-fold gene increase over invertebrates (Aparicio, 2000). Indeed, extrapolation of results from an analysis of alternatively spliced transcripts from chromosomes 22 and 19 have led to estimates that at least 59% of human genes are alternatively spliced (Lander *et al.*, 2001). This highlights the probable significance of post-transcriptional modifications such as alternative splicing as an alternative means by which to express the full phenotypic complexity of vertebrates without a very large number of genes.

A much simpler organism has given us a glimpse of the possibilities of splicing as a mechanism to generate phenotypic complexity. The drosophila homologue of the human Down syndrome cell adhesion molecule (DSCAM) has 115 exons, 20 of which are constitutively spliced and 95 of which are alternatively spliced (Schmucker *et al.*, 2000). The alternatively spliced exons are organized into four clusters, with 12 alternative versions of exon 4, 48 versions of exon 6, 33 versions of exon 9 and two versions of exon 17. These clusters of alternative exons code for 38,016 related but distinct protein isoforms!

12.4.2 Splicing Mechanisms, Human Disease and Functional Analysis

The remarkable diversity of potential proteins produced from the DSCAM gene, gives us some idea of the tight regulation of alternative splicing that must be in place to not only regulate the choice of each version of a particular exon, but also to exclude all other versions of the exon once one version has been selected. Regulation of splicing is mediated by the spliceosome, a complex network of small nuclear ribonucleoprotein (snRNP) complexes and members of the serine/arginine-rich (SR) protein family. At its most basic level, pre-mRNA splicing involves precise removal of introns to form mature mRNA with an intact open reading frame (ORF). Correct splicing requires exon recognition with accurate cleavage and rejoining at the exon boundaries designated by the invariant intronic GT and AG dinucleotides, respectively known as the splice donor and splice acceptor sites (Figure 12.2). Other more variable consensus motifs have been identified in adjacent locations to the donor and acceptor sites, including a weak exonic 'CACCAG' consensus flanking the splice donor site, an intronic polypyrimidine- (Y : C or T) rich tract flanking the splice acceptor site and a weakly conserved intronic 'YNYURAY' consensus 18–40 bp from the acceptor site, which acts as a branch site for lariat formation (Figure 12.2). Other regulatory motifs are known to be involved in splicing, including exonic splicing enhancers (ESE) and intronic splicing enhancers (ISE), both of which promote exon recognition, and exonic and intronic splicing silencers (ESS and ISS, respectively), which have an opposite action, inhibiting the recognition of exons. DNA recognition motifs for splicing enhancers and silencers are generally quite degenerate. The degeneracy of these consensus recognition motifs points to fairly promiscuous binding by SR proteins. These interactions can also explain the use of alternative and inefficient splice sites, which may be influenced by competitive binding of SR proteins and hnRNP determined by the relative ratio of hnRNP

to SR proteins in the nucleus. A natural stimulus that influences the ratio of these proteins is genotoxic stress, which can lead to the often observed phenomenon of differential splicing in tumours and other disease states (Hastings and Krainer, 2001).

Mutations affecting mRNA splicing are a common cause of Mendelian disorders, 10–15% of Mendelian disease mutations affect pre-mRNA splicing (Human Gene Mutation Database, Cardiff). These mutations can be divided into two subclasses according to their position and effect on the splicing pattern. Subclass I (60% of the splicing mutations) includes mutations in the invariant splice-site sequences, which completely abolish exon recognition. Subclass II includes mutations in the variant motifs, which can lead to both aberrantly and correctly spliced transcripts, by either weakening or strengthening exon-recognition motifs. Subclass II also includes intronic mutations, which generate cryptic donor or acceptor sites and can lead to partial inclusion of intronic sequences. These Mendelian disease mutations have helped to define our understanding of splicing mechanisms. Considering the proven complexity of splicing in the human genome (Lander *et al.*, 2001), it seems reasonable to expect splicing abnormality to play a significant role in complex diseases, but examples are rare. This is explained in part by the power of family-based mutations, the inheritance of which can be traced between affected and unaffected relatives. It is difficult to determine similar causality for a population-based polymorphism.

12.4.3 Functional Analysis of Polymorphisms in Putative Splicing Elements

If taken individually, there are many sequences within the human genome that match the consensus motifs for splice sites, but most of them are not used. In order to function, splice sites need appropriately arranged positive (ESEs and ISEs) and negative (ESSs, and ISSs) *cis*-acting sequence elements. These *cis*-acting arrangements of regulatory elements can be both activated and deactivated by DNA sequence polymorphisms. DNA polymorphism at the invariant splice acceptor (AG) and donor (GT) sites, are generally associated with severe diseases and so, are likely to be correspondingly rare. But, as we have seen, recognition motifs for some of the elements that make up the larger splice site consensus are very variable, so splice site prediction from undefined genomic sequence is still imprecise at the best of times. Bioinformatics tools can fare rather better when applied to known genes with known intron/exon boundaries — this information can be used to carry out reasonably accurate evaluations of the impact of polymorphisms in putative splice regions. There are several tools which will predict the location of splice sites in genomic sequence, all match and score the query sequence against a probability matrix built from known splice sites (see Table 12.3). These tools can be used to evaluate the effect of splice region polymorphisms on the strength of splice site prediction by alternatively running wild-type and mutant alleles. As with any other bioinformatics prediction tool it is always worth running predictions on other available tools to look for a consensus between different prediction methods. These tools can also be used to evaluate the propensity of an exon to undergo alternative splicing. For example an unusually low splice site score may indicate that aberrant splicing may be more likely at a particular exon compared to exons with higher splice site scores. The phase of the donor and acceptor sites also needs to be taken into account in these calculations. Coding exons exist in three phases 0, 1 and 2, based on the codon location of the splice sites, if alternative donor or acceptor sites are in unmatched phases then a frameshift mutation will occur.

Splice site prediction tools will generally predict the functional impact of a polymorphism within close vicinity of a splice donor or acceptor site, although they will not predict

the functional effect of polymorphisms in other elements such as lariat branch sites. Definition of consensus motifs for these elements (Figure 12.2) makes it reasonably easy to assess the potential functional impact of polymorphisms in these gene regions by simply inspecting the location of a polymorphism in relation to the consensus motif. As with all functional predictions laboratory investigation is required to confirm the hypothesis.

Other *cis*-regulatory elements, such as ESE, ESS, ISE and ISS sites are very poorly defined and may be located in almost any location within exons and introns. There are currently no available bioinformatics tools to generally predict the locations of these regulatory elements. Some specific elements, *cis*-regulatory elements, have been defined in specific genes, but these do not form a consensus sequence to search other genes. One of the only possible approaches for *in silico* analysis of such elements is to use comparative genome data to look for evolutionarily conserved regions, particularly between distant species, e.g. comparison of Human/Fugu (fish) genomes. Although there may be some value in these approaches, confirmation of *cis*-regulatory elements really needs to be achieved by laboratory methods (see D'Souza and Schellenberg (2000) for a description of such methods).

12.4.4 Polyadenylation Signals

Polyadenylation of eukaryotic mRNA occurs in the nucleus after cleavage of the precursor-RNA. Several signals are known which determine the site of cleavage and subsequent polyadenylation, the most well known is a canonical hexanucleotide (AAUAAA) signal 20–50 bp from the 3' end of the pre-RNA, this works with a downstream U/GU-rich element which is believed to regulate the complex of proteins necessary to complete 3' processing (Pauws *et al.*, 2001). The specific site of cleavage of pre-RNA is located between these regulatory elements and is determined by the nucleotide composition of the cleavage region with the following nucleotide preference A > U > C >> G. In a study of 9625 known human genes Pauws *et al.* (2001) found that 44% of human genes regularly used more than one cleavage site, resulting in the generation of slightly different mRNA species.

Mutations in the canonical AAUAAA polyadenylation signal have been shown to disrupt normal generation of polyadenylated transcripts (Bennett *et al.*, 2001). This signal is needed for both cleavage and polyadenylation in eukaryotes, and failure to polyadenylate will prevent maturation of mRNA from nuclear RNA (Wahle and Keller, 1992). The complete aggregate of elements that make up the polyadenylation signal including the U/GU-rich region may not be universally required for processing (Graber *et al.*, 1999). Single nucleotide variations in this region cannot be conclusively identified as functional although any polymorphism in this region might be considered a candidate for further consideration.

12.4.5 Analysis of mRNA Transcript Polymorphism

The potential functional effects of genetic polymorphism can extend beyond a direct effect on the genomic organization and regulation of genes. Messenger RNA is far more than a simple coded message acting as an intermediary between genes and proteins. mRNA molecules have different fates related to structural features embedded in discrete regions of the molecule. The processing, localization, translation or degradation of a given mRNA may vary considerably, depending upon the environment in which it is expressed. Figure 12.3 illustrates a simplified model of an mRNA molecule, indicating the

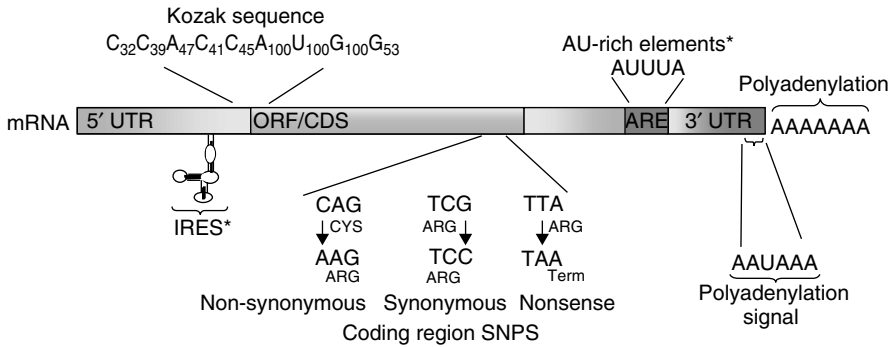


Figure 12.3 The anatomy of an mRNA transcript. This figure illustrates some of the key regulatory and structural elements that control the translation, stability and post-transcriptional processing of mRNA transcripts. Polymorphisms in these regions should be investigated for functional effects.

key features and regulatory motifs that could potentially be disrupted by polymorphism. At the most basic level an mRNA molecule consists of a protein coding, open reading frame (ORF), flanked by 5' and 3' UTR. Most polymorphism analysis in the literature has tended to focus on the coding sequence of genes, but there is evidence to suggest that UTR sequences also serve important roles in the function of mRNA. At the risk of generalizing, 5' UTR sequences are important as they are known to accommodate the translational machinery, while there is accumulating evidence that strongly implicates the 3' UTR in the regulation of gene expression. In Table 12.4 we highlight some examples of polymorphisms which impact mRNA transcripts.

12.4.6 Initiation of Translation

If a gene is known, the ORF will probably be well defined, but if a novel transcript is being studied the ORF needs to be identified. Again we refer the reader to Chapter 4 which contains details on the extension of mRNA transcripts and ORF finding procedures. The accepted convention is that the initiator codon will be the first inframe AUG encoding the largest open reading frame in the transcript. There is evidence of a scanning mechanism for initiation of translation; the initiator codon generally conforms to a 'CCACCaugG' consensus motif known as the Kozak sequence (Kozak, 1996). However, Peri and Pandey (2001) and others have recently reappraised this convention and actually found that more than 40% of known transcripts contain inframe AUG codons upstream of the actual initiator codon, some of which conform more closely to the Kozak motif than the authentic initiator codon. Their revised Kozak consensus 'C₃₂C₃₉A₄₇C₄₁C₄₅A₁₀₀U₁₀₀G₁₀₀G₅₃' was much weaker. These observations have cast some doubt on the validity of the scanning mechanism for initiation of translation, some have argued that the frequent occurrence of AUG codons upstream of the putative initiator codon, may indicate misassignment of the initiator codon or cDNA library anomalies (Kozak, 2000), others point to the empirical increase in gene expression measured in the laboratory when initiator codons conforming to the Kozak consensus are compared to other sequences. This debate may never resolve conclusively and it seems certain that the mechanism for translation initiation is still not fully understood.

TABLE 12.4 Functional Non-Coding Polymorphisms in mRNA Transcripts

Location	Gene/Disease	Mechanism
Internal ribosome entry segment (IRES)	Proto-oncogene c-myc in multiple myeloma	C–T mutation in the c-myc-IRES causes aberrant translational regulation of c-myc, enhanced binding of protein factors and enhanced initiation of translation leading to oncogenesis (Chappell <i>et al.</i> , 2000)
Kozak initiation sequence	Platelet glycoprotein Ib-alpha (GP1BA) in ischaemic stroke	C/T polymorphism at the –5 position from the initiator ATG codon of the GP1BA gene is located within the ‘Kozak’ consensus nucleotide sequence. The presence of a C at this position significantly increases the efficiency of expression of the GPIb/V/IX complex (Afshar-Kharghan <i>et al.</i> , 1999)
Anti-termination mutation and 3' UTR stability determinants	Alpha-globin in alpha-thalassemia	UAA to CAA to anti-termination mutation allows translation to proceed into the 3' UTR which masks stability determinants to substantially decrease mRNA half-life (Conne <i>et al.</i> , 2000)
UTR stability	Protein tyrosine phosphatase-1B (PTP1B)	1484insG in 3' UTR causes PTP1B over-expression leading to insulin resistance (Di Paola <i>et al.</i> , 2002)

There are some examples of polymorphisms in Kozak sequences that appear to have a direct bearing in human disease. Kaski *et al.* (1996) reported a T > C SNP with an 8–17% minor allele frequency at the –5 position from the initiator ATG codon of the GP1BA gene. This SNP is located within the most 5' (and weakest) part of the Kozak consensus sequence. The cytosine (C) allele at this position conforms more closely to the consensus and subsequent studies of the SNP found that it was associated with increased expression of the receptor on the cell membrane, both in transfected cells and in the platelets of individuals carrying the allele. The polymorphism was also associated with cardiovascular disease susceptibility (Afshar-Kharghan *et al.*, 1999).

An alternative mechanism for translation initiation has been identified which does not obey the ‘first AUG rule’, this involves cap-independent internal ribosome binding mediated by a Y-shaped secondary structure, denoted the Internal Ribosome Entry Site (IRES), located in the 5' UTR of 5–10% of human mRNA molecules (see Le and Maizel, (1997) for a review of these elements). IRES elements are complex stem loop structures, there is no reliable sequence consensus to allow prediction of the possible functional effects of polymorphisms in these elements instead this needs to be attempted by the use of RNA secondary structure prediction tools such as MFOLD (see below).

12.4.7 mRNA Secondary Structure Stability

While we have already established that nucleotide variants in mRNA can alter or create sequence elements directing splicing, processing or translation of mRNA, variants may

also influence mRNA synthesis, folding, maturation, transport and degradation. Many of these diverse biological processes are strongly dependent on mRNA secondary structure. Secondary structure is essentially determined by ribonucleotide sequence and so folding of mRNA is also likely to be influenced by SNPs and other forms of variation at any location in a transcript. Shen *et al.* (1999) studied two common silent SNPs in the coding regions of two essential genes—a U1013C transition in human alanyl tRNA synthetase (AARS) and a U1674C transition in the human replication protein A 70-kDa subunit (RPA70). The minor allele frequency was 0.49 for the AARS U allele and 0.15 for the RPA70 C allele. Using structural mapping and structure-based targeting strategies they demonstrated that both SNPs had marked effects on the structural folds of the mRNAs, suggesting phenotypic consequences of SNPs in mRNA structural motifs.

RNA stability is an intriguing disease mechanism, unfortunately beyond this and a handful of other published studies (see Conne *et al.* (2000) for a review), the true extent of detectable differences in mRNA folding caused by polymorphism is quite unknown, this may reflect the difficulties involved in studying such mutational effects *in vitro*.

There are several tools which can help to construct *in silico* secondary-structure models of polymorphic mRNA alleles. One of the best tools is MFOLD (M. Zuker, Washington University, St. Louis, MO), this is maintained on the Zuker laboratory homepage which also contains an excellent range of RNA secondary structure-related resources (<http://bioinfo.math.rpi.edu/~zukerm/rna/>). MFOLD will construct a number of possible models based on all structural permutations of a user-submitted mRNA sequence. Submission of mutant and wild-type mRNA alleles to this tool will give the user a fairly good indication of whether an allele could alter mRNA secondary structure. This can help to prioritize alleles for laboratory-based investigation of mRNA stability studies.

12.4.8 Regulatory Control of mRNA Processing and Translation

Beyond splicing and promoter based regulation, mRNAs are also tightly controlled by regulatory elements in their 5' and 3' untranslated regions (Figure 12.3). Proteins that bind to these sites are key players in controlling mRNA stability, localization and translational efficiency. Consensus motifs have been identified for many of these factors, usually corresponding to short oligonucleotide tracts, which generally fold in specific secondary structures, which are protein binding sites for various regulatory proteins. Some of these regulatory signals tend to be protein family specific, while others have a more general effect on diverse mRNAs. AU-rich elements (AREs) are the largest class of *cis*-acting 3' UTR-located regulatory molecules that control the cytoplasmic half-life of a variety of mRNA molecules. One main class of these regulatory elements consists of pentanucleotide sequences (AUUUA) in the 3' UTR of transcripts encoding oncoproteins, cytokines and growth and transcription factors. Many RNA-binding proteins, mostly members of the highly conserved ELAV family, recognize and bind AREs (Chen and Shyu, 1995). Defective functioning of AREs can lead to the abnormal stabilization of mRNA, this forms the basis of several human diseases, including mantle cell lymphoma, neuroblastoma, immune and several inflammatory diseases. Polymorphisms which disrupt AU-rich motifs in a 3' UTR sequence may be worth evaluation as potentially functional polymorphisms. Some databases to assist in the identification of these motifs are described below.

12.4.9 Tools and Databases to Assist mRNA Analysis

To assist in the analysis of diverse and often family specific regulatory elements, such as ARE elements, Pesole *et al.* (2000) have developed UTRdb, a specialized

non-redundant database of 5' and 3' untranslated sequences of eukaryotic mRNAs (<http://bighost.area.ba.cnr.it/BIG/UTRHome/>). In March 2002, UTRdb contained 39,527 non-redundant human entries; these are enriched with specialized information absent from primary databases including the presence of RNA regulatory motifs with experimental proof of a functional role. It is possible to BLAST search the database for the presence of annotated functional motifs in a query sequence.

Jacobs *et al.* (2002) have also developed Transterm, a curated database of mRNA elements that control translation (<http://uther.otago.ac.nz/Transterm.html>). This database examines the context of initiation codons for conformation with the Kozak consensus and also contains a range of mRNA regulatory elements from a broad range of species. Access is provided via a web browser in several different ways: a user-defined sequence can be searched against motifs in the database or elements can be entered by the user to search specific sections of the database (e.g. coding regions or 3' flanking regions or the 3' UTRs) or the user's sequence. All elements defined in Transterm have associated biological descriptions with references.

12.5 PSEUDOGENES AND REGULATORY MRNA

As a final word on the analysis of mRNA transcripts, it is important to be aware that not all mRNAs are intended to be translated. Some genes may produce transcripts that are truncated or retain an intron or are otherwise configured in a way that precludes translation. It is difficult to clarify the role of some of these transcripts; where a transcript has multiple premature termination codons, it is likely to be a pseudogene, others may have no obvious open reading frames, these may also be pseudogenes or they may be regulatory mRNA molecules. Several non-coding RNA (ncRNA) molecules have been described which act as riboregulators with a direct influence on post-transcriptional regulation of gene expression (see Erdmann *et al.* (2001) for a comprehensive review of the properties of regulatory mRNA). Analysis of polymorphisms in these molecules is difficult as they are very poorly defined in terms of functionality.

12.6 ANALYSIS OF NOVEL REGULATORY ELEMENTS AND MOTIFS IN NUCLEOTIDE SEQUENCES

It is very likely that our current knowledge of regulatory elements in the human genome is quite superficial. In terms of transcription factors alone, the TRANSFAC database contains a redundant set of 2263 profiles for vertebrate binding sites (Heinemeyer *et al.*, 1999), yet the first pass analysis of the human genome has identified over 4000 proteins with a putative DNA binding role (Venter *et al.*, 2001). This is likely to be an underestimate. Geneticists are working at the vanguard of efforts to close the gap between our current understanding and the full complexity of human gene regulation. Genetics has already contributed greatly to the identification of new regulatory elements by the identification of regulatory mutations and polymorphisms.

In this chapter we have reviewed a number of regulatory mechanisms and motifs in DNA sequences, including motifs in promoter regions, splice sites, introns and transcripts. Functional analysis of polymorphisms located in the consensus sequences identified for some of these elements may be an important indicator of a potential functional effect. However, despite advances in bioinformatic tools, predictive functional analysis

of sequence polymorphism is still difficult to validate without laboratory follow-up. Even with the benefit of laboratory verification, identification of deleterious alleles can be laborious and the results of analyses do not always hold true between *in vitro* and *in vivo* environments. In a sense evolution is an *in vivo* experiment on a grand scale and so Sydney Brenner (2000) and others have proposed the concept of 'inverse genetics' to cover the use of information recovered from different genomes to inform on function. Brenner suggested comparing genomes to highlight conserved areas 'in a vast sea of randomness'. This is an elegant approach for the characterization of polymorphisms. Characterization by conventional genetics demands analysis of large sample numbers, complex *in vitro* analysis or laborious transgenic approaches. In the case of inverse genetics, evolution and time have already done the work in a long-term 'experiment' which would be impossible to match in the laboratory.

Inverse genetics also has a wider application—analysis of a single promoter sequence will often identify many putative regulatory elements by chance alone. However, simultaneous analysis of many evolutionarily-related but diverse promoter sequences will clearly identify known and novel conserved motifs which are more likely to be functionally important to a particular family of genes. This approach known as phylogenetic footprinting, has been used to successfully elucidate many common regulatory modules (Gumucio *et al.*, 1996). Kleiman *et al.* (1998) used a similar approach to identify a novel potential element in the polyadenylation regulatory apparatus, a TG deletion (deltaTG) in the 3' UTR of the *HEXB* gene, 7bp upstream from the polyadenylation signal. The deltaTG *HEXB* allele, which occurred at a 10% frequency, showed 30% lower enzymatic activities compared to WT individuals. Polyacrylamide gel electrophoresis analysis of the allele revealed that the 3' UTR of the *HEXB* gene had an irregular structure. After studying a large range of eukaryotic mRNAs, including human, mouse and cat *HEXB* genes they found that the TG dinucleotide was part of a conserved sequence (TGTTTT) immersed in an A/T-rich region observed in more than 40% of mRNAs analysed. This study clearly illustrates how effective bioinformatic analysis of mRNA processing signals may require more than sequence analysis of known regulatory motifs; clearly tools are needed to identify novel regulatory elements. The web-based TRES tool is an example of a tool to assist in the identification of such novel elements.

12.6.1 TRES (<http://bioportal.bic.nus.edu.sg/tres/>)

TRES can be used to compare as many as 20 nucleotide sequences. The tool is multi-functional, it can either be used to identify conserved sequence motifs between submitted sequences or alternatively it can be used to identify known transcription factor binding sites shared between sequences using nucleotide frequency distribution matrices described in the TRANSFAC database (Heinemeyer *et al.*, 1999). This approach is not just applicable to evolutionarily-related sequences it can also be used to study unrelated sequences which may share similar regulatory cues, such as genes which show similar patterns of gene expression.

TRES also has another versatile search mode which allows detection of palindromic motifs or inverted repeats shared between sequences. These have unique features of dyad symmetry which can form hairpins or loops to facilitate protein binding in homo- or heterodimer form. Many transcription factors have palindromic recognition sequences and bind as dimmers; these motifs may be important to allow greater regulatory diversity from a limited number of transcription factors (Lamb and McKnight, 1991).

Although TRES is generally focused on the identification of transcription factor binding sites and promoter elements, the sequence motif identification facilities of the tool also

make it suitable for the identification of other motifs in non-coding sequences including UTR sequences and intronic sequences.

12.6.2 Improbizer

Improbizer was developed at the UCSC; the tool searches for motifs in DNA or RNA sequences that occur with an improbable frequency; that is greater than might be expected to occur by chance alone. Probabilities are estimated using the expectation maximization (EM) algorithm (Jim Kent, personal communication; for more details see <http://www.soe.ucsc.edu/~kent/improbizer/improbizer.html>).

Improbizer is available as a web interface, this allows the analysis of multiple sequences (up to 100 can be entered) for common motifs between sequences. Improbizer has also been used to annotate a large number of predicted promoter regions in the UCSC human genome browser (see Chapter 5). This data is presented as the so-called 'golden triangle' track. Kent and colleagues adopted this name to describe the process they called 'Regulatory region Triangulation' (J. Kent and D. Haussler, personal communication). This approach combines cDNA, genomic DNA and microarray data to locate and characterize regulatory regions in the human genome. The method identified a large set of putative transcription start sites by aligning G-cap selected ESTs (which represent 5' ends of transcripts) and other cDNA data to the human genome using BLAT. This data was compared with regions conserved between the human and mouse genomes with BLASTZ. Finally to complete the 'triangulation' process, they clustered Affymetrix microarray data to find co-regulated clusters of genes; once identified the promoter sequences were analysed using Improbizer. The highly novel data generated by this analysis is a valuable resource for the evaluation of polymorphisms in regulatory regions.

12.7 FUNCTIONAL ANALYSIS ON NON-SYNONYMOUS CODING POLYMORPHISMS

The huge diversity of protein molecules makes it very difficult to provide a generic model of a protein. Returning to our decision tree for polymorphism analysis (Figure 12.1), the consequences of an amino acid substitution are first and foremost defined by the environment in which the amino acid exists. Different cellular locations can have very different chemical environments which can have diverse effects on the properties of amino acids. The cellular location of proteins can be divided at the simplest level between intracellular, extracellular or transmembrane environments. The latter location is the most complex as amino acids in transmembrane proteins can be exposed to all three cellular environments, depending upon the topology of the protein and the location of the particular amino acid. Environments will also differ in extracellular and intracellular proteins, depending on the location of the residue within the protein. Amino acid residues may be buried in a protein core or exposed on the protein surface. Once the environment of an amino acid has been defined, different matrices are available to evaluate and score amino acid changes. For reference we have provided four amino acid substitution matrices in Appendix II. These matrices can be used to evaluate amino acid changes in extracellular, intracellular and transmembrane proteins; where the location of the protein is unknown, a matrix for 'all proteins' is also available. Preferred (conservative) substitutions have positive scores, neutral substitutions have a zero score and unpreferred (non-conservative) substitutions are scored negatively. These matrices are another application of 'inverse genetics' and are constructed by observing the propensity for exchange of one amino acid for another based on

Arg184Cys
y

JAG1MAN	:	ESRQQTLEKQNTGVSHHEBYQICMTDDEYVGFSGCNKPCRPDRDDFPGHYACDONGNRICK	:	221
JAG1MUS	:	ESRQQTLEKQNTGISHHEBYQIRVTCDDHYVGFSGCNKPCRPDRDDFPGHYACDONGNRICK	:	221
JAG1RAT	:	ESRQQTLEKQNTGISHHEBYQIRVTCDDHYVGFSGCNKPCRPDRDDFPGHYACDONGNRICK	:	221
JAG1FISH	:	ENRQQVYKHNQFVACEFYQIRVTCDEHYVGFSGCNKPCRPDRDDFPGHYTCDDHNGNRICK	:	222
JAG2HAN	:	EDRWRKSEFSGSHVHLELQIRVRCDENYVSAICNKPCCRPDRDDFPGHYTCDDYGNRACK	:	232
JAG2MUS	:	EDRWRKSEFSGSHVHLELQIRVRCDENYVSAICNKPCCRPDRDDFPGHYTCDDYGNRACK	:	232
JAG2RAT	:	EDRWRKSEFSGSHVHLELQIRVRCDENYVSAICNKPCCRPDRDDFPGHYTCDDYGNRACK	:	186
JAG2FISH	:	ESDHWQSIKHFSGITSHIYRIRVRCDENYVSSKCNKCCRPDRDDFPGHYRCDEPSNIVVCL	:	224
JAG3FISH	:	ENRQQRLLTHNGFVACEFYQIRVTCLEHYVGFSGCNKPCRPDRDDFPGHYTCDDONGNRICK	:	210
CSER2CHICK	:	EDRWRKTLQFNGFVSHHEBYQIRVRCDENYVSAICNKPCCRPDRDDFPGHYTCDDONGNRACK	:	202
XSER1PROG	:	ESRQQTLEKQNTGAGTYPFYQIRVTCDEHYVGFSGCNKPCRPDRDDFPGHYTCDDONGNRICK	:	217
CSER1CHICK	:	ESRQQTLEKHNAGSHHEBYQIRVTCDEHYVGFSGCNKPCRPDRDDFPGHYTCDDONGNRICK	:	195
SERRATEFLY	:	ESFEKRTLDHIGRMRITVYRVRVCAVTVYNTTCTTECRPRDDCGSHYACSSSEKCKVCL	:	275
		E N L a e qirV Cd Yy CnkfCcrErdDffgHy Cd Gnk C		

Figure 12.4 Functional evaluation of an Arg184Cys mutation in the Jagged protein family. Arg184Cys causes Alagille syndrome (OMIM 118450). Alignment of the mutated human amino acid sequence with vertebrate and invertebrate orthologues and homologues in the Jagged family identifies the Arg184 residue in a highly conserved position throughout this gene family. A mutation to a cysteine at this position would be expected to lead to the aberrant formation of disulphide bonds with other cysteine residues in the Jagged protein, this is likely to have a disruptive effect on the structure of the Jagged1 protein.

TABLE 12.5 Tools for Functional Analysis of Amino Acid Polymorphisms

Sequence manipulation and translation

Sequence Manipulation Suite <http://www.bioinformatics.org/sms/>

Amino acid properties

Properties of amino acids <http://www.russell.embl-heidelberg.de/aas/>

Secondary structure prediction

TMPRED http://www.ch.embnet.org/software/TMPRED_form.html

SOSUI <http://sosui.proteome.bio.tuat.ac.jp/sosuiframe0.html>

TMHMM <http://www.cbs.dtu.dk/services/TMHMM/>

PREDICTPROTEIN <http://www.embl-heidelberg.de/predictprotein/>

GPCRdb 7TM plots (Snake plots for most 7TMs) <http://www.gpcr.org/7tm/seq/snakes.html>

Tertiary structure prediction and visualization

Swiss-Model <http://expasy.hcuge.ch/swissmod/SWISS-MODEL.html>

SCOP <http://scop.mrc-lmb.cam.ac.uk/scop/>

Identification of functional motifs

INTERPRO <http://www.ebi.ac.uk/interpro/scan.html>

PROSITE <http://www.ebi.ac.uk/searches/prosite.html>

PFAM <http://www.sanger.ac.uk/Software/Pfam/>

NetPhos (serine, threonine and tyrosine phosphorylation) <http://www.cbs.dtu.dk/services/NetPhos/>

NetOGlyc (O-glycosylation) <http://www.cbs.dtu.dk/services/NetOGlyc/>

NetNGlyc (N-glycosylation) <http://www.cbs.dtu.dk/services/NetNGlyc/>

SIGNALP (signal peptide prediction) <http://www.cbs.dtu.dk/services/SignalP/>

Swissprot (functional annotation) <http://www.expasy.ch/cgi-bin/prot-search-ful>

comparison of very large sets of related proteins (see Chapter 14 and www.russell.embl-heidelberg.de/aas for more details). Defining the environment of an amino acid may be relatively straightforward if the protein is known, by looking at existing protein annotation or better still a known tertiary structure. Beyond the cellular environment of a variant there are many other important characteristics of an amino acid that need to be evaluated. These include the context of an amino acid within known protein features and the conservation of the amino acid position in an alignment of related proteins. Figure 12.4 shows an example of an evaluation of a mutation in Jagged1, a ligand for the Notch receptor family. Krantz *et al.* (1998) identified an Arg184Cys missense mutation in patients with Alagille syndrome (OMIM 118450). In terms of amino acid substitutions, Arg > Cys is very non-conservative (the extracellular substitution matrix score for this change is -5). Alignment of the mutated human amino acid sequence with vertebrate and invertebrate orthologues and homologues in the Jagged family identifies the Arg184 residue as a highly conserved position throughout this gene family. A mutation to a cysteine at this position would be expected to lead to the aberrant formation of disulphide bonds with other cysteine residues in the Jagged protein, this is likely to have a disruptive effect on the structure of the Jagged1 protein, presumably leading to the Alagille syndrome phenotype (see Chapter 14 for a description of the effects of inappropriate disulphide bond formation).

There are many different sources of protein annotation and tools to evaluate the impact of substitutions in known and predicted protein features, some of the best are listed in Table 12.5. The protein analysis approaches underlying these tools are comprehensively reviewed in Chapter 14.

12.8 A NOTE OF CAUTION ON THE PRIORITIZATION OF *IN SILICO* PREDICTIONS FOR FURTHER LABORATORY INVESTIGATION

Just as the complexity of genes, transcripts and proteins are virtually limitless, so too are the possibilities for developing functional hypotheses. If every aspect of the analyses explored in this chapter were examined in any single polymorphism, it would probably be possible to assign a *potential* deleterious function to almost every one. But clearly the human genome does not contain millions of potentially deleterious mutations (thousands maybe, but not millions!), so it is important to treat *in silico* predictions with caution. If a polymorphism shows genetic association with a phenotype it is important to first consider if the polymorphism is causal or in LD with a causal mutation. Hypotheses need to be constructed and tested in the laboratory. For example if a polymorphism is predicted to impact splicing, then *in vitro* analysis methods need to be employed to investigate evidence for alternative transcripts.

12.9 CONCLUSIONS

In this chapter we have taken an overview of some of the approaches for predictive functional analysis of polymorphisms in genes, proteins and regulatory regions. These methods can be applied equally at the candidate identification stage or at later stages to assist in the progression of associated genes to disease genes. The chapter has also examined the role of bioinformatics in the formulation of laboratory-based investigation for confirmation of functional predictions. As we have shown there are very few tools specifically designed

to evaluate the impact of polymorphisms on gene and protein function. Instead functional prediction of the potential impact of variation requires a very good grasp of the full gamut of bioinformatics tools used for predicting the properties and structure of genes, proteins and regulatory regions. This huge range of applications makes polymorphism analysis one of the most difficult bioinformatics activities to get right. The complexity of some analysis areas are worthy of special attention, particularly the analysis of polymorphisms in gene regulatory regions and protein sequences. To address some of these highly specialized analysis issues, Tom Werner presents a detailed examination of gene regulatory sequence analysis (Chapter 13) and Rob Russell and Matthew Betts present on tools and principles of protein analysis (Chapter 14).

REFERENCES

- Afshar-Kharghan V, Li CQ, Khoshnevis-Asl M, Lopez JA. (1999). Kozak sequence polymorphism of the glycoprotein (GP) Ib-alpha gene is a major determinant of the plasma membrane levels of the platelet GP Ib-IX-V complex. *Blood* **94**: 186–191.
- Aparicio SA. (2000). How to count human genes. *Nature Genet* **B25**: 129–130.
- Bennett CL, Brunkow ME, Ramsdell F, O'Briant KC, Zhu Q, Fuleihan RL, *et al.* (2001). A rare polyadenylation signal mutation of the FOXP3 gene (AAUAAA→AAUGAA) leads to the IPEX syndrome. *Immunogenetics* **53**: 435–439.
- Brenner S. (2000). Inverse genetics. *Curr Biol* **10**: R649.
- Chakravarti A. (1999). Population genetics — making sense out of sequence. *Nature Genet* **21** (Suppl.): 56–60.
- Chappell SA, LeQuesne JP, Paulin FE, deSchoolmeester ML, Stoneley M, Soutar RL, *et al.* (2000). A mutation in the c-myc-IRES leads to enhanced internal ribosome entry in multiple myeloma: a novel mechanism of oncogene de-regulation. *Oncogene* **19**: 4437–4440.
- Chen CY, Shyu AB. (1995). AU-rich elements: characterization and importance in mRNA degradation. *Trends Biochem Sci* **20**: 465–470.
- Conne B, Stutz A, Vassalli JD. (2000). The 3' untranslated region of messenger RNA: a molecular 'hotspot' for pathology? *Nature Med* **6**: 637–641.
- Di Paola R, Frittitta L, Miscio G, Bozzali M, Baratta R, Centra M, *et al.* (2002). A variation in 3prime prime or minute UTR of hPTP1B increases specific gene expression and associates with insulin resistance. *Am J Hum Genet* **70**: 806–812.
- D'Souza I, Schellenberg GD. (2000). Determinants of 4-repeat tau expression. Coordination between enhancing and inhibitory splicing sequences for exon 10 inclusion. *J Biol Chem* **275**: 17700–17709.
- Erdmann VA, Barciszewska MZ, Hochberg A, de Groot N, Barciszewski J. (2001). Regulatory RNAs. *Cell Mol Life Sci* **58**: 960–977.
- Graber JH, Cantor CR, Mohr SC, Smith TF. (1999). *In silico* detection of control signals: mRNA 3'-end-processing sequences in diverse species. *Proc Natl Acad Sci USA* **96**: 14055–14060.
- Gumucio DL, Shelton DA, Zhu W, Millinoff D, Gray T, Bock JH, *et al.* (1996). Evolutionary strategies for the elucidation of cis and trans factors that regulate the developmental switching programs of the beta-like globin genes. *Mol Phylogenet Evol* **5**: 18–32.
- Hastings ML, Krainer AR. (2001). Pre-mRNA splicing in the new millennium. *Curr Opin Cell Biol* **13**: 302–309.

- Heinemeyer T, Chen X, Karas H, Kel AE, Kel OV, Liebich I, *et al.* (1999). Expanding the TRANSFAC database towards an expert system of regulatory molecular mechanisms. *Nucleic Acids Res* **27**: 318–322.
- Ishii S, Nakao S, Minamikawa-Tachino R, Desnick RJ, Fan JQ. (2002). Alternative splicing in the alpha-galactosidase A gene: increased exon inclusion results in the Fabry cardiac phenotype. *Am J Hum Genet* **70**: 994–1002.
- Jacobs GH, Rackham O, Stockwell PA, Tate W, Brown CM. (2002). Transterm: a database of mRNAs and translational control elements. *Nucleic Acids Res* **30**: 310–311.
- Jo EK, Kanegane H, Nonoyama S, Tsukada S, Lee JH, Lim K, *et al.* (2001). Characterization of mutations, including a novel regulatory defect in the first intron in Bruton's tyrosine kinase gene from seven Korean X-linked agammaglobulinemia families. *J. Immunol* **167**: 4038–4045.
- Kaski S, Kekomaki R, Partanen J. (1996). Systemic screening for genetic polymorphism in human platelet glycoprotein Ib-alpha. *Immunogenetics* **44**: 170–176.
- Kleiman FE, Ramirez AO, Dodelson de Kremer R, Gravel RA, Argarana CE. (1998). A frequent TG deletion near the polyadenylation signal of the human HEXB gene: occurrence of an irregular DNA structure and conserved nucleotide sequence motif in the 3' untranslated region. *Hum Mut* **12**: 320–329.
- Knight JC, Udalova I, Hill AV, Greenwood BM, Peshu N, Marsh K, *et al.* (1999). A polymorphism that affects OCT-1 binding to the TNF promoter region is associated with severe malaria. *Nature Genet* **22**: 145–150.
- Kozak M. (1996). Interpreting cDNA sequences: some insights from studies on translation. *Mamm Genome* **7**: 563–574.
- Kozak M. (2000). Do the 5' untranslated domains of human cDNAs challenge the rules for initiation of translation (or is it vice versa)? **70**: 396–406.
- Krantz ID, Colliton RP, Genin A, Rand EB, Li L, Piccoli DA, *et al.* (1998). Spectrum and frequency of Jagged1 (JAG1) mutations in Alagille syndrome patients and their families. *Am J Hum Genet* **62**: 1361–1369.
- Lamb P, McKnight SL. (1991). Diversity and specificity in transcription regulation: the benefits of heterotypic dimerization. *Trends Biochem Sci* **16**: 417–422.
- Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, *et al.* (2001). Initial sequencing and analysis of the human genome. *Nature* **409**: 860–921.
- Le SY, Maizel JV Jr, (1997). A common RNA structural motif involved in the internal initiation of translation of cellular mRNAs. *Nucleic Acids Res* **25**: 362–369.
- Liu H, Chao D, Nakayama EE, Taguchi H, Goto M, Xin X, *et al.* (1999). Polymorphism in RANTES chemokine promoter affects HIV-1 disease progression. *Proc Natl Acad Sci USA* **96**: 4581–4585.
- Liu HX, Cartegni L, Zhang MQ, Krainer AR. (2001). A mechanism for exon skipping caused by nonsense or missense mutations in BRCA1 and other genes. *Nature Genet* **27**: 55–58.
- Lynch KW, Weiss A. (2001). A CD45 polymorphism associated with multiple sclerosis disrupts an exonic splicing silencer. *J Biol Chem* **276**: 24341–24347.
- Moller LB, Tumer Z, Lund C, Petersen C, Cole T, Hanusch R, *et al.* (2000). Similar splice site mutations of the ATP7A gene lead to different phenotypes: classical Menkes disease or occipital horn syndrome. *Am J Hum Genet* **66**: 1211–1220.
- Ohler U. (2000). Promoter prediction on a genomic scale: The Adh experience. *Genome Res* **10**: 539–542.

- Pauws E, van Kampen AH, van de Graaf SA, de Vijlder JJ, Ris-Stalpers C. (2001). Heterogeneity in polyadenylation cleavage sites in mammalian mRNA sequences: implications for SAGE analysis. *Nucleic Acids Res* **29**: 1690–1694.
- Peri S, Pandey A. (2001). A reassessment of the translation initiation codon in vertebrates. *Trends Genet* **17**: 685–687.
- Pesole G, Grillo G, Larizza A, Liuni S. (2000). The untranslated regions of eukaryotic mRNAs: structure, function, evolution and bioinformatic tools for their analysis. *Brief Bioinform* **3**: 236–249.
- Pitarque M, von Richter O, Oke B, Berkkan H, Oscarson M, Ingelman-Sundberg M. (2001). Identification of a single nucleotide polymorphism in the TATA box of the CYP2A6 gene: impairment of its promoter activity. *Biochem Biophys Res Commun* **284**: 455–460.
- Reese MG, Hartzell G, Harris NL, Ohler U, Abril JF, Lewis SE. (2000). Genome annotation assessment in *Drosophila melanogaster*. *Genome Res* **10**: 483–501.
- Reich DE, Lander ES. (2001). On the allelic spectrum of human disease. *Trends Genet* **17**: 502–510.
- Sakurai K, Seki N, Fujii R, Yagui K, Tokuyama Y, Shimada F, *et al.* (2000). Mutations in the hepatocyte nuclear factor-4 α gene in Japanese with non-insulin-dependent diabetes: a nucleotide substitution in the polypyrimidine tract of intron 1b. *Horm Metab Res* **32**: 316–320.
- Schmucker D, Clemens JC, Shu H, Worby CA, Xiao J, Muda M, *et al.* (2000). *Drosophila* Dscam is an axon guidance receptor exhibiting extraordinary molecular diversity. *Cell* **101**: 671–684.
- Shen LX, Basilion JP, Stantoon VP Jr. (1999). Single-nucleotide polymorphisms can cause different structural folds of mRNA. *Proc Natl Acad Sci USA* **96**: 7871–7876.
- Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, Sutton GG, *et al.* (2001). The sequence of the human genome. *Science* **291**: 1304–1351.
- Vervoort R, Gitzelmann R, Lissens W, Liebaers I. (1998). A mutation (IVS8 + 0.6kbpdelTC) creating a new donor splice site activates a cryptic exon in an Alu-element in intron 8 of the human beta-glucuronidase gene. *Hum Genet* **103**: 686–693.
- Wahle E, Keller W. (1992). The biochemistry of 3-end cleavage and polyadenylation of messenger RNA precursors. *Annu Rev Biochem* **61**: 419–440.