Bioinformatics for Geneticists. Edited by Michael R. Barnes and Ian C. Gray
Copyright © 2003 John Wiley & Sons, Ltd.
ISBNs: 0-470-84393-4 (HB); 0-470-84394-2 (PB)

SECTION 4

BIOLOGICAL SEQUENCE ANALYSIS AND CHARACTERIZATION

Bioinformatics for Geneticists. Edited by Michael R. Barnes and Ian C. Gray
Copyright © 2003 John Wiley & Sons, Ltd.
ISBNs: 0-470-84393-4 (HB); 0-470-84394-2 (PB)

CHAPTER 12

Predictive Functional Analysis of Polymorphisms: An Overview

MICHAEL R. BARNES

Genetic Bioinformatics
GlaxoSmithKline Pharmaceuticals
Harlow, Essex, UK

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 $\varepsilon 4$	16% (14%)	Alzheimer's and cardiovascular disease	107741
Factor V$^{\text{leiden}}$ 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

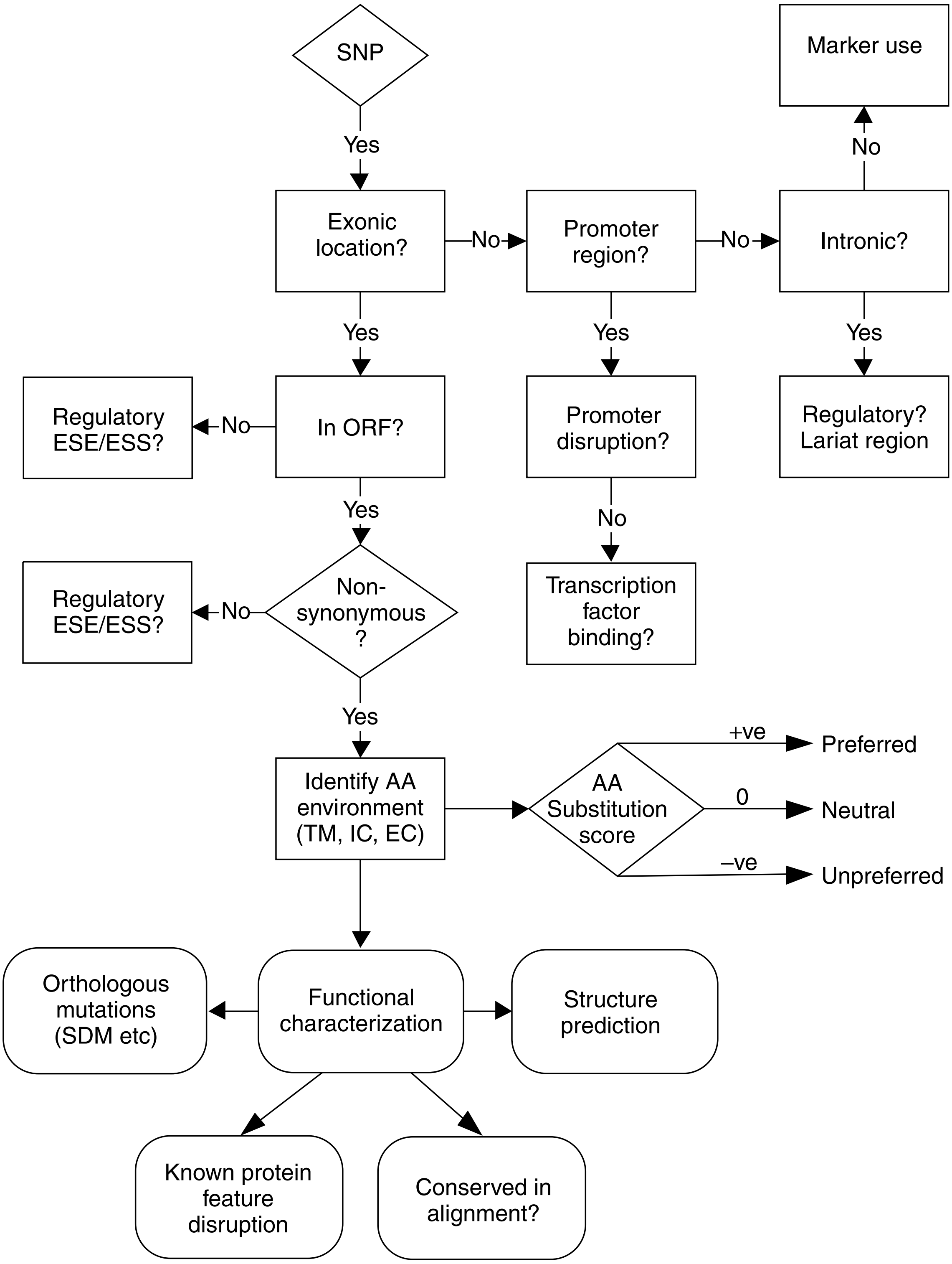

Figure 12.1 A decision tree for polymorphism analysis.

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 *et al.*, 1999)
Promoter	CYP2D6	Common — 48T > G substitution disrupts the TATA box of the CYP2D6 promoter, causing 50% reduction in expression. (Pitarque *et al.*, 2001)
Promoter	RANTES in HIV progression	−28G mutation increases transcription of the RANTES gene slowing HIV-1 disease progression (Liu *et al.*, 1999)
cis-regulatory element	Bruton's tyrosine kinase in X-linked agammaglobulinemia	+5G/A (intron 1) shows reduced BTK transcriptional activity, suggesting a novel *cis*-acting element, involved in BTK downregulation but not splicing (Jo *et al.*, 2001)
Lariat region	HNF-4alpha	NIDDM-associated C/T substitution in polypyrimidine tract in intron 1b in an important *cis*-acting element directing intron removal (lariat region) (Sakurai *et al.*, 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 *et al.*, 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 *et al.*, 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 *et al.*, 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 *et al.*, 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 *et al.*, 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://wwwiti.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

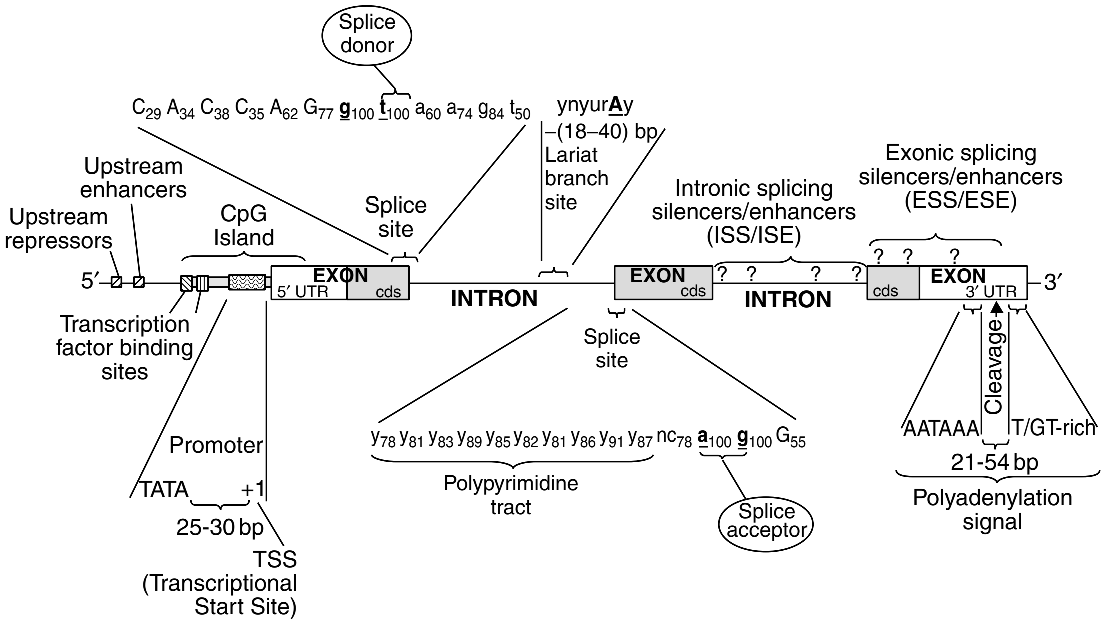
Splice donor
$C_{29}\ A_{34}\ C_{38}\ C_{35}\ A_{62}\ G_{77}\ \underline{\mathbf{g}}_{100}\ \underline{\mathbf{t}}_{100}\ a_{60}\ a_{74}\ g_{84}\ t_{50}$
ynyur**A**y
–(18–40) bp
Lariat branch site
Upstream enhancers
Upstream repressors
CpG Island
Splice site
Intronic splicing silencers/enhancers (ISS/ISE)
Exonic splicing silencers/enhancers (ESS/ESE)
5′
EXON
5′ UTR
cds
INTRON
EXON
cds
? ? ? ?
INTRON
? ? ?
cds
EXON
3′ UTR
3′
Transcription factor binding sites
Splice site
Cleavage
Promoter
TATA
+1
25-30 bp
TSS (Transcriptional Start Site)
$y_{78}\ y_{81}\ y_{83}\ y_{89}\ y_{85}\ y_{82}\ y_{81}\ y_{86}\ y_{91}\ y_{87}\ nc_{78}\ \underline{\mathbf{a}}_{100}\ \underline{\mathbf{g}}_{100}\ G_{55}$
Polypyrimidine tract
Splice acceptor
AATAAA
T/GT-rich
21-54 bp
Polyadenylation signal

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 'YNYUR**A**Y' 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

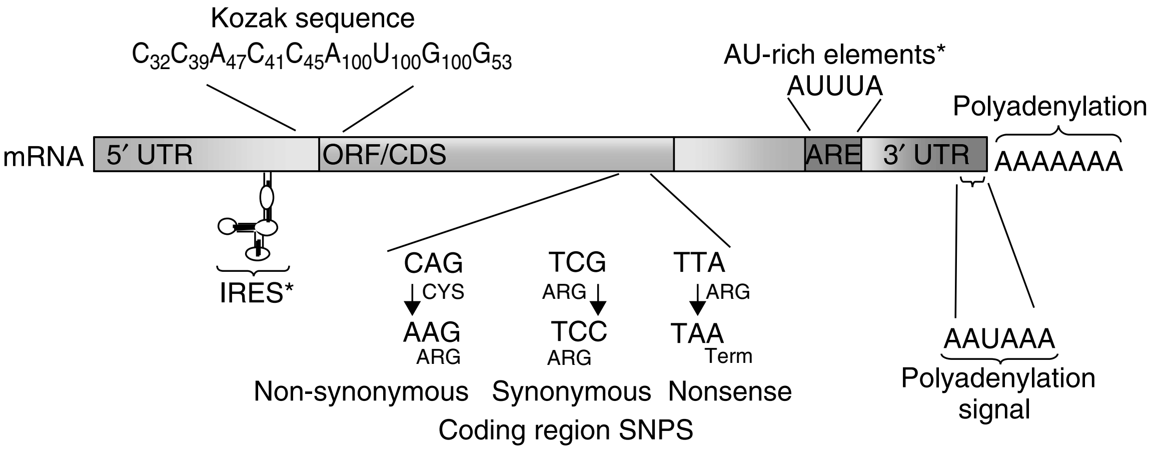

Figure 12.3 The anatomy of an mRNA transcript. This figure illustrates some of the key regulatory and structural elements that control the translation, stability and post-transcriptional processing of mRNA transcripts. Polymorphisms in these regions should be investigated for functional effects.

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 '$C_{32}C_{39}A_{47}C_{41}C_{45}A_{100}U_{100}G_{100}G_{53}$' 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 *et al.*, 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 *et al.*, 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 *et al.*, 2000)
UTR stability	Protein tyrosine phosphatase-1B (PTP1B)	1484insG in 3′ UTR causes PTP1B over-expression leading to insulin resistance (Di Paola *et al.*, 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, 7 bp 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 multifunctional, 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

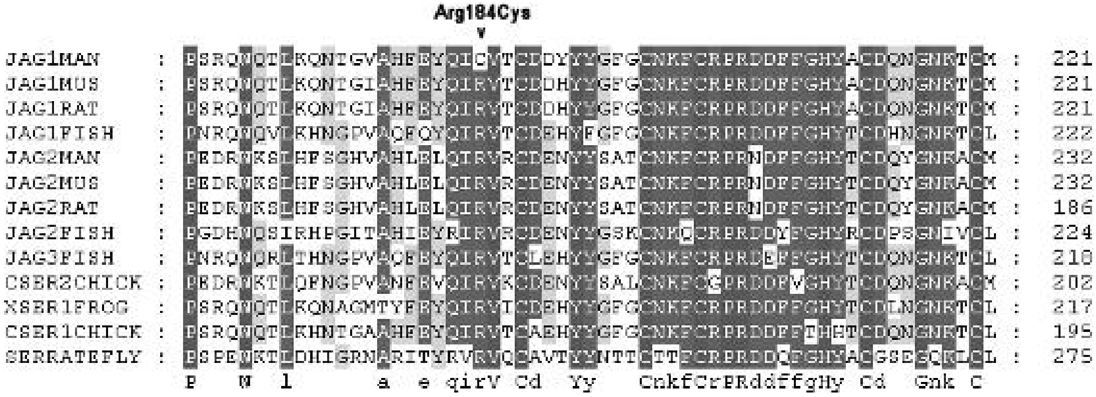
Arg184Cys
JAG1MAN : PSRQWQTLKQNTGVAHFEYQICVTCDDYYYGFGCNKFCRPRDDFFGHYACDQNGNKTCM : 221
JAG1MUS : PSRQWQTLKQNTGIAHFEYQIRVTCDDHYYGFGCNKFCRPRDDFFGHYACDQNGNKTCM : 221
JAG1RAT : PSRQWQTLKQNTGIAHFEYQIRVTCDDHYYGFGCNKFCRPRDDFFGHYACDQNGNKTCM : 221
JAG1FISH : PNRQWQVLKHNGPVAQFQYQIRVTCDEHYFGFGCNKFCRPRDDFFGHYTCDHNGNKTCL : 222
JAG2MAN : PEDRWKSLHFSGHVAHLELQIRVRCDENYYSATCNKFCRPRNDFFGHYTCDQYGNKACM : 232
JAG2MUS : PEDRWKSLHFSGHVAHLELQIRVRCDENYYSATCNKFCRPRNDFFGHYTCDQYGNKACM : 232
JAG2RAT : PEDRWKSLHFSGHVAHLELQIRVRCDENYYSATCNKFCRPRNDFFGHYTCDQYGNKACM : 186
JAG2FISH : PGDHWQSIRHPGITAHIEYRIRVRCDENYYGSKCNKQCRPRDDYFGHYRCDPSGNIVCL : 224
JAG3FISH : PNRQWQRLTHNGPVAQFEYQIRVTCLEHYYGFGCNKFCRPRDEFFGHYTCDQNGNKTCL : 210
CSER2CHICK : PEDRWKTLQFNGPVANFEVQIRVKCDENYYSALCNKFCGPRDDFVGHYTCDQNGNKACM : 202
XSER1FROG : PSRQWQTLKQNAGMTYFEYQIRVTCDEHYYGFGCNKFCRPRDDFFGHYTCDLNGNKTCL : 217
CSER1CHICK : PSRQWQTLKHNTGAAHFEYQIRVTCAEHYYGFGCNKFCRPRDDFFTHHTCDQNGNKTCL : 195
SERRATEFLY : PSPEWKTLDHIGRNARITYRVRVQCAVTYYNTTCTTFCRPRDDQFGHYACGSEGQKLCL : 275
P W l a e qirV Cd Yy CnkfCrPRddffgHy 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/sprot-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-4alpha 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 + 0.6kbdelTC) 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.

Bioinformatics for Geneticists. Edited by Michael R. Barnes and Ian C. Gray
Copyright © 2003 John Wiley & Sons, Ltd.
ISBNs: 0-470-84393-4 (HB); 0-470-84394-2 (PB)

CHAPTER 13

Functional *In Silico* Analysis of Non-coding SNPs

THOMAS WERNER

Genomatix Software GmbH
Munich, Germany

13.1 INTRODUCTION

The total amount of nucleotides within the human genome was found to be well within the expected range of about 3 billion base pairs. That was no big surprise since the physical size of the genome could already be measured fairly accurately by biophysical means well before sequencing became possible. However, the number of genes turned out to be surprisingly low, especially after the gene counts of *Drosophila melanogaster* and *Caenorhabtis elegans* were released (Adams *et al.*, 2000; The *C. elegans* Sequencing

Consortium, 1998). It was common sense that humans should have at least double or triple the amount of genes as compared to those much simpler organisms. However, despite high expectations and correspondingly high initial estimates, the number of genes to be expected within the human genome decreased constantly. There is still no final answer but current estimates converge somewhere between 30,000 and 40,000 (Venter *et al.*, 2001). This leads to a lot of questions as to where the huge differences between species will be found in the genomes, if not in gene numbers. It is also quite obvious from those numbers that only about 2–3% of the human genome is expected to encode proteins. Even disregarding the 40% repetitive sequences present in the human genomic sequence, this leaves more than half of the genomic sequence in search of a function.

Of course, encoding proteins is just one of the many known functions of the genome. There are three very prominent additional tasks that must be fulfilled by the genome. The first one is to maintain some physical ordered structure of the genomic sequence, which is a prerequisite for everything else. A hopeless tangle of 3 billion base pairs would most likely interfere severely with gene expression as well as with DNA replication.

The second task that has to be faithfully fulfilled over a lifetime in any organism is the correct replication of the genomic information to allow cell divisions. And last but not least, gene expression itself involves much more than synthesis of an RNA copy of the coding parts of the genome. The correct regulation both of transcription as well as DNA replication in space and time is probably the most crucial part of life for any organism. No cell let alone a multicellular organism, can develop or survive without perfect control over gene expression (control of replication is just one of the consequences of controlled gene expression).

Here the genome has to fulfil a formidable task. The information encoded within the genome can be regarded as invariant regardless of the few mutations that occur continuously within a living cell (most are either repaired or eliminated by selection). This view also includes Single Nucleotide Polymorphisms (SNPs) because most SNPs are frozen in evolution and very few arise during the lifespan of an individual organism. Survival of such mutations becomes most prominently visible in allelic differences where there appears to be more than one solution for a functional sequence. Development and differentiations are examples of extremely complex and linked programmes that have to be fulfilled in an exact time-frame. Nevertheless, this is the easy part for the genome as both of these processes are deterministic and every step is clear from the very beginning with very little variation included.

In contrast, every organism encounters a variety of unexpected environmental stimuli (availability of food resources, climate conditions, interactions with other organisms such as predators or competing species). The static genome must provide *a priori* all information suitable to react appropriately to such external challenges. This requires an enormous amount of ‘conditional programming’ within the genetic code, most of which is not directly manifest in protein sequences. This is probably the major reason why regulatory sequences appear to occupy almost five to 10 times more genomic sequence than coding regions (this estimate includes all regulatory sequences not only transcriptional regulation). For the very same reason most of this chapter will focus on regulatory aspects. The allelic differences (something in the region of one nucleotide in 1000) also called Single Nucleotide Polymorphisms SNPs (pronounced ‘Snips’) may have no effect under normal conditions but can make a huge difference in the case of changing conditions. Such changes do not need to be external. A developing tumour can bring a dramatic change to the physiology of an organism and genetic predisposition is linked to a large extent to allelic differences. However, the effect of any individual SNP or sets of SNPs (e.g. haplotypes) depends critically on the local context within which the SNP is located. Therefore, functional analysis or estimation of SNPs requires detailed understanding of the functional features and sequence regions within the genome.

13.2 GENERAL STRUCTURE OF CHROMATIN-ASSOCIATED DNA

DNA is complexed with histone proteins forming nucleosomes which are distributed along most of the genomic DNA like beads on a string. This structure is then organized in chromosomal loops and such loops are known to form solenoids (Daban and Bermudez, 1998). Solenoids in turn form the chromosomal fibres already visible by microscopy. For the purpose of this chapter the most relevant structures are all within a chromosomal loop. Therefore, the schematic organization of a chromosomal loop will be used as a framework for the explanation of all further components.

Repetitive DNA of retroposon origin is ubiquitously found throughout the genome. As we learned from the first published chromosomal sequence (chromosome 22, Dunham *et al*., 1999) about 40% of the human genomic DNA consists of repetitive DNA, most prominent among these are ALU repeat sequences. They have been named after the restriction enzyme (ALU I) which generates a characteristic satellite band in digests of genomic DNA. ALUs belong to the class of short interspersed elements (SINEs) which are short retroposon sequences of only about 300 nucleotides in length. Another class of repetitive DNA is the long interspersed elements (LINEs) which reach up to 7 kb in length and include retroviral sequences (Smit, 1999).

A chromatin loop is the region of chromosomal DNA located between two contact points of the DNA with a protein framework within the nucleus, the so-called nuclear matrix. These contact points are marked in the genomic DNA as Matrix/Scaffold Attachment Regions (S/MARs). Association of DNA with this nuclear matrix is a prerequisite for transcription of nucleosomal DNA (Bode *et al*., 2000). S/MARs are themselves complex structures not yet fully understood at the molecular level. There is an excellent review on chromatin domains and prediction of MAR sequences by Boulikas (1995) explaining S/MARs and their elements in detail. There are currently two methods to detect S/MAR elements in genomic sequences, the first is MARFinder by Kramer (1996) (http://www.ncgr.org/MAR-search/). The second method is SMARTest developed by Genomatix Software GmbH, Munich and available for academic researchers free of charge from http://www.genomatix.de (Frisch *et al*., 2002).

Enhancers are regulatory regions found within chromosomal loops that can significantly boost the level of transcription from a responsive promoter regardless of their orientation and distance with respect to the promoter within the same chromatin loop. Currently, there is no way to detect enhancers in general by *in silico* methods. However, at least a subclass of enhancers is organized in a very similar manner to that of promoters, i.e. they also contain frameworks of transcription factors (Gailus-Durner *et al*., 2000). In cases where enhancers share modules with promoters it is possible to find them via the module. However, since an isolated module match is no proof of either a promoter or an enhancer, experimental verification is still mandatory. Silencers are basically identical to enhancers and follow the same requirements but exert a negative effect on promoter activities. Enhancers and silencers often show a similar internal organization as promoters (Werner, 1999).

13.3 GENERAL FUNCTIONS OF REGULATORY REGIONS

The biological functionality of regulatory regions is generally not a property evenly spread over the regulatory region in total. Functional units are usually defined by a combination of defined stretches that can be delimited and possess an intrinsic functional property (e.g. binding of a protein or a curved DNA structure). Several functionally similar types of these stretches of DNA are already known and will be referred to as *elements*. Those elements are neither restricted to regulatory regions nor individually sufficient for the regulatory function of a promoter or enhancer. The function of the complete regulatory region is composed of the functions of the individual elements either in an additive manner (independent elements) or by synergistic effects (modules) (Werner, 1999). With respect to SNPs it is important to view regulatory DNA in a similar way as we regard coding genes: there are short stretches of immediate functional importance (e.g. exons or regulatory elements) and there are much larger regions with either unknown or more implicit functions (introns, ‘spacer’ DNA in regulatory regions). As a direct consequence of this sort of discontinuous organization of functional sequences the potential effects of SNPs can vary dramatically (from none to lethal) depending which part of the sequence the SNP is located in.

13.4 TRANSCRIPTION FACTOR BINDING SITES (TF-SITES)

Binding sites for specific proteins are most important among regulatory elements. They consist of about 10 to 30 nucleotides, not all of which are equally important for protein binding, a reminder of the somewhat fractal properties of genomic sequences, i.e. a binding site looks like a tiny copy of a promoter, which in turn looks like a small copy of a gene, etc. Individual protein binding sites may vary in part of their sequence, even if they bind to the same protein. There are nucleotides which are in contact with the protein in a sequence-specific manner (‘recognition exons’), which usually represent the best-conserved areas of a binding site. Different nucleotides are involved in more non-specific contacts to the DNA backbone (i.e. not sequence specific as they do not involve the bases A, G, C or T), and there are internal ‘spacers’ (‘introns’) which are not in contact with the protein at all. All in all, protein binding sites exhibit enough sequence conservation to allow for the detection of candidates by a variety of sequence similarity-based approaches. There have been many attempts to collect TF binding sites (Wingender *et al*., 2001), as well as several developments in the location of TF binding sites in genomic sequences (Chen *et al*., 1995; Prestridge, 1996; Quandt *et al*., 1995). However, potential binding sites can be found almost anywhere in the genome and are not restricted to regulatory regions. Quite a number of binding sites outside regulatory regions are also known to bind their respective binding proteins (e.g. Kodadek, 1998). Therefore the abundance of predicted binding sites is not just a shortcoming of the detection algorithms but reflects biological reality although currently hard to interpret in functional terms.

13.5 STRUCTURAL ELEMENTS

Secondary structures are mostly known for RNAs and proteins but they also play important roles in promoters (e.g. Bates *et al*., 2001). Potential secondary structures can be easily determined. For an excellent start point see Michael Zuckers homepage (Table 13.1).

TABLE 13.1 Useful URLs for Regulatory SNP Analysis

RNA secondary structure prediction	
Homepage of M. Zucker	http://bioinfo.math.rpi.edu/□zukerm/
Pattern definition:	
CoreSearch (ftp)	ariane.gsf.de/pub/unix/coresearch_1.2.tar.Z
CONSENSUS	http://bioweb.pasteur.fr/seqanal/interfaces/consensus-simple.html
S/MAR detection	
MARFinder	http://www.ncgr.org/MAR-search/
SMARTest	http://www.genomatix.de
UTR analysis	
UTR database	http://bighost-area.ba.cnr.it/BIG/BioWWW/#UTRdb
Genomatix tools	
(free to academic users)	
SMARTest	http://www.genomatix.de/free_services/
PromoterInspector	
MatInspector professional	
GEMS Launcher	
ELDorado	
Sequence tools	

There is a plethora of tools now available on the web. However, Michael Zucker has been one of the most important pioneers in the field, so starting from his page would be a good choice. Secondary structures are also often not conserved in primary nucleotide sequence but are subject to strong positional correlation within the structure. There is also always a trade-off between best and fastest structure prediction. Some algorithms dive deep into energy calculations to provide the best possible structure for one RNA while others do a much more rudimentary analysis, which can be applied to many sequences within the same time-frame as a single in-depth analysis would take. It is impossible to decide in advance which approach will be most suitable for any problem. A few experiments using different methods will be called for.

13.6 ORGANIZATIONAL PRINCIPLES OF REGULATORY REGIONS

Regulatory regions are not just statistical collections of the regulatory elements introduced above. Therefore, it is necessary to understand at least some basic organizational features of regulatory regions in order to understand the different consequences SNPs can have. Eukaryotic polymerase II promoters will serve as examples as they appear to be the currently best-studied regulatory regions. The TF-sites within promoters (and likewise most other regulatory sequences) do not show any obvious general patterns with respect to location and orientation within the promoter sequences. TF binding sites can be found virtually everywhere in promoters but in individual promoters possible locations are much more restricted. A closer look reveals that the function of a TF binding site often depends on the relative location and especially on the sequence context of the binding site.

The context of a TF-site is one of the major determinants of its role in transcription control. However, the context is not merely a few nucleotides around the binding site, which would be more like an extension of the binding site rather than a context. More important is the context of other TF binding sites located at some distance that are often grouped together and such functional groups have been described in many cases. A systematic attempt to collect synergistic or antagonistic pairs of TF binding sites has been made with the COMPEL database (Kel *et al.*, 1995, Heinemeyer *et al.*, 1998). In many cases, a specific promoter function (e.g. a tissue-specific silencer) will require more than two sites simultaneously. Such groups of promoter subunits consisting of several TF binding sites that carry a specific function independent of the promoter, will be referred to as *promoter modules*. This is a definition at the molecular level, which is more specific than the definition recently given by Arnone and Davidson (1997) requiring only the presence of the sites within a loosely defined DNA region. Within a molecular promoter module both sequential order and distance can be crucial for function indicating that these modules may be the critical determinants of a promoter rather than individual binding sites. However, promoters can contain several modules that may use overlapping sets of binding sites. Therefore, the conserved context of a particular binding site cannot be determined from the primary sequence without additional information about the modular structure (Figure 13.1).

The peculiar property of promoter modules to function only as intact units has an important consequence for the effects SNPs can have in promoters. A SNP inactivating a single TF binding site can in fact destroy the function of a complete module, which will not be obvious from inspection of the individual binding site in which the SNP was detected. A SNP affecting another binding site of the same factor within the same promoter may have a quite different effect if this binding site is either part of another module or has no direct function at all. Therefore, identification of binding sites affected by SNPs is not enough to estimate the functional consequences of regulatory SNPs.

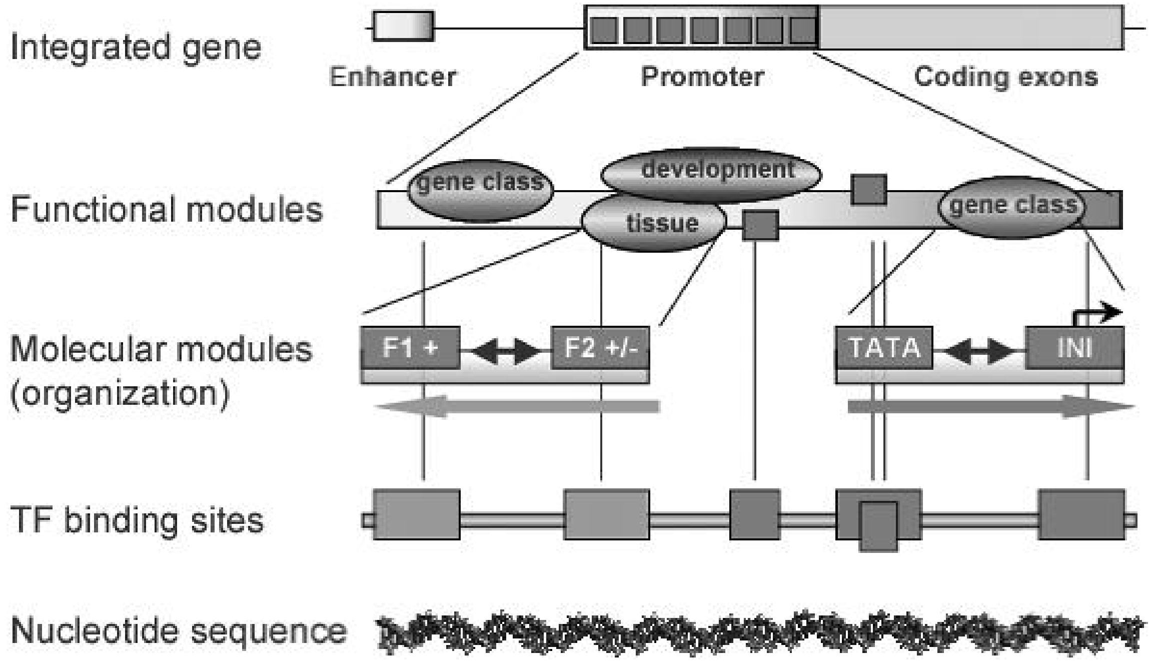

Figure 13.1 Hierarchical structure of a polymerase II promoter (schematic). Oval shapes indicate modules without defined internal structure; rectangular boxes in the lower part of the figure indicate transcription factor binding sites. Note that a direct assignment of individual binding sites to functional modules is only possible for molecular modules. The arrows below molecular modules indicate strand orientation of the modules.

Similar logic holds for insertions/deletions in promoter sequences (e.g. polymorphic microsatellite sequences). The quite variable distances found between elements in functionally related promoters could be interpreted as the result of insertion/deletion events, although it is hard to find clear evidence for this. Insertions or deletions affecting the organization of modules may well interfere with function. However, as most of the promoter functions are crucial for the function of the whole gene, such deleterious mutations are most likely selected against in evolution. Therefore, the variability in spacing seen in present time sequences should be considered neutral with respect to function unless there is direct evidence to the contrary.

13.7 RNA PROCESSING

Genomic regulation does not stop at the level of the genomic DNA sequence. RNAs contain important regulatory signals of their own, among them are transport signals that direct the RNAs to specific subcellular locations and RNA instability signals that mark RNAs for rapid destruction unless specifically protected. Most of these signals reside within the 5′ and 3′ untranslated terminal regions of the mRNAs (UTRs), which have been collected in a specialized database (Pesole *et al.*, 1996, see Table 13.1 for URL). This database is an excellent representation of the knowledge about UTRs as reported in the literature. However, there were no efforts made to complete 5′ or 3′ UTRs in case they were reported incomplete. Because RNAs are faithful copies of the genomic sequences (except for RNA editing) all of these signals can also be directly studied in the genomic DNA.

Removal of intronic sequences is one of the most important processing steps of primary transcripts (splicing). As became quite clear in recent years splicing is governed by complex and discontinuous signals located within exons as well as introns (Kramer, 1996). The set-up of complete splice signals resembles the general set-up of promoters and enhancers to a large extent. There are splice enhancers, splice donor and acceptor sites, branch point sequences as well as some less well defined accessory sequences within introns. Again the organizational context of splicing elements is most likely the most important factor determining biological function.

13.8 SNPs IN REGULATORY REGIONS

SNPs are to be found all over the genome and as a mere consequence of the amount of sequence a considerable number of SNPs are expected to be located within regulatory sequences. As discussed above the potential effects of SNPs on gene regulation depend on the location of SNPs with respect to the regulatory elements. SNPs located in non-functional spacer DNA (if anything like that exists) will not affect regulation in all likelihood, whereas a SNP destroying the binding site of a crucial transcription factor can alter transcription of a gene quite dramatically.

13.8.1 Examples for Regulatory SNPs

SNPs can influence TF binding sites in three different ways. A binding site can be destroyed by loss of binding affinity due to the SNP. The opposite effect is also possible, that is, generation of a new binding site within a regulatory sequence. A combination of these two events would result in an altered binding site that might have switched specificity to another protein. There are examples with well-established effects on gene regulation for TF binding sites being deleted as well as created by SNPs.

The RANTES gene encodes a chemokine involved in immune signalling. Unfortunately, RANTES expression is also involved in supporting HIV-1 infection, which of course is detrimental for the individual. There is one mutation (SNP) known in the RANTES promoter that destroys a potential c-myb binding site immediately upstream of the TATA box. This mutation has the astonishing effect of delaying the CD4 depletion by HIV-1 infection although it has no effect on the infection itself. However, it was found that this mutation increases the transcription of the RANTES chemokine gene, demonstrating nicely that SNPs *per se* are not determined to be positive or negative (Lui *et al.*, 1999). The positive effect of this SNP only becomes apparent upon HIV-1 infection.

Cystic fibrosis is a devastating disease caused by a defective protein which results in a dramatically shortened lifetime for the sufferers. In one case a SNP generated a new binding site for the transcription factor YY1 in the promoter of the gene already affected by a mutation in exon 11, which caused the disease. The effect of this new YY1 binding site was over-expression of the (not completely) defective protein via attracting additional protein(s) to the promoter complex, which reduced the symptoms of the disease via a gene dosage effect (Romey *et al.*, 2000).

The ‘bottom line’ of all these examples is that SNPs affected transcriptional control elements that were actually involved in the gene transcription of the respective genes.

13.9 EVALUATION OF NON-CODING SNPs

In a case where a SNP is located within the coding sequence of a gene it is very simple to find out whether this is a silent exchange or not, just from the triplet code. In the case where there is an amino acid exchange it is much less obvious whether the exchange will affect protein function or not. If there is no known example for the particular exchange there is no way to predict the functional consequences solely from the sequence.

SNPs in regulatory regions are always difficult to assess for two reasons: the first is simply to find out whether a SNP is located inside regulatory regions at all. Given that locating regulatory regions is much more difficult than locating coding sequences, this is no trivial task. However, even if this prerequisite can be satisfied there remains the question of whether the SNP affects any regulatory elements. This requires knowledge or at least a well-supported hypothesis about the regulatory elements relevant for the regulation of the gene in question. Since this information cannot be directly derived from the nucleotide sequence of a promoter for example, this requires additional efforts to locate relevant regulatory elements.

Once it has been established that a SNP is located within a putative regulatory element, it is possible to evaluate the primary effect of the SNP on this regulatory element. The primary effect is the change in binding affinity or specificity of a TF binding site or the change in the stability of a secondary structure. This kind of information can be derived from a comparison of the wild-type sequence with the SNP-containing sequence and usually a fair estimate of the resulting changes can be made.

Unfortunately, even that information is only part of the answer. The real question is whether the SNP-induced change has functional consequences on regulation, which is not necessarily a consequence of a single altered regulatory element. Therefore, it is also necessary to ascertain the relevant context of the affected elements, e.g. whether a TF binding site is part of a transcriptional module or not.

13.10 SNPs AND REGULATORY NETWORKS

Although SNPs are always necessarily located within one gene or one regulatory region and they can have pleiotropic effects. SNPs affecting the protein sequence or the regulation of transcription factors can influence the expression of many target genes of this particular factor and lead to pronounced systemic effects. This is achieved via the regulatory networks in which the affected transcription factor participates. A functional correlation of a SNP with an observed phenotype can be established generally by epidemiological studies. However, this kind of correlation does not reveal any data about the molecular mechanisms behind the correlation. The molecular link between the regulatory SNP and the observed phenotype is finally established on the level of regulatory networks by tracking relevant transcriptional modules. However, reconstruction of regulatory networks on the molecular level is far from easy with current tools and may turn out to be (still) a futile effort in many cases. Where it works, it provides the final answer not only to the effect of the SNP but also reveals the molecular mechanisms behind the phenotype. Due to that enormous gain in insight as well as the therapeutic possibilities, it is well justified to invest a significant effort into the elucidation of the pertinent regulatory networks even if the chances of success are sometimes slim.

13.11 SNPs MAY AFFECT THE EXPRESSION OF A GENE ONLY IN SPECIFIC TISSUES

In addition, SNPs are always necessarily present in all tissues and their effects may vary dramatically. As outlined above many promoter modules are active under specific conditions only, such as when they are stimulated by a signalling pathway or only in particular cell or tissue types. Consequently, a SNP affecting a specific module will only show an effect under conditions where the corresponding module is active. Therefore, lack of association of a promoter SNP with an observable phenotype in cell culture experiments only excludes a functional effect in the particular cell type under the specific conditions used in the experiment. The very same ‘silent’ SNP may have a clear effect either in the same cells under different conditions or in other cells/tissues which have not been tested. Therefore, *in vitro* results are only conclusive in the case of positive results, while negative results are of limited value.

13.12 *IN SILICO* DETECTION AND EVALUATION OF REGULATORY SNPs

So far we have established the basic factors and requirements of how to attack the problem of regulatory SNPs. Now it is time to detail the strategies that will help to elucidate the different levels of SNP-caused effects in a practical approach.

Figure 13.2 shows an overview of the general strategy to analyse SNPs for potential regulatory effects. In brief SNPs are first mapped onto predetermined regulatory regions (in this example, promoters). The gene annotation is used to identify promoters of interest as well as to include pre-existing knowledge about functional aspects of these promoters, e.g. promoter elements already known to be involved in promoter function. Then relevant transcriptional elements are identified either from knowledge databases or by comparative analyses of sets of functionally related promoters (detailed below). At this point it is

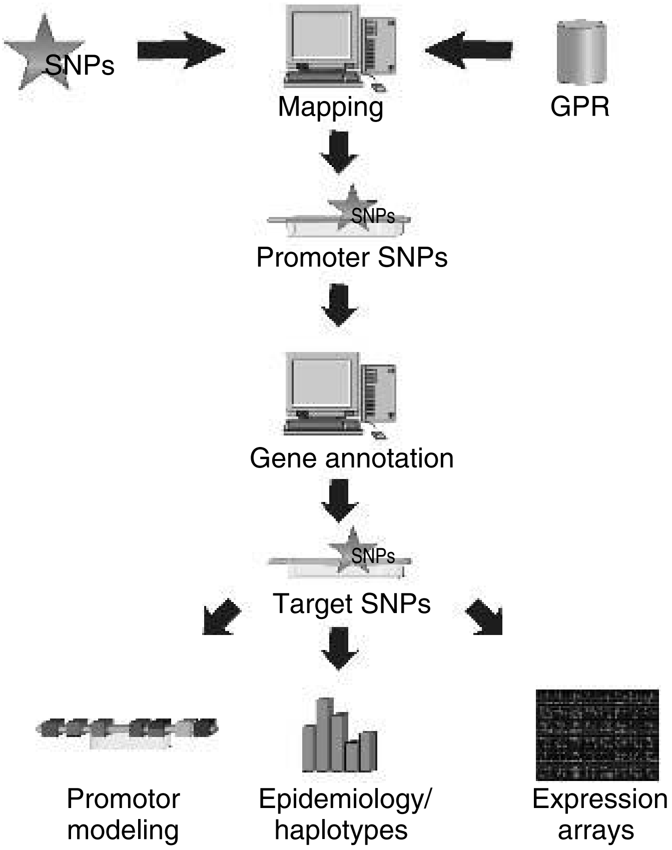

Figure 13.2 Strategy to qualify genomic SNPs as relevant regulatory SNPs. GPR (Genomatix Promoter Resource now part of ElDorado®). The transparent light grey box symbolizes a promoter region; darker boxes indicate transcription factor binding sites. (See note added in proof).

possible to determine whether the SNP is located within a relevant transcription factor binding site and the effect on binding affinity can be calculated. The comparative sequence analysis may also have revealed promoter modules facilitating estimation of regulatory effects. However, since a promoter can contain several independent functional modules it may be necessary to re-analyse the promoter in another functional context to focus on the effect of a particular SNP.

How to implement such a general strategy for real life application? This will be shown in the example below of Genomatix sequence analysis tools that were especially developed to facilitate this kind of approach.

13.13 GETTING PROMOTER SEQUENCES

Promoter sequences can be derived from the literature (about 10 to 20% of genomic promoters), by promoter prediction (about 50% of the genomic promoters) or by mapping of 5′-complete mRNAs (up to the Transcription Start Site, TSS) to the genomic sequence. In the last case the promoter is the sequence containing about 100 bp of the mRNA and a region of about 500–600 nucleotides immediately upstream of the TSS. So far there are also only about 10% of the mRNAs available as 5′-complete sequence. In summary it is possible to obtain about 50 to 70% of the human promoters by a combination of these approaches.

Of course, there are resources as well as methods claiming to be able to provide up to 90% of human promoters. However, the problem is that higher sensitivity is always achieved at the cost of lower specificity and a promoter collection containing all true promoters burdened by a high amount of false positives is absolutely useless for the regulatory SNP analysis, while half of the promoters with few false positives are sufficient to obtain useful results, although only for a subset of regulatory SNPs.

There is no best way to deal with the results of promoter prediction. For example, while it is quite popular to let the user play with some sort of cryptic scoring/quality Genomatix PromoterInspector does not have such a parameter. A promoter cannot be predicted with more or less scoring. The point is that every method will also produce false positives at any threshold. Any specificity given is only valid using exactly the corresponding parameters. A much better way to strengthen the likelihood of a true prediction is to check for additional evidence, e.g. a gene annotation or prediction that indicates the same region as the potential promoter. However, for a significant part of the human genome there is no alternative to prediction at this time.

Genomatix provides the Genomatix Promoter Resource (GPR) for this purpose (containing predictions for about 50% of all human promoters with a false positive rate of less than 15%, Scherf *et al.*, 2000, 2001) complemented by mRNA mapping as well as promoters extracted from the experimental literature. This is by no means the most complete collection of promoters available but most likely the one with the least false positives. GPR is a product of Genomatix that requires licensing. However, the software tool used to generate GPR (PromoterInspector) is available (with some restrictions) to academic scientists free of charge (URL see Table 13.1).

This resource can be used to map SNPs to the promoter regions (Genomatix software can do this automatically high throughput). Based on the gene annotation corresponding to these promoters, genes of interest with SNPs inside the promoter regions can be selected for further analysis.

At this point there are two possible strategies to evaluate the functional importance of the SNPs in question. A straightforward approach involves epidemiological data connecting the SNP to a phenotype by statistical coupling as can be seen in haplotype studies (e.g. Judson *et al.*, 2000, Stephens *et al.*, 2001). If such data are available the analysis can directly proceed to the identification of the transcription factor binding sites affected and the consequences for binding affinities (see below).

13.14 IDENTIFICATION OF RELEVANT REGULATORY ELEMENTS

If no such epidemiological data are available it is necessary to first select binding sites which are likely to be involved in promoter function, because direct analysis for potential binding sites usually yields about a 10 times excess of potential binding sites. Selection can be carried out with Genomatix software (GEMS Launcher) and is based on the principle of evolutionary conservation of functional binding sites in promoters. There are two ways to assess such functional conservation. The first is to compare promoters from orthologous genes in several species (e.g. man, mouse and dog or another non-rodent mammal). Of course mouse, rat and hamster might be related too closely to reveal a useful pattern. Such an analysis usually results in a conserved framework of about three to eight binding sites, which can be directly used for further evaluation. An example of such an analysis was the determination of the general mammalian actin promoter model (Frech *et al.*, 1998).

Another approach is horizontal conservation derived from sets of genes within the same organisms that are coupled functionally, e.g. by co-expression. The strategy has already been outlined in detail in Werner (2001).

Although not immediately evident, the difference between co-regulation and mere co-expression is of great importance for this strategy. Co-regulated genes usually share partial promoter features, so-called modules responsible for the observed co-regulation. Co-expressed genes which just show up at the same time but are not co-regulated do not necessarily share such modules. Therefore, they may interfere with comparative sequence analysis and should be removed first. There are several possible ways to focus co-regulated rather than co-expressed genes.

The basic idea of transcription event-oriented clustering is to include additional information beyond the mere expression level into the clustering process. One way to do this is to use multiple time points. Different pathways might resemble each other in expression level of genes for a limited period of time but separate at other time points. Clustering of genes based on time profiles of gene expression is leading more directly towards identification of the underlying mechanisms than clustering based on expression levels at a single time point. Groups of genes suitable for clustering can be derived from pathway information or directly from expression arrays. Again the GPR can be used to locate the corresponding promoters for human genes (or PromoterInspector to analyse other mammalian genomic sequences for promoters, Scherf *et al*., 2000).

Comparative sequence analysis of such a set of co-expressed genes usually reveals the promoter module responsible for the observed co-expression and not a complete model. If the same promoter is analysed in combination with different sets of expression-related genes, different modules may be found. For example, analysis of a set of promoters expressed during glucose starvation may reveal a different module within the promoters than analysis of one of these promoters in a context of growth factor-induced genes.

13.15 ESTIMATION OF FUNCTIONAL CONSEQUENCES OF REGULATORY SNPs

The selection of conserved binding sites either from the analysis of orthologous promoters or from sets of co-expressed genes can be directly evaluated for SNP-induced differences in binding affinities, e.g. by applying the MatInspector program (also integrated in the GEMS Launcher, Frech *et al*., 1997; Quandt *et al*., 1995). If a site from this selected set is affected, GEMS Launcher can directly determine the resulting change in binding affinity for the protein at least in a qualitative manner.

If promoters from co-expressed genes were used, there are also indications about module structures, indicating which signal response might be affected by the SNP. In this manner it is possible to formulate a detailed hypothesis about the functional consequence of a particular SNP. However, final proof will only come from an experiment. The bioinformatics analysis will provide exact guiding of how to set up a decisive experiment.

The analysis for conserved binding sites as well as the evaluation of binding affinity changes will be fully automatic in a new software package Genomatix is releasing in 2002. So far, these steps remain interactive and may require a substantial amount of interactive work in some cases.

13.15.1 Limitations of this Approach

The up-side of the approach is that the tedious work of defining the promoter framework has only to be done once. After that an unlimited number of SNPs hitting this promoter can be evaluated automatically.

Unfortunately, this strategy has more limitations than the amount of work required. If a SNP hits the promoter outside of any binding site belonging to an identified framework, there is no guarantee that this will be a silent mutation as promoter frameworks are usually incomplete and there is always the possibility that a binding site, unknown so far, may be affected. In such cases additional experimental validation of the SNP is required or orthologous or co-expressed promoters have to be analysed for unknown but conserved patterns that might correspond to new binding sites. There is also software available to carry out such analyses (e.g. Stormo and Hartzell, 1989, Wolfertstetter *et al.*, 1996) (CONSENSUS and CoreSearch, see Table 13.1 for URLs).

13.16 CONCLUSION

In summary, SNPs located within known elements of frameworks can be evaluated for potential functional consequences, while SNPs located in a region not assigned to any functional framework remain unresolved in the absence of additional data. Fortunately, the information about regulatory sequences which accumulates during the analysis of individual SNPs remains valid for additional SNP analyses. Therefore, this new kind of genomic analysis may have a steep learning curve in the beginning but will gradually develop into a very powerful high throughput system in the near future.

The Evaluation of regulatory SNPs described in 13.12 to 13.15 and in Figure 13.2 down to the level of target SNPs has been carried out genome-wide in the meantime and is all available as part of the ElDorado system.

REFERENCES

Adams MD, Celniker SE, Holt RA, Evans CA, Gocayne JD, Amanatides PG, *et al.* (2000). The genome sequence of *Drosophila melanogaster*. *Science* **287**: 2185–2195.

Arnone MI, Davidson EH. (1997). The hardwiring of development: organization and function of genomic regulatory systems. *Development* **124**: 1851–1864.

Bates MD, Schatzman LC, Harvey RP, Potter SS. (2001). Two CCAAT boxes in a novel inverted repeat motif are required for Hlx homeobox gene expression. *Biochim Biophys Acta* **1519**: 96–105.

Bode J, Benham C, Knopp A, Mielke C. (2000). Transcriptional augmentation: modulation of gene expression by scaffold/matrix-attached regions (S/MAR elements). *Crit Rev Eukaryot Gene Expr* **10**: 73–90.

Boulikas T. (1995). Chromatin domains and prediction of MAR sequences. *Int Rev Cytol* **162A**: 279–388.

Chen QK, Hertz GZ, Stormo GD. (1995). MATRIX SEARCH 1.0: a computer program that scans DNA sequences for transcriptional elements using a database of weight matrices. *Comp Appl Biosci* **11**: 563–566.

Daban JR, Bermudez A. (1998). Interdigitated solenoid model for compact chromatin fibres. *Biochemistry* **37**: 4299–4304.

Dunham I, Shimizu N, Roe BA, Chissoe S, Hunt AR, Collins JE, *et al.* (1999). The DNA sequence of human chromosome 22. *Nature* **402**: 489–495.

Frech K, Quandt K, Werner T. (1997). Software for the analysis of DNA sequence elements of transcription. *Comp Appl Biosci* **13**: 89–97.

Frech K, Quandt K, Werner, T. (1998). Muscle actin genes: a first step towards computational classification of tissue specific promoters. *In Silico Biol* **1**: 5.

Frisch M, Frech K, Klingenhoff A, Quandt K, Liebich I, Werner T. (2002). *In Silico* prediction of matrix attachment regions in large genomic sequences. *Genome Res* **12**: 349–354.

Gailus-Durner V, Scherf M, Werner T. (2000). Experimental data of a single promoter can be used for *in silico* detection of genes with related regulation in absence of sequence similarity. *Mammal Genome* **12**: 67–72.

Heinemeyer T, Wingender E, Reuter I, Hermjakob H, Kel AE, Kel OV, *et al.* (1998). Databases on transcriptional regulation: TRANSFAC, TRRD and COMPEL. *Nucleic Acids Res* **26**: 362–367.

Judson R, Stephens JC, Windemuth A. (2000). The predictive power of haplotypes in clinical response. *Pharmacogenomics* **1**: 5–16.

Kel OV, Romaschenko AG, Kel AE, Wingender E, Kolchanov NA. (1995). A compilation of composite regulatory elements affecting gene transcription in vertebrates. *Nucleic Acids Res* **23**: 4097–4103.

Kodadek T. (1998). Mechanistic parallels between DNA replication, recombination and transcription. *Trends Biochem Sci* **23**: 79–83.

Kramer A. (1996). The structure and function of proteins involved in mammalian pre-mRNA splicing. *Annu Rev Biochem* **65**: 367–409.

Kramer JA, Singh GB, Krawetz SA. (2000). Computer-assisted search for sites of nuclear matrix attachment *Genomics* **35**: 273.

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.

Pesole G, Grillo G, Liuni S. (1996). Databases of mRNA untranslated regions for metazoa. *Comput Chem* **20**: 141–144.

Prestridge DS. (1996). SIGNAL SCAN 4.0: additional databases and sequence formats. *Comp Appl Biosci* **12**: 157–160.

Quandt K, Frech K, Karas H, Wingender E, Werner T. (1995). Matlnd and Matlnspector: new fast and versatile tools for detection of consensus matches in nucleotide sequence data. *Nucleic Acids Res* **23**: 4878–4884.

Romey MC, Pallares-Ruiz N, Mange A, Mettling C, Peytavi R, Demaille J, *et al.* (2000). A naturally occurring sequence variation that creates a YY1 element is associated with increased cystic fibrosis transmembrane conductance regulator gene expression. *J Biol Chem* **275**: 3561–3567.

Scherf M, Klingenhoff A, Werner T. (2000). Highly specific localization of promoter regions in large genomic sequences by PromoterInspector — a novel context analysis approach. *J Mol Biol* **297**: 599–606.

Scherf M, Klingenhoff A, Frech K, Quandt K, Schneider R, Grote K, *et al.* (2001). First pass annotation of promoters on human chromosome 22. *Genome Res* **11**, 333–340.

Smit AF. (1999). Interspersed repeats and other mementos of transposable elements in mammalian genomes. *Curr Opin Genet Dev* **6**: 657–663.

Stephens JC, Schneider JA, Tanguay DA, Choi J, Acharya T, Stanley SE, *et al.* (2001). Haplotype variation and linkage disequilibrium in 313 human genes. *Science* **293**: 489–493.

Stormo GD, Hartzell III GW. (1989). Identifying protein-binding sites from unaligned DNA fragments. *Proc Natl Acad Sci USA* **86**: 1183–1187.

The *C. elegans* Sequencing Consortium. (1998). Genome sequence of the nematode *C. elegans*: a platform for investigating biology. *Science* **282**: 2012–2018.

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.

Werner T. (1999). Identification and characterization of promoters in eukaryotic DNA sequences. *Mammal Genome* **10**: 168–175.

Werner T. (2001). Cluster analysis and promoter modelling as bioinformatics tools for the identification of target genes from expression array data. *Pharmacogenomics* **2**: 25–36.

Wingender E, Chen X, Fricke E, Geffers R, Hehl R, Liebich I, *et al.* (2001). The TRANSFAC system on gene expression regulation. *Nucleic Acids Res* **29**: 281–283.

Wolfertstetter F, Frech K, Herrmann G, Werner T. (1996). Identification of functional elements in unaligned nucleic acid sequences by a novel tuple search algorithm. *Comp Appl Biosci* **12**: 71–80.

Bioinformatics for Geneticists. Edited by Michael R. Barnes and Ian C. Gray
Copyright © 2003 John Wiley & Sons, Ltd.
ISBNs: 0-470-84393-4 (HB); 0-470-84394-2 (PB)

CHAPTER 14

Amino Acid Properties and Consequences of Substitutions

MATTHEW J. BETTS[1] and ROBERT B. RUSSELL[2]

[1] *Bioinformatics*
deCODE genetics, Sturlugötu 8
101 Reykjavık, Iceland

[2] *Structural & Computational Biology Programme*
EMBL, Meyerhofstrasse 1
69117 Heidelberg, Germany

14.1 INTRODUCTION

Since the earliest protein sequences and structures were determined, it has been clear that the positioning and properties of amino acids are key to understanding many biological processes. For example, the first protein structure, haemoglobin provided a molecular explanation for the genetic disease sickle cell anaemia. A single nucleotide mutation leads to a substitution of glutamate in normal individuals with valine in those who suffer the disease. The substitution leads to a lower solubility of the deoxygenated form of haemoglobin and it is thought that this causes the molecules to form long fibres within blood cells which leads to the unusual sickle-shaped cells that give the disease its name.

Haemoglobin is just one of many examples now known where single mutations can have drastic consequences for protein structure, function and associated phenotype. The current availability of thousands or even millions of DNA and protein sequences means that we now have knowledge of many mutations, either naturally occurring or synthetic. Mutations can occur within one species, or between species at a wide variety of evolutionary distances. Whether mutations cause diseases or have subtle or drastic effects on protein function is often unknown.

The aim of this chapter is to give some guidance as to how to interpret mutations that occur within genes that encode for proteins. Both authors of this chapter have been approached previously by geneticists who want help interpreting mutations through the use of protein sequence and structure information. This chapter is an attempt to summarize our thought processes when giving such help. Specifically, we discuss the nature of mutations and the properties of amino acids in a variety of different protein contexts. The hope is that this discussion will help in anticipating or interpreting the effect that a particular amino acid change will have on protein structure and function. We will first highlight features of proteins that are relevant to considering mutations: cellular environments, three-dimensional structure and evolution. Then we will discuss classifications of the amino acids based on evolutionary, chemical or structural principles, and the role for amino acids of different classes in protein structure and function in different contexts. Last, we will review several studies of mutations, including naturally-occurring variations, SNPs, site-directed mutations, mutations that allow adaptive evolution and post-translational modification.

14.2 PROTEIN FEATURES RELEVANT TO AMINO ACID BEHAVIOUR

It is beyond the scope of this chapter to discuss the basic principles of proteins, since this can be gleaned from any introductory biochemistry text-book. However, a number of general principles of proteins are important to place any mutation in the correct context.

14.2.1 Protein Environments

A feature of key importance is cellular location. Different parts of cells can have very different chemical environments with the consequence that many amino acids behave differently. The biggest difference is between *soluble* proteins and *membrane* proteins. Whereas soluble proteins tend to be surrounded by water molecules, membrane proteins are surrounded by lipids. Roughly speaking this means that these two classes behave in an 'inside-out' fashion relative to each other. Soluble proteins tend to have polar or hydrophilic residues on their surfaces, whereas membrane proteins tend to have hydrophobic residues on the surface that interact with the membrane.

Soluble proteins also come in several flavours. The biggest difference is between those that are *extracellular* and those that are *cytosolic* (or *intracellular*). The cytosol is quite different from the more aqueous environment outside the cell; the density of proteins and other molecules effects the behaviour of some amino acids quite drastically, the foremost among these being cysteine. Outside the cell, cysteines in proximity to one another can be *oxidized* to form disulphide bonds, sulphur–sulphur covalent linkages that are important for protein folding and stability. However, the reducing environment inside the cell makes the formation of these bonds very difficult; in fact they are so rare as to warrant special attention.

Cells also contain numerous compartments, the organelles, which can also have slightly different environments from each other. Proteins in the nucleus often interact with DNA, meaning they contain different preferences for amino acids on their surfaces (e.g. positive amino acids or those containing amides most suitable for interacting with the negatively charged phosphate backbone). Some organelles such as mitochondria or chloroplasts are quite similar to the cytosol, while others, such as lysosomes or Golgi apparati are more akin to the extracellular environment. It is important to consider the likely cellular location of any protein before considering the consequences of amino acid substitutions.

A detailed hierarchical description of cellular location is one of the three main branches of the classification provided by the Gene Ontology Consortium (Ashburner *et al.*, 2000), the others being 'molecular function' and 'biological process'. The widespread adoption of this vocabulary by sequence databases and others should enable more sophisticated investigation of the factors governing the various roles of proteins.

14.2.2 Protein Structure

Proteins themselves also contain different microenvironments. For soluble proteins, the surface lies at the interface with water and thus tends to contain more polar or charged amino acids than one finds in the core of the protein, which is more likely to comprise hydrophobic amino acids. Proteins also contain regions that are directly involved in protein function, such as active sites or binding sites, in addition to regions that are less critical to the protein function and where mutations are likely to have fewer consequences. We will discuss many specific roles for particular amino acids in protein structures in the sections below, but it is important to remember that the context of any amino acid can vary greatly depending on its location in the protein structure.

14.2.3 Protein Evolution

Proteins are nearly always members of homologous families. Knowledge about the family a protein belongs in will generally give insights into the possible function, but several things should be considered. Two processes can give rise to homologous protein families: *speciation* or *duplication*. Proteins related by speciation only are referred to as *orthologues*, and as the name suggests, these proteins have the same function in different species. Proteins related by duplications are referred to as *paralogues*. Successive rounds of speciation and intra-genomic duplication can lead to confusing situations where it becomes difficult to say whether paralogy or orthology applies.

To be maintained in a genome over time, paralogous proteins are likely to evolve different functions (or have a dominant negative phenotype and so resist decay by point mutation (Gibson and Spring, 1998)). Differences in function can range from subtle differences in substrate (e.g. malate versus lactate dehydrogenases), to only weak similarities in molecular function (e.g. hydrolases) to complete differences in cellular location and function (e.g. an intracellular signalling domain homologous to a secreted growth factor (Schoorlemmer and Goldfarb, 2001)). At the other extreme, the molecular function may be identical, but the cellular function may be altered, as in the case of enzymes with differing tissue specificities.

Similarity in molecular function generally correlates with sequence identity. Mouse and human proteins with sequence identities in excess of 85% are likely to be orthologues, provided there are no other proteins with higher sequence identity in either organism. Orthology between more distantly related species (e.g. human and yeast) is harder to assess, since the evolutionary distance between organisms can make it virtually impossible to distinguish orthologues from paralogues using simple measures of sequence similarity. An operational definition of orthology can sometimes be used, for example if the two proteins are each other's best match in their respective genomes. However there is no substitute for constructing a phylogenetic tree of the protein family, to identify which sequences are related by speciation events. Assignment of orthology and paralogy is perhaps the best way of determining likely equivalences of function. Unfortunately, complete genomes are unavailable for most organisms. Some rough rules of thumb can be used: function is often conserved down to 40% protein sequence identity, with the broad functional class being conserved to 25% identity (Wilson *et al.*, 2000).

When considering a mutation, it is important to consider how conserved the position is within other homologous proteins. Conservation across all homologues (paralogues and orthologues) should be considered carefully. These amino acids are likely to play key structural roles or a role in a common functional theme (i.e. catalytic mechanism). Other amino acids may play key roles only in the particular orthologous group (i.e. they may confer specificity to a substrate), thus meaning they vary when considering all homologues.

14.2.4 Protein Function

Protein function is key to any understanding of the consequences of amino acid substitution. Enzymes, such as trypsin (Figure 14.1), tend to have highly conserved active sites involving a handful of polar residues. In contrast, proteins that function primarily only to interact with other proteins, such as fibroblast growth factors (Figure 14.2), interact over a large surface, with virtually any amino acid being important in mediating the interaction (Plotnikov *et al.*, 1999). In other cases, multiple functions make the situation even more confusing, for example a protein kinase (Hanks *et al.*, 1988) can both catalyse a phosphorylation event and bind specifically to another protein, such as cyclin (Jeffrey *et al.*, 1995).

It is not possible to discuss all of the possible functional themes here, but we emphasize that functional information, if known, should be considered whenever studying the effects of substitution.

14.2.5 Post-translational Modification

Although there are only 20 possible types of amino acid that can be incorporated into a protein sequence upon translation of DNA, there are many more variations that can occur

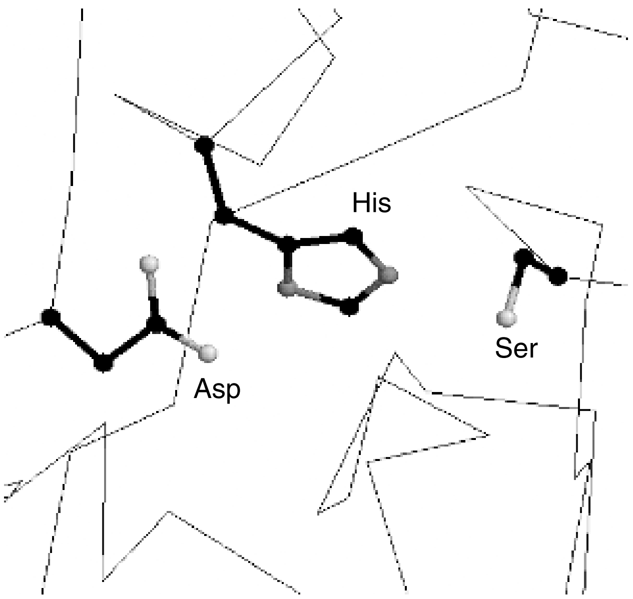

Figure 14.1 RasMol (Sayle and Milner-White, 1995) figure showing the catalytic Asp-His-Ser triad in trypsin (PDB code 1mct; Berman *et al.*, 2000).

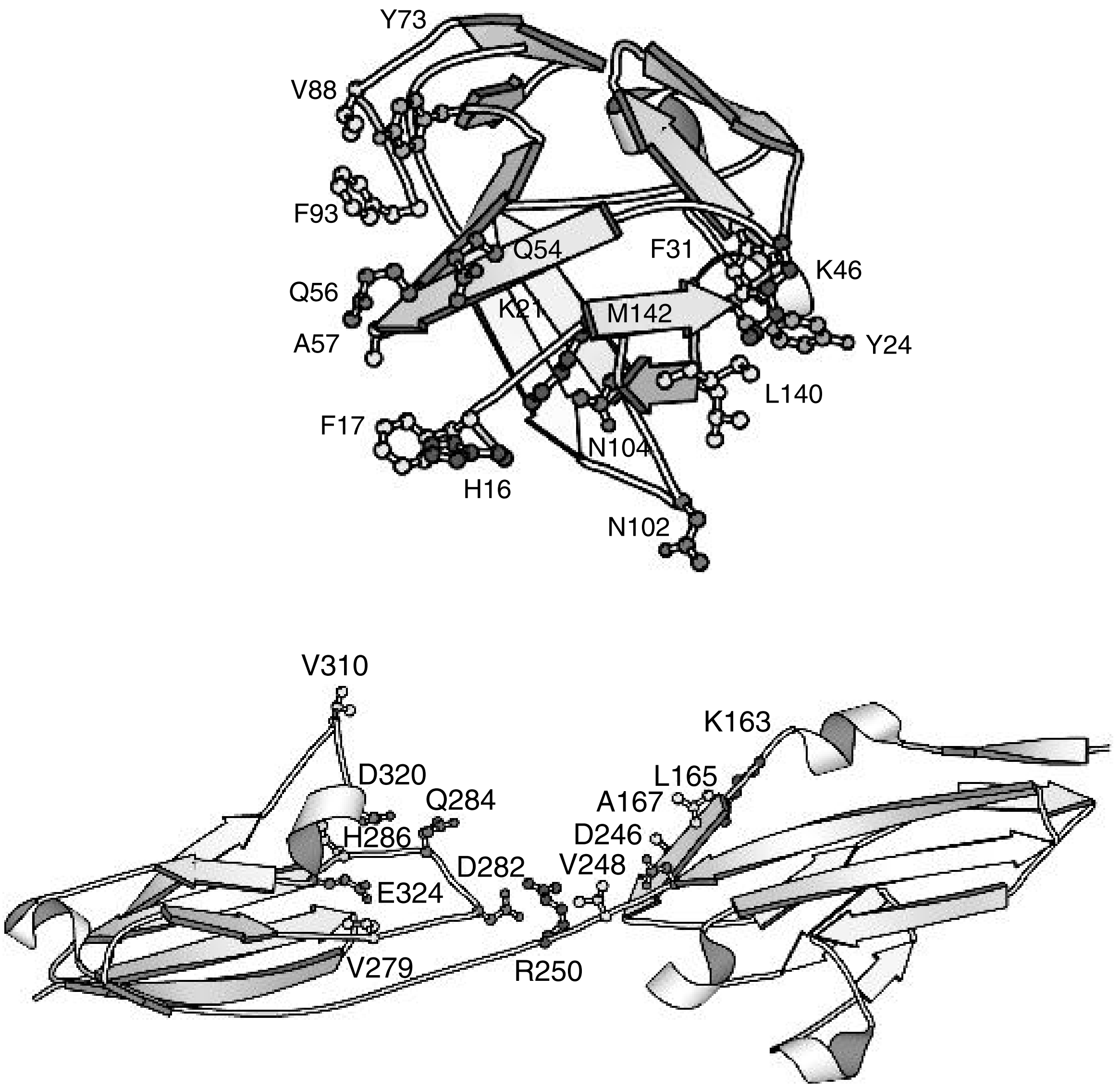

Figure 14.2 Molscript (Kraulis, 1991) figure showing fibroblast growth factor interaction with its receptor (code 1cvs; Plotnikov *et al.*, 1999). Residues at the interface are labelled. The two molecules have been pulled apart for clarity.

through subsequent modification. In addition, the gene-specified protein sequence can be shortened by proteolysis, or lengthened by addition of amino acids at either terminus.

Two common modifications, phosphorylation and glycosylation, are discussed in the context of the amino acids where they most often occur (tyrosine, serine, threonine and asparagine; see below). We direct the reader to the review by Krishna for more information on many other known types and specific examples (Krishna and Wold, 1993). The main conclusion is that modifications are highly specific, with specificity provided by primary, secondary and tertiary protein structure, although with detailed mechanisms being obscure. The biological function of the modified proteins is also summarized, from the reversible phosphorylation of serine, threonine and tyrosine residues that occurs in signalling through to the formation of disulphide bridges and other cross-links that stabilize tertiary structure, and on to the covalent attachment of lipids that allows anchorage to cell membranes. More detail on biological effects is given by Parekh and Rohlff (1997), especially where it concerns possible therapeutic applications. Many diseases arise by abnormalities in post-translational modification, and these are not necessarily apparent from genetic information alone.

14.3 AMINO ACID CLASSIFICATIONS

Humans have a natural tendency to classify, as it makes the world around us easier to understand. As amino acids often share common properties, several classifications have been proposed. This is useful, but a little bit dangerous if over-interpreted. Always remember that, for the reasons discussed above, it is very difficult to put all amino acids of the same type into an invariant group. A substitution in one context can be disastrous in another. For example, a cysteine involved in a disulphide bond would not be expected to be mutatable to any other amino acid (i.e. it is in a group on its own), one involved in binding to zinc could likely be substituted by histidine (group of two) and one buried in an intracellular protein core could probably mutate to any other hydrophobic amino acid (a group of 10 or more). We will discuss other examples below.

14.3.1 Mutation Matrices

One means of classifiying amino acids is a mutation matrix (or substitution or exchange matrix). This is a set of numbers that describe the propensities of exchanging one amino acid for another (for a comprehensive review and explanation see Durbin *et al*., 1998). These are derived from large sets of aligned sequences by counting the number of times that a particular substitution occurs and comparing this to what would be expected by chance. High values indicate that a substitution is seen often in nature and so is favourable, and vice versa. The values in the matrix are usually calculated using some model of evolutionary time, to account for the fact that different pairs of sequences are at different evolutionary distances. Probably the best known matrices are the Point Accepted Mutation (PAM) matrices of Dayhoff *et al.* (Dayhoff *et al*., 1978) and BLOSUM matrices (Henikoff and Henikoff, 1992).

Mutation matrices are very useful as rough guides for how good or bad a particular change will be. Another useful feature is that they can be calculated for different datasets to account for some of the protein features that effect amino acid properties, such as cellular locations (Jones *et al*., 1994) or different evolutionary distances (e.g. orthologues or paralogues; Henikoff and Henikoff, 1992). Several mutation matrices are reproduced in Appendix II.

14.3.2 Classification by Physical, Chemical and Structural Properties

Although mutation matrices are very useful for protein sequence alignments, especially in the absence of known three-dimensional structures, they do not precisely describe the likelihood and effects of particular substitutions at particular sites in the sequence. Position-specific substitution matrices can be generated for the family of interest, such as the profile-HMM models generated by HMMER (Eddy, 1998) and provided by Pfam (Bateman *et al*., 2000), and those generated by PSI-BLAST (Altschul *et al*., 1997). However, these are automatic methods suited to database searching and identification of new members of a family, and as such do not really give any qualitative information about the chemistry involved at particular sites.

Taylor presented a classification that explains mutation data through correlation with the physical, chemical and structural properties of amino acids (Taylor, 1986). The major factor is the size of the side chain, closely followed by its hydrophobicity. Effects of differerent amino acids on protein structure can account for mutation data when these physico-chemical properties do not. For example, hydrophobicity and size differ widely between glycine, proline, aspartic acid and glutamic acid. However, they are still closely related in mutation matrices because they prefer sharply turning regions on the surface of the protein; the phi and psi bonds of glycine are unconstrained by any side chain, proline forces a sharp turn because its side chain is bonded to the backbone nitrogen as well as to carbon, and aspartate and glutamate prefer to expose their charged side chains to solvent.

The Taylor classification is normally displayed as a Venn diagram (Figure 14.3). The amino acids were positioned on this by multidimensional scaling of Dayhoff's mutation matrix, and then grouped by common physico-chemical properties. Size is subcategorized into small and tiny (with large included by implication). Affinity for water is described by several sets: polar and hydrophobic, which overlap, and charged, which is divided into positive and negative. Sets of aromatic and aliphatic amino acids are also marked. These properties were enough to distinguish between most amino acids. However, properties such as hydrogen-bonding ability and the previously mentioned propensity for sharply turning regions are not described well. Although these factors are less important on average, and would confuse the effects of more important properties if included on the diagram, the dangers of relying on simple classifications are apparent. This can be overcome somewhat by listing all amino acids which belong to each subset (defined as an intersection or union of the sets) in the diagram, for example 'small and non-polar', and including extra subsets to describe important additional properties. These subsets can be used to give qualitative descriptions of each position in a multiple alignment, by associating the positions with the smallest subset that includes all the amino acids found at that position.

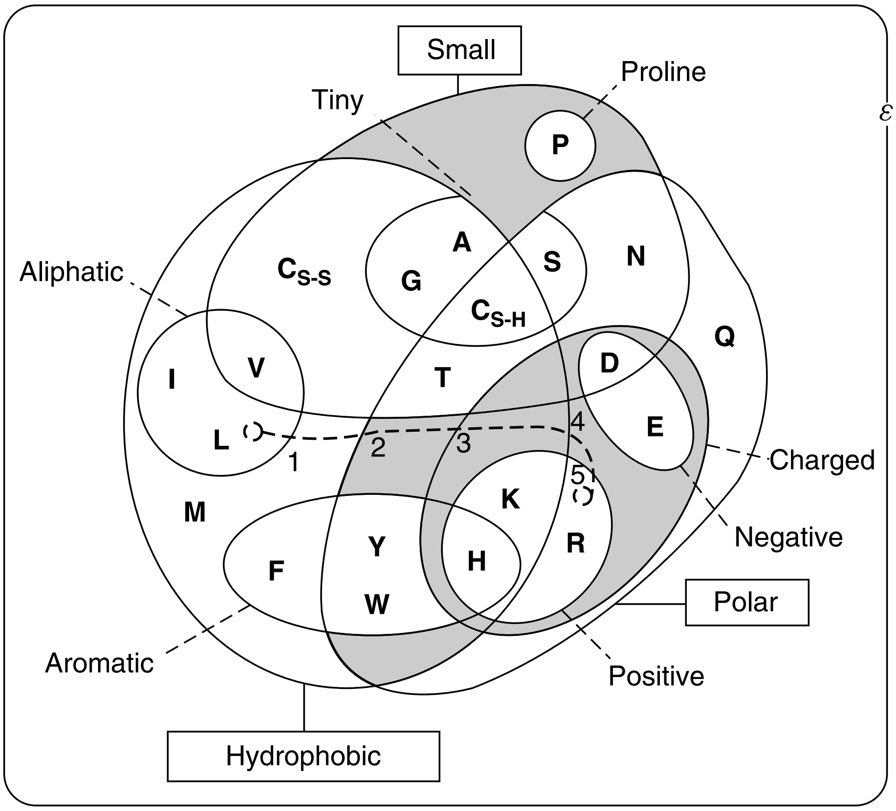

Figure 14.3 Venn diagram illustrating the properties of amino acids.

This may suggest alternative amino acids that could be engineered into the protein at each position.

14.4 PROPERTIES OF THE AMINO ACIDS

The sections that follow will first consider several major properties that are often used to group amino acids together. Note that amino acids can be in more than one group, and that sometimes properties as different as 'hydrophobic' and 'hydrophilic' can be applied to the same amino acids.

14.4.1 Hydrophobic Amino Acids

Probably the most common broad division of amino acids is into those that prefer to be in an aqueous environment (hydrophilic) and those that do not (hydrophobic). The latter can be divided according to whether they have *aliphatic* or *aromatic* side chains.

14.4.1.1 Aliphatic Side Chains

Strictly speaking aliphatic means that the side chain contains only hydrogen and carbon atoms. By this strict definition, the amino acids with aliphatic side chains are *alanine, isoleucine, leucine, proline* and *valine*. Alanine's side chain, being very short, means that it is not particularly hydrophobic and proline has an unusual geometry that gives it special roles in proteins as we shall discuss below. Although it also contains a sulphur atom, it is often convenient to consider *methionine* in the same category as isoleucine, leucine and valine. The unifying theme is that they contain largely non-reactive and flexible side chains that are ideally suited for packing in the protein interior.

Aliphatic side chains are very non-reactive, and are thus rarely involved directly in protein function, although they can play a role in substrate recognition. In particular, hydrophobic amino acids can be involved in binding/recognition of hydrophobic ligands such as lipids.

Several other amino acids also contain aliphatic regions. For example, arginine, lysine, glutamate and glutamine are *amphipathic*, meaning that they contain hydrophobic and polar areas. All contain two or more aliphatic carbons that connect the protein backbone to the non-aliphatic portion of the side chain. In some instances it is possible for such amino acids to play a dual role, with part of the side chain being buried in the protein and another being exposed to water.

14.4.1.2 Aromatic Side Chains

A side chain is aromatic when it contains an aromatic ring system. The strict definition has to do with the number of electrons contained within the ring. Generally, aromatic ring systems are planar and electrons are shared over the whole ring structure. *Phenylalanine and tryptophan* have very hydrophobic aromatic side chains, whereas *tyrosine* and *histidine* are less so. The latter two can often be found in positions that are somewhere between buried and exposed. The hydrophobic aromatic amino acids can sometimes substitute for aliphatic residues of a similar size, for example phenylalanine to leucine, but not tryptophan to valine.

Aromatic residues have also been proposed to participate in 'stacking' interactions (Hunter *et al.*, 1991) (Figure 14.4). Here, numerous aromatic rings are thought to stack

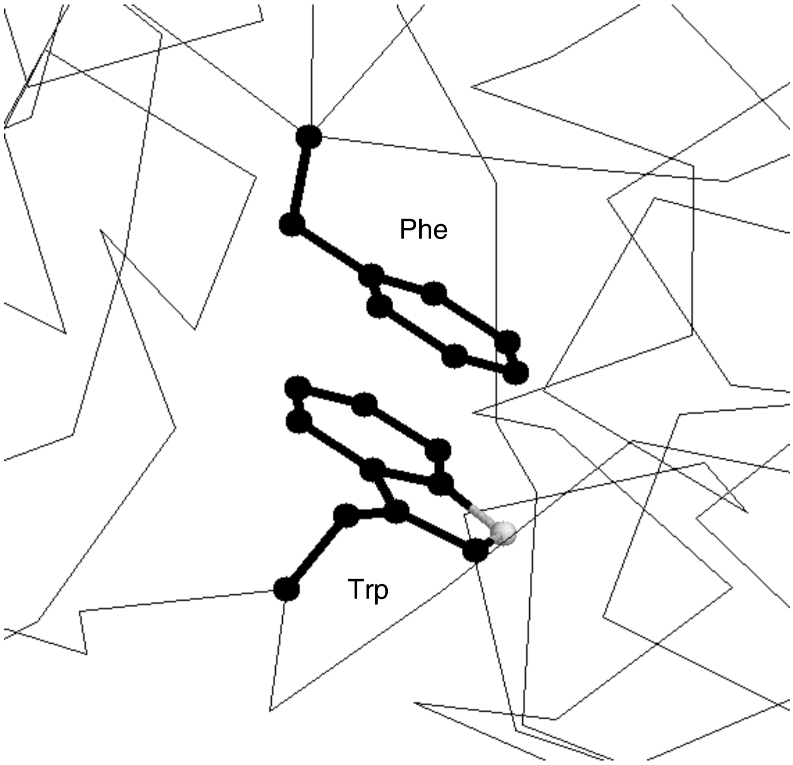

Figure 14.4 Example of aromatic stacking.

on top of each other such that their PI electron clouds are aligned. They can also play a role in binding to specific amino acids, such as proline. SH3 and WW domains, for example, use these residues to bind to their polyproline-containing interaction partners (Macias *et al.*, 2002). Owing to its unique chemical nature, histidine is frequently found in protein active sites as we shall see below.

14.4.2 Polar Amino Acids

Polar amino acids prefer to be surrounded by water. Those that are buried within the protein usually participate in hydrogen bonds with other side chains or the protein main-chain that essentially replace the water. Some of these carry a charge at typical biological pHs: *aspartate* and *glutamate* are negatively charged; *lysine* and *arginine* are positively charged. Other polar amino acids, *histidine, asparagine, glutamine, serine, threonine* and *tyrosine*, are neutral.

14.4.3 Small Amino Acids

The amino acids *alanine*, *cysteine*, *glycine*, *proline*, *serine* and *threonine* are often grouped together for the simple reason that they are all small in size. In some protein structural contexts, substitution of a small side chain for a large one can be disastrous.

14.5 AMINO ACID QUICK REFERENCE

In the sections that follow we discuss each amino acid in turn. For each we will briefly discuss general preferences for substitutions and important specific details regarding their possible structure and functional roles. More information is found on the WWW site that accompanies this chapter (www.russell.embl-heidelberg.de/aas). This website also features amino acid substitution matrices for transmembrane, extracellular and intracellular proteins. These can be used to numerically score an amino acid substitution, where unpreferred mutations are given negative scores, preferred substitutions are given positive scores and neutral substitutions are given zero scores.

14.5.1 Alanine (Ala, A)

14.5.1.1 Substitutions

Alanine can be substituted by other small amino acids.

14.5.1.2 Structure

Alanine is probably the dullest amino acid. It is not particularly hydrophobic and is non-polar. However, it contains a normal Cβ carbon, meaning that it is generally as hindered as other amino acids with respect to the conformations that the backbone can adopt. For this reason, it is not surprising to see alanine present in just about all non-critical protein contexts.

14.5.1.3 Function

The alanine side chain is very non-reactive, and is thus rarely directly involved in protein function, but it can play a role in substrate recognition or specificity, particularly in interactions with other non-reactive atoms such as carbon.

14.5.2 Isoleucine (Ile, I)

14.5.2.1 Substitutions

Isoleucine can be substituted by other hydrophobic, particularly aliphatic, amino acids.

14.5.2.2 Structure

Being hydrophobic, isoleucine prefers to be buried in protein hydrophobic cores. However, isoleucine has an additional property that is frequently overlooked. Like valine and threonine it is Cβ branched. Whereas most amino acids contain only one non-hydrogen substituent attached to their Cβ carbon, these three amino acids contain two. This means that there is a lot more bulkiness near to the protein backbone and this means that these amino acids are more restricted in the conformations the main chain can adopt. Perhaps the most pronounced effect of this is that it is more difficult for these amino acids to adopt an α-helical conformation, although it is easy and even preferred for them to lie within β-sheets.

14.5.2.3 Function

The isoleucine side chain is very non-reactive and is thus rarely directly involved in protein functions like catalysis, although it can play a role in substrate recognition. In particular, hydrophobic amino acids can be involved in binding/recognition of hydrophobic ligands such as lipids.

14.5.3 Leucine (Leu, L)

14.5.3.1 Substitutions

See Isoleucine.

14.5.3.2 Structure

Being hydrophobic, leucine prefers to be buried in protein hydrophobic cores. It also shows a preference for being within alpha helices more so than in beta strands.

14.5.3.3 Function

See Isoleucine.

14.5.4 Valine (Val, V)

14.5.4.1 Substitutions

See Isoleucine.

14.5.4.2 Structure

Being hydrophobic, valine prefers to be buried in protein hydrophobic cores. However, valine is also Cβ branched (see Isoleucine).

14.5.4.3 Function

See Isoleucine.

14.5.5 Methionine (Met, M)

14.5.5.1 Substitutions

See Isoleucine.

14.5.5.2 Structure

See Isoleucine.

14.5.5.3 Function

The methionine side chain is fairly non-reactive, and is thus rarely directly involved in protein function. Like other hydrophobic amino acids, it can play a role in binding/recognition of hydrophobic ligands such as lipids. However, unlike the proper aliphatic amino acids, methionine contains a sulphur atom, that can be involved in binding to atoms such as metals. However, whereas the sulphur atom in cysteine is connected to a hydrogen atom making it quite reactive, methionine's sulphur is connected to a methyl group. This means that the roles that methionine can play in protein function are much more limited.

14.5.6 Phenylalanine (Phe, F)

14.5.6.1 Substitutions

Phenylalanine can be substituted with other aromatic or hydrophobic amino acids. It particularly prefers to exchange with tyrosine, which differs only in that it contains an hydroxyl group in place of the ortho hydrogen on the benzene ring.

14.5.6.2 Structure

Phenylalanine prefers to be buried in protein hydrophobic cores. The aromatic side chain can also mean that phenylalanine is involved in stacking (Figure 14.4) interactions with other aromatic side chains.

14.5.6.3 Function

The phenylalanine side chain is fairly non-reactive, and is thus rarely directly involved in protein function, although it can play a role in substrate recognition (see Isoleucine). Aromatic residues can also be involved in interactions with non-protein ligands that themselves contain aromatic groups via stacking interactions (see above). They are also common in polyproline binding sites, for example in SH3 and WW domains (Macias *et al*., 2002).

14.5.7. Tryptophan (Trp, W)

14.5.7.1 Substitutions

Trytophan can be replaced by other aromatic residues, but it is unique in terms of chemistry and size, meaning that often replacement by anything could be disastrous.

14.5.7.2 Structure

See Phenylalanine.

14.5.7.3 Function

As it contains a non-carbon atom (nitrogen) in the aromatic ring system, tryptophan is more reactive than phenylalanine although it is less reactive than tyrosine. Tryptophan can play a role in binding to non-protein atoms, but such instances are rare. See also Phenylalanine.

14.5.8 Tyrosine (Tyr, Y)

14.5.8.1 Substitutions

Tyrosine can be substituted by other aromatic amino acids. See Phenylalanine.

14.5.8.2 Structure

Being partially hydrophobic, tyrosine prefers to be buried in protein hydrophobic cores. The aromatic side chain can also mean that tyrosine is involved in stacking interactions with other aromatic side chains.

14.5.8.3 Function

Unlike the very similar phenylalanine, tyrosine contains a reactive hydroxyl group, thus making it much more likely to be involved in interactions with non-carbon atoms. See also Phenylalanine.

A common role for tyrosines (and serines and threonines) within intracellular proteins is phosphorylation. Protein kinases frequently attach phosphates to these three residues as part of a signal transduction process. Note that in this context, tyrosine will rarely substitute for serine or threonine since the enzymes that catalyse the reactions (i.e. the protein kinases) are highly specific (i.e. tyrosine kinases generally do not work on serines/threonines and vice versa (Hanks *et al*., 1988)).

14.5.9 Histidine (His, H)

14.5.9.1 Substitutions

Histidine is generally considered to be a polar amino acid, however it is unique with regard to its chemical properties, which means that it does not substitute particularly well with any other amino acid.

14.5.9.2 Structure

Histidine has a pK_a near to that of physiological pH, meaning that it is relatively easy to move protons on and off of the side chain (i.e. changing the side chain from neutral to positive charge). This flexibility has two effects. The first is ambiguity about whether it prefers to be buried in the protein core or exposed to solvent. The second is that it is an ideal residue for protein functional centres (discussed below). It is false to presume that histidine is always protonated at typical pHs. The side chain has a pK_a of approximately 6.5, which means that only about 10% of molecules will be protonated. The precise pK_a depends on local environment.

14.5.9.3 Function

Histidines are the most common amino acids in protein active or binding sites. They are very common in metal binding sites (e.g. zinc), often acting together with cysteines or other amino acids (Figure 14.5; Wolfe *et al.*, 2001). In this context, it is common to see histidine replaced by cysteine.

The ease with which protons can be transferred on and off of histidines makes them ideal for charge relay systems such as those found within catalytic triads and in many cysteine and serine proteases (Figure 14.1). In this context, it is rare to see histidine exchanged for any amino acid at all.

14.5.10 Arginine (Arg, R)

14.5.10.1 Substitutions

Arginine is a positively-charged, polar amino acid. It thus most prefers to substitute for the other positively-charged amino acid, lysine, although in some circumstances it will also tolerate a change to other polar amino acids. Note that a change from arginine to lysine is not always neutral. In certain structural or functional contexts, such a mutation can be devastating to function (see below).

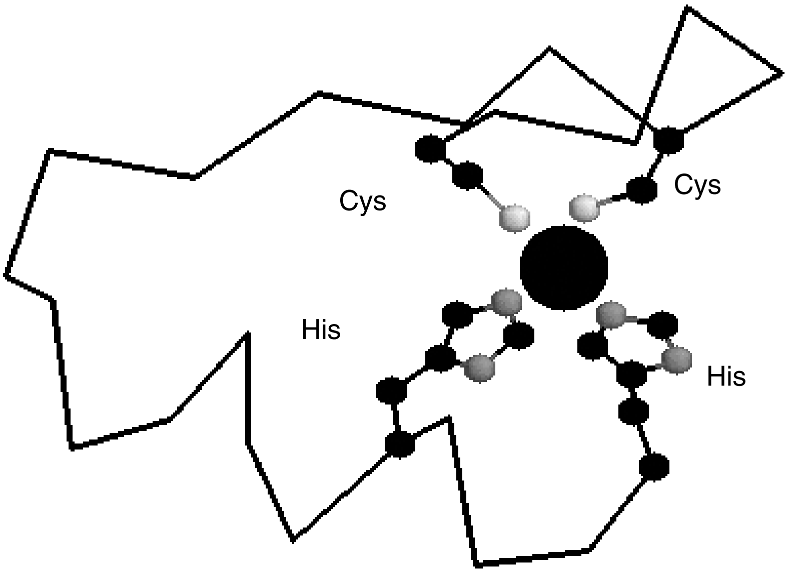

Figure 14.5 Example of a metal binding site coordinated by cysteine and histidine residues (code 1g2f; Wolfe *et al.*, 2001).

14.5.10.2 Structure

Arginine generally prefers to be on the surface of the protein, but its amphipathic nature can mean that part of the side chain is buried. Arginines are also frequently involved in salt-bridges where they pair with a negatively charged aspartate or glutamate to create stabilizing hydrogen bonds that can be important for protein stability (Figure 14.6).

14.5.10.3 Function

Arginines are quite frequent in protein active or binding sites. The positive charge means that they can interact with negatively-charged non-protein atoms (e.g. anions or carboxylate groups). Arginine contains a complex guanidinium group on its side chain that has a geometry and charge distribution that is ideal for binding negatively-charged groups on phosphates (it is able to form multiple hydrogen bonds). A good example can be found in the src homology 2 (SH2) domains (Figure 14.7; Waksman *et al.*, 1992). The two arginines shown in the figure make multiple hydrogen bonds with the phosphate. In this context arginine is not easily replaced by lysine. Although lysine can interact with phosphates, it contains only a single amino group, meaning it is more limited in the number of hydrogen bonds it can form. A change from arginine to lysine in some contexts can thus be disastrous (Copley and Barton, 1994).

14.5.11 Lysine (Lys, K)

14.5.11.1 Substitutions

Lysine can be substituted by arginine or other polar amino acids.

14.5.11.2 Structure

Lysine frequently plays an important role in structure. First, it can be considered to be somewhat amphipathic as the part of the side chain nearest to the backbone is long, carbon-containing and hydrophobic, whereas the end of the side chain is positively charged. For

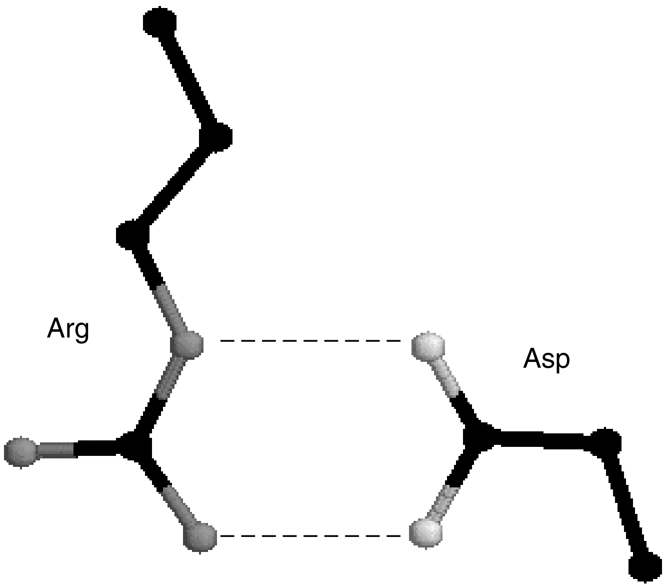

Figure 14.6 Example of a salt-bridge (code 1xel).

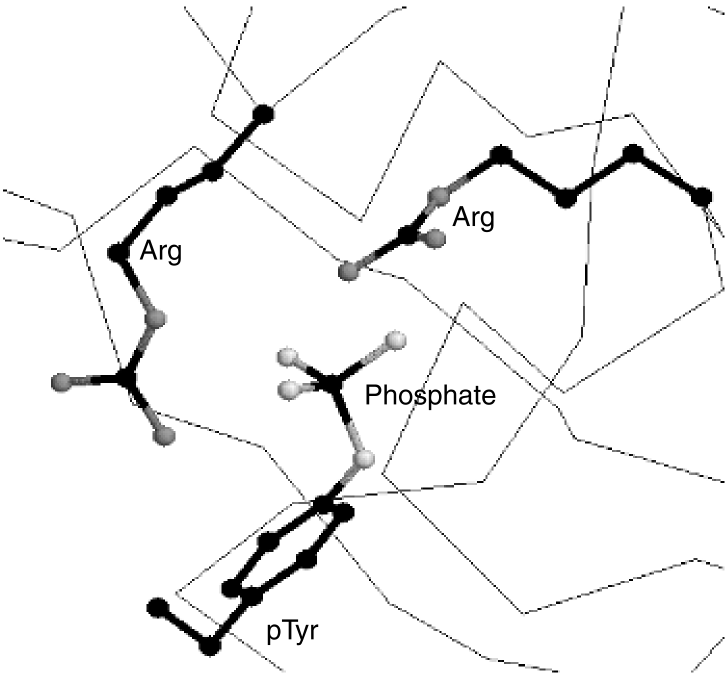

Figure 14.7 Interaction of arginine residues with phosphotyrosine in an SH2 domain (code 1sha; Waksman *et al.*, 1992).

this reason, one can find lysines where part of the side chain is buried and only the charged portion is on the outside of the protein. However, this is by no means always the case and generally lysines prefer to be on the outside of proteins. Lysines are also frequently involved in salt-bridges (see Arginine).

14.5.11.3 Function

Lysines are quite frequent in protein active or binding sites. Lysine contains a positively-charged amino group on its side chain that is sometimes involved in forming hydrogen bonds with negatively-charged non-protein atoms (e.g. anions or carboxylate groups).

14.5.12 Aspartate (Asp, D)

14.5.12.1 Substitutions

Aspartate can be substituted by glutamate or other polar amino acids, particularly asparagine, which differs only in that it contains an amino group in place of one of the oxygens found in aspartate (and thus also lacks a negative charge).

14.5.12.2 Structure

Being charged and polar, aspartates generally prefer to be on the surface of proteins, exposed to an aqueous environment. Aspartates (and glutamates) are frequently involved in salt-bridges (see Arginine).

14.5.12.3 Function

Aspartates are quite frequently involved in protein active or binding sites. The negative charge means that they can interact with positively-charged non-protein atoms, such as cations like zinc. Aspartate has a shorter side chain than the very similar glutamate meaning that is slightly more rigid within protein structures. This gives it a slightly stronger preference to be involved in protein active sites. Probably the most famous example of aspartate being involved in an active site is found within serine proteases such as trypsin, where it functions in the classical Asp-His-Ser catalytic triad (Figure 14.1). In this context, it is quite rare to see aspartate exchange for glutamate, although it is possible for glutamate to play a similar role.

14.5.13 Glutamate (Glu, E)

14.5.13.1 Substitutions

Substitution can be by aspartate or other polar amino acids, in particular glutamine, which is to glutamate what asparagine is to aspartate (see above).

14.5.13.2 Structure

See Aspartate.

14.5.13.3 Function

Glutamate, like aspartate, is quite frequently involved in protein active or binding sites. In certain cases, they can also perform a similar role to aspartate in the catalytic site of proteins such as proteases or lipases.

14.5.14 Asparagine (Asn, N)

14.5.14.1 Substitutions

Asparagine can be substituted by other polar amino acids, especially aspartate (see above).

14.5.14.2 Structure

Being polar asparagine prefers generally to be on the surface of proteins, exposed to an aqueous environment.

14.5.14.3 Function

Asparagines are quite frequently involved in protein active or binding sites. The polar side chain is good for interactions with other polar or charged atoms. Asparagine can play a similar role to aspartate in some proteins. Probably the best example is found in certain cysteine proteases, where it forms part of the Asn-His-Cys catalytic triad. In this context, it is quite rare to see asparagine exchange for glutamine.

Asparagine, when occurring in a particular motif (Asn-X-Ser/Thr) can be *N*-glycosylated (Gavel and von Heijne, 1990). Thus in this context it is impossible to substitute it with any amino acid at all.

14.5.15 Glutamine (Gln, Q)

14.5.15.1 Substitutions

Glutamine can be substituted by other polar amino acids, especially glutamate (see above).

14.5.15.2 Structure

See Asparagine.

14.5.15.3 Function

Glutamines are quite frequently involved in protein active or binding sites. The polar side chain is good for interactions with other polar or charged atoms.

14.5.16 Serine (Ser, S)

14.5.16.1 Substitutions

Serine can be substituted by other polar or small amino acids in particular threonine which differs only in that it has a methyl group in place of a hydrogen group found in serine.

14.5.16.2 Structure

Being a fairly indifferent amino acid, serine can reside both within the interior of a protein, or on the protein surface. Its small size means that it is relatively common within tight turns on the protein surface, where it is possible for the serine side chain hydroxyl oxygen to form a hydrogen bond with the protein backbone, effectively mimicking proline.

14.5.16.3 Function

Serines are quite common in protein functional centres. The hydroxyl group is fairly reactive, being able to form hydrogen bonds with a variety of polar substrates.

Perhaps the best known role for serine in protein active sites is exemplified by the classical Asp-His-Ser catalytic triad found in many hydrolases (e.g. proteases and lipases; Figure 14.1). Here, a serine, aided by a histidine and an aspartate act as a nucleophile to hydrolyse (effectively cut) other molecules. This three-dimensional 'motif' is found in many non-homologous (i.e. unrelated) proteins and is a classic example of molecular convergent evolution (Russell, 1998). In this context, it is rare for serine to exchange with threonine, but in some cases, the reactive serine can be replaced by cysteine, which can fulfil a similar role.

Intracellular serines can also be phosphorylated (see Tyrosine). Extracellular serines can also be *O*-glycosylated where a carbohydrate is attached to the side chain hydroxyl group (Gupta *et al.*, 1999).

14.5.17 Threonine (Thr, T)

14.5.17.1 Substitutions

Threonine can be substituted with other polar amino acids, particularly serine (see above).

14.5.17.2 Structure

Being a fairly indifferent amino acid, threonine can reside both within the interior of a protein or on the protein surface. Threonine is also Cβ branched (see Isoleucine).

14.5.17.3 Function

Threonines are quite common in protein functional centres. The hydroxyl group is fairly reactive, being able to form hydrogen bonds with a variety of polar substrates. Intracellular threonines can also be phosphorylated (see Tyrosine) and in the extracellular environment they can be *O*-glycosylated (see Serine).

14.5.18 Cysteine (Cys, C)

14.5.18.1 Substitutions

In the case of cysteine there is no general preference for substitution with any other amino acid, although it can tolerate substitutions with other small amino acids. Cysteine has a role that is very dependent on cellular location, making substitution matrices dangerous to interpret (e.g. Barnes and Russell, 1999).

14.5.18.2 Structure

The role of cysteines in structure is very dependent on the cellular location of the protein in which they are contained. Within extracellular proteins, cysteines are frequently involved in disulphide bonds, where pairs of cysteines are oxidized to form a covalent bond. These bonds serve mostly to stabilize the protein structure and the structure of many extracellular proteins is almost entirely determined by the topology of multiple disulphide bonds (e.g. Figure 14.8).

The reducing environment inside cells makes the formation of disulphide bonds very unlikely. Indeed, instances of disulphide bonds in the intracellular environment are so rare that they almost always attract special attention. Disulphide bonds are also rare within the membrane, although membrane proteins may contain disulphide bonds within extracellular domains. Disulphide bonds are such that cysteines must be paired. If one half of a disulphide bond pair is lost, then the protein may not fold properly.

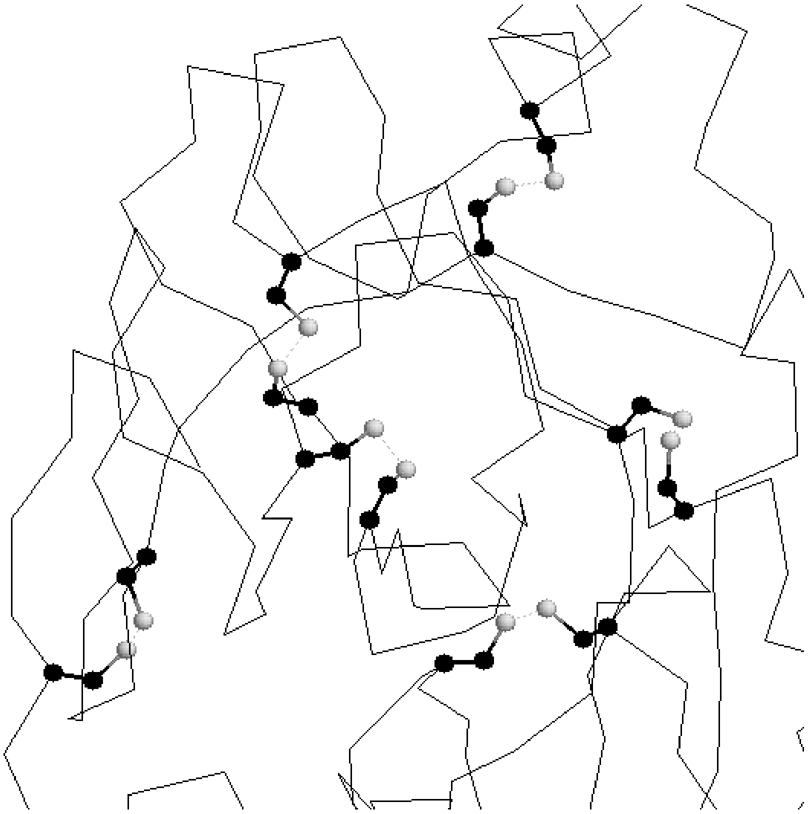

Figure 14.8 Example of a small, disulphide-rich protein (code 1tfx).

In the intracellular environment cysteines can still play a key structural role. Their sulphydryl side chain is excellent for binding to metals, such as zinc, meaning that cysteines (and other amino acids such as histidines) are very common in metal binding motifs such as zinc fingers (Figure 14.5). Outside of this context within the intracellular environment and when it is not involved in molecular function, cysteine is a neutral, small amino acid and prefers to substitute with other amino acids of the same type.

14.5.18.3 Function

Cysteines are also very common in protein active and binding sites. Binding to metals (see above) can also be important in enzymatic functions (e.g. metal proteases). Cysteine can also function as a nucleophile (i.e. the reactive centre of an enzyme). Probably the best known example of this occurs within the cysteine proteases, such as caspases or papains, where cysteine is the key catalytic residue, being helped by a histidine and an asparagine.

14.5.19 Glycine (Gly, G)

14.5.19.1 Substitutions

Glycine can be substituted by other small amino acids, but be warned that even apparently neutral mutations (e.g. to alanine) can be forbidden in certain contexts (see below).

14.5.19.2 Structure

Glycine is unique as it contains a hydrogen as its side chain (rather than a carbon as is the case for all other amino acids). This means that there is much more conformational flexibility in glycine and as a result of this it can reside in parts of protein structures that are forbidden to all other amino acids (e.g. tight turns in structures).

14.5.19.3 Function

The uniqueness of glycine also means that it can play a distinct functional role, such as using its backbone (without a side chain) to bind to phosphates (Schulze-Gahmen *et al.*, 1996). This means that if one sees a conserved glycine changing to any other amino acid, the change could have a drastic impact on function.

A good example is found among the protein kinases. Figure 14.9 shows a region around the ATP binding site in a protein kinase; the ATP is shown to the right of the figure and part of the protein to the left. The glycines in this loop are part of the classic ‘Gly-X-Gly-X-X-Gly’ motif present in the kinases (Hanks *et al.*, 1988). These three glycines are almost never mutated to other residues; only glycines can function to bind to the phosphates of the ATP molecule using their main chains.

14.5.20 Proline (Pro, P)

14.5.20.1 Substitutions

Proline can sometimes substitute for other small amino acids, although its unique properties mean that it does not often substitute well.

14.5.20.2 Structure

Proline is unique in that it is the only amino acid where the side chain is connected to the protein backbone twice, forming a five-membered ring. Strictly speaking, this makes proline an imino acid (since in its isolated form, it contains an NH^{2+} rather than an NH^{3+} group, but this is mostly just pedantic detail). This difference is very important as

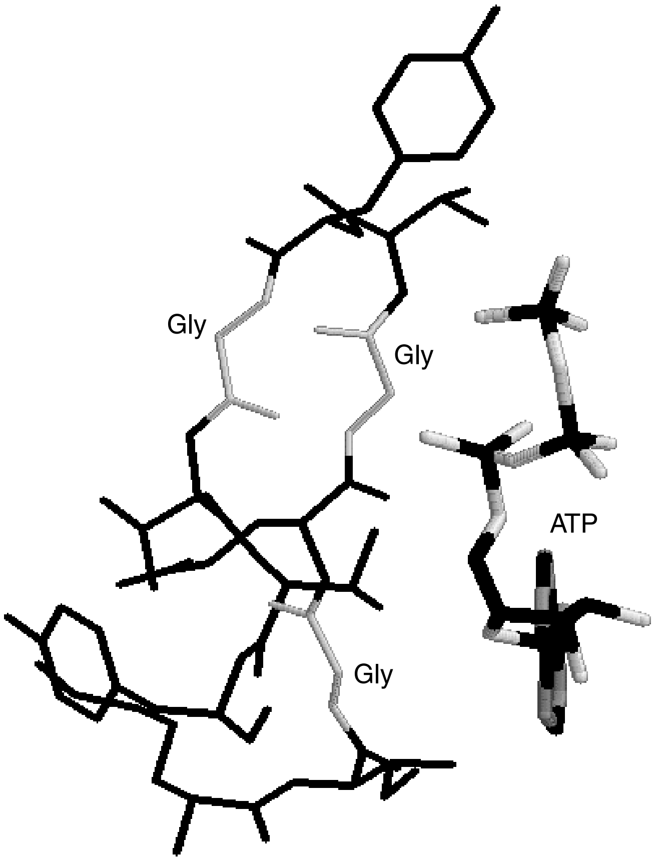

Figure 14.9 Glycine-rich phosphate binding loop in a protein kinase (code 1hck; Schulze-Gahmen *et al.*, 1996).

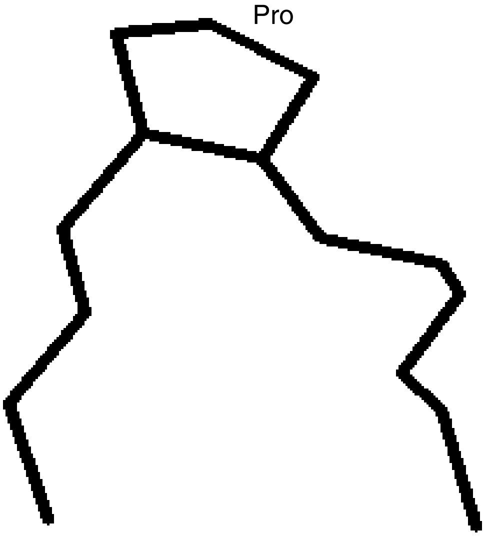

Figure 14.10 Example of proline in a tight protein turn (code 1ag6).

it means that proline is unable to occupy many of the main-chain conformations easily adopted by all other amino acids. In this sense, it can be considered to be an opposite of glycine, which can adopt many more main-chain conformations. For this reason proline is often found in very tight turns in protein structures (i.e. where the polypeptide chain must change direction; Figure 14.10). It can also function to introduce kinks into α-helices, since it is unable to adopt a normal helical conformation. Despite being aliphatic the preference for turn structure means that prolines are usually found on the protein surface.

14.5.20.3 Function

The proline side chain is very non-reactive. This, together with its difficulty in adopting many protein main-chain conformations means that it is very rarely involved in protein active or binding sites.

14.6 STUDIES OF HOW MUTATIONS AFFECT FUNCTION

Several studies have been carried out previously in an attempt to derive general principles about the relationship between mutations, structure, function and diseases. We review some of these below.

14.6.1 Single Nucleotide Polymorphisms (SNPs)

A SNP is a point mutation that is present at a measurable frequency in human populations. They can occur either in coding or non-coding DNA. Non-coding SNPs may have effects on important mechanisms such as transcription, translation and splicing. However, the effects of coding SNPs are easier to study and are potentially more damaging, and so they have received considerably more attention. They are also more relevant to this chapter. Coding SNPs can be divided into two main categories, synonymous (where there is no change in the amino acid coded for), and non-synonymous. Non-synonymous SNPs tend to occur at lower frequencies than synonymous SNPs. Minor allele frequencies also tend to be lower in non-synonymous SNPs. This is a strong indication that these replacement polymorphisms are deleterious (Cargill *et al.*, 1999).

To examine the phenotypic effects of coding SNPs, Sunyaev *et al.* (2000) studied the relationships between non-synonymous SNPs and protein structure and function. Three sets of SNP data were compared: disease causing susbtitutions, substitutions between orthologues and those represented by human alleles. Disease-causing mutations were more common in structurally and functionally important sites than were variations between orthologues, as might be expected. Allelic variations were also more common in these regions than were those between orthologues. Minor allele frequency and the level of occurrence in these regions were correlated, another indication of evolutionary selection of phenotype. The most damaging allelic variants affect protein stability, rather than binding, catalysis, allosteric response or post-translational modification (Sunyaev *et al.*, 2001). The expected increase in the number of known protein structures will allow other analyses and refinement of the details of the phenotypic effects of SNPs.

Wang and Moult (2001) developed a description of the possible effects of missense SNPs on protein structure and used it to compare disease-causing missense SNPs with a set from the general population. Five general classes of effect were considered: protein stability, ligand binding, catalysis, allosteric regulation and post-translational modification. The disease and population sets of SNPs contain those that can be mapped onto known protein structures, either directly or through homologues of known structure. Of the disease-causing SNPs, 90% were explained by the description, with the majority (83%) being attributed to effects on protein stability, as reported by Sunyaev *et al.* (2001). The 10% that are not explained by the description may cause disease by effects not easily identified by structure alone. Of the SNPs from the general population, 70% were predicted to have no effect. The remaining 30% may represent disease-causing SNPs previously unidentified as such, or molecular effects that have no significant phenotypic effect.

14.6.2 Site-directed Mutagenesis

Site-directed mutagenesis is a powerful tool for discovering the importance of an amino acid in the function of the protein. Gross changes in amino acid type can reveal sites that are important in maintaining the structure of the protein. Conversely, when investigating functionally interesting sites it is important to choose replacement residues that are unlikely to affect structure dramatically, for example by choosing ones of a similar size to the original. Peracchi (2001) reviews the use of site-directed mutagenesis to investigate mechanisms of enzyme catalysis, in particular those studies involving mutagenesis of general acids (proton donors), general bases (proton acceptors) and catalytic nucleophiles in active sites. These types of amino acid could be considered to be the most important to enzyme function as they directly participate in the formation or cleavage of covalent bonds. However, studies indicate that they are often important but not essential — rates are still higher than the uncatalysed reaction even when these residues are removed, because the protein is able to use an alternative mechanism of catalysis. Also, direct involvement in the formation and cleavage of bonds is only one of a combination of methods that an enzyme can use to catalyse a reaction. Transition states can be stabilized by complementary shape and electrostatics of the binding site of the enzyme and substrates can be precisely positioned, lowering the entropy of activation. These factors can also be studied by site-directed mutagenesis, with consideration of the physical and chemical properties of the amino acids again guiding the choice of replacements, along with knowledge of the structure of the protein.

14.6.3 Key Mutations in Evolution

Golding and Dean (1998) reviewed six studies that demonstrate the insight into molecular adaptation that is provided by combining knowledge of phylogenies, site-directed mutagenesis and protein structure. These studies emphasize the importance of protein structure when considering the effects of amino acid mutations.

Many changes can occur over many generations, with only a few being responsible for changes in function. For example, the sequences of lactate dehydrogenase (LDH) and malate dehydrogenase (MDH) from *Bacillus stearothermophillus* are only about 25% identical, but their tertiary structures are highly similar. Only one mutation, of uncharged glutamine 102 to positive arginine in the active site, is required to convert LDH into a highly specific MDH. The arginine is thought to interact with the carboxylate group which is the only difference between the substrate/products of the two enzymes (Figure 14.11; Wilks *et al.*, 1988).

Thus amino acid changes that appear to be radical or conservative from their scores in mutation matrices or amino acid properties may be the opposite when their effect on protein function is considered; glutamine to arginine has a score of 0 in the PAM250 matrix, meaning that it is neutral. The importance of the mutation at position 102 in LDH and MDH could not be predicted using this information alone.

Another study showed that phylogeny and site-directed mutagenesis can identify key amino acid changes that would likely be overlooked if only structure was considered; the reconstruction of an ancestral ribonuclease showed that the mutation that causes most of the five-fold loss in activity towards double-stranded RNA is of Gly38 to Asp, more than 5 Å from the active site (Golding and Dean, 1998).

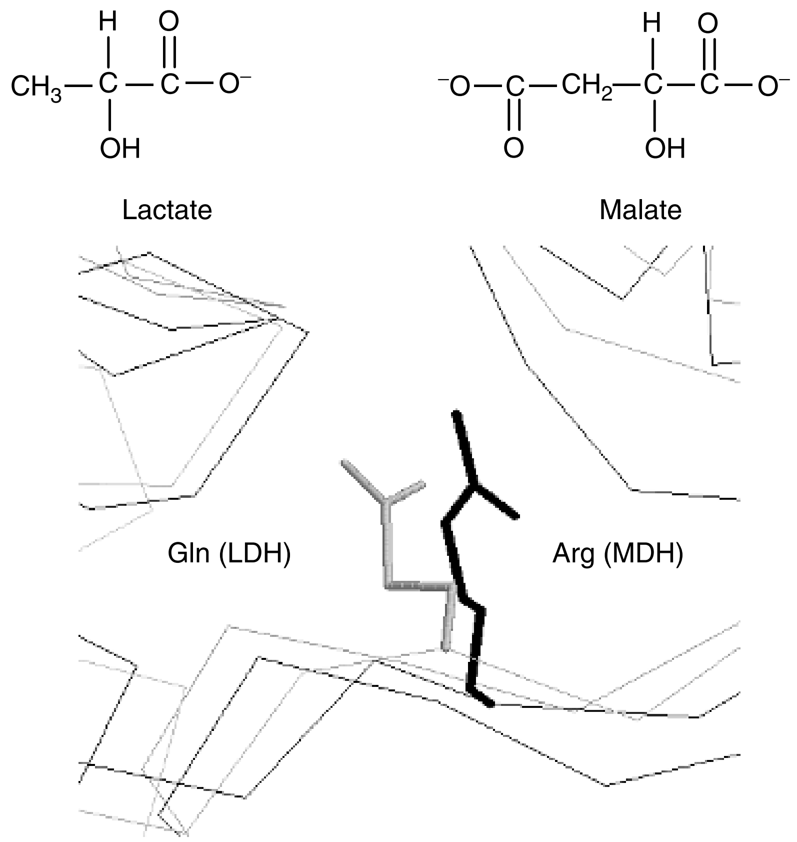

Figure 14.11 Lactate and malate dehydrogenase specificity (codes 9ltd and 2cmd; Wilks *et al.*, 1988).

A third study showed that knowledge of structure can be important in understanding the effects of mutations. Two different mutations in different locations in the haemoglobin genes of the bar-headed goose and Andean goose give both species a high affinity for oxygen. Structural studies showed that both changes remove an important van der Waals contact between subunits, shifting the equilibrium of the haemoglobin tetramer towards the high-affinity state. The important point in all these studies is that no single approach, such as phylogeny alone or structural studies alone, is enough to understand the effects of all amino acid mutations.

14.7 A SUMMARY OF THE THOUGHT PROCESS

It is our hope that this chapter has given the reader some guidelines for interpreting how a particular mutation might affect the structure and function of a protein. Our suggestion would be that you ask the following questions:

First about the protein:

1. What is the cellular environment?
2. What does it do? Is anything known about the amino acids involved in its function?
3. Is there a structure known or one for a homologue?
4. What protein family does it belong to?
5. Are any post-translational modifications expected?

Then about a particular amino acid:

1. Is the position conserved across orthologues? Across paralogues?
2. If a structure is known: is the amino acid on the surface? Buried in the core of the protein?
3. Is it directly involved in function or near (in sequence or space) to other amino acids that are?
4. Is it an amino acid that is likely to be critical for function? For structure?

Once these questions have been answered it should be possible to make a rational guess or interpretation of effects seen by an amino acid substitution and select logical amino acids for mutagenesis experiments.

REFERENCES

Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, *et al.* (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* **25**: 3389–3402.

Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, *et al.* (2000). Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nature Genet* **25**: 25–29.

Barnes MR, Russell RB. (1999). A lipid-binding domain in Wnt: a case of mistaken identity? *Curr Biol* **9**: R717–R719.

Bateman A, Birney E, Durbin R, Eddy SR, Howe KL, Sonnhammer EL. (2000). The Pfam protein families database. *Nucleic Acids Res* **28**: 263–266.

Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, *et al.* (2000). The Protein Data Bank. *Nucleic Acids Res* **28**: 235–242.

Cargill M, Altshuler D, Ireland J, Sklar P, Ardlie K, Patil N, *et al.* (1999). Characterization of single-nucleotide polymorphisms in coding regions of human genes. *Nature Genet* **22**: 231–238.

Copley RR, Barton GJ. (1994). A structural analysis of phosphate and sulphate binding sites in proteins. Estimation of propensities for binding and conservation of phosphate binding sites. *J Mol Biol* **242**: 321–329.

Dayhoff MO, Schwartz RM, Orcutt BC. (1978). A model of evolutionary change in proteins. In Dayhoff MO. (Ed.), *Atlas of Protein Sequence and Structure*, Vol. 5. National Biomedical Research Foundation: Washington DC, pp. 345–352.

Durbin R, Eddy S, Krogh A, Mitchison G. (1998). *Biological Sequence Analysis. Probabalistic Models of Proteins and Nucleic Acids*. Cambridge University Press: Cambridge.

Eddy, SR (1998). Profile hidden Markov models. *Bioinformatics* **14**: 755–763.

Gavel Y, von Heijne G. (1990). Sequence differences between glycosylated and non-glycosylated Asn-X-Thr/Ser acceptor sites: implications for protein engineering. *Protein Eng* **3**: 433–442.

Gibson TJ, Spring J. (1998). Genetic redundancy in vertebrates: polyploidy and persistence of genes encoding multidomain proteins. *Trends Genet* **14**: 46–49; discussion 49–50.

Golding GB, Dean AM. (1998). The structural basis of molecular adaptation. *Mol Biol Evol* **15**: 355–369.

Gupta R, Birch H, Rapacki K, Brunak S, Hansen JE. (1999). O-GLYCBASE version 4.0: a revised database of O-glycosylated proteins. *Nucleic Acids Res* **27**: 370–372.

Hanks SK, Quinn AM, Hunter T. (1988). The protein kinase family: conserved features and deduced phylogeny of the catalytic domains. *Science* **241**: 42–52.

Henikoff S, Henikoff JG. (1992). Amino acid substitution matrices from protein blocks. *Proc Natl Acad Sci USA* **89**: 10915–10919.

Hunter CA, Singh J, Thornton JM. (1991). Pi–pi interactions: the geometry and energetics of phenylalanine–phenylalanine interactions in proteins. *J Mol Biol* **218**: 837–846.

Jeffrey PD, Russo AA, Polyak K, Gibbs E, Hurwitz J, Massague J, *et al.* (1995). Mechanism of CDK activation revealed by the structure of a cyclinA–CDK2 complex. *Nature* **376**: 313–320.

Jones DT, Taylor WR, Thornton JM. (1994). A mutation data matrix for transmembrane proteins. *FEBS Lett* **339**: 269–275.

Kraulis PJ. (1991). MOLSCRIPT: a program to produce both detailed and schematic plots of protein structures. *J Appl Crys* **24**: 946–950.

Krishna RG, Wold, F. (1993). Post-translational modification of proteins. *Adv Enzymol Relat Areas Mol Biol* **67**: 265–298.

Macias MJ, Wiesner S, Sudol M. (2002). WW and SH3 domains, two different scaffolds to recognize proline-rich ligands. *FEBS Lett* **513**: 30–37.

Parekh RB, Rohlff C. (1997). Post-translational modification of proteins and the discovery of new medicine. *Curr Opin Biotechnol* **8**: 718–723.

Peracchi A. (2001). Enzyme catalysis: removing chemically 'essential' residues by site-directed mutagenesis. *Trends Biochem Sci* **26**: 497–503.

Plotnikov AN, Schlessinger J, Hubbard SR, Mohammadi M. (1999). Structural basis for FGF receptor dimerization and activation. *Cell* **98**: 641–650.

Russell RB. (1998). Detection of protein three-dimensional side chain patterns: new examples of convergent evolution. *J Mol Biol* **279**: 1211–1227.

Sayle RA, Milner-White EJ. (1995). RASMOL: biomolecular graphics for all. *Trends Biochem Sci* **20**: 374.

Schoorlemmer J, Goldfarb M. (2001). Fibroblast growth factor homologous factors are intracellular signaling proteins. *Curr Biol* **11**: 793–797.

Schulze-Gahmen U, De Bondt HL, Kim SH. (1996). High-resolution crystal structures of human cyclin-dependent kinase 2 with and without ATP: bound waters and natural ligand as guides for inhibitor design. *J Med Chem* **39**: 4540–4546.

Sunyaev S, Lathe W III, Bork P. (2001). Integration of genome data and protein structures: prediction of protein folds, protein interactions and 'molecular phenotypes' of single nucleotide polymorphisms. *Curr Opin Struct Biol* **11**: 125–130.

Sunyaev S, Ramensky V, Bork P. (2000). Towards a structural basis of human non-synonymous single nucleotide polymorphisms. *Trends Genet* **16**: 198–200.

Taylor WR. (1986). The classification of amino acid conservation. *J Theor Biol* **119**: 205–218.

Waksman G, Kominos D, Robertson SC, Pant N, Baltimore D, Birge RB, *et al.* (1992). Crystal structure of the phosphotyrosine recognition domain SH2 of v-src complexed with tyrosine-phosphorylated peptides. *Nature* **358**: 646–653.

Wang Z, Moult J. (2001). SNPs, protein structure, and disease. *Hum Mutat* **17**: 263–70.

Wilks HM, Hart KW, Feeney R, Dunn CR, Muirhead H, Chia WN, *et al.* (1988). A specific, highly active malate dehydrogenase by redesign of a lactate dehydrogenase framework. *Science* **242**: 1541–1544.

Wilson CA, Kreychman J, Gerstein M. (2000). Assessing annotation transfer for genomics: quantifying the relations between protein sequence, structure and function through traditional and probabilistic scores. *J Mol Biol* **297**: 233–249.

Wolfe SA, Grant RA, Elrod-Erickson M, Pabo CO. (2001). Beyond the ‘recognition code’: structures of two Cys2His2 zinc finger/TATA box complexes. *Structure (Camb)* **9**: 717–723.

APPENDIX: TOOLS

Protein sequences

http://www.expasy.ch/
http://www.ncbi.nlm.nih.gov/

Amino acid properties

http://russell.embl-heidelberg.de/aas/

Domain assignment/sequence search tools

http://www.ebi.ac.uk/interpro/
http://www.sanger.ac.uk/Software/Pfam/
http://smart.embl-heidelberg.de/
http://www.ncbi.nlm.nih.gov/BLAST/
http://www.ncbi.nlm.nih.gov/COG/
http://www.cbs.dtu.dk/TargetP/

Protein structure

Databases of 3D structures of proteins
http://www.rcsb.org/pdb/
Structural classification of proteins
http://scop.mrc-lmb.cam.ac.uk/scop/

Protein function

http://www.geneontology.org/