CHAPTER 2

Insertion and Deletion Events, Their Molecular Mechanisms, and Their Impact on Sequence Alignments

LIAM J. MCGUFFIN
University of Reading, United Kingdom

The alignment of biological sequences allows us to infer the evolutionary relationships between different genes and proteins. Most new genes and proteins will evolve either through insertions or deletions of sets of subsequences, or through point mutations, where one amino acid is replaced with another. Therefore, we can judge the evolutionary

A)

```
Yeast: FTKENVRILESWFAKNIENPYLDTKGLENLMKNTSLSRIQIKNWVSNRRRKEK
Human: YSKGQLRELEREYAAN---KFITKDKRRKISAATSLSERQITIWFQNRRVKEK
```

B)

Figure 2.1. (A) A sequence alignment between the mat alpha2 Homeodomain protein from yeast (Protein Data Bank (PDB) code *1k61*) and Homeobox protein Hox-B13 from Human (PDB code *2cra*). (B) A structural alignment between *1k61* and *2cra* carried out using TM-align (Zhang and Skolnick 2005).

distance between related organisms by scoring the differences occurring between their protein and DNA sequences. In this chapter, we will focus on insertion and deletion events and explain how these events affect how we carry out and score sequence alignments.

We will begin by looking at a sequence alignment in order to illustrate the problem of handling insertions and deletions in related sequences. Figure 2.1A shows a local alignment carried out for similar proteins from two different eukaryotic organisms separated by more than a billion years of evolution.

It is apparent that despite the fact that these proteins have a similar DNA binding function and similar structure (Figure 2.1B), their respective amino acid sequences are quite different. Many mutations have occurred to produce the changes in sequences, most of which may be point mutations; however, the two proteins are also different in length, which means that a gap has been introduced into the alignment.

Given further evidence that the yeast sequence was closer to the common ancestral sequence, one might argue that a deletion event (or a number of deletion events) had probably occurred. However, without such information, we are unable to make assumptions about whether an insertion event had occurred in the yeast sequence or whether a deletion event had occurred in the human sequence. For this reason, when we carry out an alignment between sequences, we often treat insertion events and deletion events the same. Thus, insertions and deletions are often grouped together and collectively referred to as *indels*.

In fact, it could be said that all mutation events occurring to alter protein sequences could be explained by indel events. For example, the occurrence of point mutations may be accounted for by the deletion of one base or amino acid and the insertion of another. However, in the case of sequence alignments, the molecular mechanisms of how such an amino acid substitution occurs is less important. What is more important, for the scoring of protein sequence alignments in particular, is which type of substitution has occurred at a particular position in the sequence.

The effectiveness of the alignment of protein sequences can be improved by the use of an amino acid substitution matrix. A plethora of matrices has been developed, such as PAM (Dayhoff et al. 1978), GCB (Gonnet et al. 1992), JTT (Jones et al. 1992), BLOSUM62 (Henikoff and Henikoff 1992), and, more recently, OPTIMA (Kann et al. 2000), which are used to score the alignment of different pairs of amino acids with different weightings. These weightings account for the different physical, chemical, and structural properties shared by each pair of amino acids; a leucine–isoleucine match within an alignment may be scored higher than a leucine–tryptophan match. Similarly, when DNA sequences are aligned, there are various substitution models; for example, *transitions* (e.g., C-T, T-C, A-G, G-A) will often be weighted higher than *transversions* (e.g., T-A, A-T, C-G, G-C).

However, mutations considered as "true" indel events are those that cause changes in the lengths of biological sequences, and these have a more drastic individual effect on the scoring of sequence alignments. To account for the insertions or deletions of stretches of subsequences, gaps must be introduced into the alignment, and this incurs a penalty.

Traditionally, there are two types of penalty imposed when accounting for gaps within an alignment: gap opening penalties and gap extension penalties. The cost of opening a new gap is generally far weightier

than the cost of extending an existing gap. However, the weighting that should be used for gap penalties often depends on the context and type of sequences.

There is still no perfect system for scoring biological sequence alignments. However, most of the systems that have been developed have attempted, in one way or another, to take into account the different types of mutation and the probability that they may occur, the inferred mechanisms involved, and their evolutionary consequences. Each scoring scheme will be appropriate for each varying situation. It is important to use the appropriate gap penalties as well as choosing an appropriate substitution matrix for the sequences you are aligning. In this chapter, we will review the different events that may lead to insertions and deletions in biological sequences and how these events help us to better understand sequence alignments and their scoring.

MUTATIONS

Before we can begin to explain the mechanisms of indel events, we must first review the different types of mutation that may occur at the DNA level and the resulting effects that these different types have on the translated protein sequence. Novel protein sequences will normally evolve through the mutation of existing sequences via indel events or point mutations. In order to become accepted, the mutations occurring within the sequence must be either neutral or advantageous.

If a mutation leads to an alteration in the protein structure that is sufficient to inactivate the protein, then it is unlikely that the mutation will become accepted. The vast majority of mutations in protein sequences are deleterious and are therefore eliminated through natural selection. However, in some cases, mutations can be neutral; for example, an amino acid substitution that does not alter the structure or function of the resulting protein and is neither damaging nor beneficial. In rare cases, a mutation may be advantageous and as a result may propagate throughout the population (Alberts et al. 1994).

Point Mutations

Point mutations occur when one nucleotide is exchanged for another. When these mutations occur within coding regions of DNA, they can be classified as one of three types: *silent*, *nonsense*, or *missense* (Weaver and Hedrick 1992).

Silent Mutations When a silent mutation occurs, the codon is altered such that it results in the translation of the same amino acid. Therefore, no effect is seen at the protein sequence level and such mutations can be seen as neutral.

Nonsense Mutations A nonsense mutation occurs when the base change results in the production of a stop codon. This leads to the translation of truncated proteins and is often deleterious. However, a nonsense mutation toward the beginning of the sequence will have a potentially more dramatic effect than one that occurs near the end of the sequence.

Missense Mutations In this case, the codon is altered such that it results in the translation of a different amino acid. These mutations can be deleterious, effectively neutral if they lead to no change in biological activity, or, in rare cases, advantageous.

Point accepted missense mutations are accounted for in protein sequence alignments by using mutation matrices. These matrices are essentially tables of scores that weigh changes in amino acids according to the differences in their physical, chemical, or structural properties.

Certain mutations that appear to be point mutations could be considered to have occurred through the deletion of one amino acid followed by the replacement of another. Intuitively, it would seem that these events will be far less likely than a true point mutation. True indel events have a greater effect on the gene products and can significantly alter the amino acid sequences. Often, in the case where a frameshift has occurred, indels may lead to a deleterious nonsense mutation.

Insertions and Deletions (Indels)

For the purpose of a sequence alignment, we do not differentiate between insertion and deletion events; however, there are subtle differences as to how they may occur and potentially major differences in their effects.

Simply, an insertion occurs when one or more nucleotides are added to the DNA sequence. A deletion occurs when one or more nucleotides are removed from the DNA sequence. Insertions and deletions can occur through errors during DNA replication followed by failure of repair, unequal crossover during recombination, or the introduction

of transposable elements. Insertions and deletions observed in protein sequence alignments may also have occurred through alternative splicing, where different combinations of exons from the same gene are translated into different protein sequences.

Mutations caused by insertions of DNA subsequences can be reverted, or back-mutated, for example, through correctly functioning DNA repair mechanisms or through the excision of the transposable element. Similarly, a deletion may be reverted through the reintroduction of the lost subsequence. However, such reversions are highly unlikely to occur naturally (Weaver and Hedrick 1992). The mechanisms of indel events are discussed in more detail later in the chapter.

Frameshift Mutations Insertions and deletions of bases may cause alterations to the reading frame of the gene. Such alterations are known as *frameshift mutations,* and they may have a drastic effect on the coding regions. The insertion or deletion of one or two nucleotides will alter every codon downstream of the mutation, which can lead to translation of a different amino acid sequence. This may also lead to a nonsense mutation occurring downstream, which causes the premature termination of translation (Streisinger et al. 1966). Figure 2.2a illustrates how the insertion of a single base leads to a frameshift mutation.

Triplet indels may have milder consequences because the reading frame of the gene downstream of the mutation will be unaffected. Figure 2.2b illustrates how the insertion of a triplet maintains the reading frame downstream of the mutation. Multiple triplet indels may also occur, which may lead to severe consequences. For example, variable numbers of trinucleotide repeats occur in human genetic disorders such as Huntington's disease and Fragile X (Weaver and Hedrick 1992).

Splice Site Mutations In eukaryotic organisms the coding regions of DNA (*exons*) are interrupted by noncoding regions (*introns*). Prior to translation, the introns are cut out and exons are stitched back together in a process known as *RNA splicing*.

Indels can lead to the disruption of the specific sequences denoting the sites at which splicing takes place. These mutations can result in one or more introns remaining in the mature messenger RNA or lead to one or more exons being spliced out. This has severe implications for the resulting protein sequence.

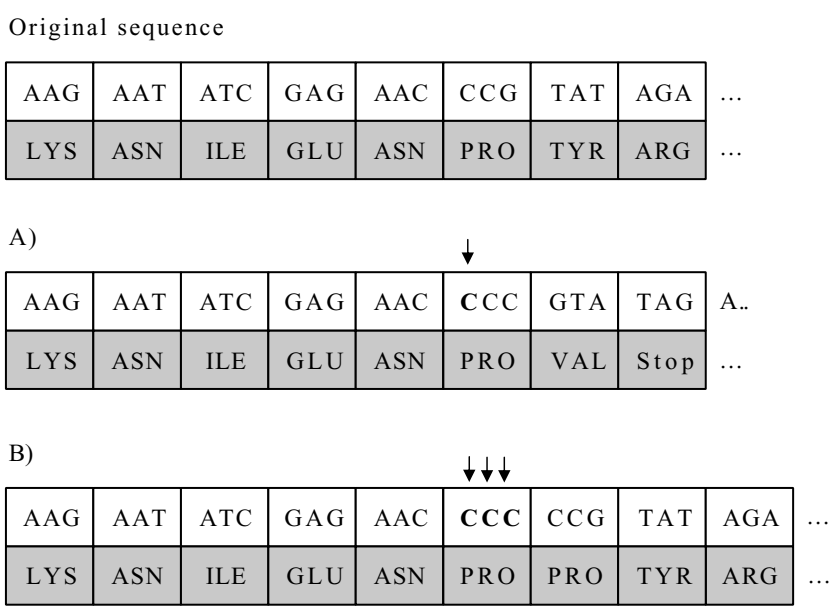

Figure 2.2. Insertions leading to frameshift mutations. Original sequence: A hypothetical DNA sequence is divided into codons with the amino acids translated below. (A) Insertion of a single cytosine nucleotide (in bold) leading to a frameshift and nonsense mutation. (B) Insertion of three cytosine nucleotides leading to additional proline inserted into the protein sequence. The reading frame downstream of the mutation is preserved.

Alternative splicing can also be a mechanism explaining the occurrence of insertions and deletions in eukaryotic proteins. The mechanisms of alternative splicing are discussed in later in the chapter.

PROPOSED MECHANISMS OF INDEL FORMATION

There are many different molecular events that can lead to the formation of indels in biological sequences. Errors can occur during DNA replication; these may be missed during proofreading by polymerases and DNA repair mechanisms. Alternatively, insertions can occur in DNA due to transposable genetic elements such as simple insertion sequences and more complex transposons and retroviruses. At the RNA level, alternative splicing of genes can give rise to many different protein sequences encoded from various combinations of exons. Larger

insertions and deletions may also be explained by DNA recombination mechanisms occurring during meiosis.

Errors in DNA Replication

The main molecular mechanism proposed that causes frameshift mutations at the DNA level involves errors occurring within the replication machinery (Levinson and Gutman 1987). When a gene is inaccurately copied, one or many base pairs may be introduced or omitted from the original sequence.

Streisinger and Owen (1985) developed a model for the mechanisms of frameshift mutations; the idea for the model arose from their studies on the lysozyme gene sequences in bacteriophage T4. The mutations were found to occur most frequently in regions of repeated sequences, and their model provided an explanation for this observation.

In the model, the insertion or deletion of bases is explained by the slippage of one strand occurring during DNA synthesis, which creates a loop of one or many bases. The loop is stabilized by pairing that occurs at an alternative position within the repetitive sequence. An insertion occurs when the loop is in the newly synthesized strand, and a deletion occurs when the loop is in the template strand. Figure 2.3 shows a simplified view of Streisinger and Owen's model and illustrates how the DNA replication machinery can "slip a cog" occasionally.

If the proofreading stages and the DNA repair mechanisms fail to correct the insertion or deletion of a subsequence, then it will remain in the sequence. In some cases, after slippage of the replication machinery has caused a frameshift, the DNA repair mechanism may attempt to revert the frameshift by inserting or deleting several extra bases, thereby incorporating new insertions or deletions (Levinson and Gutman 1987).

Transposons

Transposons are mobile genetic elements that move around the genome of an organism, causing both insertion and deletion events in DNA sequences. The discovery of transposable elements by Barbara McClintock, for which she won a Nobel Prize in the early 1980s, has had a major impact on our understanding of genetics. Transposable elements are considered to be a major contributor to natural genetic indel variation among humans (Mills, Bennett, et al. 2006). In fact, it is estimated that up to approximately 44% of the human genome is made

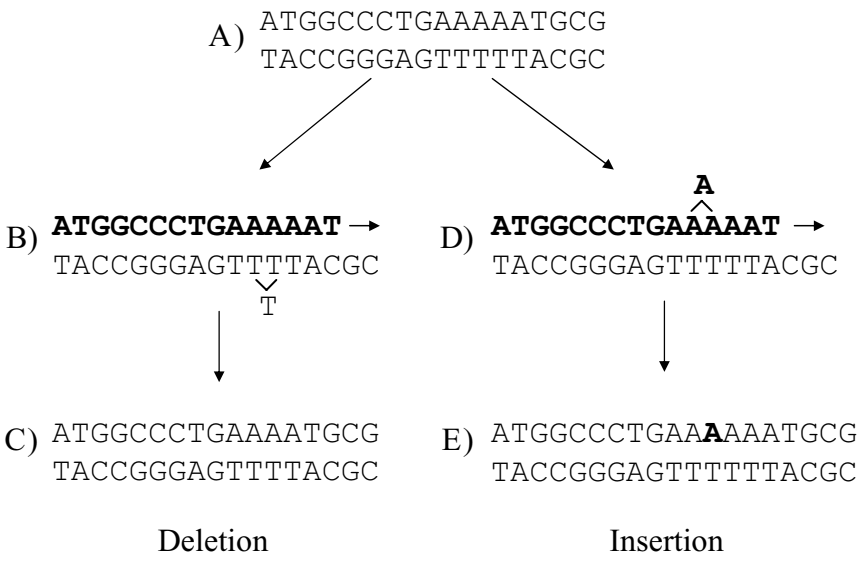

Figure 2.3. Model for frameshift mutation proposed by Streisinger and Owen (1985). (A) Original DNA duplex with adenine repeat. (B) Slippage occurs during replication causing the template strand to loop out. (C) Single-base deletion occurs after another round of replication. (D) Newly synthesized strand loops out during replication. (E) Single-base insertion occurs after another round of replication. Figure adapted from Weaver and Hedrick (1992).

up of transposons or remnants of transposons (International Human Genome Sequencing Consortium 2001). Transposons are a major cause of mutations and variation in the amount of DNA within a genome.

Transposons are generally classified into two categories based on the mechanisms of their transposition. First, conservative or simple transposition occurs when both strands of DNA are conserved as they move from place to place; that is, the DNA is "cut" from one location and "pasted" into another. The second mechanism is replicative transposition, where the DNA is "copied" from one location and "pasted" to another. All transposons generate direct repeats within the DNA sequence at their point of insertion (Weaver and Hedrick 1992).

Class I Transposons Class I transposons, which are also known as retrotransposons, generally use a replicative "copy and paste" mechanism. An RNA copy of the transposon is made, which is then reverse transcribed into DNA and inserted back into the genome. Retrotransposons use a

mechanism similar to that of retroviruses (such as HIV) and are very probably their evolutionary ancestors. The retrotransposon often carries the gene encoding the reverse transcriptase enzyme required to facilitate transposition. Long interspersed nuclear elements (LINEs) and short interspersed nuclear elements (SINEs) are highly abundant types of retrotransposons found in mammalian genomes (Weiner 2000).

Class II Transposons Class II transposons mainly transpose via a "cut and paste" mechanism. This requires an enzyme known as a *transposase*, which is often encoded within the transposon itself. In general, the transposase acts by binding to DNA and making a staggered cut at the target site to produce sticky ends. The transposon is then cut out and ligated into the target site, leaving gaps at each end, which are then filled by a DNA polymerase. Some Class II transposons may also transpose using a "copy and paste" mechanism. The two proposed mechanisms for transposition of Class II transposons are shown in Figure 2.4.

Alternative Splicing

In eukaryotes, alternative splicing is an important mechanism that allows for the production of many different protein sequences from the same gene. This is achieved through splicing of different combinations of exons (Breitbart et al. 1987). Using this process, eukaryotic organisms can achieve more efficient data storage at the DNA level, and theoretically faster evolution of new proteins.

Alternative splicing either involves the substitution of one subsequence encoded by an exon for another, which is known as substitution alternative splicing, or results in the insertion or deletion of a subsequence, which is known as length-dependent alternative splicing (Kondrashov and Koonin 2003). Figure 2.5 illustrates how length-dependent alternative splicing may lead to the insertion or deletion of a new subsequence within a protein.

Computational studies have been carried out to investigate the structural and functional influences of alternative splicing leading to a potentially vast repertoire of proteins. Alternative splicing is thought to be a key process in modulating gene function and influencing protein networks through the generation of structures with many varying conformations (Yura et al. 2006).

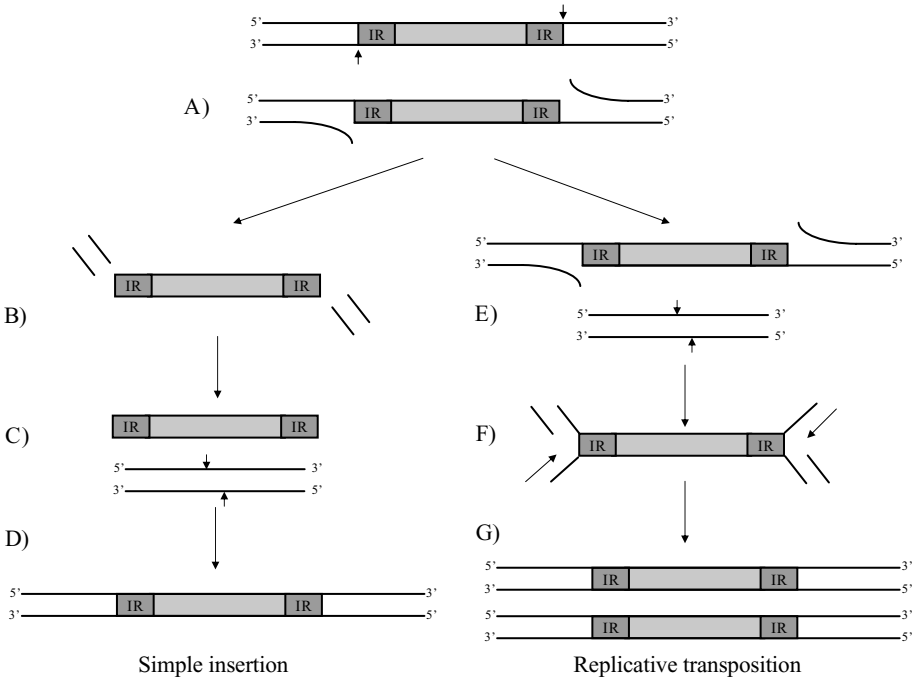

Figure 2.4. Mechanisms of transposition adapted from Tavakoli and Derbyshire (2001). (A) Transposable element with inverted repeat regions at either end. The first stage involves nicking occurring at the 3′ ends. (B) With simple insertion, the transposon is cut from the original location and pasted into the new location; this requires cleavage at the 5′ ends. (C) The transposon interacts with the target sequence which has a staggered cut. (D) The transposon is inserted into the new location. (E) With replicative transposition, the transposon is copied and pasted, which occurs via strand transfer. The 5′ ends of the transposon remain intact, and the free ends interact with the nicked target sequence. (F) The free ends in the target sequence serve as primers for DNA replication. (G) The result is a cointegrate structure and replication of the transposon.

Unequal Crossover and Chromosomal Translocation

Recombination events can create indels at the chromosomal level. If sister chromatids misalign during meiosis, for example, due to repeated regions within the sequence, unequal crossover may occur during recombination. This can lead to the duplication or deletion of thousands of base pairs within chromosomes. Figure 2.6 illustrates unequal crossover leading to an insertion in one chromosome and a deletion in the other.

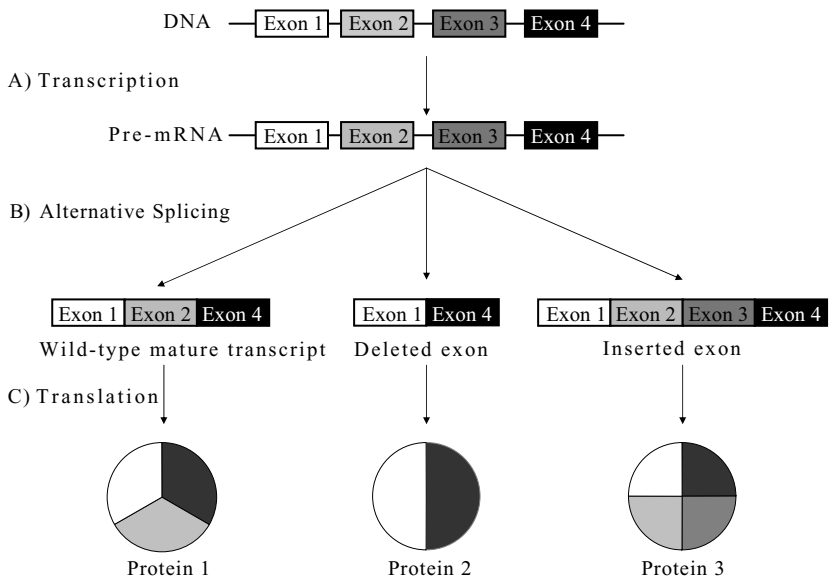

Figure 2.5. Length-dependent alternative splicing. (A) Pre-mRNA is transcribed from the DNA template including introns and several possible exons. (B) Alternative splicing leads to alternative mature mRNA molecules made up of different combinations of exons. (C) In this hypothetical example, the wild-type protein 1 is translated from exons 1, 2, and 4. Deletions and insertions may occur due to alternative splicing, which will lead to different protein sequences, such as protein 2, where exon 2 has been deleted, and protein 3, where exon 3 has been inserted.

Unequal crossover during meiosis leading to chromosomal insertions and deletions has been shown to be an important factor in disease. For example, neurofibromatosis is caused by microdeletions, a result of a homologous recombination between misaligned repeat regions (Lopez-Correa et al. 2000). In addition, the CATCH-22 group of syndromes and Williams–Beuren syndrome are correlated with high levels of unequal meiotic crossover occurring at the site of chromosomal deletions (Baumer et al. 1998).

Indels caused by unequal crossover lead to variation in microsatellite and minisatellite repeat regions. Variable number of tandem repeats (VNTRs) are useful for gauging genetic variation among populations and are widely exploited for DNA fingerprinting (Harding et al. 1992).

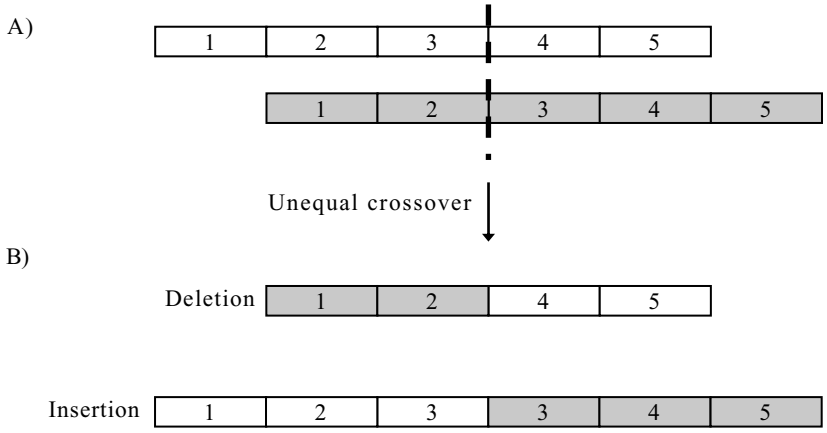

Figure 2.6. Unequal crossover during recombination leading to insertions and deletions, adapted from Alkan et al. (2002). (A) Sister chromatids misalign during meiosis due to repeat units. The site of crossover is indicated by the dashed line. (B) Strand breaks on sister chromatids lead to different numbers of repeat units. A deletion occurs in one chromatid and an insertion occurs in the other.

Indels at the chromosome level may also occur due to chromosomal breaks followed by translocation events. Several widely studied diseases in humans are known to be caused by chromosomal translocations. For example, translocation of the long arm of human chromosome 21 to chromosome 14 can lead to Down syndrome (Weaver and Hedrick 1992).

IMPACT OF INDELS ON SEQUENCE ALIGNMENTS

Whatever their scale and underlying mechanisms, insertions and deletions are indicated by gaps in sequence alignments. These gaps pose a significant problem and greatly affect the accuracy of the alignment. The way we treat the scoring of these gaps is therefore crucial to effective sequence alignment and is dependent on the context of the insertion or deletions. The penalty incurred by a gap should reflect the expected frequency of the indel event. Therefore, it is crucial to know the type of sequence in which the gap is occurring and the likelihood that a gap will occur in that context. For example, a large gap in an important protein that causes a major conformational change may be less likely to occur than indel events leading to VNTRs.

Global and Local Alignments

Perhaps the most widely used algorithm for alignment of two sequences is dynamic programming. Needleman and Wunsch (1970) were the first to describe the use of dynamic programming to perform a global alignment between pairs of biological sequences. Smith and Waterman (1981b) later refined the algorithm to allow for optimal local alignments; the refined algorithm proved to be more accurate for aligning repetitive regions in sequences. (See Chapter 3 for further discussion about global and local alignments.)

Dynamic programming is a computationally efficient way of optimally aligning sequences and is dependent on a scoring scheme that is based on both the substitution of bases or amino acids and the creation and extension of gaps. Methods to improve sequence alignments have often focused on the optimization of substitution matrices and gap penalties.

Gap Penalties

Most programs will implement a cost for introducing a gap when aligning two biological sequences. This cost contributes to the overall score of the alignment and is often closely dependent on the substitution matrix used. For example, when determining the cost of introducing a gap, we must also consider the costs associated with aligning dissimilar bases or amino acids. The placement of gaps in a sequence alignment is a difficult task, and a number of solutions to the problem have been proposed.

Affine Versus Linear Gap Penalties The most basic gap penalty is the linear gap penalty, where the cost of introducing a new gap in a sequence is the same as the cost of extending an existing gap. For example, the introduction of three gaps each of one amino acid in length is seen as equally likely as the introduction of one large gap of three amino acids in length and therefore incurs the same cost.

The affine gap penalty, however, differentiates between the opening of a gap and extension of a gap. The opening of a new gap is considered to be of greater cost than extending the gap. This is based on the conjecture that one long gap cause by a single indel event is more likely to occur than several shorter gaps caused by several different indel events.

```
A)  GATCGCGCGCGCGCGCATGC
    GATC--G--C--G--CATGC

B)  GATCGCGCGCGCGCGCATGC
    GATCGCG-------CATGC
```

Figure 2.7. Affine gap penalties versus linear gap penalties. (A) Alignment of two hypothetical DNA sequences. Using a linear penalty, all gaps are scored equally, which could result in a "gappy" alignment. (B) The affine gap penalty scheme involves two scores: one for opening a gap and another for extending the gap. The cost of opening a new gap is higher than that for extending a gap, which results in a more accurate alignment. A single indel event indicated by a single long gap is assumed to be more likely than a number of separate events indicated by a series of smaller gaps.

Thus, according to the affine gap penalty scheme, gaps are scored using the formula $o + e \times l$, where o is the cost for opening a gap, e is the cost for extending the gap, and l is the length of the gap. Figure 2.7 shows a hypothetical example highlighting the advantage of the affine gap penalty scheme.

The parameters representing the costs of affine gap penalties were rigorously optimized by Barton and Sternberg (1987a). Since then several groups have attempted to further improve or generalize the scoring of the affine gap penalty (Altschul 1998; Mott 1999). Despite being an obvious improvement over a linear gap penalty, the affine gap penalty is still an imprecise treatment of real gaps. It has been criticized for having no sound theoretical basis and no real supporting experimental evidence (Goonesekere and Lee 2004).

Improving Gap Penalties by Understanding the Context of Indel Events

The context and mechanisms of indel events have been recognized as important factors for determining optimal gap penalty functions. There have been many attempts to deduce gap penalties empirically based on observation of patterns of indel in aligned sequences, such as the studies carried out by Benner et al. (1993), Reese and Pearson (2002), and Chang and Benner (2004).

The gap penalties may be manually adjusted when using common sequence alignment tools, such as ClustalW (Thompson et al. 1994), to account for rules determined by protein structure, for example. Indels may have a greater impact on the protein structure, whereas point mutations usually have less effect. It may be preferable to allow gaps only in coil regions, which are often more variable, rather than introduce a gap in a strand or helix.

Recent work on the production of improved gap penalties for homology modeling has focused on the structural context. Goonesekere and Lee (2004) suggest a simple modification of the affine gap penalty based on their observations of the patterns of gaps occurring in a database of structurally aligned proteins. More recently, Madhusudhan et al. (2006) have benchmarked a new gap penalty, which automatically varies depending on the structural context, against the standard affine gap penalty scheme. This novel variable gap penalty is reported to significantly increase the number of correctly aligned residues.

CONCLUSION

It is important to understand the context of insertions and deletions and their mechanisms when carrying out sequence alignments. Some indel events can cause major variations in sequences, which will have a great effect on the inferred relationship. Other events may have a less severe impact on the variation and should be taken into account when the sequence alignment is scored. Thus, gap penalties should be dependent on where the gap occurs, and the scoring should be appropriate to the situation in which the indel has arisen. Gap penalties incurred for one sequence alignment may not be appropriate for another.

We have begun to develop different scoring schemes for sequence alignments, which take into account the context of indels. However, the scoring of sequence alignments can be further improved, perhaps through our improved understanding of the mechanisms causing indel events.