Chapter 8
Status of Research on Insertion and Deletion Variations in the Human Population

Liqing Zhang, Mingming Liu, and Layne T. Watson

Abstract Insertion and deletion (indel) variants comprise a major proportion of human genetic variation. However, little is known about their effect on humans. The void of understanding is largely due to the lack of both biological data and computational resources. Thanks to the progress made by many large-scale genomic projects, a substantial amount of data is now available, enabling the prediction of functional elements in the genome. In this work, we review the impact of indel variants on human biology, evolution, and health, and examine the currently available computational resources for predicting the functional effects of indels and their limitations. We then present a newly developed program for indel effect prediction using a hidden Markov model-based framework and discuss future work for better understanding the effects of indel variants on human biology and health.

8.1 Indel Effects on Human Biology, Health, and Evolution

Indel is the Second Most Common Type of Genetic Variation in Humans The rapid development of sequencing technologies has made possible cataloging the entire set of genetic variants harbored in human populations. The recent pilot study conducted by the 1000 Genome Project Consortium has revealed that there are about 15 million single nucleotide polymorphisms (SNPs), one million short insertions and deletions (indels), and 20,000 structural variants (SVs) harbored by the populations they studied [1]. Thus, indel ranks as the second most common type of genetic variation in humans.

Indel Variants Have Profound Functional Impact on Human-Specific Evolution and Adaptation Comparison of the human genome and several other closely related species' genomes has shown that approximately 0.8 million human-specific

L. Zhang (✉) · M. Liu · L.T. Watson
Department of Computer Science, Virginia Tech, Blacksburg, VA, USA
e-mail: lqzhang@vt.edu

L.T. Watson
Department of Mathmatics, Virginia Tech, Blacksburg, VA, USA

C. Chauve et al. (eds.), *Models and Algorithms for Genome Evolution*,
Computational Biology 19, DOI 10.1007/978-1-4471-5298-9_8,

indels affect more than 7000 genes, and these genes may have contributed to human traits via changes at the RNA and protein levels [2]. In addition, indels have been found to contribute to about 5 % of the human–chimpanzee divergence, much higher than the 1.5 % nucleotide divergence, suggesting that indels might have played an even bigger role than nucleotide divergence in human-chimpanzee differentiation [20]. Some human-specific indels show evidence of positive selection and might have played important roles in human adaptation both at the species level and the subpopulation level [3]. The importance of indels to the evolution of genomes is further supported by a study that has found an increased rate of mutations (higher levels of SNPs) near the regions with indels [7] at both the species and population levels.

Indels May Hold the Key to Understanding Human Diseases [17] Depending on the locations of the indels, indels can potentially lead to frame shift and thus influence proteins by changing protein sequences, gene expression patterns (by affecting promoter regions, introns, or UTRs), and exon splice patterns. A well-known case of indel effects is cystic fibrosis, a genetic disease frequently caused by a 3-bp deletion within the coding region of CFTR [5]. Although it is in-frame, the deletion leads to abnormal protein folding and protein degradation. Indels in noncoding regions can also cause human diseases. For example, when indels occur within the promoter region of the FMR1 gene, they can change the promoter methylation pattern and thus the gene expression pattern of FMR1, resulting in fragile X syndrome [5]. Mill et al. [17] have shown that about 42 % of the nearly two million indels they identified are mapped to human genes and more than 2000 indels affect coding exons and likely disrupt protein function and cause phenotypic changes in humans. Their analysis of the experimental data in mice shows that 83 % of the coding indels yield abnormal phenotypes. Moreover, the indels tend to have strong linkage disequilibrium (LD) with the SNPs identified in genome wide association studies (GWAS). Diseases with an indel genetic basis might have been mistakenly determined as SNP related, only because of strong LD. In these cases, accurate indel effect prediction is the only way to improve our understanding of these diseases.

8.2 Current Research on Indel Variants

Despite all the evidence suggesting the importance of indels, their research has lagged behind studies of other variant types such as SNPs and SVs. From a biological point of view, it is time consuming and difficult to experimentally characterize the impact of indels on genes or protein function. Computationally, there are two main problems: the lack of specialized database resources for indel curation and function annotation, and the lack of computational methods/programs to predict the effect of indel variants.

The Lack of Specialized Database Resources for Indel Curation and Function Annotation Currently, indel polymorphisms are loosely stored in dbSNP, where simple annotation based mostly on location of indel variants with respect to

genes is provided. The current dbSNP (build 135) contains 6,312,022 nonredundant or reference indels, which are clustered from 7,806,204 indels submitted by various researchers, with a major proportion generated by a few large-scale studies [1, 11, 12, 16, 19]. Indels are roughly annotated to categories including introns, intergenic regions, UTRs, and frameshift indels. Evidently, annotation of indel variants by dbSNP is so coarse-grained and overly simplistic that it does not help researchers prioritize and choose from the sea of indels the strongest candidate indels for traits or diseases of interest. Other large data servers, such as the UCSC Genome Browser and Ensembl, import indel annotation directly from dbSNP. Hence, there is no dedicated computational resource and database for fine-grained annotation of indel effect. It must be noted that indels cannot be simply taken as repeats or mini- or microsatellites as the majority (70 %) of them are nonrepetitive [16].

The need for a database dedicated to indels is further emphasized by several recent studies that demonstrate the far-from-completeness of our current catalog of indel variants in humans. A 2011 study shows that more than 63 % of the nearly two million indels identified in the 79 diverse human genomes are novel [17], compared to the ones in dbSNP. Most recently (August 2012), sequencing and analysis of an Indian female's genome reveals that about 84 % of this person's indels are unique, i.e., not documented in any of the sequenced genome databases, in contrast to less than 3 % of the SNPs being unique [9]. Thus compared to SNPs, the research on cataloging indel variants is still in its infancy and intense effort is needed in order to have a complete inventory. A specialized database devoted to indels would greatly facilitate this task and thus take an important step towards understanding their effect on human traits and diseases.

The Lack of Computational Methods/Programs to Predict the Effect of Indel Variants A survey of the tools for predicting SNP variant effect shows that there are a few dozen computer programs and web servers devoted to such a purpose [13]. In contrast, the computational resources devoted to indel effect prediction is very limited and nearly nonexistent. At the time of this writing, only three studies were found that propose computational methods for predicting the functional effect of indel variants. The first recent study proposed an evolutionary conservation-based approach to score and predict the effect of indel variants for both coding and noncoding regions [21]. Although the results are encouraging, there is no readily available source program. The provided online web server has several major limitations. First, it has limited prediction power, restricted to only one indel on one sequence per analysis. Ideally, the user should have the freedom to upload an input file for batch analyses. Second, although the paper has predictions for both coding and noncoding indels, the web server does not have noncoding indel prediction. Third, the prediction score indicates the deleteriousness of an indel, but does not have any information on what functions are likely affected. Finally, the online server has bugs, returning randomly truncated amino acid sequences in some tests. Another recent study proposed SIFT-Indel that uses a simple decision tree approach to classify the effect of indel variants [10]. For indel effect prediction, four features are extracted for each indel: fraction of affected conserved DNA bases, indel location relative to

the transcript, fraction of affected conserved amino acids, and minimum distance of the indel to the exon boundary of all the affected transcripts. Though easy to interpret due to the nature of a decision tree, the predictive power of SIFT-Indel is rather limited due to two major drawbacks. First, the method only applies to frameshift indels, which account for a tiny proportion (~ 0.05 %) of indel variants [18]. Second, it can only make coarse-grained qualitative predictions, that is, an indel can be either "gene-damaging" or "neutral". However, a computational method or program that can produce quantitative ranking of variant effect is much more useful for indel filtering and prioritization than qualitative assessment [6].

The third latest study introduced an alignment-based score to predict the effect of genetic variants, including single SNPs, indels, and multiple mutations [4]. The corresponding program PROVEAN also uses an evolutionary conservation-based method to evaluate the deleteriousness of variants. Though promising, the program is only applicable to in-frame indels. However, frameshift indels are expected to be more deleterious and thus are also an important type of indels that require function effect prediction. To address the limitations of the current programs, the authors recently developed HMMvar [15], a program using a hidden Markov model (HMM)-based scoring method to predict the effect of indels. The following section gives an overview of the program and some results on its application.

8.3 The Hidden Markov Model-Based Scoring Method for Predicting Indel Effects

The HMM-based method to score the effect of indel variants incorporates hypothesis testing naturally and formally into a probabilistic framework. A profile HMM can be used to describe the probabilities of multiple sequences generated from the HMM model, thus representing a family of proteins. Briefly, a profile HMM, named for the characteristic output "profile" of a particular hidden Markov model, is a finite state machine consisting of a series of nodes, each of which corresponds roughly to a position (column) in the alignment from which it was built. Most of the previous prediction methods are based on the principle that important amino acids will be conserved in the protein family, and so mutations occurring at well-conserved positions tend to be deleterious to the functions of the protein. This principle can be reflected exactly by the profile HMMs. Basically, a HMM profile is a probabilistic description of the consensus of a multiple sequence alignment. Thus it is reasonable to use a profile HMM to gauge how far mutations take the original sequence away from the set of sequences represented by the HMM. The further away from the representation, the more likely the mutation is deleterious.

Figure 8.1 shows the flowchart of profile HMM-based prediction, or the workflow of the HMMvar program. The pipeline consists of five steps: (1) find "seed" proteins that are associated with indels; (2) for each seed protein, find homologous sequences from a database; (3) do multiple sequence alignment (MSA) for each set of homologous sequences; (4) build a profile HMM based on each MSA; (5) predict

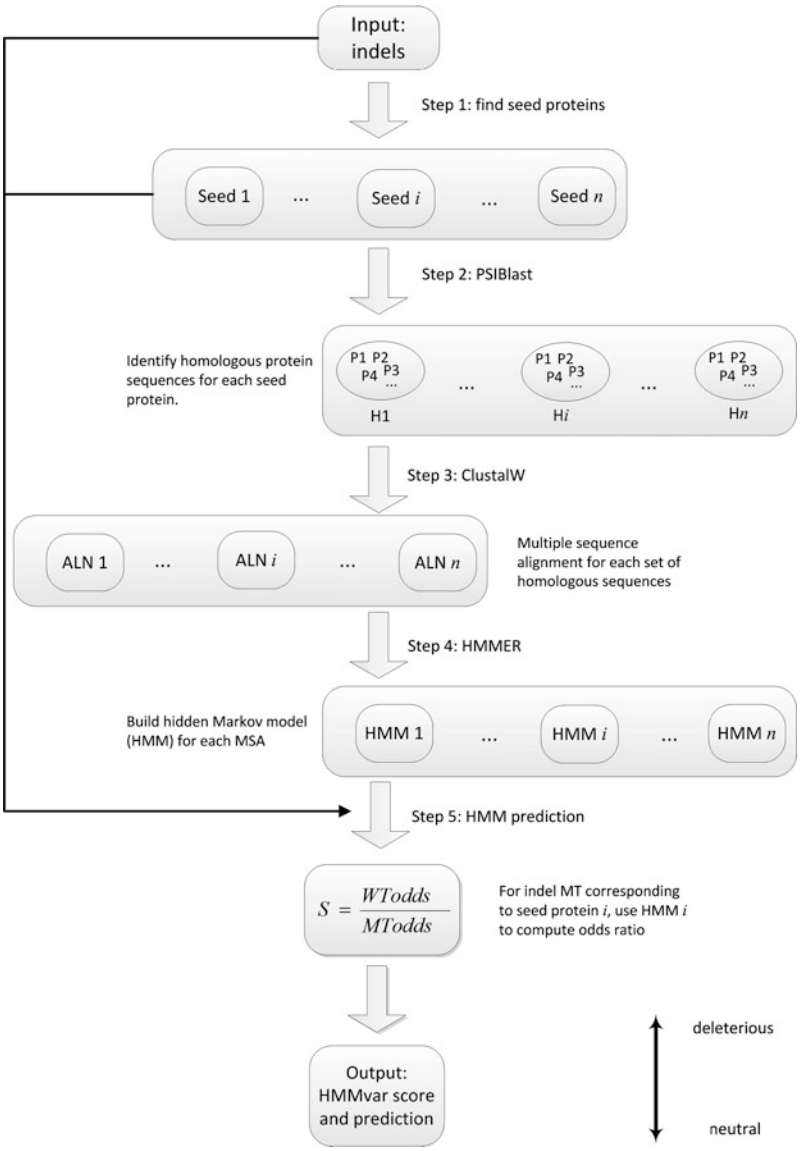

Fig. 8.1 An overview of profile HMM-based variant prediction

the functional effects of indels using the profile HMMs; precisely, for each mutated protein with the indel (MT, mutant type) and corresponding seed protein i (WT, wild type), use the ith HMM to compute the odds ratio (odds that the HMM could have generated the WT sequence)/(odds that the HMM could have generated the MT sequence)—this odds ratio is the HMM indel score.

Table 8.1 Numbers of different indel types with or without the LSDB records

Indel types	LSDB	NonLSDB	Total
Nonsense	112	15	127
Missense	0	56	56
Frameshift	2519	1387	3906
Total	2631	1458	4089

The bit scores calculated from the HMMs are used to quantitatively evaluate the effect of indels. Specifically, the bit score from HMMER3 [8] measures the similarity of a query sequence with the set of homologous sequences used to define the profile HMM. The HMMER3 bit score is a base 2 logarithm of a ratio of probabilities (homology hypothesis over the null hypothesis),

$$B = \log_2 \frac{P(o_1 o_2 \ldots o_n|\text{HMM})}{P(o_1 o_2 \ldots o_n|\text{NULL})}, \tag{8.1}$$

where $o_1 o_2 \ldots o_n$ is the observed protein sequence and "HMM" is the trained profile HMM. "NULL" is the "null model", which is a one-state HMM configured to generate "random" sequences of the same length as the target sequence, with each residue drawn from a background frequency distribution (in HMMER3, for proteins, the frequencies of the 20 amino acids are set to the amino acid composition of SWISS-PROT 34). Since this logarithm score has no direct statistical interpretation, the constituent probabilities are extracted and used to define the odds ratio of the HMM probabilities,

$$S = \frac{P_w/(1 - P_w)}{P_m/(1 - P_m)}, \tag{8.2}$$

where $P_w(P_m)$ is the probability that the wild type (mutated type) protein sequence could have been generated by the profile HMM trained on a seed protein homologous sequence set (i.e., the numerator in B). The greater S is, the more likely that the mutation is deleterious. Usually S is expected to be greater than 1 as most mutations tend to be deleterious. When S is less than 1, it suggests that the mutant sequence better fits the HMM profile and the mutation may lead to amino acids that are more compatible than the wild type proteins. Experiments were done to set a threshold for the odds ratio, S_t, below which, the indel is considered as neutral, otherwise deleterious.

To demonstrate the effectiveness of HMMvar in predicting the effect of indels, indel variant data was obtained from dbSNP, and the indel effects scored using HMMvar. There are three types of indel variants in dbSNP, nonsense, missense, and frameshift indels. Missense indels refer to the indels that add or remove amino acids to or from the original protein sequence. Nonsense indels refer to indels that cause a stop codon where the indel occurs. Frameshift indels refer to indels that are not a multiple of three base pairs, thus change the reading frame of the original protein. Note, these categories of indels are mutually exclusive, that is, an indel can be either

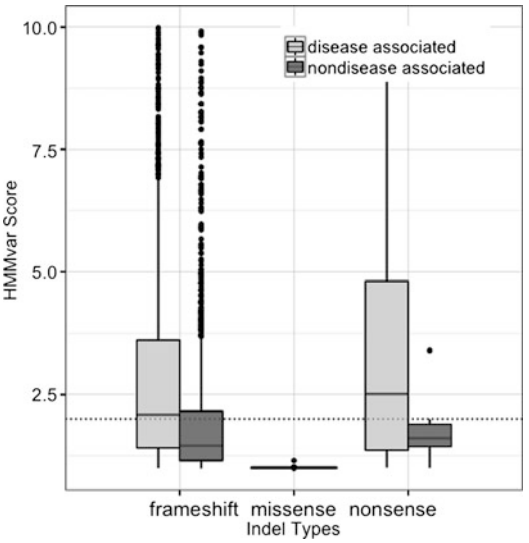

Fig. 8.2 Distributions of HMM scores for different types of indel variants. The *dotted line* shows the HMM score cutoff ($S_t = 2.0$) for determining whether an indel is deleterious or not

frameshift or in-frame, and if in-frame, it can be either missense or nonsense, but not both. The data contains altogether 4089 indels, among which 127, 56, 3906 are nonsense, missense, and frameshift indels, respectively (Table 8.1). These indels are further classified into two groups, indels that have locus-specific mutation database (LSDB) [14] annotation, which are expected to be disease associated and have more harmful effects, and indels that do not have LSDB annotation, which are expected to be nondisease (or unknown) associated and have less harmful effects (Table 8.1). The indels were fed into HMMvar and the odds ratio of the HMM probabilities were computed for each class of indels. Figure 8.2 shows the distributions of the HMM scores for three types of indel variants, frameshift indels, nonsense indels, and missense indels. The most remarkable feature is that the score of missense indels is much lower than the scores of the other two types, consistent with the notion that missense mutations tend to be less deleterious than frameshift indels and nonsense mutations. In each type of indel, the median of the nondisease associated group is lower than the median of the disease associated group, demonstrating that the HMM score is effective in evaluating the deleteriousness of indel mutations. Further comparison shows that the HMM odds ratio score has comparable performance to PROVEAN, with the added advantage of having smaller variance in the predicted scores, a desirable property for a scoring metric.

8.4 Future Directions

As an increasing amount of sequence data shows the prevalence and dominance of indels as the second most common type of mutation in human populations, much effort is required in order to fully understand indel effect on human biology and health.

Future research needs to focus on designing new, or improving existing, algorithms for predicting indel effects, by a combination of methods such as evolutionary-based approaches and sophisticated machine learning algorithms. Integration with diverse data and analysis results promises to provide a complete picture of indel effects on various aspects such as protein function, gene splicing, and gene expression.

References

1. 1000 Genomes Project Consortium: A map of human genome variation from population-scale sequencing. Nature **467**(7319), 1061–1073 (2010)
2. Chen, F.-C., Chen, C.-J., Li, W.-H., Chuang, T.-J.: Human-specific insertions and deletions inferred from mammalian genome sequences. Genome Res. **17**(1), 16–22 (2007)
3. Chen, C.-H., Chuang, T.-J., Liao, B.-Y., Chen, F.-C.: Scanning for the signatures of positive selection for human-specific insertions and deletions. Genome Biol. Evol. **1**, 415–419 (2009)
4. Choi, Y., Sims, G.E., Murphy, S., Miller, J.R., Chan, A.P.: Predicting the functional effect of amino acid substitutions and indels. PLoS ONE **7**(10), e46688 (2012)
5. Collins, F.S., Drumm, M.L., Cole, J.L., Lockwood, W.K., Vande Woude, G.F., Iannuzzi, M.C.: Construction of a general human chromosome jumping library, with application to cystic fibrosis. Science **235**(4792), 1046–1049 (1987)
6. Cooper, G.M., Shendure, J.: Needles in stacks of needles: finding disease-causal variants in a wealth of genomic data. Nat. Rev. Genet. **12**(9), 628–640 (2011)
7. De, S., Madan Babu, M.: A time-invariant principle of genome evolution. Proc. Natl. Acad. Sci. USA **107**(29), 13004–13009 (2010)
8. Finn, R.D., Clements, J., Eddy, S.R.: HMMER web server: interactive sequence similarity searching. Nucleic Acids Res. **39**(Web Server issue), W29–W37 (2011)
9. Gupta, R., Ratan, A., Rajesh, C., Chen, R., Lim Kim, H., Burhans, R., Miller, W., Santhosh, S., Davuluri, R.V., Butte, A.J., Schuster, S.C., Seshagiri, S., Thomas, G.: Sequencing and analysis of a South Asian–Indian personal genome. BMC Genomics **13**, 440 (2012)
10. Hu, J., Ng, P.C.: Predicting the effects of frameshifting indels. Genome Biol. **13**(2), R9 (2012)
11. International HapMap Consortium: The international HapMap project. Nature **426**(6968), 789–796 (2003)
12. International HapMap Consortium: A haplotype map of the human genome. Nature **437**(7063), 1299–1320 (2005)
13. Karchin, R.: Next generation tools for the annotation of human SNPs. Brief. Bioinform. **10**(1), 35–52 (2009)
14. Kato, S., Han, S.-Y., Liu, W., Otsuka, K., Shibata, H., Kanamaru, R., Ishioka, C.: Understanding the function-structure and function-mutation relationships of p53 tumor suppressor protein by high-resolution missense mutation analysis. Proc. Natl. Acad. Sci. USA **100**(14), 8424–8429 (2003)
15. Liu, M., Watson, Layne.T., Zhang, L.: HMMvar: Predicting the functional effects of indels and SNPs based on HMM profiles. BMC Bioinform. (under review)
16. Mills, R.E., Luttig, C.T., Larkins, C.E., Beauchamp, A., Tsui, C., Stephen Pittard, W., Devine, S.E.: An initial map of insertion and deletion (INDEL) variation in the human genome. Genome Res. **16**(9), 1182–1190 (2006)
17. Mills, R.E., Stephen Pittard, W., Mullaney, J.M., Farooq, U., Creasy, T.H., Mahurkar, A.A., Kemeza, D.M., Strassler, D.S., Ponting, C.P., Webber, C., Devine, S.E.: Natural genetic variation caused by small insertions and deletions in the human genome. Genome Res. **21**(6), 830–839 (2011)
18. Mullaney, J.M., Mills, R.E., Pittard, W.S., Devine, S.E.: Small insertions and deletions (INDELs) in human genomes. Hum. Mol. Genet. **19**, R131–R136 (2010)

19. Siva, N.: 1000 Genomes Project. Nat. Biotechnol. **26**(3), 256 (2008)
20. Wetterbom, A., Sevov, M., Cavelier, L., Bergstrom, T.F.: Comparative genomic analysis of human and chimpanzee indicates a key role for indels in primate evolution. J. Mol. Evol. **63**(5), 682–690 (2006)
21. Zia, A., Moses, A.M.: Ranking insertion, deletion and nonsense mutations based on their effect on genetic information. BMC Bioinform. **12**, 299 (2011)