ORIGINAL ARTICLE Open Access

ISSN 2278-1404

International Journal of Fundamental & Applied Sciences

Predicting Pathogenic Missense Mutations in the Human *c-MET* Oncogene Using a Nucleotide Scoring Function - An approach using mammalian

Padmini Arunkumar[1], Kumar Sankaran[2], Shivangi Naik[2] and Prashantha Karunakar[3*]

[1]Department of Biotechnology, PES Institute of Technology, Bengaluru- 560085

[2]Leucine Rich Bio Pvt. Ltd. 283 M. K. Puttalingaih Road, Padmanabhanagar, Bengaluru - 560070

[3]Department of Biotechnology, PES University, Bengaluru - 560085

Manuscript: received 16 October, 2018 revised 06 Dec, 2018 accepted 09 Dec, 2018

Abstract

Background: Many nucleotide variations in the human genome remain uncharacterized even more than ten years after the first draft was published. **Objective**: A three-parameter nucleotide-based scoring function was designed to predict the possible pathogenicity of 163 uncharacterized but validated missense mutations of the c-MET oncogene. **Methodology:** The parameters used by the scoring function were: surrounding consensus regions in a multiple sequence alignment of the human c-*MET* oncogene and five orthologous variants, inter-species allele frequency of each human missense mutation, and the nature of the mutation. **Results:** Out of 163 variants of unknown significance, 99 and 52 mutations were characterized as likely pathogenic and pathogenic respectively. An analysis of nine variants of known significance revealed that this scoring function could classify six out of nine (66.67%) variants with a reasonable accuracy.

***Keywords*:** Cancer pathogenicity; scoring function; MET; *c-MET*; oncogene; pathogenic; pathogenicity prediction

1. Introduction

A. Single Nucleotide Polymorphisms and Missense Mutations

Single nucleotide polymorphisms (SNPs) are variations in the genetic code of an organism, which occur at any single position in the genome and do not result in a frame-shift across the entire genomic sequence [1]. In the human genome, which is approximately 3 billion base pairs long, they occur at an average rate of 1 per 1,000 base pairs [1] [2].

Missense mutations are non-synonymous point mutations that result in the substitution of a different amino acid in the corresponding protein. A very high degree of conservation has been observed in functionally significant regions of oncogenes [3]. Such regions normally contain low numbers of mutations [4]. When occurring in oncogenes, missense mutations have shown strong associations with many types of cancers [5].

B. The Problem of Pathogenicity Assessment

Out of 179 missense variants of the human *c-MET* oncogene deposited in the NCBI database of Single Nucleotide Polymorphisms (dbSNP) (build 142), the clinical significance is known (i.e. listed as benign or pathogenic) for only nine variants [1]. The remaining variants' clinical significances are listed as 'other', 'unknown' or 'uncertain' [1], and there is a need to know the effect of each variant on the development and/or continuance of cancer. The same problem exists for other oncogenes with multiple genetic variations.

Missense mutations are of particular interest for three reasons: firstly, silent mutations are not expected to have an observable effect at the protein level; secondly, since nonsense nucleotide variants result in the truncation of proteins, there is a likelihood of negative selection pressure at work against them [4]; and thirdly, the *c-MET* oncogene contains only eight nonsense mutations [1].

Therefore, to predict the degree of pathogenicity of these missense variants of unknown significance, a novel, three -parameter nucleotide-based scoring function was designed. The scoring function presented in this research paper uses the property of evolutionary conservation of the gene (and therefore the protein product) across mammalian species. The parameters used were as follows: i) Surrounding consensus regions in an alignment of mammalian *MET* nucleotide reference sequences, ii) inter-species allele frequency in the same

Corresponding author
Full Address :
Department of Biotechnology, PES University, Bengaluru - 560085
E-mail: *prashanthakarunakar@pes.edu*

sequence alignment, and iii) nature of the human missense mutation (transition or transversion)

C. The c-MET Oncogene

Located on the 7th chromosome in humans, the *c-MET* oncogene (alternatively known as the *MET* oncogene) has two transcript variants of length 6,695 and 6,641 base pairs, which code for two isoforms of the hepatocyte growth factor receptor (HGFR). These proteins - HGFR isoforms *a* and *b* - are 1,408 and 1,390 amino acids long, respectively [6][7].

The HGFR, a tyrosine kinase receptor, is an active component of many cell signalling processes. The HGFR binds to the hepatocyte growth factor (HGF) ligand as part of its normal functioning. This binding activity is a necessary component of ordinary tissue and organ growth, development and regeneration throughout the human lifespan [8][9]. It is also involved in the cellular regeneration cascade pathway of the liver [8].

The c-*MET* proto-oncogene's pathogenicity is seen in three ways: an over-expression of HGFR in cancer cells, the occurrence of mutated HGF receptors due to non-synonymous mutations, and fluctuations in kinase activity [8]. *MET* missense mutations have been found to play a role in papillary renal cell carcinoma [10][11] lung cancer [12] and hepatocellular carcinoma [13].

2. Materials and methods

A. SNP Dataset Retrieval and Filtration

All human missense mutations of the *MET* proto-oncogene were retrieved from dbSNP in the XML file format. Build 142 was in effect at the time of retrieval. Redundant SNPs were filtered out using a Perl script. The filtered dataset of 179 missense variants was stored as a tab-delimited, single line separated text file.

B. Gene Sequence Retrieval and Alignment

Six validated mammalian RefSeq mRNA sequences of the c-*MET* oncogene were downloaded from GenBank [6]. Four primate (*Homo sapiens,* transcript variant 2; *Pan troglodytes*; *Macaca mulatta*; *Pongo abelii*) and two non-primate mRNA sequences (*Mus musculus* and *Rattus norvegicus*) were downloaded and saved in the FASTA format. A combined multi-FASTA file was created.

The nucleotide sequences were aligned using the online version of EBI's ClustalW2 multiple sequence alignment program [16]. The multi-FASTA file created in the previous step was uploaded to the ClustalW2 interface. All parameters were left intact from the defaults, except for the 'DNA Weight Matrix' parameter which was changed to 'ClustalW.' The alignment program was allowed to run. The generated alignment file was downloaded and saved upon completion.

The output file of ClustalW2 (*.clustalw) was opened in BioEdit [17], a stand-alone offline tool to work with nucleotide or protein sequence alignments. The true sequence positions were exported as a tab-delimited text file.

C. Scoring Function Design and Implementation

i) Design

a) Surrounding consensus regions

The number of positions in the alignment with '*' (indicating complete consensus) before and after each mutation were considered. A 9-position cut-off was decided based on the fact that 9 nucleotides code for 3 amino acids. The minimum and maximum number of total consensus positions which can occur (with the present cutoff value of nine on either side) are 0 and 18. The raw score generated for each mutation was scaled to a score between 0 and 1.

b) Inter-species allele frequency in the alignment

At each position, a minimum of two and a maximum of six nucleotides can be the same (since the alignment contains six sequences and only four nucleotides exist). To obtain a non-zero normalized score, the minimum and maximum were taken as 0 and 9. The highest possibility was given a higher weight since the similarity is seen across both primates and non-primates. Other possibilities were given a weight of +1. The scores were scaled such that the normalized values were between 0 and 1.

c) Nature of the missense mutation

Taking the 2:1 transition/transversion evolutionary bias across the whole genome [18][19] and the observations of the greater numbers of pathogenic transitions of the *MET* gene into consideration, transitions were given a lower score of 0.25 and transversions were given a higher score of 0.75. Post-scaling, the normalized scores occurred at two extremes for transitions and transversions: 0 and 1 respectively.

d) Final score calculation and classification

The normalized scores of all three parameters were summed up. The raw score obtained for each mutation was scaled between 0 and 1 by setting a minimum value of 0 and a maximum value of 3. In the final score, 0 indicates benignity while 1 indicates pathogenicity. The pathogenicity classification presented below in TABLE I. follows the American College of Medical Genetics (ACMG) guidelines [20].

Table I: Pathogenicity Classification Employed

Score	Classification [a]
< =0.25	Benign
0.25-<0.5	Likely benign
=0.5	Uncertain
0.5-0.75	Likely pathogenic
>=0.75	Pathogenic

[a] According to the ACMG Guidelines [20]

ii) Implementation

Perl scripts were implemented for each of the above steps. Detailed algorithms, flowcharts and code can be found in the supplementary materials subsections S-I through S-VIII.

3. Results and discussion

i) Predictions of the Scoring System

TABLE II. indicates the number of variants classified by this scoring function into each category. These numbers are inclusive of variants with known clinical significance.

Table II: Pathogenicity Classification Employed

Category	No. of variants
Benign	0
Likely benign	9
Uncertain	3
Likely pathogenic	111
Pathogenic	56
Total	179

ii) Comparison with dbSNP Variants of Known Clinical Significance

More data regarding TABLE III. can be found in the supplementary materials (Supplementary Table IV).

TABLE III. Comparison with dbSNP

#	rsID	Residue	Final Score	Predic-tion	dbSNP Signifi-cance
1	rs33917957	N375S	0.6481	Likely patho-genic	Benign
2	rs121913243	H1094R	0.6481	Likely patho-genic	Patho-genic
3	rs121913668	M1131T	0.6481	Likely patho-genic	Patho-genic
4	rs121913670	V1220I	0.6481	Likely patho-genic	Patho-genic
5	rs121913671	D1228N	0.6481	Likely patho-genic	Patho-genic
6	rs121913246	Y1230C	0.6296	Likely patho-genic	Patho-genic
7	rs121913669	V1188L	0.9629	Patho-genic	Patho-genic
8	rs77523018	M362T	0.5925	Likely patho-genic	Benign
9	rs35284565	R218S	0.9259	Patho-genic	Benign

Out of six variants classified as pathogenic by dbSNP, only one was classified as pathogenic by the current scoring function. SNP cluster rs121913669 (missense mutation G3749T; V1188L) had a 6/6 inter-species allele frequency, i.e. it was conserved across both primates and non-primates. Six out of nine (66.67%) variants roughly matched in their predictions and known significance.

The other five dbSNP pathogenic variants were classified as 'likely pathogenic' by the algorithm, with scores ranging from 0.6296 to 0.64815. Four out of these five variants (rs121913243, rs121913668, rs121913670 and rs121913671) had a score of 0.64815 while rs121913246 (missense mutation A3876G; Y1230C) had a score of 0.6296. This suggests that while the basis of the scoring system is correct, more nucleotide-based parameters are necessary to refine it further.

4. Conclusion

Determining the pathogenicity of nucleotide variants remains a challenge due to their sheer numbers - 1 per 1000 base pairs in a genome which 3.2 billion base pairs in length [2], in humans alone. Existing predictive scoring functions such as Sorting Intolerant From Tolerant (SIFT) [21], PolyPhen [22] and PROtein Variation Effect ANalyzer (PROVEAN) [23] approach this problem from the perspective of change in amino acids.

The scoring function presented in this paper has attempted to approach the pathogenicity problem from an evolutionary conservation perspective. By aligning orthologous mammalian variants of the same oncogene (c-*MET*), and using human missense SNPs of the same gene sourced from dbSNP, a three-step scoring function was designed.

The scoring scheme can be further refined by using more mammalian gene sequences as and when they are validated, and employing more refined nucleotide-based parameters. Its efficiency can be verified across the whole genome, but a restriction to primate sequences would be advisable.

The ten SNPs which have generated the highest scores can be clinically tested for disease associations in specific populations (in this case, cancer patients). This can be accomplished by genome-wide association studies (GWAS).

Acknowledgment

This paper was written to publicise the findings of Ms. Padmini Arunkumar's M. Tech Bioinformatics dissertation. The protocol was designed by Ms. Padmini Arunkumar, Mr. Kumar Sankaran and Dr. Prashantha Karunakar. Ms. Shivangi Naik was involved in the programmatic implementation of the scoring function.

Conflict of interest

The author's declares none.

References

[1] S. T. Sherry, M. H. Ward, M. Kholodov, J. Baker, L. Phan, E. M. Smigielski, and K. Sirotkin, "dbSNP: the NCBI database of genetic variation," *Nucleic Acids Res.*, vol. 29, no. 1, pp. 308 –311, 2001.

[2] R. Sachidanandam, D. Weissman, S. C. Schmidt, J. M. Kakol, L. D. Stein, G. Marth, S. Sherry, J. C. Mullikin, B. J. Mortimore, D. L. Willey, S. E. Hunt, C. G. Cole, P. C. Coggill, C. M. Rice, Z. Ning, J. Rogers, D. R. Bentley, P. Y. Kwok, E. R. Mardis, R. T. Yeh, B. Schultz, L. Cook, R. Davenport, M. Dante, L. Fulton, L. Hillier, R. H. Waterston, J. D. McPherson, B. Gilman, S. Schaffner, W. J. Van Etten, D. Reich, J. Higgins, M. J. Daly, B. Blumenstiel, J. Baldwin, N. Stange-Thomann, M. C. Zody, L. Linton, E. S. Lander, and D. Altshuler, "A map

of human genome sequence variation containing 1.42 million single nucleotide polymorphisms," *Nature*, vol. 409, no. 6822, pp. 928–933, 2001.

[3] M. P. Hare and S. R. Palumbi, "High intron sequence conservation across three mammalian orders suggests functional constraints," *Mol. Biol. Evol.*, vol. 20, no. 6, pp. 969–978, 2003.

[4] L. Cartegni, S. L. Chew, and A. R. Krainer, "Listening to silence and understanding nonsense: exonic mutations that affect splicing," *Nat. Rev. Genet.*, vol. 3, no. 4, pp. 285–298, 2002.

[5] M. S. Greenblatt, W. P. Bennett, M. Hollstein, and C. C. Harris, "Mutations in the p53 tumor suppressor gene: Clues to cancer etiology and molecular pathogenesis," *Cancer Research*, vol. 54, no. 18. pp. 4855–4878, 1994.

[6] D. A. Benson, I. Karsch-Mizrachi, D. J. Lipman, J. Ostell, and D. L. Wheeler, "GenBank.," *Nucleic Acids Res.*, vol. 34, no. Database issue, pp. D16–D20, 2006.

[7] M. Zhu, R. Tang, S. Doshi, K. S. Oliner, S. Dubey, Y. Jiang, R. C. Donehower, T. Iveson, E. Y. Loh, and Y. Zhang, "Exposure-response analysis of rilotumumab in gastric cancer: the role of tumour MET expression," *Br J Cancer*, vol. 112, no. 3, pp. 429–437, Feb. 2015.

[8] R. Salgia, "Role of c-Met in Cancer: Emphasis on Lung Cancer," *Semin. Oncol.*, vol. 36, no. SUPPL. 1, 2009.

[9] C. M. Ho-Yen, J. L. Jones, and S. Kermorgant, "The clinical and functional significance of c-Met in breast cancer: a review.," *Breast Cancer Res.*, vol. 17, no. 1, p. 52, Jan. 2015.

[10] L. Schmidt, F. M. Duh, F. Chen, T. Kishida, G. Glenn, P. Choyke, S. W. Scherer, Z. Zhuang, I. Lubensky, M. Dean, R. Allikmets, A. Chidambaram, U. R. Bergerheim, J. T. Feltis, C. Casadevall, A. Zamarron, M. Bernues, S. Richard, C. J. Lips, M. M. Walther, L. C. Tsui, L. Geil, M. L. Orcutt, T. Stackhouse, J. Lipan, L. Slife, H. Brauch, J. Decker, G. Niehans, M. D. Hughson, H. Moch, S. Storkel, M. I. Lerman, W. M. Linehan, and B. Zbar, "Germline and somatic mutations in the tyrosine kinase domain of the MET proto-oncogene in papillary renal carcinomas," *Nat. Genet.*, vol. 16, no. 1, pp. 68–73, May 1997.

[11] L. Schmidt, K. Junker, N. Nakaigawa, T. Kinjerski, G. Weirich, M. Miller, I. Lubensky, H. P. Neumann, H. Brauch, J. Decker, C. Vocke, J. A. Brown, R. Jenkins, S. Richard, U. Bergerheim, B. Gerrard, M. Dean, W. M. Linehan, and B. Zbar, "Novel mutations of the MET proto-oncogene in papillary renal carcinomas," *Oncogene*, vol. 18, no. 14, pp. 2343–2350, Apr. 1999.

[12] P. C. Ma, R. Jagadeeswaran, S. Jagadeesh, M. S. Tretiakova, V. Nallasura, E. A. Fox, M. Hansen, E. Schaefer, K. Naoki, A. Lader, W. Richards, D. Sugarbaker, A. N. Husain, J. G. Christensen, and R. Salgia, "Functional expression and mutations of c-Met and its therapeutic inhibition with SU11274 and small interfering RNA in non-small cell lung cancer," *Cancer Res.*, vol. 65, no. 4, pp. 1479–1488, Feb. 2005.

[13] W. S. Park, S. M. Dong, S. Y. Kim, E. Y. Na, M. S. Shin, J. H. Pi, B. J. Kim, J. H. Bae, Y. K. Hong, K. S. Lee, S. H. Lee, N. J. Yoo, J. J. Jang, S. Pack, Z. Zhuang, L. Schmidt, B. Zbar, and J. Y. Lee, "Somatic mutations in the kinase domain of the Met/ hepatocyte growth factor receptor gene in childhood hepatocellular carcinomas," *Cancer Res.*, vol. 59, no. 2, pp. 307–310, 1999.

[14] C. Greenman, P. Stephens, R. Smith, G. L. Dalgliesh, C. Hunter, G. Bignell, H. Davies, J. Teague, A. Butler, C. Stevens, S. Edkins, S. O'Meara, I. Vastrik, E. E. Schmidt, T. Avis, S. Barthorpe, G. Bhamra, G. Buck, B. Choudhury, J. Clements, J. Cole, E. Dicks, S. Forbes, K. Gray, K. Halliday, R. Harrison, K. Hills, J. Hinton, A. Jenkinson, D. Jones, A. Menzies, T. Mironenko, J. Perry, K. Raine, D. Richardson, R. Shepherd, A. Small, C. Tofts, J. Varian, T. Webb, S. West, S. Widaa, A. Yates, D. P. Cahill, D. N. Louis, P. Goldstraw, A. G. Nicholson, F. Brasseur, L. Looijenga, B. L. Weber, Y.-E. Chiew, A. DeFazio, M. F. Greaves, A. R. Green, P. Campbell, E. Birney, D. F. Easton, G. Chenevix-Trench, M.-H. Tan, S. K. Khoo, B. T. Teh, S. T. Yuen, S. Y. Leung, R. Wooster, P. A. Futreal, and M. R. Stratton, "Patterns of somatic mutation in human cancer genomes.," *Nature*, vol. 446, no. 7132, pp. 153–158, Mar. 2007.

[15] G. M. Stella, S. Benvenuti, D. Gramaglia, A. Scarpa, A. Tomezzoli, P. Cassoni, R. Senetta, T. Venesio, E. Pozzi, A. Bardelli, and P. M. Comoglio, "MET mutations in cancers of unknown primary origin (CUPs).," *Hum. Mutat.*, vol. 32, no. 1, pp. 44–50, Jan. 2011.

[16] M. A. Larkin, G. Blackshields, N. P. Brown, R. Chenna, P. A. Mcgettigan, H. McWilliam, F. Valentin, I. M. Wallace, a. Wilm, R. Lopez, J. D. Thompson, T. J. Gibson, and D. G. Higgins, "Clustal W and Clustal X version 2.0," *Bioinformatics*, vol. 23, no. 21, pp. 2947–2948, 2007.

[17] T. Hall, "BioEdit: a user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT," *Nucleic Acids Symposium Series*, vol. 41. pp. 95–98, 1999.

[18] A. Siepel, G. Bejerano, J. S. Pedersen, A. S. Hinrichs, M. Hou, K. Rosenbloom, H. Clawson, J. Spieth, L. W. Hillier, S. Richards, G. M. Weinstock, R. K. Wilson, R. A. Gibbs, W. J. Kent, W. Miller, and D. Haussler, "Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes.," *Genome Res.*, vol. 15, no. 8, pp. 1034–1050, Aug. 2005.

[19] Z. Zhang and M. Gerstein, "Patterns of nucleotide substitution, insertion and deletion in the human genome inferred from pseudogenes.," *Nucleic Acids Res.*, vol. 31, no. 18, pp. 5338–5348, Sep. 2003.

[20] S. Richards, N. Aziz, S. Bale, D. Bick, S. Das, J. Gastier-Foster, W. W. Grody, M. Hegde, E. Lyon, E. Spector, K. Voelkerding, and H. L. Rehm, "Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology.," *Genet. Med.*, vol. 17, no. 5, pp. 405–423, May 2015.

[21] P. C. Ng and S. Henikoff, "SIFT: predicting amino acid changes that affect protein function," *Nucleic Acids Res.* , vol. 31 , no. 13 , pp. 3812–3814, Jul. 2003.

[22] Y. Choi, G. E. Sims, S. Murphy, J. R. Miller, and A. P. Chan, "Predicting the Functional Effect of Amino Acid Substitutions and Indels," *PLoS One*, vol. 7, no. 10, p. e46688, Oct. 2012.

[23] I. Adzhubei, D. M. Jordan, and S. R. Sunyaev, "Predicting functional effect of human missense mutations using PolyPhen-2.," *Curr. Protoc. Hum. Genet.*, vol. Chapter 7, p. Unit7.20, Jan. 2013.