Masso, Majid; **Vaisman, Iosif I.**

Knowledge-based computational mutagenesis for predicting the disease potential of human non-synonymous single nucleotide polymorphisms. (English) Zbl 1407.92082

J. Theor. Biol. 266, No. 4, 560-568 (2010).

Summary: Certain genetic variations in the human population are associated with heritable diseases, and single nucleotide polymorphisms (SNPs) represent the most common form of such differences in DNA sequence. In particular, substantial interest exists in determining whether a non-synonymous SNP (nsSNP), leading to a single residue replacement in the translated protein product, is neutral or disease-related. The nature of protein structure-function relationships suggests that nsSNP effects, either benign or leading to aberrant protein function possibly associated with disease, are dependent on relative structural changes introduced upon mutation. In this study, we characterize a representative sampling of 1790 documented neutral and disease-related human nsSNPs mapped to 243 diverse human protein structures, by quantifying environmental perturbations in the associated proteins with the use of a computational mutagenesis methodology that relies on a four-body, knowledge-based, statistical contact potential. These structural change data are used as attributes to generate a vector representation for each nsSNP, in combination with additional features reflecting sequence and structure of the corresponding protein. A trained model based on the random forest supervised classification algorithm achieves 76% cross-validation accuracy. Our classifier performs at least as well as other methods that use significantly larger datasets of nsSNPs for model training, and the novelty of our attributes differentiates the model as an orthogonal approach that can be utilized in conjunction with other techniques. A dedicated server for obtaining predictions, as well as supporting datasets and documentation, is available at http://proteins.gmu.edu/automute.

MSC:

92D10 Genetics and epigenetics
92D20 Protein sequences, DNA sequences
68T05 Learning and adaptive systems in artificial intelligence
62P10 Applications of statistics to biology and medical sciences; meta analysis

Keywords:

protein structure; computational geometry; statistical potential; variant; machine learning

Software:

AUTO-MUTE; Cell-PLoc; GPCR-GIA; nsSNPAnalyzer; PMut; Qhull; Quat-2L; SNPs3D; UniProt

Full Text: DOI

References:

[1] Apweiler, R.; Bairoch, A.; Wu, C.H.; Barker, W.C.; Boeckmann, B.; Ferro, S.; Gasteiger, E.; Huang, H.; Lopez, R.; Magrane, M.; Martin, M.J.; Natale, D.A.; O'Donovan, C.; Redaschi, N.; Yeh, L.S., Uniprot: the universal protein knowledgebase, Nucleic acids res., 32, D115-D119, (2004)

[2] Arbiza, L.; Duchi, S.; Montaner, D.; Burguet, J.; Pantoja-Uceda, D.; Pineda-Lucena, A.; Dopazo, J.; Dopazo, H., Selective pressures at a codon-level predict deleterious mutations in human disease genes, J. mol. biol., 358, 1390-1404, (2006)

[3] Baldi, P.; Brunak, S.; Chauvin, Y.; Andersen, C.A.; Nielsen, H., Assessing the accuracy of prediction algorithms for classification: an overview, Bioinformatics, 16, 412-424, (2000)

[4] Bao, L.; Cui, Y., Prediction of the phenotypic effects of non-synonymous single nucleotide polymorphisms using structural and evolutionary information, Bioinformatics, 21, 2185-2190, (2005)

[5] Bao, L.; Zhou, M.; Cui, Y., Nssnpanalyzer: identifying disease-associated nonsynonymous single nucleotide polymorphisms, Nucleic acids res., 33, W480-W482, (2005)

[6] Barber, C.B.; Dobkin, D.P.; Huhdanpaa, H.T., The quickhull algorithm for convex hulls, ACM trans. math. software, 22, 469-483, (1996) · Zbl 0884.65145

[7] Barenboim, M.; Masso, M.; Vaisman, I.I.; Jamison, D.C., Statistical geometry based prediction of nonsynonymous SNP

functional effects using random forest and neuro-fuzzy classifiers, Proteins, 71, 1930-1939, (2008)

[8] Berman, H.M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T.N.; Weissig, H.; Shindyalov, I.N.; Bourne, P.E., The protein data bank, Nucleic acids res., 28, 235-242, (2000)

[9] Bowie, J.U.; Luthy, R.; Eisenberg, D., A method to identify protein sequences that fold into a known three-dimensional structure, Science, 253, 164-170, (1991)

[10] Breiman, L., Random forests, Mach. learn., 45, 5-32, (2001) · Zbl 1007.68152

[11] Bromberg, Y.; Rost, B., SNAP: predict effect of non-synonymous polymorphisms on function, Nucleic acids res., 35, 3823-3835, (2007)

[12] Capriotti, E.; Calabrese, R.; Casadio, R., Predicting the insurgence of human genetic diseases associated to single point protein mutations with support vector machines and evolutionary information, Bioinformatics, 22, 2729-2734, (2006)

[13] Chen, C.; Chen, L.; Zou, X.; Cai, P., Prediction of protein secondary structure content by using the concept of Chou's pseudo amino acid composition and support vector machine, Protein pept. lett., 16, 27-31, (2009)

[14] Chou, K.C.; Zhang, C.T., Prediction of protein structural classes, Crit. rev. biochem. mol. biol., 30, 275-349, (1995)

[15] Chou, K.C.; Shen, H.B., Recent progress in protein subcellular location prediction, Anal. biochem., 370, 1-16, (2007)

[16] Chou, K.C.; Shen, H.B., Cell-ploc: a package of web servers for predicting subcellular localization of proteins in various organisms, Nat. protoc., 3, 153-162, (2008)

[17] Collins, F.S.; Brooks, L.D.; Chakravarti, A., A DNA polymorphism discovery resource for research on human genetic variation, Genome res., 8, 1229-1231, (1998)

[18] Conde, L.; Vaquerizas, J.M.; Dopazo, H.; Arbiza, L.; Reumers, J.; Rousseau, F.; Schymkowitz, J.; Dopazo, J., Pupasuite: finding functional single nucleotide polymorphisms for large-scale genotyping purposes, Nucleic acids res., 34, W621-W625, (2006)

[19] Dayhoff, M.O.; Schwartz, R.M.; Orcut, B.C., A model for evolutionary change in proteins, (), 345-352

[20] De Gobbi, M.; Viprakasit, V.; Hughes, J.R.; Fisher, C.; Buckle, V.J.; Ayyub, H.; Gibbons, R.J.; Vernimmen, D.; Yoshinaga, Y.; de Jong, P.; Cheng, J.F.; Rubin, E.M.; Wood, W.G.; Bowden, D.; Higgs, D.R., A regulatory SNP causes a human genetic disease by creating a new transcriptional promoter, Science, 312, 1215-1217, (2006)

[21] Deutsch, C.; Krishnamoorthy, B., Four-body scoring function for mutagenesis, Bioinformatics, 23, 3009-3015, (2007)

[22] Ding, H.; Luo, L.; Lin, H., Prediction of cell wall lytic enzymes using Chou's amphiphilic pseudo amino acid composition, Protein pept. lett., 16, 351-355, (2009)

[23] Esmaeili, M.; Mohabatkar, H.; Mohsenzadeh, S., Using the concept of Chou's pseudo amino acid composition for risk type prediction of human papillomaviruses, J. theor. biol., 263, 203-209, (2010) · Zbl 1406.92455

[24] Feng, Y.; Kloczkowski, A.; Jernigan, R.L., Four-body contact potentials derived from two protein datasets to discriminate native structures from decoys, Proteins, 68, 57-66, (2007)

[25] Ferrer-Costa, C.; Orozco, M.; de la Cruz, X., Characterization of disease-associated single amino acid polymorphisms in terms of sequence and structure properties, J. mol. biol., 315, 771-786, (2002)

[26] Ferrer-Costa, C.; Gelpi, J.L.; Zamakola, L.; Parraga, I.; de la Cruz, X.; Orozco, M., PMUT: a web-based tool for the annotation of pathological mutations on proteins, Bioinformatics, 21, 3176-3178, (2005)

[27] Frank, E.; Hall, M.; Trigg, L.; Holmes, G.; Witten, I.H., Data mining in bioinformatics using weka, Bioinformatics, 20, 2479-2481, (2004)

[28] Frazer, K.A.; Ballinger, D.G.; Cox, D.R.; Hinds, D.A.; Stuve, L.L.; Gibbs, R.A.; Belmont, J.W.; Boudreau, A.; Hardenbol, P.; Leal, S.M.; Pasternak, S.; Wheeler, D.A.; Willis, T.D.; Yu, F.; Yang, H.; Zeng, C.; Gao, Y.; Hu, H.; Hu, W.; Li, C.; Lin, W.; Liu, S.; Pan, H.; Tang, X.; Wang, J.; Wang, W.; Yu, J.; Zhang, B.; Zhang, Q.; Zhao, H.; Zhou, J.; Gabriel, S.B.; Barry, R.; Blumenstiel, B.; Camargo, A.; Defelice, M.; Faggart, M.; Goyette, M.; Gupta, S.; Moore, J.; Nguyen, H.; Onofrio, R.C.; Parkin, M.; Roy, J.; Stahl, E.; Winchester, E.; Ziaugra, L.; Altshuler, D.; Shen, Y.; Yao, Z.; Huang, W.; Chu, X.; He, Y.; Jin, L.; Liu, Y.; Sun, W.; Wang, H.; Wang, Y.; Xiong, X.; Xu, L.; Waye, M.M.; Tsui, S.K.; Xue, H.; Wong, J.T.; Galver, L.M.; Fan, J.B.; Gunderson, K.; Murray, S.S.; Oliphant, A.R.; Chee, M.S.; Montpetit, A.; Chagnon, F.; Ferretti, V.; Leboeuf, M.; Olivier, J.F.; Phillips, M.S.; Roumy, S.; Sallee, C.; Verner, A.; Hudson, T.J.; Kwok, P.Y.; Cai, D.; Koboldt, D.C.; Miller, R.D.; Pawlikowska, L.; Taillon-Miller, P.; Xiao, M.; Tsui, L.C.; Mak, W.; Song, Y.Q.; Tam, P.K.; Nakamura, Y.; Kawaguchi, T.; Kitamoto, T.; Morizono, T.; Nagashima, A.; Ohnishi, Y.; Sekine, A.; Tanaka, T.; Tsunoda, T., A second generation human haplotype map of over 3.1 million SNPs, Nature, 449, 851-861, (2007)

[29] Fredman, D.; Siegfried, M.; Yuan, Y.P.; Bork, P.; Lehvaslaiho, H.; Brookes, A.J., Hgvbase: a human sequence variation database emphasizing data quality and a broad spectrum of data sources, Nucleic acids res., 30, 387-391, (2002)

[30] Hinds, D.A.; Stuve, L.L.; Nilsen, G.B.; Halperin, E.; Eskin, E.; Ballinger, D.G.; Frazer, K.A.; Cox, D.R., Whole-genome patterns of common DNA variation in three human populations, Science, 307, 1072-1079, (2005)

[31] Karchin, R.; Diekhans, M.; Kelly, L.; Thomas, D.J.; Pieper, U.; Eswar, N.; Haussler, D.; Sali, A., LS-SNP: large-scale annotation of coding non-synonymous SNPs based on multiple information sources, Bioinformatics, 21, 2814-2820, (2005)

[32] Kimchi-Sarfaty, C.; Oh, J.M.; Kim, I.W.; Sauna, Z.E.; Calcagno, A.M.; Ambudkar, S.V.; Gottesman, M.M., A "Silent" polymorphism in the MDR1 gene changes substrate specificity, Science, 315, 525-528, (2007)

[33] Lee, S.; Kasif, S.; Weng, Z.; Cantor, C.R., Quantitative analysis of single nucleotide polymorphisms within copy number variation, Plos one, 3, e3906, (2008)

[34] Lin, W.Z.; Xiao, X.; Chou, K.C., GPCR-GIA: a web-server for identifying G-protein coupled receptors and their families with grey incidence analysis, Protein eng. des. sel., 22, 699-705, (2009)

[35] Masso, M.; Vaisman, I.I., Accurate prediction of enzyme mutant activity based on a multibody statistical potential, Bioinformatics, 23, 3155-3161, (2007)

[36] Masso, M.; Vaisman, I.I., Accurate prediction of stability changes in protein mutants by combining machine learning with structure based computational mutagenesis, Bioinformatics, 24, 2002-2009, (2008)

[37] Masso, M.; Lu, Z.; Vaisman, I.I., Computational mutagenesis studies of protein structure – function correlations, Proteins, 64, 234-245, (2006)

[38] McCarroll, S.A., Extending genome-wide association studies to copy-number variation, Hum. mol. genet., 17, R135-R142, (2008)

[39] Ng, P.C.; Henikoff, S., Accounting for human polymorphisms predicted to affect protein function, Genome res., 12, 436-446, (2002)

[40] Oinonen, C.; Tikkanen, R.; Rouvinen, J.; Peltonen, L., Three-dimensional structure of human lysosomal aspartylglucosaminidase, Nat. struct. biol., 2, 1102-1108, (1995)

[41] Ramensky, V.; Bork, P.; Sunyaev, S., Human non-synonymous SNPs: server and survey, Nucleic acids res., 30, 3894-3900, (2002)

[42] Sachidanandam, R.; Weissman, D.; Schmidt, S.C.; Kakol, J.M.; Stein, L.D.; Marth, G.; Sherry, S.; Mullikin, J.C.; Mortimore, B.J.; Willey, D.L.; Hunt, S.E.; Cole, C.G.; Coggill, P.C.; Rice, C.M.; Ning, Z.; Rogers, J.; Bentley, D.R.; Kwok, P.Y.; Mardis, E.R.; Yeh, R.T.; Schultz, B.; Cook, L.; Davenport, R.; Dante, M.; Fulton, L.; Hillier, L.; Waterston, R.H.; McPherson, J.D.; Gilman, B.; Schaffner, S.; Van Etten, W.J.; Reich, D.; Higgins, J.; Daly, M.J.; Blumenstiel, B.; Baldwin, J.; Stange-Thomann, N.; Zody, M.C.; Linton, L.; Lander, E.S.; Altshuler, D., A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms, Nature, 409, 928-933, (2001)

[43] Shastry, B.S., SNPs in disease gene mapping, medicinal drug development and evolution, J. hum. genet., 52, 871-880, (2007)

[44] Sherry, S.T.; Ward, M.H.; Kholodov, M.; Baker, J.; Phan, L.; Smigielski, E.M.; Sirotkin, K., Dbsnp: the NCBI database of genetic variation, Nucleic acids res., 29, 308-311, (2001)

[45] Singh, R.K.; Tropsha, A.; Vaisman, I.I., Delaunay tessellation of proteins: four body nearest-neighbor propensities of amino acid residues, J. comput. biol., 3, 213-221, (1996)

[46] Stitziel, N.O.; Binkowski, T.A.; Tseng, Y.Y.; Kasif, S.; Liang, J., Toposnp: a topographic database of non-synonymous single nucleotide polymorphisms with and without known disease association, Nucleic acids res., 32, D520-D522, (2004)

[47] Stranger, B.E.; Forrest, M.S.; Dunning, M.; Ingle, C.E.; Beazley, C.; Thorne, N.; Redon, R.; Bird, C.P.; de Grassi, A.; Lee, C.; Tyler-Smith, C.; Carter, N.; Scherer, S.W.; Tavare, S.; Deloukas, P.; Hurles, M.E.; Dermitzakis, E.T., Relative impact of nucleotide and copy number variation on gene expression phenotypes, Science, 315, 848-853, (2007)

[48] Sunyaev, S.; Ramensky, V.; Bork, P., Towards a structural basis of human non-synonymous single nucleotide polymorphisms, Trends genet., 16, 198-200, (2000)

[49] Taylor, T.; Rivera, M.; Wilson, G.; Vaisman, I.I., New method for protein secondary structure assignment based on a simple topological descriptor, Proteins, 60, 513-524, (2005)

[50] Vaisman, I.I.; Tropsha, A.; Zheng, W., Compositional preferences in quadruplets of nearest neighbor residues in protein structures: statistical geometry analysis, Proc IEEE symp intel and syst, 163-168, (1998)

[51] Xiao, X., Wang, P., Chou, K.C., 2010. Quat-2L: a web-server for predicting protein quaternary structural attributes. Mol. Divers., doi: 10.1007/s11030-010-9227-8.

[52] Xie, L.; Bourne, P.E., A robust and efficient algorithm for the shape description of protein structures and its application in predicting ligand binding sites, BMC bioinform., 8, Suppl. 4, S9, (2007)

[53] Yip, Y.L.; Scheib, H.; Diemand, A.V.; Gattiker, A.; Famiglietti, L.M.; Gasteiger, E.; Bairoch, A., The swiss-prot variant page and the modsnp database: a resource for sequence and structure information on human protein variants, Hum. mutat., 23, 464-470, (2004)

[54] Yue, P.; Melamud, E.; Moult, J., SNPs3D: candidate gene and SNP selection for association studies, BMC bioinform., 7, 166, (2006)