

topoSNP: A Topographic Database of Non-Synonymous Single Nucleotide
Polymorphisms With and Without Known Disease Association

Nathan O. Stitzel¹, T. Andrew Binkowski¹, Yan Yuan Tseng¹, Simon Kasif², Jie Liang^{1*}

¹Department of Bioengineering
University of Illinois at Chicago
M/C 063, 851 S. Morgan St.
Chicago, IL 60607

and

²Department of Biomedical Engineering
Boston University
Boston, MA 02215

* To whom correspondence should be addressed: jliang@uic.edu, Ph: (312) 355-1789, Fax: (312) 996-5921

Abstract

The database of topographic mapping of Single Nucleotide Polymorphism (topoSNP) provides an online resource for analyzing non-synonymous SNPs (nsSNPs) that can be mapped onto known three-dimensional structures of proteins. These include disease associated nsSNPs derived from the Online Mendelian Inheritance in Man (OMIM) database and other nsSNPs derived from dbSNP, a resource at the National Center for Biotechnology Information that catalogs SNPs. TopoSNP further classifies each nsSNP site into three categories based on their geometric locations: those located in a surface pocket or an interior void of the protein, those on a convex region or a shallow depressed region, and those that are completely buried in the interior of the protein structure. Current release also includes relative entropy of SNPs calculated from multiple sequence alignment as obtained from the PFAM database (a database of protein families and conserved protein motifs) as well as manually adjusted multiple alignments obtained from CLUSTALW. These structural and conservation data can be useful for studying whether nsSNPs in coding regions are likely to lead to phenotypic changes. TopoSNP includes an interactive structural visualization web interface, as well as downloadable batch data. The database will be updated at regular intervals and can be accessed at: <http://gila.bioengr.uic.edu/snp/toposnp>.

Introduction

Non-synonymous single nucleotide polymorphisms (nsSNPs) have been implicated in numerous disease processes because they may alter protein function (1), alter splice sites (2), destabilize protein core structure, and reduce protein solubility (3). However, not all nsSNPs are associated with disease, and it is useful to explore general structural and conservational features of disease-associated nsSNPs versus non disease-associated nsSNPs. For example, solvent accessibility and other features such as experimental B-factors are found to be indicative of functional changes accompanying nsSNPs (4-6). Additionally, sequence conservation has been shown to be useful for predicting when a nsSNP is likely to have deleterious effects(7-9). In another recent study, it was found that disease-associated nsSNPs are more likely to be located in surface pockets or interior voids when compared to control nsSNPs (10). In addition, disease-associated nsSNPs buried in the protein interior are more likely to occur at conserved residue sites, whereas disease-associated nsSNPs located in surface pockets or interior voids do not have such propensity (10). Two data sets of nsSNPs were used in this study, one for disease-associated SNPs, which is derived from the Online Mendelian Inheritance in Man (OMIM) database (11), and one for non disease-associated or control nsSNPs, which is derived from dbSNP (dbSNP is a resource at the National Center for Biotechnology Information that catalogs SNPs as well as other genetic differences) (12). Although results obtained using this approach alone cannot lead to prediction of deleterious effects of nsSNP sites, these studies illustrate that structural characterization of nsSNP sites and their sequence conservation as measured by entropy scores are useful information that can be incorporated in studies addressing the fundamental problem of predicting when nsSNPs are likely to cause disease or significant phenotypic changes.

Here we make the structural mapping of both disease and non-disease associated nsSNPs available through a web accessible database topoSNP (<http://gila.bioengr.uic.edu/snp/toposnp>). In addition, we also provide structural characterization and entropy measurement of these nsSNP sites. TopoSNP also allows for convenient visualization of both disease-associated and non disease-associated nsSNPs.

Methods

Disease-associated nsSNPs were extracted from the Online Mendelian Inheritance in Man (OMIM) database (<http://www.ncbi.nlm.nih.gov/omim>) (11). The control nsSNPs were extracted from dbSNP (12) (release 103) (<ftp://ftp.ncbi.nlm.nih.gov/snp/human>). While not a perfect source of negative control nsSNPs, it is reasonable to expect that a much smaller fraction of the nsSNPs from dbSNP would be associated with disease. Geometric locations and entropy calculations were performed as previously described (10).

Database Access

The database can be accessed at <http://gila.bioengr.uic.edu/snp/toposnp>. This site hosts an interactive session based on the CHIME plug-in (freely available from <http://www.mdlchime.com>, a plug-in that interactively displays three-dimensional molecules), allowing for the visualization of the mutation along with its classification of geometric location and entropy score. After selecting a gene and a specific protein structure to explore, the three-dimensional representation of the protein is displayed, along with a list of known nsSNPs. Selecting a SNP will highlight its position in the protein as well as its corresponding pocket or void if appropriate (see Figure 1). Selecting a SNP will also bring up its specific assignment of geometric class and relative entropy score which are displayed below the protein visualization.

There is an online help page as well as a walkthrough example for familiarization with the database. The entire database is also downloadable in tar format from the topoSNP website (<http://gila.bioengr.uic.edu/snp/toposnp>).

Database status and future work

The database currently contains 27,417 nsSNP mappings (26,859 from the disease-associated nsSNPs as derived from the OMIM database and 558 control nsSNPs as derived from the dbSNP database) that correspond to 770 protein structures. These 770 protein structures are derived from 421 gene loci. This is a much larger dataset than was previously published (10) because redundant and homologous protein structures are now included. The database will be updated at regular intervals.

Discussion

We present here a resource for accessing both geometric location information and conservation information from a study of disease-associated nsSNPs and control nsSNPs. Conservation is assessed here by entropy calculated using a Hidden Markov Model (HMM). There are other approaches for assessing conservation at SNP sites. For example, recent studies demonstrated that position specific scoring matrices (PSSMs) are quite effective for SNP analysis (7). A comparison of PSSM and HMM methods for remote homology detection showed that these two methods often obtain comparable results (13), although it is unclear exactly to what extent these two approaches differ when assessing conservation at individual sites. In addition, phylogenetic information ideally should be incorporated when assessing sequence conservation, for example, into a codon or amino acid substitution model, together with a maximum likelihood estimator, or a Bayesian estimator (14-16). However, these methods are very involved and are computationally demanding, and therefore are unsuitable for large scale calculation of many proteins. Entropy calculation in this case provides an efficient and rapid, albeit less precise, assessment of sequence conservation.

In this study, we do not differentiate missense mutations causing Mendelian type diseases from nsSNPs associated with complex disease phenotypes. Our purpose is to

examine nsSNPs that are clearly associated with disease. It is possible that these two sub-populations of nsSNPs may have different characteristics.

Acknowledgements

This work is supported by grants from National Science Foundation (CAREER DBI0133856, DBI0078270, and MCB998008) and National Institute of Health (GM68958). N.S. was supported in part by an NIH/NIDDK-funded predoctoral training program (T32 DK007739) in "Signal Transduction and Cellular Endocrinology."

REFERENCES

1. Yoshida, A., Huang, I.Y. and Ikawa, M. (1984) Molecular abnormality of an inactive aldehyde dehydrogenase variant commonly found in Orientals. *Journal of Molecular Evolution*, **81**, 258-261.
2. Jaruzelska, J., Abadie, V., d'Aubenton-Carafa, Y., Brody, E., Munnich, A. and Marie, J. (1995) In vitro splicing deficiency induced by a C to T mutation at position -3 in the intron 10 acceptor site of the phenylalanine hydroxylase gene in a patient with phenylketonuria. *Human Molecular Genetics*, **270**, 20370-20375.
3. Proia, R.L. and Neufeld, E.F. (1982) Synthesis of beta-hexosaminidase in cell-free translation and in intact fibroblasts: an insoluble precursor alpha chain in a rare form of Tay-Sachs disease. *Journal of Biological Chemistry*, **79**, 6360-6364.
4. Sunyaev, S., Ramensky, V. and Bork, P. (2000) Towards a structural basis of human non-synonymous single nucleotide polymorphisms. *Nature Reviews Genetics*, **16**, 198-200.
5. Chasman, D. and Adams, R.M. (2001) Predicting the functional consequences of non-synonymous single nucleotide polymorphisms: Structure-based assessment of amino acid variation. *Journal of Molecular Evolution*, **307**, 683-706.
6. Ramensky, V., Bork, P. and Sunyaev, S. (2002) Human non-synonymous SNPs: server and survey. *Nucleic Acids Research*, **30**, 3894-3900.
7. Ng, P.C. and Henikoff, S. (2001) Predicting deleterious amino acid substitutions. *Journal of Molecular Evolution*, **11**, 863-874.
8. Ng, P.C. and Henikoff, S. (2002) Accounting for human polymorphisms predicted to affect protein function. *Journal of Molecular Evolution*, **12**, 436-446.
9. Ng, P.C. and Henikoff, S. (2003) SIFT: Predicting amino acid changes that affect protein function. *Nucleic Acids Research*, **31**, 3812-3814.
10. Stitzel, N.O., Tseng, Y.Y., Pervouchine, D., Goddeau, D., Kasif, S. and Liang, J. (2003) Structural location of disease-associated single-nucleotide polymorphisms. *Nucleic Acids Research*, **327**, 1021-1030.
11. Wheeler, D.L., Church, D.M., Federhen, S., Lash, A.E., Madden, T.L., Pontius, J.U., Schuler, G.D., Schriml, L.M., Sequeira, E., Tatusova, T.A. (2003) Database resources of the National Center for Biotechnology. *Nucleic Acids Research*, **31**, 28-33.
12. Sherry, S.T., Ward, M.H., Kholodov, M., Baker, J., Phan, L., Smigielski, E.M. and Sirotkin, K. (2001) dbSNP: the NCBI database of genetic variation. *Nucleic Acids Research*, **29**, 308-311.
13. Madera, M. and Gough, J. (2002) A comparison of profile hidden Markov model procedures for remote homology detection. *Bioinformatics*, **30**, 4321-4328.


14. Swofford, D.L., Olsen, G.L., Waddell, P.J. and Hillis, D.L. (1996) In Hillis, D. M., Moritz, C. and Mable, B. (eds.), *Molecular systematics*. Sinauer Associates, Sunderland, Massachusetts, pp. 407-514.
15. Yang, Z. (2001) In Balding, D., Bishop, M. and Cannings, C. (eds.), *Evolutionary genetics*. Wiley, London, pp. 327-350.
16. Huelsenbeck, J.P. and Ronquist, F. (2001) MRBAYES: Bayesian inference of phylogenetic trees. *Systematic Biology*, 50, 17, 754-755.

FIGURE LEGENDS


Fig 1. Example of topoSNP visualization for a nsSNP from alcohol dehydrogenase (PDB code 1htb). The R47 mutation is highlighted along with the surface pocket where it is located.

snp :: toposnp :: 1htb - Microsoft Internet Explorer

Address <http://gila.bioengr.uic.edu/snp/toposnp/?mode=query&pdb=1htb>
























































UIC
BIOINFORMATICS



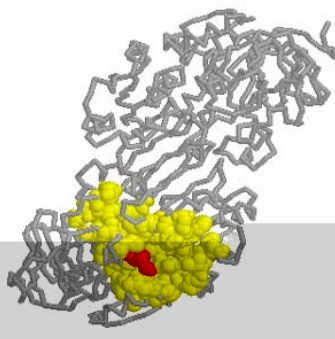
Welcome to the topoSNP database. This site allows for visualization of disease and non-disease associated non-synonymous single nucleotide polymorphisms (SNPs) and displays geometric and relative entropy calculations as described in Stitzel et. al. (JMB 327 (5):1021-1030).

The structure of your query should be viewable in the Chime window to the right. The variants from the topoSNP database that correspond to this structure are listed under the appropriate header to the right of the structure. Clicking on the "View" button for a variant will highlight the SNP in red, and if it is located in a pocket or void, the pocket or void will be highlighted in yellow.

When you move your mouse over the "View" button, the geometric assignment and entropy score are displayed in the box below the structure.



For help,                                                     

1htb - OXIDOREDUCTASE
CRYSTALLIZATION OF HUMAN BETA3 ALCOHOL DEHYDROGENASE 1HTB
32 (10 MG/ML) IN 100 MM SODIUM PHOSPHATE (PH 7.5), 7.5 MM 1HTB
43 NAD+ AND 1 MM 4-IODOPYRAZOLE AT 25 C 1HTB 5



MDL

Disease-associated nsSNPs

Swissprot	Chain	Residue	Original	Mutation	View
P00325	B	47	R	H	
P00325	B	369	R	C	

Non disease-associated nsSNPs

There are no non disease-associated nsSNPs for this entry

Variant information:

Location	Entropy	Score
Pocket	Conserved position	4.1e-148

PDB entry at RCSB: [1htb](#)

Figure 1