Published online 5 October 2007 *Nucleic Acids Research, 2008, Vol. 36, Database issue* ***D409–D413***
doi:10.1093/nar/gkm801

coliSNP database server mapping nsSNPs on protein structures

Hidetoshi Kono[1,2,3,*], Tomo Yuasa[4], Shinya Nishiue[5] and Kei Yura[3,6]

[1]Computational Biology Group, Quantum Beam Science Directorate, Japan Atomic Energy Agency, 8-1 Umemidai, Kizugawa, Kyoto 619-0215, [2]PRESTO, Japan Science and Technology Agency, 4-1-8 Kawaguchi, Saitama, 332-0012, [3]Research Unit for Quantum Beam Life Science Initiative, Quantum Beam Science Directorate, Japan Atomic Energy Agency, 8-1 Umemidai, Kizugawa, Kyoto 619-0215, [4]Bioinformatics Department, Mitsubishi Space Software CO. LTD, 5-4-36 Tsukaguchi-honmachi, Amagasaki, Hyogo 661-0001, [5]Department of Bioengineering, Nagaoka University of Technology, Nagaoka, Niigata 940-2188 and [6]Quantum Bioinformatics Team, Center for Computational Science and Engineering, Japan Atomic Energy Agency, 8-1 Umemidai, Kizugawa, Kyoto 619-0215 Japan

Received August 15, 2007; Revised September 13, 2007; Accepted September 17, 2007

ABSTRACT

We have developed coliSNP, a database server (http://yayoi.kansai.jaea.go.jp/colisnp) that maps non-synonymous single nucleotide polymorphisms (nsSNPs) on the three-dimensional (3D) structure of proteins. Once a week, the SNP data from the dbSNP database and the protein structure data from the Protein Data Bank (PDB) are downloaded, and the correspondence of the two data sets is automatically tabulated in the coliSNP database. Given an amino acid sequence, protein name or PDB ID, the server will immediately provide known nsSNP information, including the amino acid mutation caused by the nsSNP, the solvent accessibility, the secondary structure and the flanking residues of the mutated residue in a single page. The position of the nsSNP within the amino acid sequence and on the 3D structure of the protein can also be observed. The database provides key information with which to judge whether an observed nsSNP critically affects protein function and/or stability. As far as we know, this is the only web-based nsSNP database that automatically compiles SNP and protein information in a concise manner.

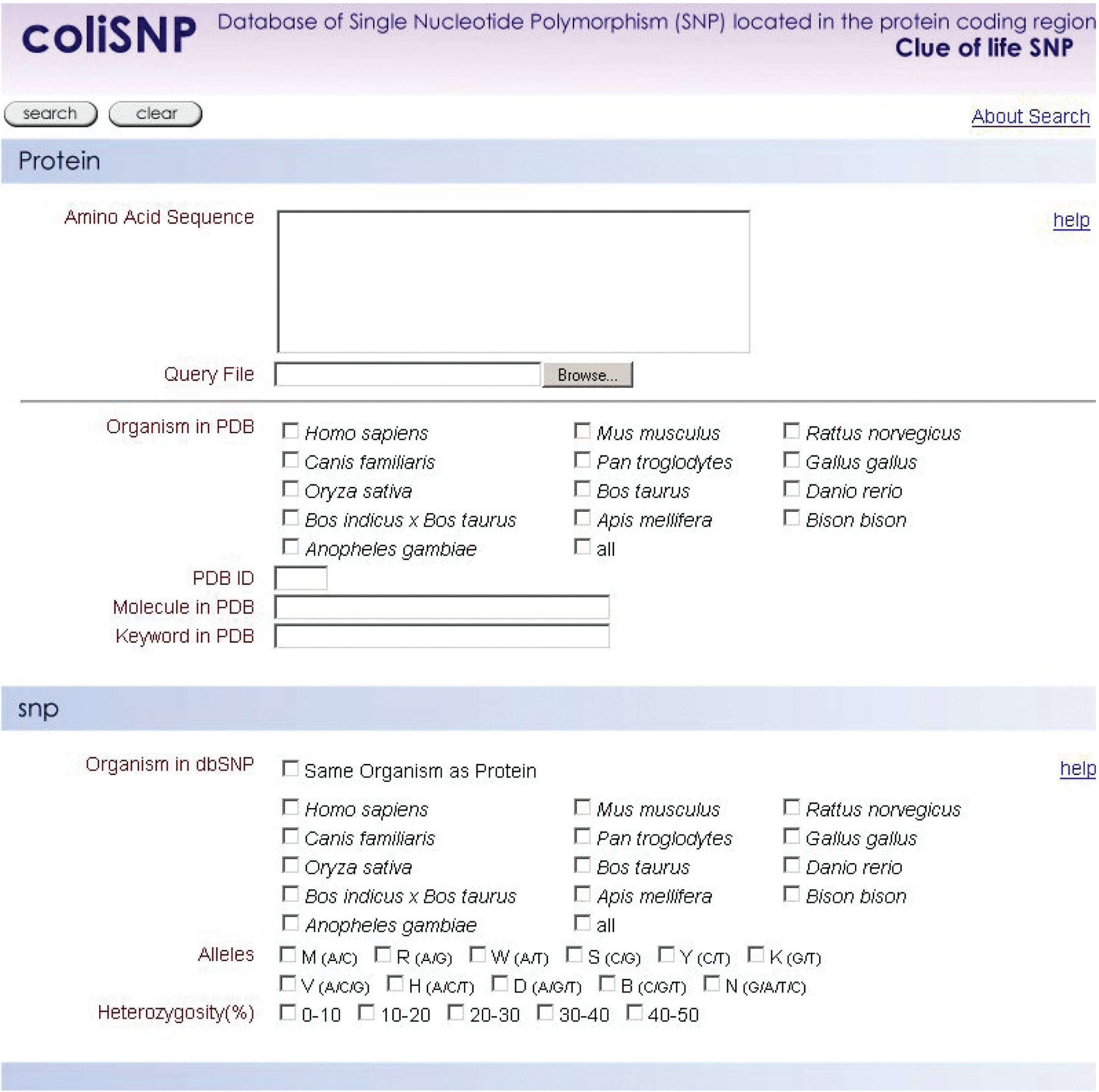
INTRODUCTION

Single nucleotide polymorphisms (SNPs) have the potential to affect gene function, especially when they are located in coding or regulatory regions. Among the many types of SNPs, non-synonymous SNPs (nsSNPs) are believed to have the greatest impact on protein function because they often lead to mutation of the encoded amino acids, which can have a deleterious effect on the structure and/or function of the proteins. Such nsSNPs are often associated with disease-modifying alleles that have been compiled, for example, in the OMIM database (http://www.ncbi.nlm.nih.gov/omim/) (1).

Disease-associated SNPs were often interpreted solely on the basis of their sequences, mainly with respect to sequence conservation; however, thanks to structural genomics projects in the USA, Canada, Europe and Japan, which now make available more than 40 000 protein structures (2), it is possible to interpret the effects of a large number of SNPs on three-dimensional (3D) protein structures. To further investigate the possible causes of disease at the molecular level, we have started mapping nsSNPs on 3D protein structures and developed a database named coliSNP (Clue of Life SNP), which provides users with both the protein sequence and structural information on nsSNPs, enabling them to gain significant insight into the effects of nsSNPs at the molecule level.

To date, several databases, including SAAP (3), PolyDoms (4), topoSNP (5), SNPeffect (6), SNPs3D (7), MutDB (8,9) and LS-SNP (10), have been developed to provide links between SNPs and protein sequence/structure data and/or cellular processes such as localization, phosphorylation and glycosylation. Among these, topoSNP, SNPs3D, MutDB and LS-SNP have a direct link to 3D protein structures from nsSNP locations within nucleotide sequences. Apparently, however, these databases are no longer actively maintained or are updated only about once a year, at best. The coliSNP we have launched is automatically updated every week and contains the up-to-date nsSNP data mapped on the 3D protein structures. Moreover, coliSNP enables visualization of 3D protein structures directly using Jmol or RasMol by downloading the coordinates attached to a RasMol script. Both of these features are unique to

*To whom correspondence should be addressed. Tel: +81-774-71-3465; Fax: +81-774-71-3460; Email: kono.hidetoshi@jaea.go.jp



coliSNP Database of Single Nucleotide Polymorphism (SNP) located in the protein coding region
Clue of life SNP

Protein

Amino Acid Sequence

Query File

Organism in PDB
□ *Homo sapiens* □ *Mus musculus* □ *Rattus norvegicus*
□ *Canis familiaris* □ *Pan troglodytes* □ *Gallus gallus*
□ *Oryza sativa* □ *Bos taurus* □ *Danio rerio*
□ *Bos indicus x Bos taurus* □ *Apis mellifera* □ *Bison bison*
□ *Anopheles gambiae* □ all

PDB ID

Molecule in PDB

Keyword in PDB

snp

Organism in dbSNP □ Same Organism as Protein
□ *Homo sapiens* □ *Mus musculus* □ *Rattus norvegicus*
□ *Canis familiaris* □ *Pan troglodytes* □ *Gallus gallus*
□ *Oryza sativa* □ *Bos taurus* □ *Danio rerio*
□ *Bos indicus x Bos taurus* □ *Apis mellifera* □ *Bison bison*
□ *Anopheles gambiae* □ all

Alleles □ M (A/C) □ R (A/G) □ W (A/T) □ S (C/G) □ Y (C/T) □ K (G/T)
□ V (A/C/G) □ H (A/C/T) □ D (A/G/T) □ B (C/G/T) □ N (G/A/T/C)

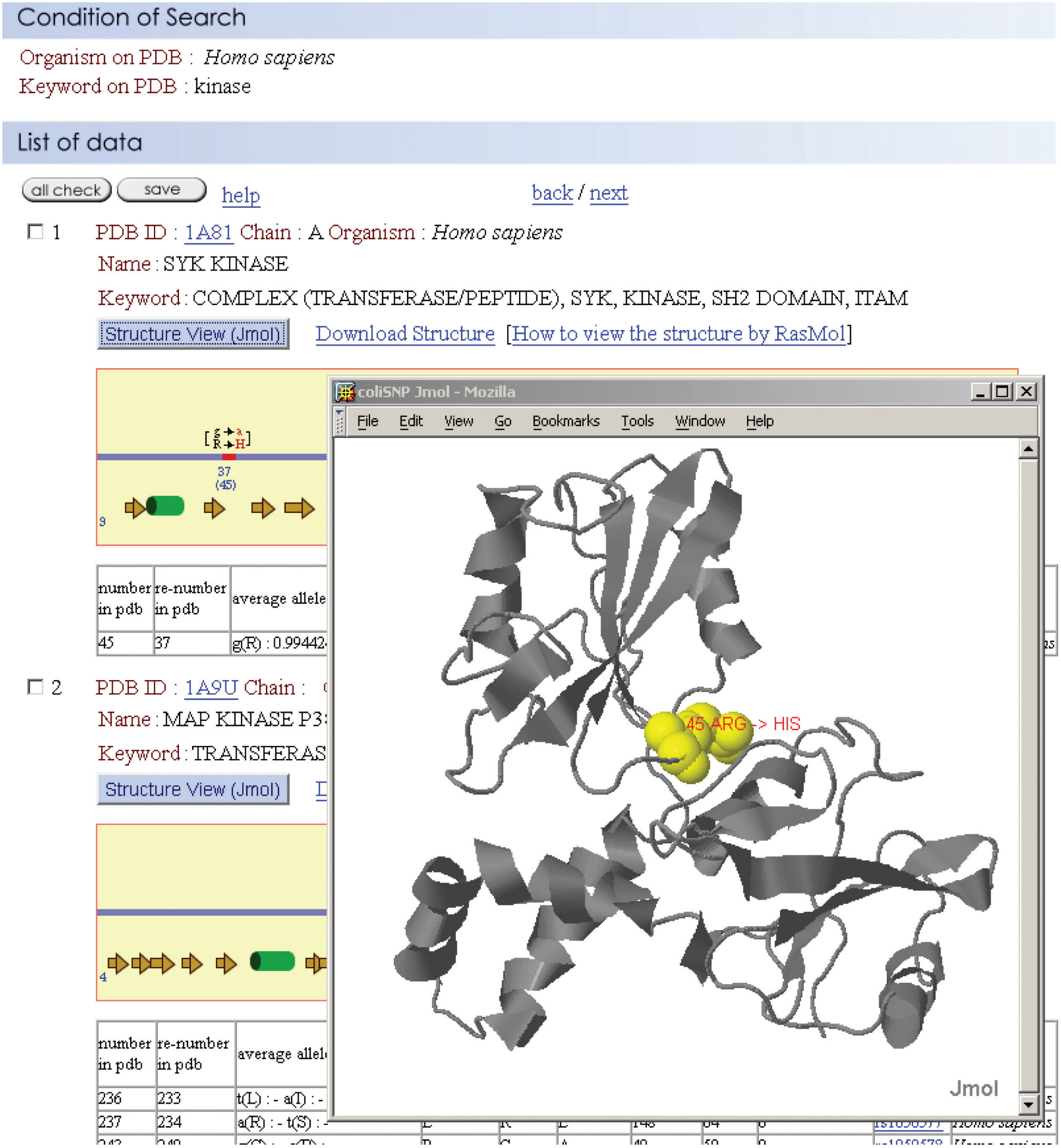
Heterozygosity(%) □ 0-10 □ 10-20 □ 20-30 □ 30-40 □ 40-50

Figure 1. The coliSNP search interface. The user can use the protein section, SNP section or both to set the search conditions.

coliSNP and enable one to easily observe the locations of mutations caused by nsSNPs, even when the nsSNPs/protein structures have only been very recently identified/determined. The regularly updated nsSNP information, combined with 3D protein structures, represents an invaluable resource for evaluating the effect of the mutation on protein function and stability.

Data sources and integrated information

To develop the coliSNP database, we integrated three publicly available databases: RefSeq, for a comprehensive, integrated, non-redundant set of sequences (11); Protein Data Bank (PDB), for 3D protein structures (2); and dbSNP, for SNP information (12). We first compared dbSNP and RefSeq and built a temporary database, RefSeqSNP, which was a subset of RefSeq that only contained amino acid sequences encoded by genes with nsSNPs. We then used Basic Local Alignment Search Tool (BLAST) (13) with default parameters to search for PDB entries that matched each of the sequences in RefSeqSNP. We collected the PDB entries whose amino acid sequence identity against the query was >95% over the region aligned by BLAST if the length of the aligned region



Condition of Search

Organism on PDB : *Homo sapiens*
Keyword on PDB : kinase

List of data

□ 1 PDB ID : 1A81 Chain : A Organism : *Homo sapiens*
Name : SYK KINASE
Keyword : COMPLEX (TRANSFERASE/PEPTIDE), SYK, KINASE, SH2 DOMAIN, ITAM

number in pdb	re-number in pdb	average allele
45	37	g(R) : 0.99442

□ 2 PDB ID : 1A9U Chain :
Name : MAP KINASE P3
Keyword : TRANSFERAS

number in pdb	re-number in pdb	average allele
236	233	t(L) : - a(I) : -
237	234	a(R) : - t(S) : -

Figure 2. A typical search result. nsSNP information is provided with structural information on the mutated amino acid residue—e.g. the secondary structure and solvent accessibility.

was ≥ 30 amino acids. At this point, we removed similar amino acid sequences by limiting the PDB entries to those that were ranked at the top in each of the 90% sequence identity clusters and compiled in the 'clusters90.txt' file provided by PDB. Our reasoning was that sequences with such a high identity would assume very similar tertiary structures, close enough to assess the impact of mutations.

Data access and the search interface

coliSNP can be accessed at http://yayoi.kansai.jaea.go.jp/colisnp. In the search form (Figure 1), a user can provide several search keys on the protein and/or the SNPs. In the protein section, the user can use as query terms the amino acid sequence, PDB ID, molecule name and keyword in PDB. The user can also limit the scope of the 3D protein

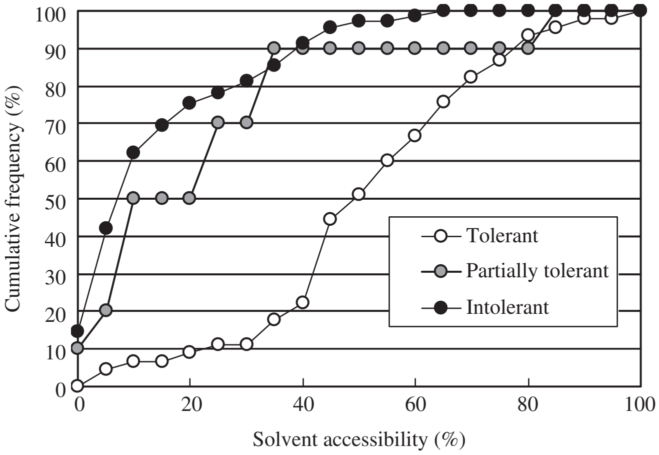


Figure 3. Cumulative plots of tolerant, partially tolerant and intolerant sites in Lac repressor against the solvent accessibility. In the experiment (17), 12 or 13 mutations (depending on the identity of the wild-type residue) were tested at 124 sites. We defined the tolerance at each site as follows: tolerant, <5 of the mutations cause loss of function (45 sites); intolerant, >8 of the mutations cause loss of function (69 sites); and partially tolerant, 5–8 of the mutations cause loss of function (10 sites). The solvent accessibility was calculated using the program ASC (19) with a protein–DNA complex form (PDB:1EFA) or a tetrameric form (PDB:1LBI), depending on the site considered.

structures to be mapped for SNPs by specifying the organisms that the proteins were derived from. In the SNP section, the user can give the organism, allele type and heterozygosity as queries. The user can also use both the protein and SNP sections to narrow the search.

As shown in Figure 2, the search result emerges with information about the mutation, the flanking amino acids, the secondary structure and the solvent accessibility of the mutated residue, allele frequency and heterozygosity. The output page also has a link to the original dbSNP database, enabling more detailed information to be obtained. If desired, the information can be saved in a flat text format for further analysis. The user can also easily observe the location of an nsSNP on the 3D protein structure by clicking either the 'Download Structure' link or the 'Structure View' box. We adopted two graphics programs, RasMol (http://openrasmol.org) and Jmol (http://jmol.sourceforge.net), for 3D protein visualization. The former is one of the most widely used software packages for visualizing the 3D structures of proteins and has a number of handy operations. The latter displays 3D structures in a Java-implemented browser.

nsSNPs location and its impact on 3D protein structure

One of the unique features of the coliSNP database is that it gives the solvent accessibility of a wild-type residue that has been mutated by an nsSNP. We found that the solvent accessibility is the best indicator of the impact of a mutation on protein function. Other properties that we evaluated include the secondary structure where the mutation occurred, changes in hydrogen bonding, and the chemical properties of the affected residue. The correlation between the effect substituting a single residue and its solvent accessibility has long been discussed (14–16). To provide a quantitative limit for the solvent accessibility of residues able to tolerate mutation caused by nsSNPs, we collected experimental data showing the relationship between point mutations and the activities of proteins with known 3D structures. The point mutation studies on Lac repressor (17) and T4 lysozyme (18), in particular, provided us with sufficient data to determine that limit of solvent accessibility. The solvent accessibility was calculated with ASC(19). We then re-examined the relationship between solvent accessibility and viability of the organism for these two proteins. Figure 3 shows the loss of function rate plotted against the solvent accessibility of the wild-type residue. In the case of Lac repressor, about 90% of mutation-tolerant sites (see Figure 3 caption) were located at positions where the solvent accessibility of the wild-type residue was >30%, and about 80% of mutations in intolerant or partially tolerant sites were located at positions where the solvent accessibility was $\leq$30%. Based on this observation, we decided to provide the solvent accessibility value of the mutated residue together with the 3D structure of the protein in the database, and the residues with solvent accessibility of $\leq$ 30% were marked in yellow in the 3D structures. We believe that these data enable one to evaluate possible effect of nsSNPs on protein stability and function. For instance, the nsSNP in human SYK kinase shown in Figure 2 results in the substitution of Arg45 with His, and the solvent accessibility is 8%. Because of the degree to which this residue is buried, it is highly likely that the substitution will have a deleterious effect on the protein's function and/or stability, as suggested by Figure 3. In fact, the residue forms one of the loops for domain association and is located relatively close to the phosphorylated Tyr of the target peptide (20). Both of these pieces of information are easily retrieved from coliSNP and may add medically important annotation to the SNP site. It is worth noting that the impact of a mutation should also be evaluated based on sequence conservation. Disease-associated nsSNPs tend to be located at highly conserved sites (21). This information will be incorporated in the coliSNP database in the near future.

Database status and future work

As of 26 July 2007, coliSNP contains 4470 nsSNPs, which are mapped on 1559 distinct protein structures, mostly from *Homo sapiens* (4216 out of 4470). A process to include all of the data in dbSNP, which is derived from 22 organisms, is ongoing. Modification of 3D protein structure visualization tools to accept the new PDB format (http://www.wwpdb.org/docs.html and see Remediation Documentation) is also ongoing. In the near future, coliSNP will also provide SNPs located in the gene regulatory region together with potential target regions of the regulatory proteins.

Availability

The coliSNP database can be accessed freely at http://yayoi.kansai.jaea.go.jp/colisnp.

ACKNOWLEDGEMENTS

The authors would like to thank Ms. Atsuko Doi for careful inspection of the initial data. This project is supported in part by a Grant-in-Aid for Scientific Research 15710144 (H.K.) from the Ministry of Education, Culture, Sports, Science and Technology in Japan. Funding to pay the Open Access publication charges for this article was provided by Japan Atomic Energy Agency (JAEA).

Conflict of interest statement. None declared.

REFERENCES

1. McKusick,V.A. (1998) Online Mendelian Inheritance in Man, OMIM (TM). McKusick-Nathans Institute of Genetic Medicine. Johns Hopkins University, Baltimore, MD and National Center for Biotechnology Information, National Library of Medicine, Bethesda, MD.
2. Berman,H., Henrick,K., Nakamura,H. and Markley,J.L. (2007) The worldwide Protein Data Bank (wwPDB): ensuring a single, uniform archive of PDB data. *Nucleic Acids Res.*, **35**, D301–D303.
3. Cavallo,A. and Martin,A.C. (2005) Mapping SNPs to protein sequence and structure data. *Bioinformatics*, **21**, 1443–1450.
4. Jegga,A.G., Gowrisankar,S., Chen,J. and Aronow,B.J. (2007) PolyDoms: a whole genome database for the identification of non-synonymous coding SNPs with the potential to impact disease. *Nucleic Acids Res.*, **35**, D700–D706.
5. Stitziel,N.O., Binkowski,T.A., Tseng,Y.Y., Kasif,S. and Liang,J. (2004) topoSNP: a topographic database of non-synonymous single nucleotide polymorphisms with and without known disease association. *Nucleic Acids Res.*, **32**, D520–D522.
6. Reumers,J., Schymkowitz,J., Ferkinghoff-Borg,J., Stricher,F., Serrano,L. and Rousseau,F. (2005) SNPeffect: a database mapping molecular phenotypic effects of human non-synonymous coding SNPs. *Nucleic Acids Res.*, **33**, D527–D532.
7. Yue,P., Melamud,E. and Moult,J. (2006) SNPs3D: candidate gene and SNP selection for association studies. *BMC Bioinformatics*, **7**, 166.
8. Dantzer,J., Moad,C., Heiland,R. and Mooney,S. (2005) MutDB services: interactive structural analysis of mutation data. *Nucleic Acids Res.*, **33**, W311–W314.
9. Mooney,S.D. and Altman,R.B. (2003) MutDB: annotating human variation with functionally relevant data. *Bioinformatics*, **19**, 1858–1860.
10. Karchin,R., Diekhans,M., Kelly,L., Thomas,D.J., Pieper,U., Eswar,N., Haussler,D. and Sali,A. (2005) LS-SNP: large-scale annotation of coding non-synonymous SNPs based on multiple information sources. *Bioinformatics*, **21**, 2814–2820.
11. Pruitt,K.D., Tatusova,T. and Maglott,D.R. (2007) NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Res.*, **35**, D61–D65.
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 Res.*, **29**, 308–311.
13. Altschul,S.F., Gish,W., Miller,W., Myers,E.W. and Lipman,D.J. (1990) Basic local alignment search tool. *J. Mol. Biol.*, **215**, 403–410.
14. Ramensky,V., Bork,P. and Sunyaev,S. (2002) Human non-synonymous SNPs: server and survey. *Nucleic Acids Res.*, **30**, 3894–3900.
15. Chasman,D. and Adams,R.M. (2001) Predicting the functional consequences of non-synonymous single nucleotide polymorphisms: structure-based assessment of amino acid variation. *J. Mol. Biol.*, **307**, 683–706.
16. Stitziel,N.O., Tseng,Y.Y., Pervouchine,D., Goddeau,D., Kasif,S. and Liang,J. (2003) Structural location of disease-associated single-nucleotide polymorphisms. *J. Mol. Biol.*, **327**, 1021–1030.
17. Kleina,L.G. and Miller,J.H. (1990) Genetic studies of the lac repressor XIII. Extensive amino acid replacements generated by the use of natural and synthetic nonsense suppressors. *J. Mol. Biol.*, **212**, 295–318.
18. Rennell,D., Bouvier,S.E., Hardy,L.W. and Poteete,A.R. (1991) Systematic mutation of bacteriophage T4 lysozyme. *J. Mol. Biol.*, **222**, 67–88.
19. Eisenhaber,F., Lijnzaad,P., Argos,P. and Sander,C. (1995) The double cubic lattice method: efficient approaches to numerical integration of surface area and volume and to dot surface contouring of molecular assemblies. *J. Comp. Chem.*, **16**, 273–284.
20. Futterer,K., Wong,J., Grucza,R.A., Chan,A.C. and Waksman,G. (1998) Structural basis for Syk tyrosine kinase ubiquity in signal transduction pathways revealed by the crystal structure of its regulatory SH2 domains bound to a dually phosphorylated ITAM peptide. *J. Mol. Biol.*, **281**, 523–537.
21. Ng,P.C. and Henikoff,S. (2003) SIFT: Predicting amino acid changes that affect protein function. *Nucleic Acids Res.*, **31**, 3812–3814.