Amino Acids (2008) 35:615–626
DOI 10.1007/s00726-008-0087-9

ORIGINAL ARTICLE

Incorporating the amino acid properties to predict the significance of missense mutations

Tze-Chuen Lee · Ann S. G. Lee · Kuo-Bin Li

Received: 18 January 2008 / Accepted: 26 February 2008 / Published online: 16 April 2008
© Springer-Verlag 2008

Abstract Determining if missense mutations are deleterious is critical for the analysis of genes implicated in disease. However, the mutational effects of many missense mutations in databases like the Breast Cancer Information Core are unclassified. Several approaches have emerged recently to determine such mutational effects but none have utilized amino acid property indices. We modified a previously described phylogenetic approach by first classifying benign substitutions based on the assumption that missense mutations that are maintained in orthologs are unlikely to affect function. A consensus conservation score based on 16 amino acid properties was used to characterize the remaining substitutions. This approach was evaluated with experimentally verified T4 lysozyme missnese mutations and is shown to be able to sieve out putative biochemical and structurally important residues. The use of amino acid properties can enhance the prediction of biochemical and structurally important residues and thus also predict the significance of missense mutations.

T.-C. Lee and A. S. G. Lee contributed equally to this work.

T.-C. Lee
Bioinformatics Institute, 30 Biopolis Street,
Singapore 138671, Singapore
e-mail: leetc@bii-sg.org

A. S. G. Lee
National Cancer Center, 11 Hospital Drive,
Singapore 169610, Singapore
e-mail: dmslsg@nccs.com.sg

K.-B. Li (✉)
Center for Systems and Synthetic Biology,
National Yang-Ming University, Taipei 11221, Taiwan
e-mail: kbli@ym.edu.tw

Keywords Missense mutation · T4 lysozyme · *BRCA1* · Non-synonymous SNP · Physico-chemical properties

Introduction

Determining if missense mutations are deleterious is critical for the analysis of genes implicated in disease. Mutation databases like the Breast Cancer Information Core (BIC) database (Szabo et al. 2000) are unable to determine the mutational effects of about 88% of *BRCA1* (MIM# 113705) missense or non-synonymous single nucleotide polymorphism (nsSNP) mutations. Recently, various in silico methods that attempt to predict the phenotype of missense mutations have emerged (Chasman and Adams 2001; Ramensky et al. 2002; Herrgard et al. 2003; Krishnan and Westhead 2003; Mooney and Altman 2003; Ng and Henikoff 2003; Cai et al. 2004; Lau and Chasman 2004; Verzilli et al. 2005; Xiao et al. 2005). These include approaches that predict the degree of structural changes the mutation will cause and evolutionary approaches that identify and measure the degree of conservation. However, there may be discrepancies between different prediction methods (Tchernitchko et al. 2004). Besides, mutagenesis study is also an important approach for structure-based drug design (see, e.g., Chou 1993, 1996; Chou et al. 2003; Du et al. 2004, 2005a, b, 2007; Wei et al. 2005, 2006a, b, 2007; Zhang et al. 2006, 2007; Gao et al. 2007; Li et al. 2007a, b; Wang et al. 2007a, b, c, d; Wu and Yan 2007; and particularly some comprehensive reviews in this area: Chou 2004, 2006; Chou et al. 2006).

Both evolutionary and structural approaches have been used to classify the phenotype of missense mutations. Evolutionary approaches have different schemes measuring

the degree of conservation in the multiple sequence-aligned regions. The majority of current approaches have not taken into account important aspects of an amino acid residue such as hydrophobicity, surface area, charge, and volume. Santibanez Koref and coworkers (Santibanez Koref et al. 2003) had proposed using amino acid properties but eventually adopted an approach (I_{20} parameter set) that did not take amino acid properties into account when assessing the effect of missense mutations. The I_{20} parameter set regards a conservative mutation to be as equally damaging or benign as a non-conservative mutation. In this paper, however, we investigated whether the use of amino acid properties, together with evolutionary distances and positional residue frequency, could improve the assessment of the degree of conservation in a set of evolutionary aligned amino acid residues.

We present here an evolutionary approach that is an extension of the method by Santibanez Koref and coworkers (2003) for the functional prediction of missense mutation. Santibanez Koref and coworkers (2003) used a phylogenetic approach to quantify the variability of an amino acid change and applied it to p53 (*TP53*; MIM#191170). Their results showed that a parameter that represents the identity of amino acids could be used to classify mutants with an accuracy of 95.6%. In our approach, benign substitutions were first classified based on the assumption that missense mutations that are maintained in orthologs are unlikely to affect the function of the protein. Then, taking into account the evolutionary distance between species and their relative selection pressure, a conservation score based on 16 amino acid properties was used to characterize the remaining substitutions. We tested our predictions on the experimentally verified set of bacteriophage T4 lysozyme annotations (Rennell et al. 1991) and applied the method to some of the most common missense mutations of the *BRCA1* gene found in the BIC database. Our approach compares well with existing methods and may potentially be used to identify biochemical and structurally important amino acids and also be used to assess the significance of missense mutations documented in mutational databases.

Materials and methods

Analysis

Bacteriophage T4 lysozyme and human *BRCA1* amino acid and nucleotide sequences were extracted from Genbank. Amino acid sequences were aligned using CLUSTALX 1.83 (Jeanmougin et al. 1998). The nucleotide sequences were aligned to amino acid sequences and were checked visually. Residues and their corresponding codons that created gaps in the reference sequences (T4 lysozyme or human *BRCA1*) were removed. Missense mutations that were maintained in any one of the aligned orthologous sequences were considered as benign (B). This criterion is also called "Deleterious Change Not Present" (DCNP) in the work by Santibanez Koref and coworkers (2003).

For mutations that were not found in any of the orthologous positions, a Z-score based on the work by Santibanez Koref and coworkers (2003) was calculated. Firstly, an S-score that calculates the degree of variability of a mutation is defined. This S-score takes into account amino acid properties such as hydrophobicity, molecular weight, volume of residues, etc., and compares their variation between the mutant residue m with all the other residues at a particular position in a set of aligned sequences. A high S-score means that a property (or a linear combination of properties) exists for which there is little variation among the residues in the wild-type sequences but a large difference between these and the mutant residue. The S-score can be defined as the maximum of the following expression adapted from Santibanez Koref and coworkers (2003)

$$Q(e) = \frac{\left(v_m^* - \overline{v^*}\right)^2}{1 + \frac{1}{n+1}\sum_{x=1}^{n}\left(v_x^* - \overline{v^*}\right)^2}. \quad (1)$$

Here, m is the mutant residue, x is a residue in the n aligned sequences, $\bar{v}^* = \frac{1}{n}\sum_{x=1}^{n} v_x^*$ and vector v^* is the linear combination of amino acid properties that maximizes the expression. With a given nucleotide mutation rate and evolutionary distance from the reference sequence, a simulated alignment could be generated to represent amino acid residues that do not undergo evolutionary selective pressure. For each mutation, 1,000 of such simulations were performed and 1,000 S*-scores were generated from the simulated alignments. Since the missense mutants are defined with respect to the human wild type sequence, the observed human codon was used as the starting point to simulate the sequences from the remaining species (Santibanez Koref et al. 2003). Finally, for each mutation, a Z-score that computes the probability of the S-score being less than or equal to S*-score is defined. A small Z-score indicates a mutation with a higher tendency to affect the function of protein, and thus having the tendency to be deleterious. The detailed description of this numerical scheme can be found in the study by Santibanez Koref and coworkers (2003).

In our paper, the following modifications to the method described by Santibanez and coworkers (2003) were made. Z-scores were calculated from 16 sets of amino acid properties. Sets of properties used were hydropathy (Kyte and Doolittle 1982), hydrophilicity (Hpl) (Kuhn et al. 1995), mean polarity (Pol), accessible surface area (Sa)

from (Radzicka et al. 1988), net charge (Chg) (Klein et al. 1984), coil formation (Coil) (Charton and Charton 1983), volume (Vol) (Grantham 1974), number of hydrogen bond donors (Hbd) (Fauchere et al. 1988), molecular weight (Mw) (Fasman 1976), isoelectric point (Ispt) (Zimmerman et al. 1968), normalized frequency of alpha-helix (Phelx), beta-sheet (Pbeta), beta-turn (Pturn) from (Chou and Fasman 1978) and PCI-III defined by (Sneath 1966) by principal component analysis of a large set of physicochemical properties. These properties were also used by (Santibanez Koref et al. 2003). All parameters were normalized prior to the calculations so that the variance across the 20 amino acids was 1. The 16 sets of Z-scores were summed up to classify the missense mutation by consensus (consensus score). Thresholds that gave the most accurate classification were used. Mutations with consensus score between 0 and 1.28 (with a limit of 0.08 for each of the 16 properties) were classified as deleterious (D). Mutations with a consensus score above 1.28 are classified as benign (B). From here forward, this summation of the 16 sets of Z-scores will be referred as the Property Consensus score.

(Rennell et al. 1991) designated four phases of phenotype for the size of bacteriophage plaques. "++" being similar to wild type, "+" being significantly smaller in size than wild type, "±" being small till the stage where discerning individual plaques become difficult, and "−" being no plaques produced at all. In our study, we simply considered "-" as deleterious, and "++" as benign. False positive rate (FP) was defined as the proportion of benign mutations that were incorrectly classified as deleterious. False negative rate (FN) was defined as the proportion of deleterious mutations that were incorrectly classified as benign. The overall accuracy was the percentage of the total predictions that were correctly predicted.

Sources of data

Bacteriophage T4 lysozyme

Lysozyme amino acid and nucleotide sequences from the following 33 strains were used: *Bacillus amyloliquefaciens* phage Morita2001 (AY030242.1), *Bacillus* phage phi 29 (X04962.1), Bacteriophage 933W (AF125520.1), Bacteriophage *Aeh1* (AY266303.2), Bacteriophage *APSE-1* (AF157835.1), Bacteriophage B103 (NC_004165.1), Bacteriophage K139 (NC_003313.1), Bacteriophage P1 (X87673.1), Bacteriophage phiKMV (NC_005045.1), Bacteriophage PZA (from *B. subtilis*) (M11813.1), *Bartonella henselae* str. Houston-1 (NC_005956.1), Bartonella *quintana* str. Toulouse (NC_005955.1), *Coxiella burnetii* RSA 493 (NC_002971.2), *Dictyostelium discoideum* (AC116986.2), Enterobacteria phage P22 (NC_002371.2), Enterobacteria phage RB69 (AY303349.1), Enterobacteria phage T4 (NC_000866.4), *Escherichia coli* K-12 MG1655 (U00096.2), *Escherichia coli* O157:H7 EDL933 (NC_002655.2), *Helicobacter pylori* strain TS142 (AY054410.1), *Lactococcus bacteriophage* phi-vML3 (X16178.1), *Magnetospirillum magnetotacticum* MS-1 Magn01_2730 (NZ_AAAP01002730.1), *Methylococcus capsulatus* str. Bath (AE017282.1), *Nitrosomonas europaea* ATCC 19718 (NC_004757.1), *Pseudomonas aeruginosa* phage PaP3 (AY078382.2), *Psychrobacter* sp. (NZ_AADI01000002.1), *Salmonella enterica* subsp., *Enterica serovar* Paratypi A str. ATCC9150 (NC_006511.1), *Salmonella enterica* subsp. *Enterica serovar* Typhi Ty2 (NC_004631.1), *Salmonella typhimurium* bacteriophage ES18 (X67137.1), *Salmonella typhimurium* bacteriophage ES18 (X67137.1), *Salmonella typhimurium* LT2 (NC_003197.1, AAL21602.1), *Xylella fastidiosa* Ann-1 Xfas001_28 (NZ_AAAM01000028.1) and *Yersinia pestis* CO92 (NC_003143.1). The regions spanning the codons 36–39, 61–74, 83–85, 162–164 had few representatives in the alignment and were excluded. Analysis was done from codon 2 to codon 161. To simulate the control sequences, a substitution rate of 4.5×10^{-9} per base per year (Ochman et al. 1999) was used with 1,000 simulations performed at each position (Santibanez Koref et al. 2003). Evolutionary distances were estimated with the program r8s (Sanderson 2003). The phylogenetic tree for the input of the r8s program was obtained from the Bootstrap Neighborhood-Joining method using CLUSTALX 1.83 package (Jeanmougin et al. 1998). The prediction results were compared with those of (Chasman and Adams 2001; Ng and Henikoff 2001; Cai et al. 2004) and also with the I_{20} parameter set of (Santibanez Koref et al. 2003).

Human BRCA1

Only codon region 225–1,365 of the human *BRCA1* sequence is available for comparison with 55 other organisms. These organisms are from all 17 eutherian orders and one marsupial (metatherian). *BRCA1* amino acid and nucleotide sequences from the following species were used: Aardvark (AF284030.1), American beaver (AF540622.1), Asiatic elephant (AF284022.1), black rhinoceros (AF284011.1), brown hare (AF284005.1), Norway rat (NM_012514.1), thick-tailed bush baby (AF019080.1), cat (AF284018.1), chimpanzee (AF019075.1), cow (NM_178573.1), Daubenton's bat (AF203746.1), dog (U50709.1), woodland dormouse (AF332046.1), dugong (AF284019.1), elephant shrew (AF284029.1), European mole (AY121756.1), European shrew (AY057828.1), southern flying squirrel (AF284003.1), Malayan flying lemur (AF019081.1), fox squirrel (AF332044.1), giant anteater (AF484232.1), Hottentot golden mole (AF284027.1), gorilla (AF019076.1), gray dolphin (AY057825.1), greater

glider (AY243455.1), northern gundi (AF540624.1), small Madagascar hedgehog (AF284025.1), western European hedgehog (AF284008.1), hippopotamus (AF284015.1), horse (AF284010.1), howler monkey (AF019079.1), human (U14680.1), hyrax (AF284024.1), Shaw's jird (AF332048.1), red kangaroo (AF284033.1), koala (AY243445.1), llama (AF284012.1), Caribbean manatee (AF284020.1), mountain tapir (AY057830.1), house mouse (NM_009764.2), nine-banded armadillo (AF484222.1), orangutan (AF019077.1), pale-throated sloth (AF284002.1), pangolin (AF284009.1), pig (AF284014.1), Cape porcupine (AF540631.1), rhesus monkey (AF019078.1), Herbert river ringtail (AY243448.1), Cape rock hyrax (AF284023.1), shrew mole (AY121758.1), sperm whale (AF284016.1), springhare (AF540637.1), tailless tenrec (AF284026.1), large tree shrew (AF284006.1), common tube-nosed fruit bat (AF447502.1), common wombat (AF284031.1), and as outgroup, African clawed frog (AF416868.1) and chicken (NM 204169.1). The analyses were based on sequences between human codons 225 and 1,365. Sequence data that fully spans this region are available for 12 eutherian mammals. Codon numbers for all sequences are based on human sequence. Sequences start at codon 241 for fox squirrel, woodland dormouse, northern gundi, 260 for Shaw's jird, 262 for springhare, 267 for Daubenton's bat, 277 for common tube-nosed fruit bat, hyrax, 279 for shrew mole, European shrew, 280 for common wombat, 281 for European mole, 282 for greater glider, koala, Herbert river ringtail, 285 for western European hedgehog, 301 for brown hare, 316 for Cape porcupine and at 274 for the remaining 26 mammals. Sequences end at 609 for red kangaroo, 1,004 for greater glider, 1,032 for Herbert river ringtail, 1,144 for koala, 1,153 for common wombat, 1,197 for common tube-nosed fruit bat, 1,201 for brown hare, 1,211 for northern gundi, 1,212 for small Madagascar hedgehog, 1,213 for shrew mole, European shrew, European mole, elephant shrew, 1,214 for Daubenton's bat, 1,215 for Asiatic elephant, fox squirrel, 1,218 for hyrax, 1,222 for Cape porcupine, 1,338 for Shaw's jird, springhare, 1,344 for American beaver and at 1,219 for the remaining 22 mammals. To simulate the control sequences, a substitution rate of 2×10^{-9} per base per year was used, with 1,000 simulations performed at each position. Evolutionary distances were estimated as described by (Kumar and Hedges 1998). The DCNP (Deleterious Change Not Present) criterion which assumes that a mutation leading to a residue being present at an orthologous position as being benign, was compared. For mutations that were not classified as benign by the DCNP criterion, the Property Consensus score was used to assess the effect of the mutation. This procedure that incorporates both the DCNP criterion and the Property Consensus score is called the Property Consensus with DCNP from here onwards (Fig. 1). Predictions derived from the Sorting Intolerant From Tolerant (SIFT, version 2) (Ng and Henikoff 2002) and Polymorphism Phenotyping (PolyPhen) (Ramensky et al. 2002) programs were also compared.

Analysis of predictions of T4 lysozyme

T4 lysozyme structures 2LZM (Weaver and Matthews 1987) and 1L10 (double-mutant T155A, T157I) were downloaded from Protein Data Bank. Structures were viewed with Swiss-PDBViewer (Guex and Peitsch 1997). T4 lysozyme residues that our approach predicted to be deleterious, four or more times, for the 13 substitutions conducted by (Rennell et al. 1991) are listed in Table 2 (see below). The Perf Software (http://kodiak.cs.cornell.edu/kddcup/software.html) was used to derive the Receiver Operator Characteristic Plot (ROC) for our prediction and the I_{20} parameter prediction.

Results

Evaluation of the property consensus with DCNP method by comparison with other prediction methods

Figure 1 describes the prediction process for a given missense mutation. The first step in the prediction process is to

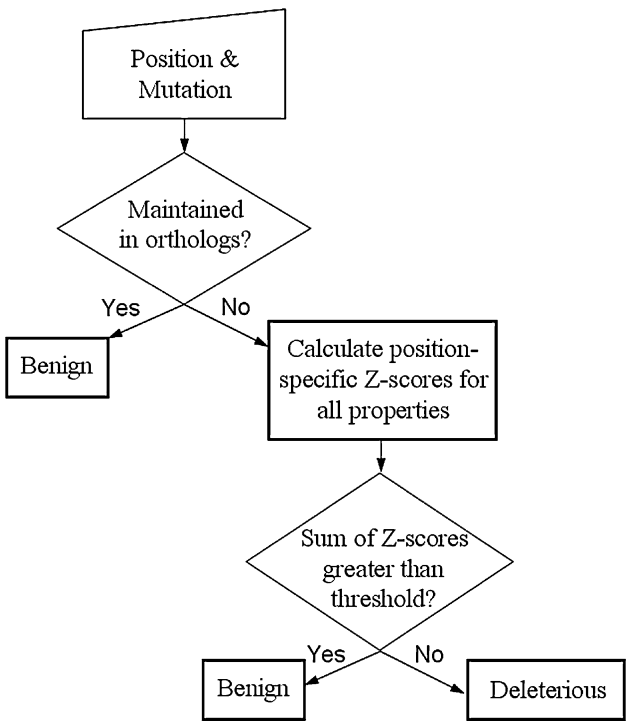

Fig. 1 Flowchart showing the steps involved in the prediction of the property consensus with DCNP method

Table 1 Comparison of prediction methods for evaluating missense mutations using the T4 lysozyme data sets

Prediction indicators	DCNP[a]	Property consensus without DCNP	I_{20} (Santibanez Koref et al. 2003)	Property consensus with DCNP (this report)	Cai et al. (2004)[b]	Ng and Henikoff (2002)[c]
FP (%)	83.8	64.7	84.2	60	2	41
FN (%)	6.8	11.9	5.8	11.8	72	28
Overall accuracy (%)	49.3	87.5	44.6	87.6	85	63
Number of data examined	1,366	1,366	1,366	1,366	NA	2,015

[a] The DCNP (Deleterious Change Not Present) criterion assumes that a mutation leading to a residue that is present at a corresponding position in another species is not deleterious

[b] From Table 4 of Cai et al. (2004)on the T4 lysozyme test set, number of data examined is not available

[c] From Table 1 of Ng and Henikoff (2002) on the T4 lysozyme test set

check if the mutation has been maintained in the corresponding position in the aligned orthologous sequences. Mutation that is not maintained in the orthologs will be subjected to the calculation of Z-scores for the prediction of its effect on the protein.

The approach used in this study is called the Property Consensus with DCNP method because it predicts the effect of a mutation based on the consensus of several amino acid properties. When predicting the effect of a mutation, each of the properties examined has an equal weight in deciding the outcome of the prediction. In order to evaluate the accuracy of the Property Consensus with DCNP method, experimentally verified T4 lysozyme mutations were analyzed with this method and other computationally derived prediction methods (Table 1). Predictions were compared in terms of the percentages of FP, FN and overall accuracy.

Table 1 shows that the Property Consensus with DCNP method, in terms of the overall accuracy, ranked first out of the five methods. The high false positive rate (60%) of the Property Consensus with DCNP approach was possibly due to an over-estimation of benign mutations, resulting from a high representation of biochemical properties among the 16 properties favoring benign prediction. When a threshold of 0.01 is used for the I_{20} parameter method by (Santibanez Koref et al. 2003), a relatively high false positive rate of 84.2% was obtained. However, it has also one of the lowest false negative rate of 5.8%. SIFT prediction also had a relatively high false positive rate of 41% (Ng and Henikoff 2001). Though Cai et al. (2004) had the lowest false positive rate of 2% among the approaches compared, their method had the highest false negative rate of 72%. From Fig. 2, the Area Under the ROC curve (AUC) for Property Consensus with DCNP has a similar AUC to that for the Property Consensus without DCNP (AUC of 0.7). In contrast, the AUC for the I_{20} parameter (0.36) is considerably lower than that for the Property Consensus methods

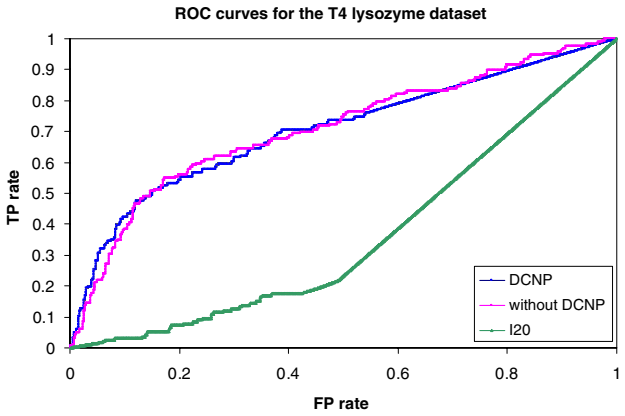

Fig. 2 ROC curves for the T4 lysozyme dataset. The area under the curve (*AUC*) for the Property Consensus with DCNP method (*blue line*, represented by DCNP in the legend) is 0.702. The AUC for Property Consensus without DCNP method (*pink line*, represented by "without DCNP" in the legend) is 0.706. The AUC for the I_{20} parameter set (*green line*, represented by "I20" in the legend) is 0.359

Analysis of the property consensus with DCNP method on T4 lysozyme mutations

In the study by (Rennell et al. 1991), each amino acid in the T4 lysozyme was substituted to 13 other amino acids. Table 2 lists all the T4 lysozyme residues where the Property Consensus with DCNP method predicted them to be deleterious, four or more times, for the 13 substitutions (Rennell et al. 1991).

These residues therefore represent amino acids that may play important structural and biochemical roles in the protein, concurring with data from published studies (Alber et al. 1987; Grutter et al. 1987; Matthews 1996; Goto et al. 2001). These include mutations at putative active sites (Glu 11, Tyr 18, Thr 26), sites creating loop turns (Gly 28, Gly 30, Gly 107), buried hydrophobic residues (Leu 7, Leu 99), polar exposed residues that may be involved in dipolar coupling (Thr 142) and which may affect the thermal-stability of the protein (Thr 151, Thr 155). A structural representation of these mutations is illustrated in Fig. 3.

Table 2 List of all the T4 lysozyme residues that the property consensus with DCNP method predicted to be deleterious, four or more times, for the 13 substitutions conducted by (Rennell et al. 1991)

Sites (Codon number)	Frequency of deleterious prediction	Role likely played by residue in T4 lysozyme
Leu 7	4	Fully buried hydrophobic residue
Glu 11	4	In active site cleft
Tyr 18	4	In active site cleft
Thr 26	5	In active site cleft
Gly 28	5	Conserved loop turn
Gly 30	6	Conserved loop turn
Leu 99	5	Cavity stabilization (Matthews 1996)
Gly 107	6	Conserved loop turn
Thr 142	6	Dipolar coupling with Thr 21 (Matsumura et al. 1989)
Thr 151	7	Temperature-sensitive stabilization (Grutter et al. 1987)
Thr 155	4	Temperature-sensitive stabilization (Grutter et al. 1987)

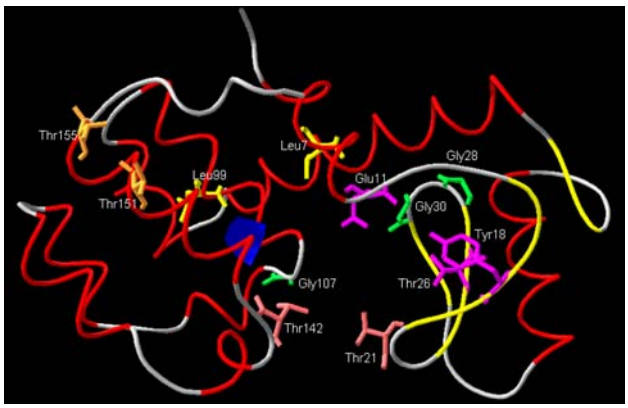

Fig. 3 Structure of T4 lysozyme (2LZM) in coil representation colored by secondary structure (*red* helices, *yellow* sheets). The SwissPDB-Viewer (Guex and Peitsch 1997) has been used to prepare these cartoons. Residues that the Property Consensus with DCNP approach predicted to be deleterious in more than 3 out of 13 mutations conducted by (Rennell et al. 1991) are represented. Conserved glycine residues are colored in green (28, 30, 107) forming loop turns, residues found in the active site cleft are colored in magenta (Glu 11, Tyr 18, Thr 26), buried leucine residues are in yellow (7, 99) with Leu 99 pointing into a cavity (represented as the blue globular structure), threonine residues at the entrance to the active site are in *light red* and threonine residues near the C-terminal are colored in *orange*

All the eleven sites from Table 2 are represented in Fig. 3. The probable importance of Thr 151 and Thr 155 in maintaining structural stability is illustrated in Fig. 4 with reference to (Alber et al. 1987; Grutter et al. 1987).

Analysis of various prediction methods on BRCA1 mutations

Twelve of the most frequent missense mutations listed in BIC between codons 225 and 1,365 were subjected to the various prediction methods (Table 3). The remaining eight most frequent missense mutations do not fall in this region of *Brca1*. The PolyPhen program predicted two possibly damaging (pD) and two probably damaging (D) mutations. The property consensus scores that are below 5.0 could be classified as intermediate (I) or possibly deleterious predictions. With such a classification, the prediction matches exactly with those predicted by PolyPhen (R1347G, R841W, Q356R, N550H). The prediction given by SIFT however, matches only marginally with PolyPhen or the Property Consensus with DCNP method (Tchernitchko et al. 2004).

Discussion

Comparison of the property consensus with DCNP method to other methods

Evaluation of the Property Consensus with DCNP method using the T4 lysozyme mutation data set ($n = 1{,}366$) had an overall accuracy of 87.6% when a threshold of 1.28 (0.08 for each of the 16 properties) was used (Table 1). There are limitations when comparing prediction methods based on percentages of false positives rates, false negatives rates and overall accuracy generated from different methods and mutational datasets. For example, if a dataset such as T4 lysozyme contains mutations that are largely tolerant with high degree of plasticity (Matthews 1996), a prediction method that has a tendency to over-predict tolerant mutations may still have a high overall accuracy. On the other hand, if the dataset went through a series of selection criteria to include only the important and well conserved residues, and when the prediction method tends to over-predict deleterious mutations, such a prediction will also have a relatively high overall accuracy.

Intuition favors the assumption that a mutation is most likely to be benign if it is found to be maintained in the same position in an alignment of orthologs. However, the Property Consensus with DCNP method did not produce a considerably better prediction than the Property Consensus method without DCNP (Fig. 2; both with AUC of 0.7). This phenomenon can be partially explained by the relatively high rate of co-evolution of residues within the bacteriophage sequence. For example, codon 10 of T4 lysozyme is an Aspartate residue and many other species has a Tyrosine maintained at codon 10. Hence, the Property Consensus with DCNP method will treat D10Y as a benign mutation. However, due to co-evolution of interacting

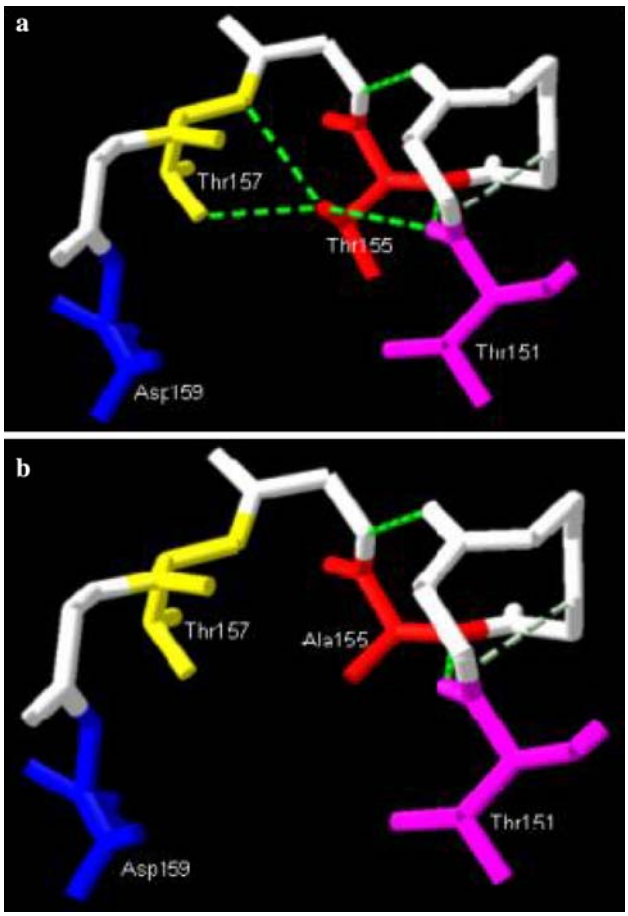

Fig. 4 Structure of residues 151–159 of 2LZM illustrating the effect of a T155A on hydrogen-bonding. **a** Residues 151–159 of wild type T4 lysozyme (PDB code: 2LZM) (Weaver and Matthews 1987). The gamma-hydroxyl group of Thr 155 (*red*) forms a hydrogen bond with Thr 157 (*yellow*). Asp 159 is colored in *blue* and Thr 151 in *magenta*. **b** Residues 151–159 of the double-mutant (T155A, T157I) T4 lysozyme (PDB code: 1L10). A mutation of Thr 155 to Ala 155 results in loss of hydrogen-bonding between residues 151 and 157

residues, where the Aspartate at codon 10 has an association with the Tyrosine residue at codon 161 (an Aspartate-Tyrosine association), the other species have Glycine maintained at 161 (i.e., a Tyrosine–Glycine association). So, in this case, a single mutation from Aspartate to Tyrosine (Tyrosine–Tyrosine association) will result in steric hinderance of two bulky side-chain residues that will cause disturbance to the structure of T4 lysozyme. Such instances will increase the rate of false negative of the prediction.

The I_{20} parameter prediction initially proposed by (Santibanez Koref et al. 2003), which did not consider amino acid properties, produced an overall accuracy of 44.6% and AUC of 0.359. In the I_{20} parameter prediction, a conservative substitution from an Aspartate to Glutamate is considered equally benign or equally deleterious when compared to a non-conservative substitution from an Aspartate to Tryptophan, independent of the position of the residue. Hence, the relatively high false positive rates could be due to its tendency of predicting a benign substitution as a deleterious substitution. In the study by (Santibanez Koref et al. 2003) on *TP53*, the assessment of the performance of the I_{20} parameter was based on a set of controls (defined as amino acid residues that could be found in the corresponding position of the aligned ortholog sequences) and mutants [defined as missense mutation found in the *TP53* database (Beroud and Soussi 1998) but were not found in the corresponding position of the aligned ortholog sequences]. From our analysis, we found that this assumption, which their study took as the set of controls, does not always hold true.

Moreover, other studies (Rennell et al. 1991; Matthews 1996) have shown that not all missense mutations, which do not fit their definition of controls, would truly affect the function of a protein. Hence, the relatively good performance of the I_{20} parameter in their study on *TP53* may, in part, be due to the possibility that *TP53* may be more sensitive to mutational changes as compared to proteins like T4 lysozyme that exhibits a higher tolerance to mutational effects (Rennell et al. 1991; Matthews 1996). However, the relatively good performance of the I_{20} parameter may also be due to their definition of deleterious mutants which, again, would have the tendency of treating benign mutations as deleterious.

When the Property Consensus with DCNP method was compared to the SIFT method (Ng and Henikoff 2001), it had a higher overall accuracy and comparable false positive and false negative rates. Although our prediction method could be applied to all proteins with sufficient number of orthologous sequences, its accuracy has not been proven in distant proteins of a different function or structural stability. However, these evolutionary-based prediction methods may be expected to predict mutations with a similar level of accuracy when they are applied to subfamilies of proteins with similarity in their structure and function.

Structural and biochemical evidence that support the property consensus with DCNP method's predictions of functionally important residues in T4 lysozyme

Table 2 lists all the residues in T4 lysozyme that the Property Consensus with DCNP method predicted to be deleterious, four or more times, for the 13 substitutions conducted by (Rennell et al. 1991). Most of these sites have been known for their structural and functional importance. The T4 lysozyme family has an important triad in the active site comprising of Glu 11, Asp 20 and Thr 26 (Kuroki et al. 1993). Included in the list in Table 2 was Glu 11 and Thr 26 but not Asp 20, which may be the least important residue in the triad as studies have shown that

Table 3 Analysis of the 12 most common missense mutations reported in breast cancer information core (BIC), in the region of codon 225 to 1,365, in descending order of frequency as recorded in BIC

Missense mutation	Frequency in BIC	SIFT (0.05)	DCNP	Property consensus with DCNP (this study)[a]	PolyPhen
R1347G	149	0.63 B	0 D	1.597 I	2.073 D
R841W	114	0.00 D	0 D	3.999 I	1.618 pD
M1008I	105	0.65 B	1 B	B	0.410 B
R496H	69	1.00 B	1 B	B	0.468 B
L246V	62	0.01 D	0 D	5.413 B	1.210 B
Q356R	56	0.08 B	0 D	3.872 I	2.007 D
V772A	42	0.02 D	1 B	B	0.882 B
F486L	37	0.37 B	1 B	B	0.815 B
E1038G	36	0.04 D	1 B	B	1.018 B
N550H	35	0.02 D	0 D	3.078 I	1.570 pD
R496C	33	0.31 B	1 B	B	0.604 B
K1183R	32	1.00 B	1 B	B	0.008 B

[a] Property consensus with DCNP predictions first classifies benign mutations by the DCNP criterion which treats mutations that were found in orthologous sequences as benign. The Property Consensus scores were conducted for mutations that were predicted to be deleterious by the DCNP criterion. Those that score below 1.28 are considered as deleterious. Mutations that score above 1.28 but below 5.0 were considered as possibly deleterious (or intermediate) while those above 5 were classified as benign. With this system of intermediate classification, the property consensus with DCNP predictions matches well with the predictions made by PolyPhen

B Benign; *pD* possibly deleterious (analogous to an intermediate prediction); *D* probably deleterious; *DCNP* deleterious change not present; *SIFT* sorting intolerant from tolerant; *PolyPhen* polymorphism phenotyping

catalytic activities could still be carried out when Asp 20 is replaced by some other amino acids (Rennell et al. 1991; Kuroki et al. 1999). All 12 substitutions of Glu 11 were deleterious and 11 out of the 13 substitutions of Thr 26 were deleterious, while in contrast, only 9 out of 13 substitution of Asp 20 were deleterious (Rennell et al. 1991). Tyr 18, which is also found in the active site cleft of T4 lysozyme, is, however, replaceable by several other amino acids. Tyr 18 was included in the list in Table 2 as one of the important residues because it is conserved in 31 out of the 33 organisms used in this study. Thus, the Property Consensus with DCNP method was able to accurately predict biochemically important residues in T4 lysozyme.

Gly 28, Gly 30 and Gly 107 are well-conserved residues responsible for the loop turns as illustrated in Fig. 3. Leu 7 and Leu 99 are buried hydrophobic amino acids and their substitutions by hydrophobic residues will most likely disrupt its structure. Leu 99 extends into a relatively large cavity within the lysozyme structure (Fig. 3). Substitutions on Leu 99 that produce an even larger cavity within the protein tend to be very destabilizing (Matthews 1996). The list in Table 2 also shows Thr 142 to be an important site.

Replacement of Thr 21 and Thr 142 to a disulphide linkage abolishes the activity of T4 lysozyme, because this link bridges across the active-site cleft, preventing the substrate from entering the active site of the enzyme (Matsumura et al. 1989). Hence, substitutions on Thr 21 or Thr 142 by charged or bulky residues may cause lysozyme to lose its catalytic activity. The experimental study by (Rennell et al. 1991) showed that substitution of Thr 142 by an Arg or Lys had a deleterious effect on the enzyme activity. Substitutions of Thr 151 and 155 disrupt part of a network of hydrogen bonds (surrounding residue Thr 157), which may result in unsatisfactory hydrogen-bonding potential that could possibly contribute to a reduction of stability (Table 2, Fig. 4). (Grutter et al. 1987) showed experimentally that disruptions to this network of hydrogen bonds were indirectly responsible for reducing the temperature of the midpoint of the reversible thermal denaturation transition by 11°C. The identification of these residues demonstrates that the Property Consensus with DCNP method was able to accurately predict structurally important residues in T4 lysozyme.

A sequence-based evolutionary approach such as the Property Consensus with the DCNP method can produce structural inferences because amino acid properties used in the study contain parameters that directly or indirectly affect structural stability. The ability of this method to sieve out structurally and experimentally important residues from a multitude of substitutions shows that this approach may be potentially useful in the discovery of biochemical and structurally important residues.

Comparison of predictions for twelve frequently occurring BRCA1 missense mutations

Previous structure-based studies on missense mutations in *BRCA1* have looked at the BRCT domains (codon region

1,640–1,729, 1,760–1,821) at the C-terminal of the protein (Williams and Glover 2003; Mirkovic et al. 2004). However, there are only a few orthologous sequences available in this region and hence this region is not viable for a sequence-comparison method such as the approach used in this study. Hence, this study can only examine the region 225–1,365 where more orthologous sequences can be found. As such, our *BRCA1* predictions can only be compared to prediction methods that can examine missense mutations within this region.

Missense mutations that occur frequently may play a significant role in disease penetrance in a given population. Of the 20 most frequently occurring missense mutations recorded in the BIC *BRCA1* database, 12 of them fall in the *BRCA1* 225–1,365 region and were analyzed with various approaches (Table 3). SIFT predicted 5 deleterious mutations (D) out of the 12 missense mutations. Only 2 of these predicted deleterious mutations overlapped with the deleterious mutations predicted by PolyPhen and the Property Consensus with DCNP method. On the other hand, all of PolyPhen's 4 possibly deleterious (pD) and D mutations were also predicted as being possibly deleterious by Property Consensus with DCNP.

Discrepancies between SIFT and PolyPhen have previously been documented, and it has been suggested that PolyPhen may perform better because, in addition to sequence information, it incorporates structural information of the protein in its prediction (Tchernitchko et al. 2004; Johnson et al. 2005). Notably, although PolyPhen's methodology differs from that of the Property Consensus with DCNP method, their prediction results were similar. This similarity in the prediction between PolyPhen and Property Consensus with DCNP could partially be attributed to the latter's incorporation of structurally related amino acid properties. Hence, results from our study are in line with previous inferences that PolyPhen may outperform SIFT in predicting the phenotype of nsSNPs. PolyPhen (polymorphism phenotyping) developed by (Ramensky et al. 2002) used empirically derived rules to assess a set of analysis that calculates the PSIC (position specific independent counts) (Sunyaev et al. 1999) profile scores for two amino acid variants. It then combines information on sequence features, the structural parameters and contacts to characterize the substitution. In line with results from previous studies (Barker et al. 1996; Durocher et al. 1996; Dong et al. 1998; Petersen et al. 1998; Fleming et al. 2003), our results also show that the *BRCA1* missense mutations R1347G, R841W and Q356R with a population frequencies of 4.3, 3.2 and 2.3%, respectively, are candidate mutations that may play important roles in assessing cancer susceptibility in a population.

Factors that affect the performance of the property consensus with DCNP method

We highlight here some of the factors that may affect the accuracy of prediction by the Property Consensus with DCNP method. Firstly, the level of confidence of its prediction will largely rely on the availability of sequences in other species. Regions where only few hits could be obtained from homology searches (in this case, the region before codon 225 and after 1,365 of *BRCA1*) will decrease the level of confidence in the prediction. However, as more orthologous genes and genomes are being sequenced, this limitation will eventually be eliminated. Secondly, changing the assumed rate of mutation by one or two orders does not substantially affect the prediction (Santibanez Koref et al. 2003). In the calculation of the predictive score, the degree of conservation between orthologs is considered as a more important factor than the degree of conservation between the controls (i.e., the simulated sequences representing genes that do not evolve under selective pressure). Thirdly, depending on the function of the protein (e.g., structural protein or enzymatic protein) and the site of the mutation in the three-dimensional conformation of the protein, the importance of amino acid properties will vary. For example, we may expect the hydrophobic character of a residue to be more important for a buried site and less important at the protein surface. Giving equal weight to all the amino acid properties may not adequately represent their in vivo conditions. Fourthly, co-evolution of binding partners or acquisition of compensatory mutations (Kondrashov et al. 2002) is also likely to increase with divergence times and higher mutation rates. Hence, the DCNP criterion may not be a good measure to sieve out benign phenotypes especially for distant orthologs or sequences with high mutation rates. Fifthly, this study has compared predictions only with the experimentally verified T4 lysozyme substitutions. Thus, its prediction accuracy for a protein from different functional or structural subclasses is not known. However, it may still be used to retrieve preliminary information of a protein but should not be used in the absence of other tests or arguments to reach conclusions for clinical assessment.

Conclusions and future work

This study shows evidence that the Property Consensus approach has the potential to provide relatively accurate predictions in phenotyping missense mutations. The use of an orthogonal set of amino acid properties (Kidera et al. 1985a, b) instead of choosing some important properties or using all of the 500 available amino acid properties (Kawashima et al. 1999) may more closely model

functionally important aspects of amino acid residues. In addition, different weights could be assigned to different properties by conducting a sequence-based characterization prior to the prediction process. For example, trans-membrane, signal region, or binding, active site predictions, based on sequence-based characterization at the region surrounding the site of substitution could be conducted, as in the case of PolyPhen. Statistical methods that measure the relative importance of various amino acid properties at a given site of substitution could also be explored. More experimentally verified data from diverse sets of proteins are necessary to understand the relative importance of this orthogonal set of amino acid properties in their respective characteristic sites.

In summary, we have shown that the use of amino acid properties (Kawashima et al. 1999) has good potential in improving the measurement of the degree of conservation, and hence the degree of importance of a point missense mutation. It may also be used by scientists who want to identify functionally important residues in a protein. Such functional prediction methods may contribute to the improvement of functional annotation of mutational and SNP databases and ultimately may improve our understanding of the role of disease susceptibility and various SNP-related complex diseases.

Acknowledgments This work was supported in part by the National Medical Research Council, Singapore and the Biomedical Research Council of A*STAR, Singapore.

References

Alber T, Sun DP, Wilson K, Wozniak JA, Cook SP, Matthews BW (1987) Contributions of hydrogen bonds of Thr 157 to the thermodynamic stability of phage T4 lysozyme. Nature 330:41–46

Barker DF, Almeida ER, Casey G, Fain PR, Liao SY, Masunaka I, Noble B, Kurosaki T, Anton-Culver H (1996) BRCA1 R841W: a strong candidate for a common mutation with moderate phenotype. Genet Epidemiol 13:595–604

Beroud C, Soussi T (1998) p53 gene mutation: software and database. Nucleic Acids Res 26:200–204

Cai Z, Tsung EF, Marinescu VD, Ramoni MF, Riva A, Kohane IS (2004) Bayesian approach to discovering pathogenic SNPs in conserved protein domains. Hum Mutat 24:178–184

Charton M, Charton BI (1983) The dependence of the Chou–Fasman parameters on amino acid side chain structure. J Theor Biol 102:121–134

Chasman D, Adams RM (2001) Predicting the functional consequences of non-synonymous single nucleotide polymorphisms: structure-based assessment of amino acid variation. J Mol Biol 307:683–706

Chou KC (1993) A vectorized sequence-coupling model for predicting HIV protease cleavage sites in proteins. J Biol Chem 268:16938–16948

Chou KC (1996) Prediction of human immunodeficiency virus protease cleavage sites in proteins. Anal Biochem 233:1–14

Chou KC (2004) Structural bioinformatics and its impact to biomedical science. Curr Med Chem 11:2105–2134

Chou KC (2006) Structural bioinformatics and its impact to biomedical science and drug discovery. In: Atta-ur-Rahman, Reitz AB (eds) Frontiers in medicinal chemistry, vol 3. Bentham Science Publishers, Bussum, The Netherlands, pp 455–502

Chou KC, Wei DQ, Du QS, Sirois S, Zhong WZ (2006) Progress in computational approach to drug development against SARS. Curr Med Chem 13:3263–3270

Chou KC, Wei DQ, Zhong WZ (2003) Binding mechanism of coronavirus main proteinase with ligands and its implication to drug design against SARS. Biochem Biophys Res Commun 308:148–151

Chou PY, Fasman GD (1978) Prediction of the secondary structure of proteins from their amino acid sequence. Adv Enzymol Relat Areas Mol Biol 47:45–148

Dong J, Chang-Claude J, Wu Y, Schumacher V, Debatin I, Tonin P, Royer-Pokora B (1998) A high proportion of mutations in the BRCA1 gene in German breast/ovarian cancer families with clustering of mutations in the 3′ third of the gene. Hum Genet 103:154–161

Du Q, Wang S, Wei D, Sirois S, Chou KC (2005a) Molecular modeling and chemical modification for finding peptide inhibitor against severe acute respiratory syndrome coronavirus main proteinase. Anal Biochem 337:262–270

Du QS, Wang SQ, Chou KC (2007) Analogue inhibitors by modifying oseltamivir based on the crystal neuraminidase structure for treating drug-resistant H5N1 virus. Biochem Biophys Res Commun 362:525–531

Du QS, Wang SQ, Jiang ZQ, Gao WN, Li Y, Wei DQ, Chou KC (2005b) Application of bioinformatics in search for cleavable peptides of SARSCoV Mpro and chemical modification of octapeptides. Med Chem 1:209–213

Du QS, Wang SQ, Zhu Y, Wei DQ, Guo H, Sirois S, Chou KC (2004) Polyprotein cleavage mechanism of SARS CoV Mpro and chemical modification of the octapeptide. Peptides 25:1857–1864

Durocher F, Shattuck-Eidens D, McClure M, Labrie F, Skolnick MH, Goldgar DE, Simard J (1996) Comparison of BRCA1 polymorphisms, rare sequence variants and/or missense mutations in unaffected and breast/ovarian cancer populations. Hum Mol Genet 5:835–842

Fasman GD (1976). Handbook of biochemistry and molecular biology. CRC, Cleveland

Fauchere JL, Charton M, Kier LB, Verloop A, Pliska V (1988) Amino acid side chain parameters for correlation studies in biology and pharmacology. Int J Pept Protein Res 32:269–278

Fleming MA, Potter JD, Ramirez CJ, Ostrander GK, Ostrander EA (2003) Understanding missense mutations in the BRCA1 gene: an evolutionary approach. Proc Natl Acad Sci USA 100:1151–1156

Gao WN, Wei DQ, Li Y, Gao H, Xu WR, Li AX, Chou KC (2007) Agaritine and its derivatives are potential inhibitors against HIV proteases. Med Chem 3:221–226

Goto NK, Skrynnikov NR, Dahlquist FW, Kay LE (2001) What is the average conformation of bacteriophage T4 lysozyme in solution? A domain orientation study using dipolar couplings measured by solution NMR. J Mol Biol 308:745–764

Grantham R (1974) Amino acid difference formula to help explain protein evolution. Science 185:862–864

Grutter MG, Gray TM, Weaver LH, Wilson TA, Matthews BW (1987) Structural studies of mutants of the lysozyme of bacteriophage T4. The temperature-sensitive mutant protein Thr157-Ile. J Mol Biol 197:315–329

Guex N, Peitsch MC (1997) SWISS-MODEL and the Swiss-PdbViewer: an environment for comparative protein modeling. Electrophoresis 18:2714–2723

Herrgard S, Cammer SA, Hoffman BT, Knutson S, Gallina M, Speir JA, Fetrow JS, Baxter SM (2003) Prediction of deleterious

functional effects of amino acid mutations using a library of structure-based function descriptors. Proteins 53:806–816
Jeanmougin F, Thompson JD, Gouy M, Higgins DG, Gibson TJ (1998) Multiple sequence alignment with Clustal X. Trends Biochem Sci 23:403–405
Johnson MM, Houck J, Chen C (2005) Screening for deleterious nonsynonymous single-nucleotide polymorphisms in genes involved in steroid hormone metabolism and response. Cancer Epidemiol Biomarkers Prev 14:1326–1329
Kawashima S, Ogata H, Kanehisa M (1999) AAindex: amino acid index database. Nucleic Acids Res 27:368–369
Kidera A, Konishi Y, Oka M, Ooi T, Scheraga HA (1985a) Statistical analysis of the physical properties of the 20 naturally occurring amino acids. J Protein Chem 4:23–55
Kidera A, Konishi Y, Ooi T, Scheraga HA (1985b) Relation between sequence similarity and structural similarity in proteins. Role of important properties of amino acids. J Protein Chem 4:265–297
Klein P, Kanehisa M, DeLisi C (1984) Prediction of protein function from sequence properties. Discriminant analysis of a data base. Biochim Biophys Acta 787:221–226
Kondrashov AS, Sunyaev S, Kondrashov FA (2002) Dobzhansky–Muller incompatibilities in protein evolution. Proc Natl Acad Sci USA 99:14878–14883
Krishnan VG, Westhead DR (2003) A comparative study of machine-learning methods to predict the effects of single nucleotide polymorphisms on protein function. Bioinformatics 19:2199–2209
Kuhn LA, Swanson CA, Pique ME, Tainer JA, Getzoff ED (1995) Atomic and residue hydrophilicity in the context of folded protein structures. Proteins 23:536–547
Kumar S, Hedges SB (1998) A molecular timescale for vertebrate evolution. Nature 392:917–920
Kuroki R, Weaver LH, Matthews BW (1993) A covalent enzyme-substrate intermediate with saccharide distortion in a mutant T4 lysozyme. Science 262:2030–2033
Kuroki R, Weaver LH, Matthews BW (1999) Structural basis of the conversion of T4 lysozyme into a transglycosidase by reengineering the active site. Proc Natl Acad Sci USA 96:8949–8954
Kyte J, Doolittle RF (1982) A simple method for displaying the hydropathic character of a protein. J Mol Biol 157:105–132
Lau AY, Chasman DI (2004) Functional classification of proteins and protein variants. Proc Natl Acad Sci USA 101:6576–6581
Li L, Wei DQ, Wang JF, Chou KC (2007a) Computational studies of the binding mechanism of calmodulin with chrysin. Biochem Biophys Res Commun 358:1102–1107
Li Y, Wei DQ, Gao WN, Gao H, Liu BN, Huang CJ, Xu WR, Liu DK, Chen HF, Chou KC (2007b) Computational approach to drug design for oxazolidinones as antibacterial agents. Med Chem 3:576–582
Matsumura M, Wozniak JA, Sun DP, Matthews BW (1989) Structural studies of mutants of T4 lysozyme that alter hydrophobic stabilization. J Biol Chem 264:16059–16066
Matthews BW (1996) Structural and genetic analysis of the folding and function of T4 lysozyme. Faseb J 10:35–41
Mirkovic N, Marti-Renom MA, Weber BL, Sali A, Monteiro AN (2004) Structure-based assessment of missense mutations in human BRCA1: implications for breast and ovarian cancer predisposition. Cancer Res 64:3790–3797
Mooney SD, Altman RB (2003) MutDB: annotating human variation with functionally relevant data. Bioinformatics 19:1858–1860
Ng PC, Henikoff S (2001) Predicting deleterious amino acid substitutions. Genome Res 11:863–874
Ng PC, Henikoff S (2002) Accounting for human polymorphisms predicted to affect protein function. Genome Res 12:436–446
Ng PC, Henikoff S (2003) SIFT: predicting amino acid changes that affect protein function. Nucleic Acids Res 31:3812–3814
Ochman H, Elwyn S, Moran NA (1999) Calibrating bacterial evolution. Proc Natl Acad Sci USA 96:12638–12643
Petersen GM, Parmigiani G, Thomas D (1998) Missense mutations in disease genes: a Bayesian approach to evaluate causality. Am J Hum Genet 62:1516–1524
Radzicka A, Pedersen L, Wolfenden R (1988) Influences of solvent water on protein folding: free energies of solvation of cis and trans peptides are nearly identical. Biochemistry 27:4538–4541
Ramensky V, Bork P, Sunyaev S (2002) Human non-synonymous SNPs: server and survey. Nucleic Acids Res 30:3894–3900
Rennell D, Bouvier SE, Hardy LW, Poteete AR (1991) Systematic mutation of bacteriophage T4 lysozyme. J Mol Biol 222:67–88
Sanderson MJ (2003) r8s: inferring absolute rates of molecular evolution and divergence times in the absence of a molecular clock. Bioinformatics 19:301–302
Santibanez Koref MF, Gangeswaran R, Santibanez Koref IP, Shanahan N, Hancock JM (2003) A phylogenetic approach to assessing the significance of missense mutations in disease genes. Hum Mutat 22:51–58
Sneath PH (1966) Relations between chemical structure and biological activity in peptides. J Theor Biol 12:157–195
Sunyaev SR, Eisenhaber F, Rodchenkov IV, Eisenhaber B, Tumanyan VG, Kuznetsov EN (1999) PSIC: profile extraction from sequence alignments with position-specific counts of independent observations. Protein Eng 12:387–394
Szabo C, Masiello A, Ryan JF, Brody LC (2000) The breast cancer information core: database design, structure, and scope. Hum Mutat 16:123–131
Tchernitchko D, Goossens M, Wajcman H (2004) In silico prediction of the deleterious effect of a mutation: proceed with caution in clinical genetics. Clin Chem 50:1974–1978
Verzilli CJ, Stallard N, Whittaker JC (2005) Bayesian modelling of multivariate quantitative traits using seemingly unrelated regressions. Genet Epidemiol 28:313–325
Wang JF, Wei DQ, Li L, Zheng SY, Li YX, Chou KC (2007a) 3D structure modeling of cytochrome P450 2C19 and its implication for personalized drug design. Biochem Biophys Res Commun 355:513–519
Wang JF, Wei DQ, Lin Y, Wang YH, Du HL, Li YX, Chou KC (2007b) Insights from modeling the 3D structure of NAD(P)H-dependent D-xylose reductase of Pichia stipitis and its binding interactions with NAD and NADP. Biochem Biophys Res Commun 359:323–329
Wang SQ, Du QS, Chou KC (2007c) Study of drug resistance of chicken influenza A virus (H5N1) from homology-modeled 3D structures of neuraminidases. Biochem Biophys Res Commun 354:634–640
Wang SQ, Du QS, Zhao K, Li AX, Wei DQ, Chou KC (2007d) Virtual screening for finding natural inhibitor against cathepsin-L for SARS therapy. Amino Acids 33:129–135
Weaver LH, Matthews BW (1987) Structure of bacteriophage T4 lysozyme refined at 1.7 A resolution. J Mol Biol 193:189–199
Wei DQ, Du QS, Sun H, Chou KC (2006a) Insights from modeling the 3D structure of H5N1 influenza virus neuraminidase and its binding interactions with ligands. Biochem Biophys Res Commun 344:1048–1055
Wei DQ, Sirois S, Du QS, Arias HR, Chou KC (2005) Theoretical studies of Alzheimer's disease drug candidate 3-[(2,4-dimethoxy)benzylidene]-anabaseine (GTS-21) and its derivatives. Biochem Biophys Res Commun 338:1059–1064
Wei DQ, Zhang R, Du QS, Gao WN, Li Y, Gao H, Wang SQ, Zhang X, Li AX, Sirois S, Chou KC (2006b) Anti-SARS drug screening by molecular docking. Amino Acids 31:73–80
Wei H, Zhang R, Wang C, Zheng H, Li A, Chou KC, Wei DQ (2007) Molecular insights of SAH enzyme catalysis and implication for inhibitor design. J Theor Biol 244:692–702

Williams RS, Glover JN (2003) Structural consequences of a cancer-causing BRCA1-BRCT missense mutation. J Biol Chem 278:2630–2635

Wu G, Yan S (2007) Prediction of mutations engineered by randomness in H5N1 hemagglutinins of influenza A virus. Amino Acids

Xiao X, Shao S, Ding Y, Huang Z, Chen X, Chou KC (2005) An application of gene comparative image for predicting the effect on replication ratio by HBV virus gene missense mutation. J Theor Biol 235:555–565

Zhang R, Wei DQ, Du QS, Chou KC (2006) Molecular modeling studies of peptide drug candidates against SARS. Med Chem 2:309–314

Zheng H, Wei DQ, Zhang R, Wang C, Wei H, Chou KC (2007) Screening for new agonists against Alzheimer's disease. Med Chem 3:488–493

Zimmerman JM, Eliezer N, Simha R (1968) The characterization of amino acid sequences in proteins by statistical methods. J Theor Biol 21:170–201