J Hum Genet (2008) 53:379–389
DOI 10.1007/s10038-008-0260-8

ORIGINAL ARTICLE

Prediction of functional nonsynonymous single nucleotide polymorphisms in human G-protein-coupled receptors

**Dan Xue · Jingyuan Yin ·
Mingfeng Tan · Junjie Yue ·
Yuelan Wang · Long Liang**

Received: 16 October 2007 / Accepted: 24 January 2008 / Published online: 26 February 2008

Abstract G-protein-coupled receptors (GPCRs) are found in a wide range of organisms and are central to a cellular signaling network that regulates many basic physiological processes. GPCRs are the focus of a significant amount of current pharmaceutical research because they play a key role in many diseases. In this paper, we predict the functional nonsynonymous single nucleotide polymorphisms (nsSNPs) in human GPCRs by defining optimal attributes and using a decision tree method. The predictive power of each attribute was evaluated. A subset of sequences with optimal attributes was obtained using the decision tree method combined with a genetic search algorithm. The subset contains both sequence-based and structure-based information, and the information for each subset consists of a conservation score, the location of the mutation, the BLOSUM62 substitution matrix score, as well as the hydrophobicity change, the solvent accessibility, and the buried charge. Seven important rules were derived from the decision tree. A total of 166 functional nsSNPs in human GPCRs from the dbSNP have been predicted using the optimal attributes subset.

D. Xue
College of Communication and Information Engineering,
Shanghai University, 200072 Shanghai, China

D. Xue · M. Tan · J. Yue · Y. Wang · L. Liang (✉)
Institute of Biotechnology, Academy of Military
Medical Sciences, 100071 Beijing, China
e-mail: ll@bioinflab.org

J. Yin
School of Life Sciences, Shanghai University,
200444 Shanghai, China

Keywords GPCRs · Functional nsSNPs · Optimal attributes · Decision tree · Bioinformatics

Introduction

G-protein-coupled receptors (GPCRs) are the largest protein superfamily in most mammalian genomes. Despite their great diversity in terms of sequence composition, all GPCRs share a common protein structure. An N-terminal extracellular domain of variable length is followed by seven transmembrane (TM) helices, connected by three intracellular loops (ICL) and three extracellular loops (ECL), one of which terminates in a C-terminal intracellular domain (Gether 2000). These receptors are plasma membrane-bound and can respond to a large number of extracellular signals from nucleotides, peptides, amines and hormones (Sakmar 1998). Upon recognition of these ligands, GPCRs act through G proteins in signaling pathways that influence almost all physiological functions. As such, pharmacologic agents that serve to antagonize GPCR-mediated signaling are common. Actually, more than one-third of all known small-molecule drugs are targeted at GPCRs (Marinissen and Gutkind 2001; Howard et al. 2001). Thus, small genomic-level differences in GPCRs may explain the different drug-response behaviors of different individuals toward a drug, and they can be used to tailor drugs based on an individual's genetic makeup (Drysdale et al. 2000; Phillips et al. 2001; Roses 2004).

Single nucleotide polymorphisms (SNPs) are defined as single base variations in sequence that occur at a frequency of at least 1% and may directly explain the pathogenesis of disease. SNPs in protein-coding exons are classified as synonymous or nonsynonymous (nsSNPs) according to whether or not they alter the protein sequence. Some nsSNPs can affect gene function through their effects on the structure or function of the encoded protein. Recently, several studies have tried to investigate how to determine whether nsSNPs are either functional or neutral using protein sequence and structural information. Empirical rules identifying detrimental nsSNPs were derived based on structural information (Wang and Moult 2001). An algorithm named SIFT, which is based on the sequence conservation and scores from position-specific scoring matrices, was developed to rationalize amino acid changes that were likely to affect the function of a protein (Ng and Henikoff 2001). Structural and sequence information was then combined and used to predict functional nsSNPs (Chasman and Adams 2001; Sunyaev et al. 2001; Ramensky et al. 2002; Saunders and Baker 2002; Krishnan and Westhead 2003; Bao and Cui 2005); however, bovine rhodopsin (BR) is the only three-dimensional structure of a GPCR that has been resolved (Palczewski et al. 2000). Balasubramanian et al. (2005) predicted disease-causing nsSNPs in GPCRs based on sequence information using logistic regression methods. However, some nsSNPs will influence protein function but will not cause inherited diseases. Here, we aimed to predict the functional nsSNPs in human GPCRs from dbSNP based on the optimal sequence and structural information using a decision tree method.

Materials and methods

Datasets

The GPCRDB has an extensive collection of point mutations that have been compiled from the literature using the MuteXt automated extraction method (Horn et al. 2003, 2004). We analyzed the functions of these mutated residues and collected those mutations that changed receptor function or structure in order to include them in our training set.

To derive a dataset of neutral mutations, proteins which have >90% sequence identity to the target GPCRs were extracted from GPCRDB. For each target protein, only one ortholog was chosen from each species based on the best match to the target protein, and multiple sequence alignments (MSA) were performed to the target protein and its homologs. Amino acid variations at any position in the MSA were considered to be neutral variations (Sunyaev et al. 2001; Balasubramanian et al. 2005). The logic behind this assumption is that variations in highly homologous sequences between species are generally neutral and are highly unlikely to be deleterious, because detrimental changes will be selectively removed during the course of evolution. There may be examples, however, where some are functional changes in one species but not in the others. In total, the training dataset contained 750 functional mutations and 1,345 neutral changes in 72 receptors.

The nsSNPs in human GPCRs with known ligands were extracted from dbSNP. First, the corresponding gene of each human GPCR was found in Swiss-Prot, and then, according to the name of each gene, nsSNPs were searched for in dbSNP. The opsins, olfactory and taste receptors are excluded since they are not drug targets. In all, 519 nsSNPs were identified.

Attributes

The sequence-based and structure-based attributes of amino acid polymorphisms that may serve as generalized predictors of effects on function were chosen from the literature. The sequence-based attributes that were used in the prediction of functional nsSNPs were the sequence conservation score at the mutated position, the physiochemical changes (mass, hydrophobicity, volume) between the wild-type residues and mutated residues, and substitution matrix scores, such as BLOSUM62 and PAM120 matrices. Three structure-based attributes, including the location of the mutation (whether or not it was in the TM regions), the solvent accessibility, and the buried charge were also considered.

The sequence conservation score was calculated in two steps using the software program AL2CO (Pei and Grishin 2001). First, an independent count-based weighting scheme was used to estimate the amino acid frequencies. The conservation score was then calculated from these frequencies based on an entropy-based method (Shannon 1948). The MSA files of the subtype families containing the target proteins were extracted from the GPCRDB. Each family has a different number of receptors, and as the number of sequences in a family varies, the level of conservation of each position changes, and thus the average conservation score changes (Armon et al. 2001). In order to diminish this effect, the conservation score was normalized to its z-score function, which was calculated by subtracting the mean conservation score from the conservation score and dividing by the standard deviation.

The hydrophobicity of the amino acids was evaluated using the Kyte–Doolittle hydrophobicity scale (Kyte and Doolittle 1982). Average residue volumes were taken from Harpaz et al. (1994). The mass, hydrophobicity and volume changes were the absolute value of the difference between the wild-type residue and the mutated residue. The TM regions of a protein were taken from the Swiss-Prot database entry for each protein. If no information was available, the TMHMM program was used to predict the TM regions of the receptors (Krogh et al. 2001). The location of the wild-type residue in the TM regions was 1; otherwise it was 0. The solvent accessibility was predicted by PHD (Rost and Sander 1993, 1994). Relative accessibility was grouped into three states: buried (<9%), intermediate (9–36%) and exposed (≥36%). The wild-type residue was deemed to be a buried charge if it was K, R, D, E or H and its solvent accessibility was in the buried state (Krishnan and Westhead 2003).

Decision tree

Decision tree learning is a means of approximating discrete-valued target functions in which the learned function is represented by a decision tree. It has been shown to perform well in homogeneous cross-validated training datasets (Krishnan and Westhead 2003). Here we used the C4.5 decision tree algorithm developed by Quinlan (1993). It was performed as a J48 decision tree classifier using a Weka machine learning workbench (Witten and Frank 2000; Frank et al. 2004). The default set of parameters and tenfold cross-validation were used in the predictions. The decision tree not only provides a prediction but also yields an estimate of the probability that a prediction from the rule is correct. Each rule was derived from the training dataset, and the estimated accuracy was used to assign a confidence level to the prediction. Rules with estimated accuracies of x% were taken to have a confidence level of x/100. Another measurement of a rule was "cover," which was the number of mutations conforming to the rule in the training dataset. If the cover of a rule was too small, it meant that only a few mutations in the training dataset met this rule, and so the rule had no representative meaning. In this paper, we used 30 as the cover threshold.

Optimize attributes set

The attributes mentioned above have been proven to be related to functional mutations. Combining of all those attributes may result in redundant descriptions of each polymorphism and cause a reduction in prediction quality (Dobson et al. 2006). Therefore, attributes selection was an indispensable step before prediction. Here, optimization means finding the best combination of attributes that maximizes the prediction accuracy. The optimized attributes subset was obtained using wrapper-based attribute selection with J48 as the learning method combined with the genetic search method with default option settings. The genetic search algorithm was initialized with a population size of 20 and then 50 generations were evaluated.

Evaluation of the prediction accuracy

The mutations are classified into "effect" or "no effect." Mutations in the "effect" class will influence the structure or function of the protein, which means that these are functional mutations. Because the training dataset contained more neutral mutations than functional mutations, Matthew's correlation coefficient (MCC) was used to evaluate the performance (Matthews 1985):

$$\mathrm{MCC} = \frac{(\mathrm{TP} \cdot \mathrm{TN} - \mathrm{FP} \cdot \mathrm{FN})}{\sqrt{(\mathrm{TN} + \mathrm{FN})(\mathrm{TN} + \mathrm{FP})(\mathrm{TP} + \mathrm{FN})(\mathrm{TP} + \mathrm{FP})}}$$

where TP is true positives, FN is false negatives, TN is true negatives and FP is false positives. When there is an obvious disparity in the number of positive samples and negative samples, MCC is usually a better evaluation criterion than the overall accuracy. MCC combines both sensitivity and specificity into one measure and the values lie in the range of −1 to 1. A value of 1 means complete prediction accuracy, while a value of 0 means that every prediction was randomly assigned.

Statistics

Statistical analysis of the distribution of each attribute for functional mutations and neutral mutations was performed using the chi-squared test.

Results

Predictive powers of individual attributes

The prediction performance of each attribute was assessed using the decision tree method. Except for the solvent accessibility, all other attributes played a role in predicting whether an nsSNP has an effect on protein function or not. When solvent accessibility was used as a single attribute in the prediction, the MCC was 0. In contrast, the conservation score, whose MCC reached 0.68, was found to be the best discriminator of functional versus neutral variations. The MCCs of the location, mass change, and volume change attributes were higher than 0.4, but these achieved less prediction accuracy than the conservation score. Other attributes, such as PAM120 and BLOSUM62 substitution matrices, hydrophobicity change, and buried charge, had poor predictive performance (Table 1).

Distribution of functional and neutral mutations

The distribution of attribute values for functional mutations was significantly different from that of neutral mutations

Table 1 Prediction performances of attributes and attribute sets obtained using the decision tree method

Attribute or attribute set	MCC
Conservation score	0.68
Location	0.51
Volume change	0.45
Mass change	0.41
PAM120 score	0.38
BLOSUM62 score	0.28
Hydrophobicity change	0.27
Buried charge	0.13
Solvent accessibility	0
Sequence-based attributes	0.73
Structure-based attributes	0.51
All nine attributes	0.78
Optimal attributes subset	0.81

(Fig. 1). Approximately 62.67% of the functional mutations had conservation scores of >0.5, whereas only 2.15% of the neutral mutations had conservation scores of >0.5. For those mutations with a conservation score <−0.5, only 17.6% were functional mutations, whereas 81.86% were neutral. For those mutations with a conservation score of between −0.5 and 0.5, functional mutations were only 3.74% more than neutral (Fig. 1a).

When the hydrophobicity value changes of wild-type residues and mutated residues were >3, 50.27% of the mutations were functional compared with 21.27% neutral mutations. When the hydrophobicity value changes were <3, there were more neutral mutations than functional (Fig. 1b). The distributions of the mass and volume changes were similar to that of the hydrophobicity. When volume changes were >60 or mass changes were >40, there were more functional mutations than neutral mutations (Fig. 1c,d). These data indicate that dramatic changes in physiochemical properties tend to change the structure or function of the protein, and thus the mutations would be functional.

The nature of the amino acid changes was assessed using BLOSUM62 and PAM120 substitution matrices, since these two matrices are widely used and robust. A total of 67.74% of the functional mutations have BLOSUM62 scores of <−1, and only 27.52% of the neutral variations have BLOSUM62 scores of <−1 (Fig. 1e). The distribution of the PAM120 substitution matrix score is similar to that of the BLOSUM62 results (Fig. 1f). For these methods, the smaller the score, the higher the probability that a mutation is functional. We found that 47.06% of the functional mutations and 14.35% of the neutral changes have a PAM120 score of <−2, while 25.46% of the functional mutations and 57.84% of the neutral changes have a PAM120 score of >0.

TM regions contained 77.87% of functional mutations and 24.98% of neutral mutations, while extracellular and intracellular domains contained 22.13% of functional and 75.02% of neutral mutations (Fig. 1g). The distributions of buried charge and solvent accessibility for functional variations and neutral mutations were significantly different ($\chi^2 = 33.78$, $P < 0.01$ and $\chi^2 = 51.49$, $P < 0.01$, respectively), although this difference was not as pronounced for the other attributes (Fig. 1h,i).

Optimal attributes subset

During the attribute selection process, six attributes were found to be the optimal attributes subset: the conservation score, the BLOSUM62 substitution matrix score, the location, the solvent accessibility, the buried charge, and the hydrophobicity change. The prediction performance of this optimal attributes subset was compared with four different attribute sets: all attributes, sequence-based attributes, structure-based attributes, and conservation score alone (Table 1). The MCC of the optimal attributes set (0.81) was the highest among them. Sequence-based attributes (even using just the conservation score) were better than the structure-based ones. When all attributes were combined, the prediction accuracy was improved compared with those of sequence-based or structure-based attributes alone.

Rules for predicting functional nsSNPs

The decision tree method can produce intelligible rules and attach a confidence level to each rule. Seven important rules with covers of >30 were obtained (Table 2). These rules were used to predict functional mutations, and they conveniently discriminate functional nsSNPs from neutral mutations. For example, according to Rule 1, if the conservation score of an nsSNP was less than or equal to −0.343, and it was located in the extracellular or intracellular domains, then the probability that this nsSNP is neutral would be 0.96.

Functional nsSNPs in human GPCRs

We collected 519 nsSNPs from dbSNP, and 166 of these (32%) were predicted to be functional using the optimal attributes set (Table 3). Analysis of these nsSNPs in GPCRs will provide the basis for assessing susceptibility to diseases and designing individualized therapy.

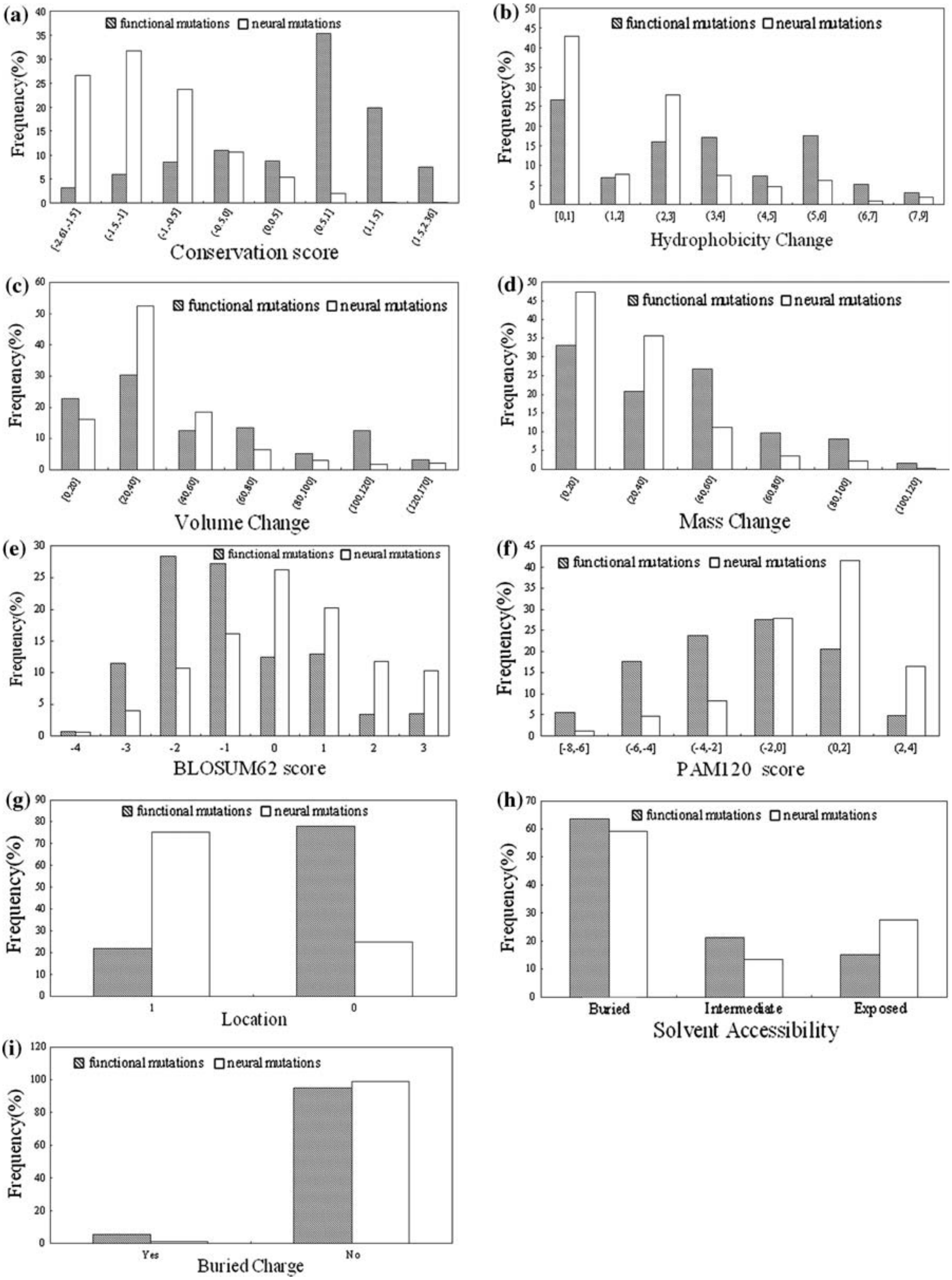

Fig. 1a–i The distribution of attributes for functional mutations and neutral mutations. The *shaded bars* represent the functional mutations and the *white bars* are neutral mutations. Attributes: **a** conservation score ($\chi^2 = 1099.05$, $P < 0.01$, 7 *df*), **b** hydrophobicity change ($\chi^2 = 208.37$, $P < 0.01$, 7 *df*), **c** volume change ($\chi^2 = 212.64$, $P < 0.01$, 6 *df*), **d** mass change ($\chi^2 = 211.83$, $P < 0.01$, 5 *df*), **e** BLOSUM62 score ($\chi^2 = 281.85$, $P < 0.01$, 7 *df*), **f** PAM120 ($\chi^2 = 314.47$, $P < 0.01$, 5 *df*), **g** location ($\chi^2 = 546.78$, $P < 0.01$, 1 *df*), **h** solvent accessibility ($\chi^2 = 51.49$, $P < 0.01$, 2 *df*), and **i** buried charge ($\chi^2 = 33.78$, $P < 0.01$, 1 *df*)

Table 2 Rules derived from the decision tree with the optimized attribute set

Rule	Cover	Confidence level
Rule 1: if conservation score ≤ -0.434 and TM = 0, then class = no effect	949	0.96
Rule 2: if conservation score > -0.434 and conservation score ≤ 0.478 and TM = 0 and BLOSUM62 score > -1, then class = no effect	72	0.92
Rule 3: if conservation score ≤ 0.478 and TM = 1 and BLOSUM62 score ≤ -2 and solvent accessibility = buried and hydrophobicity change > 0.4, then class = effect	87	0.82
Rule 4: if conservation score ≤ 0.478 and TM = 1 and BLOSUM62 score ≤ -1 and solvent accessibility = intermediate, then class = effect	32	0.91
Rule 5: If conservation score ≤ 0.285 and TM = 1 and BLOSUM62 score > -1 and solvent accessibility = buried and hydrophobicity change ≤ 2.5, then class = no effect	183	0.97
Rule 6: If conservation score > -0.285 and conservation score ≤ 0.478 and TM = 1 and BLOSUM62 score > 0 and solvent accessibility = buried and hydrophobicity change ≤ 2.5, then class = no effect	51	0.92
Rule 7: if conservation score > 0.478 and BLOSUM62 score ≤ 1, class = effect	461	0.97

Discussion

In the present study, the prediction power of both sequence-based and structure-based attributes was used to predict functional nsSNPs in human GPCRs. Since only one GPCR structure is known, we used predicted structure information instead of the actual structure. A conservation score that is based on evolutionary selection information was found to be the best single predictor for discriminating functional mutations from neutral variations. A high conservation score means that there is selective pressure to maintain these residues during evolution, and therefore these are likely to be important to the structure and function of the protein. The mutations that occur at these conserved sites are often functional mutations. The change in the physiochemical properties of residues would influence the structure or stability of the proteins and indirectly change the function, and so these attributes only have moderate prediction power. Substitution matrices, which consider only the likelihood of the substitution in all proteins at all positions, can also be useful, albeit with lower prediction accuracy (Yue and Moult 2006). We found that functional mutations are overrepresented in the TM regions and are underrepresented in the extracellular and intracellular domains. This implies that changes in TM regions may directly affect either the structure or function of the receptor. Mutations in TM regions could abrogate or diminish the activity of the protein when a ligand-binding site is affected. Alternatively, a mutation in a TM region could compromise the protein's structural integrity by having an effect on helix–helix packing interactions. In general, structure-based attributes had poorer predictive powers than sequence-based attributes. The MCC of solvent accessibility was zero, which means every prediction was randomly assigned, and this was the worst predictor among all the attributes when it was used alone.

Combining the attributes can greatly improve the prediction accuracy. Though conservation score was the most powerful predictor, the MCC increased to 0.22 when it was combined with other sequence-based attributes. When all nine attributes were used in a prediction, the accuracy was improved when compared with the sequence-based attributes alone. We also found the proposed structural information to be useful in prediction. It is likely that most mutations that affect protein function actually affect it indirectly through changes in structural stability. However, simply taking all the attributes together did not achieve the best performance. We found that the optimal attributes subset only requires six attributes—the conservation score, the BLOSUM62 substitution matrix score, the location, the solvent accessibility, the buried charge, and the hydrophobicity change. The combination of these six attributes had an MCC that was 0.03 higher than that of all nine attributes. The optimal subset includes both sequence-based and structure-based attributes. Moreover, it is interesting to see that the optimal attributes subset did not consist of the six best predictors when each was assessed by itself. The predictabilities of some inferior attributes, such as solvent accessibility and buried charge, were increased when used in combination.

Seven important rules with cover >30 were derived from the decision tree. Based on these rules, we could intuitively distinguish functional nsSNPs from neutral nsSNPs only if the attribute values of the nsSNPs are available, and there is no need for any complex training or testing processes.

In summary, combining sequence-based and structure-based information will improve the prediction performance, but the optimal attributes subset was not simply a combination of the attributes. With the optimal attributes subset, a total of 166 functional nsSNPs were predicted. Given the important roles of GPCRs in many physiological processes and their pharmaceutical relevance as drug targets, further investigation of these nsSNPs will be very

Table 3 Predicted functional nsSNPs in human GPCRs, obtained with the optimized attribute set using the decision tree method

dbSNP rs no.	Receptor	Wild-type residue	Mutated residue	Position	Confidence level
rs11773032	ACM2_HUMAN	G	S	73	0.967
rs7107481	ACM4_HUMAN	D	G	112	0.967
rs16839102	ACM3_HUMAN	L	P	431	0.967
rs8192448	ADA1B_HUMAN	V	G	51	0.816
rs1800888	ADRB2_HUMAN	T	I	164	0.783
rs2229125	ADA1A_HUMAN	I	S	200	0.967
rs238741	ADRB1_HUMAN	R	S	318	1
rs1133450	ADA2C_HUMAN	S	I	401	0.967
rs1133452	ADA2C_HUMAN	R	P	446	0.967
rs1801253	ADRB1_HUMAN	R	G	389	1
rs5327	DRD1_HUMAN	T	P	37	0.967
rs5328	DRD1_HUMAN	T	R	37	0.967
rs2227840	DRD5_HUMAN	C	S	62	0.967
rs6282	DRD5_HUMAN	L	R	88	0.967
rs2227845	DRD5_HUMAN	G	E	110	0.727
rs1800443	DRD4_HUMAN	V	G	194	0.816
rs5331	DRD1_HUMAN	S	A	199	0.967
rs2227843	DRD5_HUMAN	S	N	233	0.967
rs2227851	DRD5_HUMAN	T	P	297	0.967
rs11665084	HRH4_HUMAN	A	V	138	0.609
rs12564512	5HT1D_HUMAN	T	A	62	0.967
rs130061	5HT1B_HUMAN	F	L	219	0.609
rs3828741	5HT1E_HUMAN	A	T	208	0.967
rs8192618	TAAR1_HUMAN	R	C	23	1
rs17061399	TAAR6_HUMAN	I	T	37	0.783
rs9493386	TAAR5_HUMAN	V	L	87	0.967
rs17061401	TAAR6_HUMAN	G	S	165	0.609
rs2962857	TAAR1_HUMAN	T	A	252	0.967
rs8192625	TAAR6_HUMAN	C	Y	291	0.816
rs13095608	AGTR1_HUMAN	V	G	41	0.816
rs1064533	AGTR1_HUMAN	C	W	289	0.967
rs1042860	AGTR2_HUMAN	C	W	268	0.816
rs3729979	AGTR2_HUMAN	P	L	271	0.816
rs5234	BRS3_HUMAN	L	Q	162	0.727
rs11880097	C5AR_HUMAN	K	N	279	0.967
rs1805038	CXCR1_HUMAN	R	C	71	0.967
rs6781048	CCR1_HUMAN	Y	D	10	0.967
rs4987052	CCR2_HUMAN	L	V	45	0.967
rs5742906	CCR3_HUMAN	P	L	39	0.905
rs1799863	CCR5_HUMAN	L	Q	55	0.967
rs1800452	CCR5_HUMAN	R	Q	223	0.967
rs1800943	CCR5_HUMAN	G	V	301	0.967
rs12721498	CCR9_HUMAN	I	V	80	0.75
rs3749271	CCRL1_HUMAN	K	N	143	0.967
rs2228467	CCBP2_HUMAN	V	A	41	0.967
rs3204849	O00421_HUMAN	F	Y	167	1
rs665648	CXCR5_HUMAN	G	S	344	0.967
rs2234357	CXCR6_HUMAN	V	A	239	0.967

Table 3 continued

dbSNP rs no.	Receptor	Wild-type residue	Mutated residue	Position	Confidence level
rs3732378	CX3C1_HUMAN	T	M	280	0.967
rs1805000	GASR_HUMAN	L	F	37	0.967
rs17852056	EDNRA_HUMAN	P	H	115	0.967
rs5347	EDNRB_HUMAN	F	V	112	0.967
rs5350	EDNRB_HUMAN	T	M	244	0.727
rs1805005	MSHR_HUMAN	V	L	60	0.967
rs1805006	MSHR_HUMAN	D	E	84	0.909
rs2228479	MSHR_HUMAN	V	M	92	0.967
rs11547464	MSHR_HUMAN	R	H	142	0.967
rs1805007	MSHR_HUMAN	R	C	151	0.967
rs1110400	MSHR_HUMAN	I	T	155	0.967
rs3212365	MSHR_HUMAN	V	L	156	0.967
rs1805009	MSHR_HUMAN	D	H	294	0.967
rs28926178	ACTHR_HUMAN	P	R	27	0.967
rs28926179	ACTHR_HUMAN	G	A	90	0.967
rs28940892	ACTHR_HUMAN	Y	C	254	0.967
rs28926182	ACTHR_HUMAN	F	C	278	0.967
rs17847261	MC3R_HUMAN	C	S	311	0.967
rs13447326	MC4R_HUMAN	P	L	78	0.967
rs13447327	MC4R_HUMAN	S	R	94	0.967
rs2282556	MC4R_HUMAN	G	R	98	0.905
rs13447330	MC4R_HUMAN	I	T	121	0.727
rs13447331	MC4R_HUMAN	S	L	127	0.816
rs13447332	MC4R_HUMAN	R	W	165	0.967
rs1016862	MC4R_HUMAN	I	S	169	0.816
rs13447333	MC4R_HUMAN	G	D	181	0.967
rs13447335	MC4R_HUMAN	A	E	244	0.727
rs12075	DUFFY_HUMAN	G	D	42	0.727
rs3027017	DUFFY_HUMAN	D	V	59	0.967
rs13962	DUFFY_HUMAN	A	T	100	0.967
rs3027020	DUFFY_HUMAN	L	Q	203	0.727
rs1801397	DUFFY_HUMAN	T	K	275	0.727
rs28642215	NPY4R_HUMAN	P	T	96	0.967
rs3740868	GPR83_HUMAN	P	Q	374	0.905
rs6432225	NTR2_HUMAN	A	V	54	0.609
rs17853770	NTR2_HUMAN	R	C	142	0.967
rs13057124	SSR3_HUMAN	F	L	215	0.967
rs1065191	SSR4_HUMAN	N	T	83	0.967
rs4988474	SSR4_HUMAN	R	C	244	0.967
rs2567608	SSR4_HUMAN	F	S	321	0.967
rs4988477	SSR4_HUMAN	V	A	325	0.967
rs4988489	SSR5_HUMAN	R	C	248	0.906
rs5198	V2R_HUMAN	A	V	42	0.967
rs28935496	V2R_HUMAN	R	W	113	0.967
rs5200	V2R_HUMAN	A	V	147	0.967
rs171114	OXYR_HUMAN	A	G	63	0.967
rs8192513	GALR2_HUMAN	G	R	204	0.816
rs8192514	GALR2_HUMAN	S	R	346	0.727

Table 3 continued

dbSNP rs no.	Receptor	Wild-type residue	Mutated residue	Position	Confidence level
rs28939719	KISSR_HUMAN	L	S	148	0.967
rs350132	KISSR_HUMAN	L	H	364	0.727
rs2230849	PAR1_HUMAN	Y	N	187	0.967
rs2227799	PAR1_HUMAN	S	Y	412	0.727
rs2069700	PAR3_HUMAN	M	V	177	0.967
rs2227346	PAR4_HUMAN	F	V	296	0.967
rs17438900	QRFPR_HUMAN	F	V	61	0.783
rs13305975	UR2R_HUMAN	R	H	148	0.967
rs17851452	NMUR2_HUMAN	C	G	204	0.967
rs28928870	FSHR_HUMAN	T	I	449	0.967
rs6167	FSHR_HUMAN	S	R	524	0.727
rs28928871	FSHR_HUMAN	D	N	567	0.967
rs12480652	LSHR_HUMAN	N	S	291	0.967
rs28937584	TSHR_HUMAN	D	H	633	0.783
rs1057362	PE2R1_HUMAN	A	T	71	0.967
rs41312444	PD2R_HUMAN	G	E	198	0.816
rs41516947	PD2R_HUMAN	R	Q	231	0.967
rs41312506	PD2R_HUMAN	R	Q	332	0.967
rs12656588	PE2R4_HUMAN	A	G	183	0.967
rs35425451	PF2R_HUMAN	V	G	344	0.967
rs2229127	PI2R_HUMAN	V	M	25	0.967
rs35431373	PI2R_HUMAN	A	T	128	0.967
rs34377097	TA2R_HUMAN	R	L	60	0.967
rs5743	TA2R_HUMAN	C	S	68	0.967
rs5744	TA2R_HUMAN	V	E	80	0.967
rs5749	TA2R_HUMAN	A	T	160	0.967
rs11547176	AA1R_HUMAN	R	H	105	0.967
rs2511241	P2RY2_HUMAN	P	L	46	0.816
rs35146537	GPR35_HUMAN	A	T	25	0.609
rs3749171	GPR35_HUMAN	T	M	108	0.967
rs4151553	P2RY5_HUMAN	C	W	137	0.967
rs1466684	P2Y13_HUMAN	T	M	158	0.783
rs28933074	GNRHR_HUMAN	Y	C	284	0.967
rs4988511	GHSR_HUMAN	I	T	134	0.967
rs28383653	MTR1A_HUMAN	G	E	166	0.967
rs11542862	EDG2_HUMAN	N	S	77	0.967
rs1049843	EDG2_HUMAN	G	S	340	0.967
rs34075341	EDG2_HUMAN	R	Q	243	0.967
rs3745268	EDG5_HUMAN	R	Q	60	0.967
rs3746072	EDG5_HUMAN	R	L	365	0.727
rs35483143	EDG8_HUMAN	L	Q	318	0.727
rs34010553	CALRL_HUMAN	R	I	274	0.967
rs13306399	GIPR_HUMAN	C	S	46	0.967
rs13306402	GIPR_HUMAN	R	W	136	0.967
rs13306398	GIPR_HUMAN	G	C	198	0.967
rs5392	GIPR_HUMAN	L	V	262	0.967
rs13306403	GIPR_HUMAN	R	L	316	0.967
rs1800437	GIPR_HUMAN	E	Q	354	0.909

Table 3 continued

dbSNP rs no.	Receptor	Wild-type residue	Mutated residue	Position	Confidence level
rs6726491	SCTR_HUMAN	D	N	110	0.967
rs3731600	SCTR_HUMAN	A	P	122	0.967
rs11544196	CD97_HUMAN	T	I	85	0.967
rs897738	EMR1_HUMAN	D	N	174	0.967
rs2290635	EMR1_HUMAN	C	S	296	0.967
rs2229769	EMR1_HUMAN	F	C	691	0.967
rs2524383	EMR2_HUMAN	L	F	614	0.967
rs3752187	EMR2_HUMAN	S	F	665	0.816
rs8102646	EMR3_HUMAN	R	Q	385	0.967
rs41263977	BAI2_HUMAN	R	C	1,519	1
rs4823561	CELR1_HUMAN	C	R	1,126	0.727
rs11704506	CELR1_HUMAN	A	T	1,647	0.967
rs6008795	CELR1_HUMAN	L	P	1,994	0.727
rs41279708	CELR2_HUMAN	P	S	1,190	0.967
rs12567377	CELR2_HUMAN	G	R	1,992	0.967
rs17035649	CELR2_HUMAN	T	A	2,387	0.609
rs2171560	CELR3_HUMAN	G	A	2,104	0.967
rs17856664	MGR3_HUMAN	P	A	512	0.967
rs17078901	MGR6_HUMAN	E	K	227	0.967
rs17078898	MGR6_HUMAN	E	V	227	0.967
rs2856347	MGR6_HUMAN	S	F	666	0.816
rs1051433	MGR8_HUMAN	I	G	768	0.816

useful for elucidating disease pathogenesis mechanisms and drug efficacy issues.

References

Armon A, Graur D, Ben-Tal N (2001) Consurf: an algorithmic tool for the identification of functional regions in proteins by surface mapping of phylogenetic information. J Mol Biol 307:447–463

Balasubramanian S, Xia Y, Freinkman E, Gerstein M (2005) Sequence variation in G-protein-coupled receptors: analysis of single nucleotide polymorphisms. Nucleic Acids Res 33:1710-1721

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

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

Dobson RJ, Munroe PB, Caulfied MJ, Saqi MA (2006) Predicting deleterious nsSNPs: an analysis of sequence and structural attributes. BMC Bioinf 7:217–235

Drysdale CM, McGraw DW, Stack CB, Stephens JC, Judson RS, Nandabalan K et al. (2000) Complex promoter and coding region beta 2-adrenergic receptor haplotypes alter receptor expression and predict in vivo responsiveness. Proc Natl Acad Sci USA 97:10483–10488

Frank E, Hall M, Trigg L, Holmes G, Witten IH (2004) Data mining in bioinformatics using Weka. Bioinformatics 20:2479–2481

Gether U (2000) Uncovering molecular mechanisms involved in activation of G protein-coupled receptors. Endocr Soc 21:90–113

Harpaz Y, Gerstein M, Chothia C (1994) Volume changes on protein folding. Structure 2:641–649

Horn F, Bettler E, Oliveria L, Campagne F, Cohen FE, Vriend G (2003) GPCRDB information system for G protein-coupled receptors. Nucleic Acids Res 31:294–297

Horn F, Lau AL, Cohen FE (2004) Automated extraction of mutation data from the literature: application of MuteXt to G protein-coupled receptors and nuclear hormone receptors. Bioinformatics 20:557–568

Howard AD, McAllister G, Feighner SD, Liu Q, Nargund RP, Van der Ploeg LH et al. (2001) Orphan G-protein-coupled receptors and natural ligand discovery. Trends Pharmacol Sci 22:132–140

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

Krogh A, Larsson B, von Heijne G, Sonnhammer EL (2001) Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. J Mol Biol 305:567–580

Kyte J, Doolittle RF (1982) A simple method for displaying the hydropathic character of a protein. J Mol Biol 157:105–132

Marinissen MJ, Gutkind JS (2001) G protein-coupled receptors and signaling networks: emerging paradigms. Trends Pharmacol Sci 22:368–376

Matthews BW (1985) Comparison of the predicted and observed secondary structure of T4 phage lysozyme. Biochim Biophys Acta 405:442–451

Ng PC, Henikoff S (2001) Predicting deleterious amino acid substitutions. Genome Res 11:863–874

Palczewski K, Kumasaka T, Hori T, Behnke CA, Motoshima H, Fox BA et al. (2000) Crystal structure of rhodopsin: a G protein-coupled receptor. Science 289:739–745

Pei J, Grishin NV (2001) AL2CO: calculation of positional conservation in a protein sequence alignment. Bioinformatics 17:700–712

Phillips KA, Veenstra DL, Oren E, Lee JK, Sadee W (2001) Potential role of pharmacogenomics in reducing adverse drug reactions: a systematics review. JAMA 286:2270–2279

Quinlan JR (1993) C4.5: programs for machine learning. Morgan Kaufmann, San Francisco, CA

Ramensky V, Bork P, Sunyaev S (2002) Human non-synonymous SNPs: server and survey. Nucleic Acids Res 30:3894–3900

Roses AD (2004) Pharmacogenetics and drug development: the path to safer and more effective drugs. Nat Rev Genet 5:645–656

Rost B, Sander C (1993) Improved prediction of protein secondary structure by use of sequence profiles and neural networks. Proc Natl Acad Sci USA 90:7558–7562

Rost B, Sander C (1994) Conservation and prediction of solvent accessibility in protein families. Proteins 9:56–68

Sakmar TP (1998) Rhodopsin: a prototypical G protein-coupled receptor. Prog Nucleic Acid Res Mol Biol 59:1–34

Saunders CT, Baker D (2002) Evolutionary of structural and evolutionary contributions to deleterious mutations prediction. J Mol Biol 322:891–901

Shannon CE (1948) A mathematical theory of communication. Bell Syst Tech 27:379–423; 623–656

Sunyaev S, Ramensky V, Koch I, Lathe WIII, Kondrashov AS, Bork P (2001) Prediction of deleterious human alleles. Hum Mol Genet 10:591–597

Wang Z, Moult J (2001) SNPs, protein structure, and disease. Hum Mutat 17:263–270

Witten I, Frank E (2000) Data mining: practical machine learning tools and techniques with Java implementations. Morgan Kaufmann, San Francisco, CA

Yue P, Moult J (2006) Identification and analysis of deleterious human SNPs. J Mol Biol 356:1236–1274