2006-10-01

Pharmacogenomics: Analyzing SNPs in the CYP2D6 Gene Using Amino Acid Properties

Wesley A. Beckstead
wesb22@gmail.com

Mark J. Clement
clement@cs.byu.edu

Mark Ebbert

David McClellan

Timothy O'Connor

Pharmacogenomics: Analyzing SNPs in the CYP2D6 Gene Using Amino Acid Properties

Mark T. W. Ebbert[1,2,3], Timothy D. O'Connor[1,4], Wesley A. Beckstead[1,5], Mark J. Clement[2,6], and David A. McClellan[1,7]

[1]Dept. of Integrative Biology, Brigham Young University, Provo, UT 84602; [2]Dept. of Computer Science, Brigham Young University, Provo, UT 84602

[3]marktwe@byu.net; [4]tdoconnor@byu.edu; [5]wesb@byu.edu; [6]clement.cs.byu.edu; [7]david_mcclellan@byu.edu

Key Words: pharmacogenomics, amino acid properties, SNP evaluation

Abstract: Each year people suffer from complications of adverse drug reactions, but with pharmacogenomics there is hope to prevent thousands of these people from suffering or dying needlessly. The CYP2D6 gene is responsible for metabolizing a large portion of these drugs. Because of the gene's importance, various approaches have been taken to analyze CYP2D6 and single nucleotide polymorphisms (SNPs) throughout its sequence. This study introduces a novel method to analyze the effects of SNPs on encoded protein complexes by focusing on the biochemical properties of each non-synonymous substitution using the program TreeSAAP. We apply this technique to SNPs found in the CYP2D6 gene. Our results show four SNPs that exhibit radical changes in amino acid properties which may cause a lack of functionality in the CYP2D6 gene and contribute to a person's inability to metabolize specific drugs.

1.0 Introduction

On average, 2.2 million people are hospitalized and 100,000 people die every year due to adverse drug reactions (Stipp, 2000). There are multiple factors that determine how an individual will react to a given drug dosage, including weight, age, sex, race, and habits (Kalow, 2006). Yet what has commonly been overlooked until recent years is the individual's ability to metabolize the drug. Various genes are responsible for this metabolism, but the CYP2D6 gene is responsible for metabolizing a large portion of drugs including antidepressants, antipsychotic drugs, codeine, debrisoquin and others (Evans and Relling, 1999). If there are single nucleotide polymorphisms (SNPs) within an individual's CYP2D6 gene, functionality problems ranging from overly active metabolism to complete loss of function may result. Therefore, for the purpose of safety, it is necessary to identify these problems before prescribing drugs.

Thus far there have been various methods to analyze SNPs such as using post-mortem analysis of genes along with the person's known medical history using PCR (Levo, et al. 2003), allele-specific PCR with energy-transfer primers and capillary array electrophoresis microchips (Medentz, et al 2001), and creating electrical circuits with DNA for SNP detection (Syvänen, A., and Söderlund, H. 2002), but none have taken the approach of analyzing amino acid properties to determine the likelihood of problems occurring. Using these properties it is possible to measure the extremity of each non-synonymous SNP. This is done using TreeSAAP (Woolley 2003) which is a program that was originally intended to identify selection over evolutionary time, but has proven valuable in detecting the effects of amino acid substitutions. We introduce the technique of using TreeSAAP to identify the extreme biochemical changes that SNPs produce in the encoded proteins and apply this method to SNPs in the CYP2D6 gene to determine their effect on CYP2D6's ability to metabolize drugs.

TABLE 1: The 6 non-synonymous single nucleotide polymorphisms in the CYP2D6 gene analyzed in this study with their consequent amino acid replacements. These SNPs are found in various sequences that were collected from GenBank (Table 2).

SNP #	Base Change	Base Site	AA Change	AA Site	Sequences
1	C -> T	100	Pro -> Ser	34	CYP2D6*4A, CYP2D6*4D, CYP2D6*10B
2	C -> A	271	Leu -> Met	91	CYP2D6*4A
3	A -> G	281	His -> Arg	93	CYP2D6*4A
4	C -> T	320	Thr -> Ile	107	CYP2D6*17, CYP2D6*17V
5	T -> C	886	Cys -> Arg	296	CYP2D6*4A, CYP2D6*4D, CYP2D6*9, CYP2D6*10B
6	C -> G	1457	Thr -> Ser	486	CYP2D6*9

TABLE 2: The 8 sequences of the CYP2D6 gene used in this study. According to GenBank and Swiss-Prot 2 sequences are able to metabolize specific drugs properly and 6 are dysfunctional in this regard.

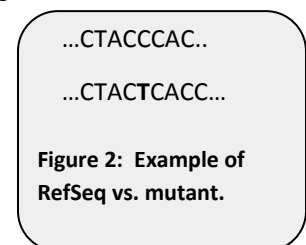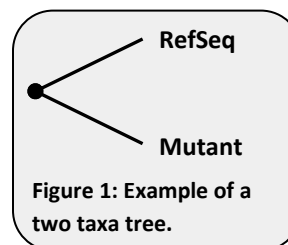| Sequence	Functional/Non-Functional
CYP2D6*1	Functional
CYP2D6*4A	Non-Functional
CYP2D6*4D	Non-Functional
CYP2D6*9	Non-Functional
CYP2D6*10B	Non-Functional
CYP2D6*17	Non-Functional
CYP2D6*17V	Non-Functional
CYP2D6 RefSeq	Functional

2.0 Materials and Methods

Table 1 shows the 6 non-synonymous SNPs that were analyzed in this study. Eight sequences containing these SNPs of the CYP2D6 gene from humans were collected from GenBank. Six are said to be non-functional (CYP2D6*4A, CYP2D6*4D, CYP2D6*9, CYP2D6*10B, CYP2D6*17, and CYP2D6*17V) and two are said to be functional (GenBank Reference Sequence and CYP2D6*1) (Table 2). By "functional" we mean that the encoded protein is able to metabolize specific drugs properly. These 6 SNPs were analyzed to determine which SNPs may be causing the lack of functionality in the 6 non-functional sequences.

TreeSAAP uses phylogenetic trees in its analysis to identify selection over evolutionary time (Woolley 2003); however, since this is population data, selection is not the interest of the study. The purpose is to analyze each individual SNP in a given sequence and identify the extremity of the mutation. It is possible to identify SNPs that are likely to cause a functional problem, even without knowing the phenotype. If the phenotype is already known to be less functional (or completely non-functional), the SNP(s) most likely to be causing the problem can be identified. Since TreeSAAP's analyses are based upon the tree given, it is important to have confidence in that foundation. However, it is essentially impossible to have confidence in any phylogenetic tree using population data because of its sequence conservation. Thus, TreeSAAP would normally not be ideal for such a study since any analyses based upon a tree produced by population data would be futile. However, there is one tree that is indisputable: a tree with two taxa as shown in Figure 1. The two taxa in Figure 1 are the reference sequence (RefSeq) from GenBank and a mutant sequence. For simplicity, every SNP in each sequence used in this study was placed individually into its own mutant sequence (this was done using the RefSeq as the base sequence) as shown in Figure 2. Thus, when the two taxa are compared in TreeSAAP there is only one SNP to be analyzed.

TreeSAAP, which implements the MM01 statistical model (McClellan and McCracken, 2001) and the baseml ancestral character-state reconstruction algorithm (Yang, 1997), is used to statistically analyze each SNP for radical biochemical shifts resulting from each amino acid

Figure 1: Example of a two taxa tree.

…CTACCCAC..

…CTAC**T**CACC…

Figure 2: Example of RefSeq vs. mutant.

replacement and identify the amino acid properties that were associated with each change. Several statistical tests are run to determine the extremity of an amino acid replacement, among which are chi-square and t-test. TreeSAAP categorizes radical changes into eight magnitude categories, 1 being the most conservative and 8 being the most radical (Woolley et al., 2003). In this analysis we only considered magnitude categories 6, 7, and 8 because they unambiguously indicate a significant change in the resulting protein (McClellan et al., 2005).

3.0 Results and Discussion

This study focused on six sequences known to be non-functional. These include CYP2D6*4A, CYP2D6*4D, CYP2D6*9, CYP2D6*10B, CYP2D6*17, and CYP2D6*17V (Table 2). Of the 6 unique non-synonymous SNPs in this study (Table 1), TreeSAAP indicated four of them, SNP numbers 1, 3, 4 and 5, that show significant changes in amino acid properties. The following is a discussion of each SNP.

SNP # 1 (C100T, P34S)

At nucleotide 100 in sequences CYP2D6*4A, CYP2D6*4D, and CYP2D6*10B there is a transitional mutation from cytosine to thymine as seen in Figure 3 which results in an amino acid substitution of proline to serine. According to TreeSAAP, in this situation there was a category 7 change in 'Power to be at the C-Terminal' and a category 6 in 'Thermodynamic transfer hydrophobicity' (Table 3). TreeSAAP indicated that the SNP has an increased propensity in both of these properties.

TABLE 3: TreeSAAP results for SNP #1. This SNP has an increased propensity in both amino acid properties: 'Thermodynamic transfer hydrophobicity' and 'Power to be at the C-Terminal'.

SNP #1 TreeSAAP Results

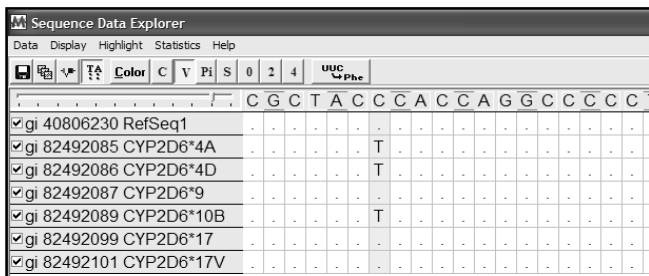| Amino Acid Property	Category	+/- Shift
Thermodynamic transfer hydrophobicity	6	+ 2.7
Power to be at the C-Terminal	7	+ 1.38

Figure 3: SNP # 1 at nucleotide 100 visualized with MEGA (Kumar, et al. 1994).

SNP # 3 (A281G, H93R)

Second, there is a mutation at nucleotide 281 which is also in sequence CYP2D6*4A shown in Figure 4. In this case it is a transitional mutation from adenine to guanine resulting in an amino acid mutation from histidine to arginine. This SNP has an increased tendency in 'Short and medium range non-bonded energy', which was found by TreeSAAP to be of category 6 (Table 4).

TABLE 4: TreeSAAP results for SNP #3. This SNP has an increased tendency in 'Short and medium range non-bonded energy'.

SNP #3 TreeSAAP Results

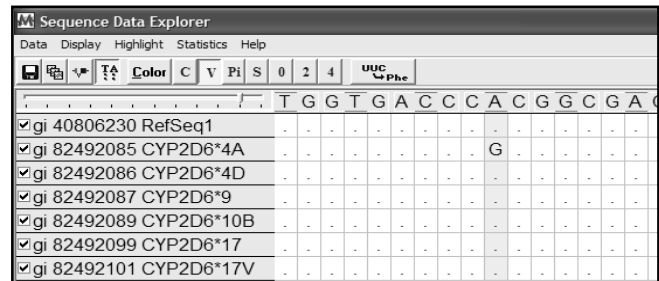| Amino Acid Property	Category	+/- Shift
Short and medium range non-bonded energy	6	+ 0.3

Figure 4: SNP # 3 at nucleotide 281 visualized with MEGA (Kumar, et al. 1994).

SNP # 4 (C320T, T107I)

The third mutation is a dramatic increase from the previous mutations. At nucleotide 320 there is a transitional mutation from cytosine to thymine (Figure 5) which results in a mutation from threonine to isoleucine. This occurs in sequences CYP2D6*17 and CYP2D6*17V. With this mutation comes four property changes of category 6 ('Average number of surrounding residues,' 'Buriedness,' 'Equilibrium constant,' 'Surrounding hydrophobicity') and two of category 7 ('Solvent accessible reduction ratio' and 'Thermodynamic transfer hydrophobicity') (Table 5). TreeSAAP indicated that this SNP has an increased propensity for the amino acid property 'Equilibrium constant (Ionization of COOH)' and decreases the other properties.

3

TABLE 5: TreeSAAP results for SNP #4. This SNP has an increased propensity for the amino acid property 'Equilibrium constant (ionization of COOH)' and a decreased propensity for the properties 'Average number of surrounding residues,' 'Buriedness,' 'Surrounding Hydrophobicity,' 'Solvent accessible reduction ratio,' and 'Thermodynamic transfer hydrophobicity'.

SNP #4 TreeSAAP Results

Amino Acid Property	Category	+/- Shift
Average number of surrounding residues	6	- 1.7
Buriedness	6	- 0.33
Equilibrium constant (Ionization of COOH)	6	+ 0.74
Surrounding hydrophobicity	6	- 3.22
Solvent accessible reduction ratio	7	- 5.09
Thermodynamic transfer hydrophobicity	7	- 3.08

Figure 5: SNP #4 at nucleotide 320 visualized with MEGA (Kumar, et al. 1994).
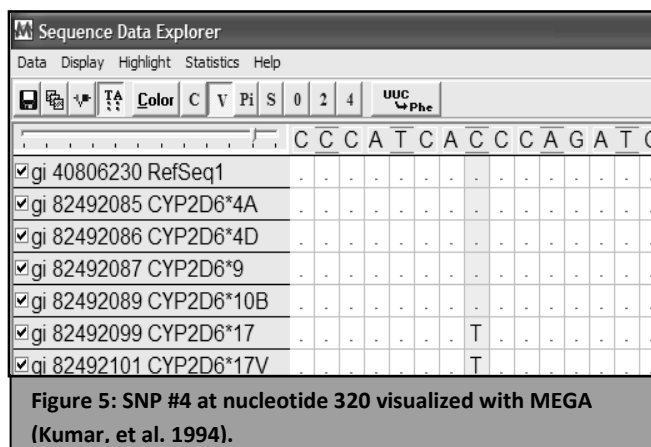
SNP # 5 (T886C, C296R)

The fourth mutation is actually the most interesting in this study. At nucleotide 886 a mutation occurs that creates a discrepancy. According to the Swiss-Prot database most analyses of the CYP2D6 gene have used the CYP2D6*1 allele as the reference sequence instead of the RefSeq at GenBank (Expasy 2006). This is important since there is a large discrepancy between the two sequences according to TreeSAAP. Both of these sequences are currently considered functional; however, at nucleotide 886 a mutation between thymine and cytosine occurs (Figure 6), resulting in an amino acid mutation of cysteine and arginine respectively. The problem is apparent in that, according to TreeSAAP, a mutation between arginine and cysteine in this situation is a very radical change. Ten properties are of a category between 6 and 8 (Table 6).

Three are category 6 ('Helical contact area,' 'Normalized consensus hydrophobicity,' 'Polarity'), three are category 7 ('Average number of surrounding residues,' 'Composition,' 'Hydropathy'), and four are category 8 changes ('Buriedness,' 'Isoelectric point,' 'Short and medium range non-bonded energy,' 'Total non-bonded energy'). Such a drastic change is very likely to render a protein non-functional. Thus, we suggest that one of the two sequences (CYP2D6*1 or RefSeq) is non-functional. Since the amino acid mutation involves a cysteine, it should not be surprising that this is considered such a radical change—cysteine being the only amino acid capable of making disulfide bonds. Two of the category 8 changes are 'Short and medium range non-bonded energy' and 'Total non-bonded energy,' which are relevant to disulfide bonds.

TABLE 6: TreeSAAP results for SNP #5. This SNP has an increased propensity for the amino acid properties: 'Normalized consensus hydrophobicity,' 'Average number of surrounding residues,' 'Composition,' 'Hydropathy,' 'Buriedness,' 'Short and medium range non-bonded energy,' and 'Total non-bonded energy,' and decreases in 'Helical contact area,' 'Polarity,' and 'Isoelectric point.'

SNP #5 TreeSAAP Results

Amino Acid Property	Category	+/- Shift
Helical contact area	6	-30
Normalized consensus hydrophobicity	6	+2.82
Polarity	6	-5
Average number of surrounding residues	7	+2.16
Composition	7	+2.1
Hydropathy	7	+7
Buriedness	8	+0.5
Isoelectric point	8	-5.71
Short and medium range non-bonded energy	8	+0.45
Total non-bonded energy	8	+0.56

Figure 6: SNP #5 at nucleotide 886 visualized with MEGA (Kumar, et al. 1994).
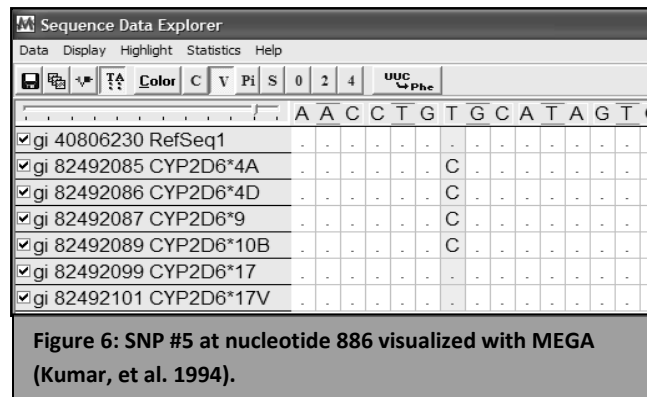
Two of the non-functional sequences in this study code for cysteine and four of them code for arginine at the SNP #5 site. We suggest that the arginine is the detrimental SNP. The two which code for cysteine (CYP2D6*17 and CYP2D6*17V) contain one other SNP with multiple radical changes in amino acid properties and CYP2D6*4A, CYP2D6*4D, and CYP2D6*10B also contain another SNP with radical changes that are likely to render the protein non-functional, which prevents us from using them in the comparison of this final SNP. However, CYP2D6*9 contains no other SNPs capable of destroying the proteins function other than the arginine at nucleotide 886, according to TreeSAAP. This would suggest that only this SNP could be the cause of CYP2D6*9 being non-functional and that any other sequence containing this SNP would also be non-functional including CYP2D6*1.

4.0 Conclusion

In this study we were able to identify four potentially detrimental SNPs in the CYP2D6 gene by analyzing the amino acid properties of each SNP using TreeSAAP. This technique brings a powerful approach that focuses on the natural environment of proteins and allows analysis of how radical a given mutation is, based on amino acid properties.

This new way to analyze SNPs is helpful for two situations: 1) identifying SNPs that are likely to cause a problem when the phenotype is unknown, 2) identifying the SNP that is causing an undesirable phenotype. Essentially it provides a new way to narrow problems down to specific SNP(s). This is especially important for pharmacogenomics in order to create a database of known detrimental SNPs and to identify persons who should not be prescribed specific drugs because of their inability to metabolize them.

References

Evans, W., and Relling, M. (1999). Pharmacogenomics: translating functional genomics into rational therapeutics. *Science. 286:487:491.*

Expasy. (2006). CP2D6_HUMAN. Retrieved May 12, 2006 from http://ca.expasy.org/cgi-bin/niceprot.pl?P10635

Kalow, W. (2006). Pharmacogenetics and pharmacogenomics: origin, status, and the hope for personalized medicine. *The Pharmacogenomics Journal. Online publication.*

Kumar, S., K. Tamura, and M. Nei. 1994. MEGA: Molecular Evolutionary Genetics Analysis software for microcomputers. *Comput. Appl. Biosci. 10:189-191.*

Levo, A., Koski, A., Ojanpera, I., Vuori. E., Sajantila, A. (2003). Post-mortem SNP analysis of CYP2D6 gene reveals correlation between genotype and opioid drug (tramadol) metabolite ratios in blood. *Forensic Science International. 135:9-15.*

McClellan, D., Palfreyman, E., Smith, M., Moss, J., Christensen, R., Sailsbery, J. (2005). Physicochemical evolution and molecular adaptation of the cetacean and artiodactyls cytochrome b proteins. *Mol. Biol. Evol., 22: 437-455.*

McClellan, D.A. and McCracken, K.G. (2001). Estimating the influence of selection on the variable amino acids sites of the cytochrome b protein functional domains. *Mol. Biol. Evol., 18:917-925*.

Medentz, I., Wong, W., Berti, L., Shiow, L., Tom, J., Scherer, J., Sensabaugh, G., Mathies, R. (2001). High-performance multiplex SNP analysis of three hemochromatosis-related mutations with capillary array electrophoresis microplates. *Genome Research 11:413-421.*

Stipp, D. (2000). A DNA tragedy. *Fortune, 142 (10), 170-178.*

Syvänen, A., and Söderlund, H. (2002). DNA sandwiches with silver and gold. *Nature Biotechnology. 20:349:350.*

Woolley, S., Johnson, J., Smith, M., Crandall, K., McClellan, D. (2003). TreeSAAP: Selection on Amino Acid Properties using phylogenetic trees. *Bioinformatics.19(5):671-672.*

Yang, Z. (1997). PAML: a program package for phylogenetic análisis by maximum likelihood. *CABIOS, 13: 555-556.*