Journal of Human Genetics (2010) 55, 175–178

www.nature.com/jhg

ORIGINAL ARTICLE

Prediction of the clinical phenotype of Fabry disease based on protein sequential and structural information

Seiji Saito[1,6], Kazuki Ohno[2,6,7], Jun Sese[3], Kanako Sugawara[4] and Hitoshi Sakuraba[4,5]

Fabry disease is a genetic disorder caused by a deficiency of α-galactosidase, exhibiting a wide clinical spectrum, from the early-onset severe 'classic' form to the late-onset mild 'variant' one. Recent screening of newborns revealed that the incidence of Fabry disease is unexpectedly high, and that the genotypes of patients with this disease are quite heterogeneous and many novel mutations have been identified in them. This suggests that a lot of Fabry patients will be found in an early clinical stage when the prognosis is obscure and a proper therapeutic schedule for them cannot be determined. Thus, it is significant to predict the clinical phenotype of this disease resulting from a novel mutation. Herein, we proposed a phenotype prediction model based on sequential and structural information. As far as we know, this is the first report of phenotype prediction for Fabry disease. First, we investigated the sequential and structural changes in the α-galactosidase molecule responsible for Fabry disease. The results showed that there are quite large differences in several properties between the classic and variant groups. We then developed a phenotype prediction model involving the decision tree technique. The accuracy of this prediction model is high (86%), and Matthew's correlation coefficient is also high (0.49). The phenotype predictor proposed in this paper may be useful for determining a proper therapeutic schedule for this disease.
Journal of Human Genetics (2010) **55,** 175–178; doi:10.1038/jhg.2010.5; published online 5 February 2010

Keywords: α-galactosidase; amino-acid substitution; Fabry disease; phenotype prediction; protein structure

INTRODUCTION

Fabry disease (MIM 301500) is a genetic disorder resulting from the deficient activity of lysosomal hydrolase α-galactosidase (GLA; EC 3.2.1.22).[1] The enzymatic defect causes the progressive accumulation of globotriaosylceramide (GL-3) in lysosomes, leading to heterogeneous phenotypes. In the classic form of Fabry disease, patients exhibit systemic manifestations, with onset of pain in the peripheral extremities, angiokeratoma, hypohidrosis and corneal opacity in childhood, followed by renal, cardiac and cerebrovascular involvement with increasing age. On the other hand, patients with the variant form of Fabry disease with late onset develop milder clinical manifestations, the main disorder being sometimes limited to the heart. The result of newborn screening revealed that the incidence of this disease is unexpectedly high, 1 in 3000–4000 male newborns,[2] and many other research groups are developing means of newborn and high-risk screening for the early diagnosis of Fabry disease.

With regard to therapy for Fabry disease, enzyme replacement therapy with recombinant human GLAs produced in Chinese hamster ovary cells and in human fibroblasts is available.[3–5] Furthermore, enzyme enhancement therapy (EET) with substrate analogs acting as pharmacological chaperones has also been developed,[6] and clinical trials have been performed. Although the number of mutants for which these chemicals are effective is limited, EET is beneficial for treating some patients with Fabry disease.

Considering this situation, prediction of the clinical outcome of this disease is becoming more and more important for determining a proper schedule for treating it. Recently, we investigated the basis of Fabry disease from the aspect of structural biology, and determined differences in the structural changes of the GLA protein between classic and variant forms.[7]

In this study, we further investigated the basis of Fabry disease by means of an improved method involving the originally developed sequential and structural analysis system, and proposed a new prediction model for the clinical phenotype of the disease using structural information.

MATERIALS AND METHODS

Data sets

So far, over 500 gene mutations have been reported in Fabry disease. In this study, we selected 210 Fabry missense mutations the phenotypes of which have been clearly described. The amino-acid substitutions, clinical phenotypes and references are listed in Supplementary Table 1. With regard to Fabry disease,

[1]Department of Biotechnology, Graduate School of Agricultural and Life Science, The University of Tokyo, Tokyo, Japan; [2]NPO for the Promotion of Research on Intellectual Property Tokyo, Tokyo, Japan; [3]Department of Computer Science, Ochanomizu University, Tokyo, Japan; [4]Department of Clinical Genetics, Meiji Pharmaceutical University, Tokyo, Japan and [5]Department of Analytical Biochemistry, Meiji Pharmaceutical University, Tokyo, Japan
[6]These authors contributed equally to this work.
[7]Current address: Astellas Pharma, 21 Miyukigaoka, Tsukuba, Ibaraki 305-8585, Japan.
Correspondence: Dr H Sakuraba, Department of Analytical Biochemistry, Meiji Pharmaceutical University, 2-522-1 Noshio, Kiyose, Tokyo 204-8588, Japan.
E-mail: sakuraba@my-pharm.ac.jp
Received 20 October 2009; revised 13 December 2009; accepted 7 January 2010; published online 5 February 2010

there are a few mutations that have been identified in both the classic and variant cases, and those were excluded in this study. The substitutions analyzed here can be divided into two groups: one comprising 196 substitutions, which cause the classic phenotype, and the other comprising 14 substitutions, which result in the variant one. Each substitution has nine selected features and the details will be given in the next section.

Features

In the data set, many physical features of every substitution are enumerated, and hence some of the features have the same structural or sequential meaning. To avoid redundancy, we selected the following nine structurally and sequentially independent features: (1) root mean squared deviation (RMSD), (2) active site pocket feature (AP), (3) ligand-binding site feature (LB), (4) dimer interface feature (DI), (5) difference in solvent-accessible surface area (ASA) values of functionally important residues between the wild type and mutant (DASA), (6) size of an amino acid in the mutant (MSIZE), (7) difference in size between wild type and mutant (DSIZE), (8) hydropathy of an amino acid in a mutant enzyme protein (MHYD) and (9) difference in hydropathy between wild type and mutant (DHYD).

RMSD. A large conformational change is generally thought to induce a significant decrease in the stability and/or catalytic activity of the enzyme, and thus would result in the classical phenotype. To evaluate conformational changes in the enzyme molecule caused by an amino-acid substitution, the RMSD values of α-carbon atoms in mutant proteins were calculated. The details of homology modeling and RMSD calculation were given in the previous paper.[7]

Functionally important residues. Mutations located in functionally important regions (the active site pocket, the ligand-binding site and the dimer surface) are deduced to decrease the catalytic activity. Therefore, we examined whether the location of a mutation is functionally important. If the mutation is located in the active site pocket, the AP is set to 1. On the other hand, if the mutation is located in a region other than the active site pocket, it is set to 0. The same procedure is used for LB and DI. The residues involved in ligand binding are defined as ones the ASA of which differs between the ligand-bound form and the ligand-unbound one. The residues comprising the dimer interface are defined as ones the ASA of which differs between the dimer form and the monomer one.

Differences in ASA values of functionally important residues. Structural changes on the surface of a functionally important region are predicted to induce a significant decrease in the catalytic activity of the enzyme, and thus would also lead to the classical phenotype. To evaluate influences of an amino-acid substitution on the surface of functionally important residues, DASA values were calculated by subtracting the ASA of the mutant from that of the wild type. The details of ASA calculations were given in the previous paper.[7]

Size of an amino-acid residue. In general, substitution of a large amino acid for a small one causes a steric conflict, which may induce a structural defect leading to a decrease in the stability of the protein molecule. On the other hand, substitution of a small amino acid for a large one causes extra-space around the mutation site, which may also induce a decrease in the stability of the mutant protein. Therefore, we determined the MSIZE, and examined the DSIZE. DSIZE was calculated by subtracting the size in the mutant from that in the wild type.

In this study, the size of an amino acid is roughly defined as the number of heavy atoms along the line from the α-carbon to β-carbon. In other words, the size of the amino acid is defined as the distance from the α-carbon to the furthest heavy atom along that line. The side chain atom of an amino acid is denoted by a Greek suffix with the atom name; that is, C-α, C-β, O-γ, O-δ, and so on. For example, the size of glycine, having no C-β, is 0. Alanine has a C-β and its size is defined as 1. Serine has an O-γ and its size is taken as 2. Leucine has a C-δ and its size is defined as 3, and so on. The size of tryptophan having a C-ζ is defined as 6. Thus, in the case of substitution from alanine to tryptophan, MSIZE and DSIZE are 6 and 5, respectively.

Hydropathy of an amino acid in a mutant enzyme protein. Hydropathy is one of the most common indices for characterizing an amino acid. In this study, it was evaluated using the Kyte and Doolittle scale.[8] We determined the MHYD, and the DHYD. DHYD was calculated by subtracting the hydropathy value of the wild type from that of the mutant.

Statistical analyses

Statistical analyses to determine the differences in the average values of each feature between the classic form and the variant one were performed by means of the F-test, followed by *t*-test, it being taken that there was a significant difference if $P<0.05$.

Construction of decision trees

Our data set consists of 196 classic cases and 14 variant ones. We call this 'data set 1'. A decision tree constructed from such an imbalanced data set would wrongly predict that all data indicate the classic form. To overcome the problem, we constructed two other data sets. First, we selected 14 classic cases as representative ones by means of clustering techniques according to their features. We call this data set consisting of the representative classic cases and the variant ones 'data set 2'. Second, we removed the classic cases the features of which were similar to those of variant cases and selected 14 new classic cases as representative ones. The details of selection are as follows: At first, we defined vector v for mutation i as follows: v(i)=(RMSD, DASA, MSIZE, DSIZE, MHYD, DHYD, AP, DI and LB). We then defined the D value between two mutations as follows: $D(i,j)=|v(i)-v(j)|$, after which we defined the S value as follows: $S(i)=\min D(i,j)$.

Here, j denotes the variant mutation, and if each feature of classic mutation i is identical to that of variant mutation j, S(i) is 0. To remove the classic cases the features of which were similar to those of variant cases, the 14 classic mutations with large S values were selected as representative ones. This data set is called 'data set 3'. We constructed decision trees for all three data sets using the 'mvpart' package of R statistical environment. This package is based on the Classification and Regression Trees algorithm.[9]

Cross-validation

We used the leave-one-out approach involving a single observation in the original data set as the validation datum, and the remaining observations as the training data. This procedure was repeated such that each observation in the data set was used once as the validation datum.

Evaluation

The accuracy of the derived model was examined for sensitivity, specificity, accuracy and for Matthew's correlation coefficient (MCC).[10] The calculation was performed according to the following formulae:

$$\text{Sensitivity} = \text{TP}/(\text{TP+FN})$$
$$\text{Specificity} = \text{TN}/(\text{FP+TN})$$
$$\text{Accuracy} = (\text{TP+TN})/(\text{TP+FP+FN+TN})$$
$$\text{MCC} = \frac{(\text{TP}\cdot\text{TN}-\text{FP}\cdot\text{FN})}{\sqrt{(\text{TN+FN})(\text{TN+FP})(\text{TP+FN})(\text{TP+FP})}}$$

TP, TN, FP and FN denote true positive (the mutation is predicted to be a classic phenotype one, and is actually a classic one), true negative (the mutation is predicted to be a variant phenotype one, and is actually a variant one), false positive (although the mutation is predicted to be a classic phenotype one, it is a variant one) and false negative (although the mutation is predicted to be a variant phenotype one, it is a classic one), respectively.

Sensitivity is the proportion of actual classic cases correctly identified as such, and specificity is the proportion of variant cases correctly identified as such. Accuracy is the proportion of correctly identified cases among all predictions. MCC is known to be a better evaluation criterion than overall accuracy. It takes into account true and false positives and negatives, and is generally regarded as a balanced measure that can be used even if the classes are of very different sizes. The values range from −1 to 1. A value of 1 means 'complete prediction', whereas a value of 0 means that every prediction was randomly assigned.

RESULTS

Differences in features between the classic and variant groups

To clarify the differences between classic and variant groups, we calculated the average values and s.d. for each feature in both classic and variant groups (Table 1 and Supplementary Figures 1a–i). As shown in the previous study,[7] there is a large difference in the RMSD value between the classic and variant groups, and the difference is statistically important. In the case of DASA and AP, there are also large differences between the classic and variant groups. These differences are also statistically important. In the cases of MSIZE, DSIZE, MHYD, DHYD and LB, although there are some differences between the classic and variant groups, they are not statistically important. In the case of DI, there is little difference between the classic and variant groups.

Predictive power of the individual features

A decision tree was generated using each feature to discriminate between the classic and variant groups, and the prediction performance of each feature was assessed (Table 2). Except for DI and LB, all other features had a role in the prediction of whether a mutation causes the classic phenotype or the variant one. When DI and LB were each used as a single feature for the prediction, MCC values were 0.00 and −0.10, respectively. In contrast, RMSD, the MCC of which reached 0.34, was found to be the best discriminator of the classic group compared with the variant one. Furthermore, accuracy, sensitivity and specificity were high (0.76, 0.76 and 0.86, respectively). The MCC values of DASA, DSIZE and DHYD were higher than 0.2, but they gave less prediction accuracy than RMSD. Other features such as MSIZE, MHYD and AP exhibited a poor predictive performance.

Table 1 Average values of the features in both the classic and variant groups

Feature	*Classic*	*Variant*	P-*value*
RMSD	0.0892 (0.0728)	0.0216 (0.0193)	8.17E-12
DASA	−3.23 (15.1)	1.82 (6.14)	0.02
MSIZE	3.33 (1.77)	2.71 (1.22)	0.11
DSIZE	0.485 (2.56)	−0.429 (1.99)	0.14
MHYD	−0.871 (2.85)	0.757 (3.06)	0.08
DHYD	−0.252 (3.61)	0.693 (3.14)	0.32
AP	0.179 (0.383)	0.00 (0.00)	6.12E-10
DI	0.148 (0.355)	0.143 (0.350)	0.96
LB	0.0969 (0.296)	0.214 (0.410)	0.33

Abbreviations: AP, active site pocket feature; DI, dimer interface feature; DASA, difference in solvent-accessible surface area values of functionally important residues between the wild type and mutant; DHYD, difference in hydropathy between wild type and mutant; DSIZE, difference in size between wild type and mutant; LB, ligand-binding site feature; MHYD, hydropathy of an amino acid in a mutant enzyme protein; MSIZE, size of an amino acid in the mutant; RMSD, root mean squared deviation.
Values are averages (s.d.).

Table 2 Prediction performance of the features obtained using the decision tree method.

Feature	*TP*	*TN*	*FP*	*FN*	*Sensitivity*	*Specificity*	*Accuracy*	*MCC*
RMSD	148	12	2	48	0.76	0.86	0.76	0.34
DASA	126	11	3	70	0.64	0.79	0.65	0.22
MSIZE	61	13	1	135	0.31	0.93	0.35	0.13
DSIZE	135	12	2	61	0.69	0.86	0.70	0.29
MHYD	143	7	7	53	0.73	0.50	0.71	0.13
DHYD	128	12	2	68	0.65	0.86	0.67	0.26
AP	35	14	0	161	0.18	1.00	0.23	0.12
DI	29	12	2	167	0.15	0.86	0.20	0.00
LB	19	11	3	177	0.10	0.79	0.14	−0.10

Abbreviations: AP, active site pocket feature; DI, dimer interface feature; DASA, difference in solvent-accessible surface area values of functionally important residues between the wild type and mutant; DHYD, difference in hydropathy between wild type and mutant; DSIZE, difference in size between wild type and mutant; FLB, ligand-binding site feature; FN, true negative; FP, false negative; MCC, Matthew's correlation coefficient; MHYD, hydropathy of an amino acid in a mutant enzyme protein; MSIZE, size of an amino acid in the mutant; RMSD, root mean squared deviation; TN, true negative; TP, true positive.

Decision tree model involving the optimal feature subsets

The prediction performance of the combined features using each data set was assessed using the decision tree method. Table 3 summarizes the accuracy of the prediction model. In the case of data set 1, the accuracy of the model is high (0.88). However, specificity is 0, and MCC is below 0. Although accuracy is high, the total prediction performance is quite poor. In contrast, in the case of data set 2, accuracy, specificity and sensitivity are 0.74, 0.93 and 0.72, respectively. The MCC value is 0.35. This is larger than that determined using each single feature. Furthermore, in the case of data set 3, sensitivity, specificity and error rate of the prediction model are 0.85, 0.93 and 0.14, respectively. The MCC value is the largest (0.49) among all prediction models.

Prediction model for Fabry disease

As shown in the previous section, the prediction model involving data set 3 is the best prediction model. Figure 1 shows the decision tree

Table 3 Prediction performance of the feature sets obtained using the decision tree method with different data sets

Data set	*TP*	*TN*	*FP*	*FN*	*Sensitivity*	*Specificity*	*Accuracy*	*MCC*
1	185	0	14	11	0.94	0.00	0.88	−0.06
2	142	13	1	54	0.72	0.93	0.74	0.35
3	167	13	1	29	0.85	0.93	0.86	0.49

Abbreviations: FN, true negative; FP, false negative; MCC, Matthew's correlation coefficient; TN, true negative; TP, true positive.

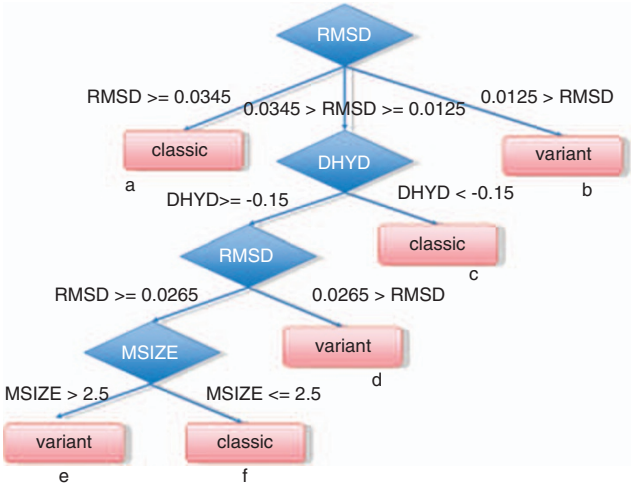

Figure 1 A flowchart of the scheme for prediction of the phenotype from sequential and structural information. This decision tree was built on the basis of data set 3. The sensitivity, specificity, accuracy and MCC for 210 missense mutations are 0.85, 0.93, 0.86 and 0.49, respectively.

involving data set 3. Four nodes were defined in this model on the basis of the following criteria: (a) classic, if RMSD $\geqslant 0.0345$, as large conformational changes are predicted to occur; (b) variant, if RMSD < 0.0125, as conformational changes induced by amino-acid substitutions are deduced to be small; (c) classic, if DHYD < -0.15, as hydrophobic residues are predicted to change to more hydrophilic ones; (d) variant, if RMSD < 0.0265, as conformational changes are deduced to be relatively small; (e) classic, if MSIZE < 2.5, as an amino acid of the mutant is small (glycine, alanine, valine, serine or cysteine); (f) variant, if MSIZE $\geqslant 2.5$, as an amino acid of the mutant is large.

DISCUSSION

The result of newborn screening among Italian people revealed that the incidence of Fabry disease is very high, especially the variant phenotype.[2] Furthermore, gene mutations causing Fabry disease are known to be quite heterogeneous.[1] Considering these results, the introduction of worldwide newborn screening will reveal a lot of Fabry disease patients with novel mutations whose prognoses are unclear at the time of diagnosis. The prediction of a clinical phenotype should help clinicians to decide a suitable treatment or a proper treatment schedule. However, we usually only know sequential information about mutations responsible for Fabry disease. Although it is very significant to predict the outcome of the disease on the basis of sequential information, there have been few studies on the prediction of phenotypes in Fabry disease.

The main purpose of this study is to establish a prediction methodology for the Fabry disease phenotype involving sequential and structural information. As far as we know, this is the first study on phenotype prediction for Fabry disease. In this study, we performed three analyses. (1) We investigated the differences in sequence and structure between the classic and variant groups carefully. (2) We examined the features that distinguish classic from variant phenotypes. (3) We constructed a decision tree-based prediction model for the phenotype groups from sequence and structure information.

In the first analysis, we examined the nine features of the sequential and structural changes induced by amino-acid substitutions to clarify the differences between classic and variant groups. With regard to RMSD, DASA and AP, there are significant differences, and they are statistically important. The results indicate that large conformational changes in any region of the molecule; changes in the surface area of a functionally important region and mutation at the active site cause a loss of enzyme activity, leading to the classic phenotype.

In the second analysis, we examined the prediction power of a single feature. RMSD is the best discriminator of the classic group compared with the variant one. The MCC of RMSD is quite high (0.34), which means that the prediction power using only RMSD information is unexpectedly high. Furthermore, DASA, DSIZE and DHYD are also good discriminators of the classic group compared with the variant one. Although prediction abilities are lower than that with RMSD, accuracy is distributed from 67 to 71%. This indicates that differences in size, hydropathy and surface area of catalytic residues between the wild type and mutant are also important factors for distinguishing between classic and variant groups.

In the last analysis, we constructed three prediction models for the phenotype of Fabry disease using three data sets. Although the prediction power of the decision tree model involving data set 1 is quite low, those of the decision tree models involving data sets 2 and 3 are high. This finding suggests that data set selection is important for phenotype prediction for this disease. In the case of data set 3, accuracy is also high (86%) and MCC is quite high (0.49), being significantly greater than that of RMSD. This means that the prediction method is significantly improved, taking not only RMSD but also other features into consideration. It is noteworthy that although data set 3 includes only 28 out of 210 mutations, 180 out of 210 mutations (86%) can be accurately predicted by the decision tree based on data set 3.

An overview of the best decision tree model is as follows: When a conformational change is predicted to be large (RMSD $\geqslant 0.0345$) or small (RMSD < 0.0125), the phenotype is predicted to be classic or variant, respectively. On the other hand, when a conformational change is predicted to be intermediate ($0.0125 \leqslant \text{RMSD} < 0.0345$), we cannot judge whether the phenotype is the classic or variant one without information on the size of the amino-acid residue substituted and/or the difference in hydropathy between wild type and mutant.

In conclusion, we investigated the sequential and structural changes in GLA responsible for Fabry disease. Results revealed that there are quite large differences between classic and variant phenotypes. We also constructed a phenotype prediction model involving sequential and structural information. Phenotype predictors may be useful for determining a proper therapeutic schedule for this disease.

ACKNOWLEDGEMENTS

This work was supported by the Japan Science and Technology Agency (HS); the Japan Society for the Promotion of Science (HS); the Ministry of Health and Welfare of Japan (HS); the High-Tech Research Center Project of the Ministry of Education, Science, Sports and Culture of Japan (HS); and by the Program for Promotion of Fundamental Studies in Health Sciences of the National Institute of Biomedical Innovation (HS).

1 Desnick, R. J., Ioannou, Y. A. & Eng, C. M. Alpha-galactosidase A deficiency: Fabry disease, in *The Metabolic and Molecular Bases of Inherited Disease* 8th edn (eds, Scriver, C.R., Beaudet, A.L., Sly, W.S., & Valle, D.), 3733–3774 (McGraw-Hill, New York, NY, 2001).
2 Spada, M., Pagliardini, S., Yasuda, M., Tukel, T., Thiagarajan., G., Sakuraba, H. *et al.* High incidence of later-onset Fabry disease revealed by newborn screening. *Am. J. Hum. Genet.* **79,** 31–40 (2006).
3 Eng, C. M., Banikazemi, M., Gordon, R. E., Goldman, M., Phelps, R., Kim, L. *et al.* A phase 1/2 clinical trial of enzyme replacement in Fabry disease: pharmacokinetic, substrate clearance, and safety studies. *Am. J. Hum. Genet.* **68,** 711–722 (2001).
4 Eng, C. M., Guffon, N., Wilcox, W. R., Germain, D. P., Lee, P., Waldek, S. *et al.* Safety and efficacy of recombinant human alpha-galactosidase A replacement therapy in Fabry's disease. *N. Engl. J. Med.* **345,** 9–16 (2001).
5 Schiffmann, R., Murray, G. J., Treco, D., Daniel, P., Sellos-Moura, M. Myers, M. *et al.* Infusion of alpha-galactosidase A reduces tissue globotriaosylceramide storage in patients with Fabry disease. *Proc. Natl Acad. Sci. USA* **97,** 365–370 (2000).
6 Yam, G. H., Bosshard, N., Zuber, C., Steinmann, B. & Roth, J. Pharmacological chaperone corrects lysosomal storage in Fabry disease caused by trafficking-incompetent variants. *Am. J. Physiol. Cell Physiol.* **290,** C1076–C1082 (2006).
7 Sugawara, K., Ohno, K., Saito, S. & Sakuraba, H. Structural characterization of mutant alpha-galactosidases causing Fabry disease. *J. Hum. Genet.* **53,** 812–824 (2008).
8 Kyte, J. & Doolittle, R. F. A simple method for displaying the hydropathic character of a protein. *J. Mol. Biol.* **157,** 105–132 (1982).
9 Breiman, L., Friedman, J. H., Olshen, R. A. & Stone, C. J. *Classification and Regression Trees* (Wadsworth Inc., Belmont, 1984).
10 Matthews, B. W. Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *Biochim. Biophys. Acta.* **405,** 442–451 (1985).

Supplementary Information accompanies the paper on Journal of Human Genetics website (http://www.nature.com/jhg)