CLINICAL STUDY

Machine learning approaches for phenotype–genotype mapping: predicting heterozygous mutations in the CYP21B gene from steroid profiles

Klaus Prank[1], Egbert Schulze[2], Olaf Eckert[3], Tim W Nattkemper[1,5], Markus Bettendorf[6], Christiane Maser-Gluth[7], Terrence J Sejnowski[8], Arno Grote[4], Erika Penner[1], Alexander von zur Mühlen[4] and Georg Brabant[4]

[1] International NRW Graduate School in Bioinformatics and Genome Research Center of Biotechnology (CeBiTec), Bielefeld University, D-33 615 Bielefeld, Germany, and Hannover Medical School, D-30 625 Hannover, Germany, [2]Molecular Genetics Laboratory Raue, Hentze, D-69 121 Heidelberg, Germany, Departments of [3]Visceral and Transplantation Surgery and [4]Clinical Enocrinology, Hannover Medical School, D-30 623 Hannover, Germany, [5]Applied Neuroinformatics Group, Faculty of Technology, Bielefeld University, D-33 615 Bielefeld, Germany, Departments of [6] Pediatrics and [7] Pharmacology, University of Heidelberg, D-69 120 Heidelberg, Germany and [8] Howard Hughes Medical Institute and Computational Neurobiology Laboratory, The Salk Institute, San Diego, California 92 186, USA

(Correspondence should be addressed to (K Prank; Email: klaus.prank@cebitec.uni-bielefeld.de)

Abstract

Objective: Non-linear relations between multiple biochemical parameters are the basis for the diagnosis of many diseases. Traditional linear analytical methods are not reliable predictors. Novel non-linear techniques are increasingly used to improve the diagnostic accuracy of automated data interpretation. This has been exemplified in particular for the classification and diagnostic prediction of cancers based on expression profiling data. Our objective was to predict the genotype from complex biochemical data by comparing the performance of experienced clinicians to traditional linear analysis, and to novel non-linear analytical methods.

Design and methods: As a model, we used a well-defined set of interconnected data consisting of unstimulated serum levels of steroid intermediates assessed in 54 subjects heterozygous for a mutation of the 21-hydroxylase gene (CYP21B) and in 43 healthy controls.

Results: The genetic alteration was predicted from the pattern of steroid levels with an accuracy of 39% by clinicians and of 64% by linear analysis. In contrast, non-linear analysis, such as self-organizing artificial neural networks, support vector machines, and nearest neighbour classifiers, allowed for higher accuracy up to 83%.

Conclusions: The successful application of these non-linear adaptive methods to capture specific biochemical problems may have generalized implications for biochemical testing in many areas. Non-linear analytical techniques such as neural networks, support vector machines, and nearest neighbour classifiers may serve as an important adjunct to the decision process of a human investigator not ‘trained’ in a specific complex clinical or laboratory setting and may aid them to classify the problem more directly.

European Journal of Endocrinology **153** 301–305

Introduction

We analyzed the genotype to phenotype relation in a series of 54 subjects heterozygous for a mutation in the 21-hydroxylase gene (CYP21B) and in 43 healthy controls. CYP21B was completely sequenced in all subjects to provide the class label, mutation or wildtype (Table 1). These genotypical data were related to the biochemical phenotype by measuring basal serum levels of the steroid intermediates 17αhydroxyprogesterone (17-OHP), 21-deoxycortisol (21-DF), dehydroepiandrosterone (DHEA), 17α-hydroxypregnenolone (17-OHPreg) and cortisol (Fig. 1).

The non-classic form of 21-hydroxylase deficiency based on heterozygous mutations of CYB21B occurs in approximately 0.2% of the general white population (1). The mutations induce a partial enzymatic block in steroid biosynthesis and alter the interrelation of steroid intermediates in the blood. Following augmentation of this altered steroid pattern by stimulation through exogenous adrenocorticotropin hormone (ACTH) approximately 80% of the heterozygous carriers may be correctly classified by the experienced clinician (2–4). However, unstimulated steroid levels show an almost complete overlap between healthy subjects and carriers heterozygous for the mutation (Fig. 1).

DOI: 10.1530/eje.1.01957
Online version via www.eje-online.org

Table 1 Frequency of mutations in the CYP21B gene.

Mutation	Number of subjects
IVS2-13A/C > G (656 -13A/C > G, intron 2)	26
Ile172Asn (1001 T > A, exon 4)	11
CYP21B deletion	9
G110fs (708_715del8, exon 3)	5
Arg356Trp (2110C > T, exon 8	3

Subjects and methods

Subjects

The first group consisted of 54 subjects (age, 27–46 years; mean, 34.1 years) with heterozygous mutations in CYP21B. These subjects were relatives of patients with homozygous CYP21B mutations leading to the classical forms of congenital adrenal hyperplasia (CAH). None of the subjects had clinical symptom of CAH or of hyperandrogenism. The geographic origin of most of the subjects with heterozygous mutations in CYP21B was Germany except for 20–30% of Turkish origin. The personal and family history of the 43 control subjects (age, 28–38 years; mean, 31.5 years) was unremarkable. Blood was sampled between 0800 and 1100 h. The study was approved by the local Committee on Medical Ethics, and all subjects gave their informed written consent.

Determination of steroid profiles and genetic testing

The serum levels of 17-OHP, 21-DF, DHEA, 17-OHPreg and cortisol were determined in both groups under basal (unstimulated) conditions (5, 6). Steroids were measured after extraction and chromatographic purification by radioimmunoassay using specific antibodies (7). The CYP21B gene of both groups was specifically amplified by PCR as described previously (8). The 10 exons, all exon–intron junctions, intron 2 and 7 and 400 bp of the promoter region of the CYP21B gene were analyzed on an automated DNA sequencer (LICOR, Lincoln, NE, USA). In a recent study we could demonstrate an exact classification of splice site mutations, the largest group of mutations in this study (9, 10).

Training and testing of classifiers{TX}

In the first step of analysis the four different classification algorithms were applied to the data: the linear discriminant analysis (LDA), the k-nearest-neighbour classifier (k-NN), the 'neural-gas' clustering algorithm and the support vector machine (SVM). The algorithms are explained breifly below. The classifiers applied to the serum levels of the steroid intermediates 17-OHP, 21-DF, DHEA, 17-OHPreg and cortisol. Since it has been reported that the diagnosis of CAH might be based on fewer parameters (11), we tried to predict mutations of CYP21B from the unstimulated levels of 17-OHP and cortisol only applying the same training and testing procedure as on the full data. Each one of the algorithms is applied in a 'leave-one-out' approach. The full data set is split into test and training data cases. In this work, the number of test cases is one. The remaining training data cases are used to train the classifier (LDA, neural gas, SVM) or to classify the test case directly (k-NN). To evaluate one algorithm, it is applied to each test/training data separation and the mean value and standard deviation of correct classification is computed as an indication of accuracy.

Linear discriminant analysis (LDA)

The linear Fisher discriminant was introduced to compute a one-dimensional projection of the data that maximizes the distance of projected class means, in regard to minimizing the result within class covariance. Based on the means and variances of the two classes the linear classification function can be computed explicitly.

k-Nearest-neighbour classifier (k-NN)

Nearest-neighbour algorithms (NNs) are ubiquitous in many research areas such as pattern recognition, machine learning and case-based reasoning (12).

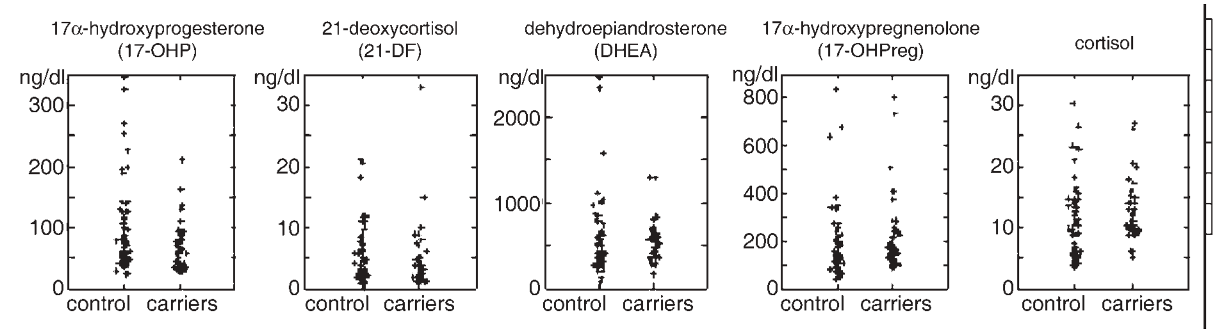

Figure 1 Profiles of steroid intermediates under basal (unstimulated) conditions in ($n = 54$) subjects with heterozygous 21-hydroxylase mutations (carriers) and in ($n = 43$) healthy controls.

The k-NN algorithm is a simple generalization of case-based reasoning, which is k-NN for $k = 1$. It maintains the set of training cases and classifies the test case according to the most frequent class among the k examples in the training set that are most similar to the test case. In practice, if there is a tie, the class assigned to the smallest distance of k-NN is chosen. Basically, the motivations for such a large diffusion of NN reside in its good performance, which can be obtained in many situations, and its ease of use, as no parameter needs to be tuned and very little experience is required on the part of the user.

Neural-gas clustering

Because of the relatively small number of profiles available, we used self-organizing artificial neural networks (13) instead of the standard multi-layer perceptrons (14). The architecture of the self-organizing artificial neural networks used in this study (neural-gas algorithm) is based on the principle of competitive learning (15, 16) and is a derivative of Kohonen's topology-conserving feature map algorithm (17, 18). In contrast to the LDA the neural-gas model is not based on the assumption of a linear relation between the input data and the predicted class or cluster. During the learning process (training) single neurons respond selective to different clusters in the sets of training patterns. After successful training each neuron (or codebook vector) is placed on the center of mass of a class, thereby representing the cases of that cluster. So the neural-gas model will elaborate differences in the profiles of steroid intermediates between healthy controls and subjects with heterozygous mutations of the 21-hydroxylase gene (CYP21B). A classification function was developed on the basis of the trained neural gas which represents the cluster structure. The steroid patterns of subjects with heterozygous mutations of the 21-hydroxylase gene or healthy controls were associated with every neuron. To allow for a stable classification we used 100 differently initialized networks for each training and testing run respectively. A majority vote was used for final classification.

Support vector machines (SVMs)

In recent years, kernel-based methods have been the object of much interest and research in the machine learning community. The success of many kernel methods is based on the concept of combining well-known linear algorithms, such as principal component analysis, with non-linear kernel functions. While the application of these functions allows more powerful non-linear solutions, the kernelized algorithms (19, 20) retain most properties of their linear versions. The most prominent algorithm among these is the support vector machine (SVM) proposed by Vapnik (21) for binary classification. The SVM is gaining popularity due to many attractive features and its superb classification performance, shown in numerous applications. It realizes pattern recognition between two classes by finding a decision function (hyperplane) determined by selected points from the training data, termed support vectors. In general this hyperplane corresponds to a non-linear decision boundary in the input space. While traditional techniques for pattern recognition are based on minimizing the empirical risk (i.e. on the attempt to optimize the performance on the training set), SVMs minimize the structural risk (i.e. the probability of yet-to-be-seen patterns to be classified correctly for an unknown probability distribution of the data). The SVM with a Gaussian kernel is trained and its parameters (width of Gaussian) and C (regularization) are optimized using the LIBSVM1 package, which combines the sequential minimal optimization (SMO) (22, 23) and the SVMLight2 algorithm. In this study, we used a SVM with Gaussian and linear kernels and ten-fold cross-validation

Diagnosis by human investigators

To compare the diagnostic results obtained from the artificial neural networks with the conventional clinical approach, four investigators known for their experience in the field were asked to classify the profiles of basal (unstimulated) levels of serum steroids of all subjects into heterozygous mutations or healthy control.

Statistical analysis

The 95% confidence intervals were calculated for the sensitivity, specificity and accuracy of the neural networks as well as for the diagnosis of each of the experts.

Results

It is clearly demonstrated in Fig. 1. that there is a huge overlap of unstimulated serum levels of all five steroids measured for the diagnostic procedure of heterozygous CYB21B mutations between healthy controls and subjects with heterozygous CYB21B mutations. Clinicians were unable to identify the genetic disorder from these unstimulated steroid levels (Fig. 1, Table 2). Comparably, multivariate LDA did not classify the genotype from unstimulated steroid levels (Table 2). Introducing three different non-linear methods (self-organizing neural networks (neural-gas algorithm), nearest-neighbour classifiers and SVMs) into the analysis dramatically enhanced diagnostic accuracy from unstimulated data only up to 83% (Table 2), a value reached by the experienced clinicians from steroid profiles under ACTH stimulation only (2–4). We further tested whether all steroid intermediates are essential for a successful classification. Reducing the steroid input

Table 2 Diagnostic performance of four computer-based classification algorithms and four human investigators using basal (unstimulated) serum levels of 17-OHP, 21-DF, DHEA, 17-OHPreg and cortisol (A) and 17-OHP and cortisol only (B).

Diagnosis method	Sensitivity (%)	Specificity (%)	Accuracy (%)
A			
Linear discriminant analysis	60 (46–71)	70 (55–81)	64 (54–73)
'Neural-gas' algorithm	81 (69–90)	79 (65–89)	80 (71–87)
k-NN classifier	74 (61–84)	63 (48–76)	69 (59–77)
Support vector machine	78 (77–79)	88 (87–89)	83 (82–83)
Expert 1	24 (15–37)	75 (60–85)	43 (34–53)
Expert 2	26 (16–39)	86 (73–93)	43 (34–53)
Expert 3	5 (2–15)	92 (80–97)	31 (23–41)
Expert 4	13 (6–24)	93 (81–98)	41 (32–51)
B			
Linear discriminate analysis	54 (41–66)	70 (55–81)	61 (51–70)
'Neural-gas' algorithm	81 (69–90)	74 (60–85)	78 (69–85)
k-NN classifier	57 (44–70)	63 (48–76)	60 (50–69)
Support vector machine	87 (86–88)	56 (55–57)	73 (73–74)

The 95% confidence interval (CI) is given in parentheses. ***P***-values are given for the comparison between the classification performance of the 'neural gas' algorithm and each of the four experts.

pattern to the two most commonly used steroid parameters, 17-OHP and cortisol, the diagnostic accuracy of non-linear analysis by neural networks and SVMs remained almost constant between 73 and 78% (Table 2). In contrast, the performance of LDA and nearest-neighbour classifier declined to values around 60% (Table 2).

In our model study, the insufficiency of linear decision boundaries is exemplified by multivariate LDA which was not able to reliably distinguish between healthy controls and heterozygous subjects (Table 2). Similar results were obtained by the clinicians, in particular in respect to sensitivity since they were focussing on detecting the genetic disorder from the subtle alterations already present in the unstimulated steroid profiles (Table 2). Introducing non-linear decision boundaries into the classification problem dramatically enhanced the predictive performance (Table 2).

Discussion

How to choose a suitable form of decision boundary? Novel non-linear adaptive techniques from machine learning have been demonstrated to be capable of selecting and locating the appropriate form of the decision boundary (24–27). These approaches are able to learn such complex non-linear decision boundaries from data sets of well-defined examples without having fixed thresholds or using a fixed rule based decision procedure. The basic methods for using these non-linear adaptive techniques are similar to those of conventional discriminant analysis. The data sets used in solving the classification problem have to be collected from a representative population of cases and the outcome must be known by some well-verified method (gold standard). These criteria are met in our study where we used a well-controlled model to test the performance of human experts, linear analysis and non-linear analysis in the prediction of the genetic disorder from specific steroid patterns.

In contrast to non-linear analysis, human experts failed to classify the disease from unstimulated steroid levels. The close similarity of their results to the results of linear analysis, suggests that the clinicians were unable to use the subtle non-linear relations between unstimulated steroid parameters for making their diagnostic decision. The predictive results of the neural network approach may be further evaluated in larger groups (28) and with formal testing of the influence of assay noise, interassay variability and the type of assay used. The observation that only two commonly assayed hormonal parameters out of the five initially included in predicting the genotype are sufficient to maintain the predictive power of non-linear analysis is in favour of a rather robust performance of the analytical procedure. However, the objective of our application of machine learning techniques is a screening test for further diagnostic refinement and not to achieve an accuracy close to 100%, which would be a requirement for genetic counselling for instance.

The present study is in line with the successful use of non-linear analytical methods such as neural networks and phase space reconstruction in the diagnosis of myocardial infarction (26), certain forms of epilepsy (27) and cancer based on expression profiling (29, 30). The efficacy of the non-linear approaches rests on their capability to delineate non-linear interactions by progressively learning from examples.

Correct classification of multivariate interconnected biochemical data is the essence of many diagnostic steps in medicine. Non-linear relations are frequent and in many instances no simple threshold levels can be defined. Thus, simple linear threshold analysis fails and the diagnostic accuracy of the experienced clinician is based on the detection of non-linear relations. In heterozygous CYP21B mutations this relation only becomes apparent following augmentation of the relative relations by ACTH stimulation.

Medical experts gather their experience from a large number of cases accumulated over a long time span by learning complex interactions of laboratory parameters, not easily evident. The novel non-linear techniques used in this study are capable of learning such non-linear relations from large data sets of well-defined examples. They are able to reduce the required time span to minutes and might help to reduce the costs of a diagnostic procedure. The successful application of these non-linear adaptive methods to capture specific biochemical problems may have generalized implications for biochemical testing in many areas. Non-linear analytical techniques such as neural networks, SVMs and nearest-neighbour classifiers may serve as an important adjunct to the decision process of a human investigator not 'trained' in a specific complex clinical or laboratory setting and may help them to classify the problem more directly.

Acknowledgements

We thank the four investigators known for their expertise in the field for their help in making diagnostic decisions based solely on the steroid profiles. We kindly thank H. Hecker for statistical assistance. The technical support of A. Rogatzki and H. P. Gensheimer in genetic analysis of CYP21B mutations and of the technicians from the Steroid Laboratory in hormone measurements is gratefully acknowledged. This study was supported by Deutsche Forschungsgemeinschaft under grant Br 915/4-4 and by a grant from the Ministry for Science and Research (MWF) of Northrhine-Westfalia (NRW) to the International NRW Graduate School in Bioinformatics and Genome Research.

References

1 Speiser PW & White PC. Congenital adrenal hyperplasia. *New England Journal of Medicine* 2003 **21** 776–788.

2 Kozower M, Veatch L & Kaplan MM. Decreased clearance of prednisolone, a factor in the development of corticosteroid side effects. *Journal of Clinical Endocrinology and Metabolism* 1974 **38** 407–412.

3 Moreira AC & Elias LL. Pituitary-adrenal responses to corticotropin-releasing hormone in different degrees of adrenal 21-hydroxylase deficiency. *Journal of Clinical Endocrinology and Metabolism* 1992 **74** 198–203.

4 New MI. 21-Hydroxylase deficiency congenital adrenal hyperplasia. *Journal of Steroid Biochemistry and Molecular Biology* 1994 **48** 15–22.

5 Fiet J, Villette JM, Galons H, Boudou P, Burthier JM, Hardy N, Soliman H, Julien R, Vexiau P, Gourmelen M & Kuttenn F. The application of a new highly-sensitive radioimmunoassay for plasma 21-deoxycortisol to the detection of steroid-21-hydroxylase deficiency. *Annals of Clinical Biochemistry* 1994 **31** 56–64.

6 Blanche H, Vexiau P, Clauin S, Le Gall I, Fiet J, Mornet E, Dausset J & Bellanne-Chantelot C. Exhaustive screening of the 21-hydroxylase gene in a population of hyperandrogenic women. *Human Genetics* 1997 **101** 56–60.

7 Grunwald K, Rabe T, Urbancsek J, Runnebaum B & Vecsei P. Normal values for a short time ACTH intravenous and intramuscular stimulation test in women in the reproductive age. *Gynecological Endocrinology* 1990 **4** 287–306.

8 Schulze E, Scharer G, Rogatzki A, Priebe L, Lewicka S, Bettendorf M, Hoepffner W, Heinrich UE & Schwabe U. Divergence between genotype and phenotype in relatives of patients with the intron 2 mutation of steroid-21-hydroxylase. *Endocrine Research* 1995 **21** 359–364.

9 Schulze E, Bettendorf M, Maser-Gluth C, Decker M & Schwabe U. Allele dropout using PCR-based diagnosis for the splicing mutation in intron 2 of the *CYP21B* gene: successful amplification with a TAQ/POW polymerase mixture. *Endocrine Research* 1998 **24** 637–641.

10 Day DJ, Speiser PW, Schulze E, Bettendorf M, Fitness J, Barany F & White PC. Identification of non-amplifying *CYP21* genes when using PCR-based diagnosis of 21-hydroxylase deficiency in congenital adrenal hyperplasia (CAH) affected pedigrees. *Human Molecular Genetics* 1996 **5** 2039–2048.

11 Azziz R. Detection of CAH heterozygotes. *Fertility and Sterility* 1997 **68** 183–184.

12 Cover TM & Hart PE. Nearest neighbour pattern classification. *IEEE Transactions on Information Theory* 1967 **13** 21–27.

13 Bishop CM. *Neural Networks for Pattern Recognition*. Oxford, UK: Clarendon Press, 1995.

14 Tarassenko L. *A Guide to Neural Computing Applications*. London: Arnold, 1998.

15 Martinetz TM, Berkovich SG & Schulten KJ. 'Neural-gas' network for vector quantization and its application to timeseries prediction. *IEEE Transactions on Neural Networks* 1993 **4** 558–569.

16 Martinetz TM & Schulten KJ. Topology representing networks. *Neural Networks* 1994 **7** 507–522.

17 Kohonen T. Self-organized formation of topographically correct feature maps. *Biological Cybernetics* 1982 **43** 59–69.

18 Kohonen T. Analysis of a simple self-organizing process. *Biological Cybernetics* 1982 **44** 135–140.

19 Schoelkopf B, Smola AJ & Mueller K-R. Kernel principal component analysis. In *Advances in Kernel Methods – SV Learning*, pp 327–352. Eds B Schoelkopf, CJC Burges & AJ Smola. Cambridge, MA: MIT Press, 1999.

20 Mika S, Raetsch G, Weston J, Schoelkopf B & Mueller K-R. Fisher discriminant analysis with kernels. In *Neural Networks for Signal Processing IX, IEEE*, pp 41–48. Eds Y-H Hu, J Larsen, E Wilson & S Douglas. New York: IEEE, 1999.

21 Vapnik V. *The Nature of Statistical Learning Theory*. Berlin: Springer, 1995.

22 Platt J. Fast training of support vector machines using sequential minimal optimization. In *Advances in Kernel Methods – Support Vector Learning*, B Schoelkopf, C Burges & A Smola. Cambridge, MA: MIT Press, 1998.

23 Keerthi SS, Shevade SK, Bhattacharyya C & Murthy KRK. *Improvements to Platt's SMO Algorithm for SVM Classifier Design*. Technical report TR CD-99-14, National University of Singapore, 1999.

24 Cross SS, Harrison RF & Kennedy RL. Introduction to neural networks. *Lancet* 1995 **346** 1075–1079.

25 Baxt WG. Application of artificial neural networks to clinical medicine. *Lancet* 1995 **346** 1135–1138.

26 Baxt WG & Skora J. Prospective validation of artificial neural network trained to identify acute myocardial infarction. *Lancet* 1996 **347** 12–15.

27 Martinerie J, Adam C, Le Van Quyen M, Baulac M, Clemenceau S, Renault B & Varela FJ. Epileptic seizures can be anticipated by non-linear analysis. *Nature Medicine* 1998 **4** 1173–1176.

28 Wilson RC, Mercado AB, Cheng KC & New MI. Steroid 21-hydroxylase deficiency: genotype may not predict phenotype. *Journal of Clinical Endocrinology and Metabolism* 1995 **80** 2322–2329.

29 Khan J, Wei JS, Ringner M, Saal LH, Ladanyi M, Westermann F, Berthold F, Schwab M, Antonescu CR, Peterson C & Meltzer PS. Classification and diagnostic prediction of cancers using gene-expression profiling and artificial neural networks. *Nature Medicine* 2001 **7** 673–679.

30 Dhanasekaran SM, Barrette TR, Ghosh D, Shah R, Varambally S, Kurachi K, Pienta KJ, Rubin MA & Chinnaiyan AM. Delineation of prognostic biomarkers in prostate cancer. *Nature* 2001 **412** 822–826.

Received 6 September 2004
Accepted 3 May 2005