

# **Applying Artificial Neural Networks in Pedigree Analysis of Complex Genetic Diseases**

*Wentian Li*

Lab of Statistical Genetics  
Rockefeller University

July 1, 1999

(notes from July 2000 can be found  
in the last page)

The context: It has been highly successful for the last ten years to locate chromosome regions that contain the disease gene for *Mendelian diseases*, even though very little may be known about the disease.

How was it done? It was done by "LINKAGE ANALYSIS".

First, families with a genetic disease are identified.

(Examples of Mendelian diseases: Cystic fibrosis [1985,1989], Duchenne muscular dystrophy [1982,1987], Retinitis pigmentosa type 4 [1989, 1990], Early-onset Alzheimer disease [1987, 1991], Fragile X syndrome [1983,1991], Huntington disease [1983,1993])

Second, DNA samples from members of these families are collected. Genetic markers on all chromosomes are “typed” (whole-genome “genotyping”).

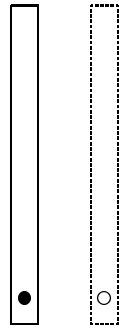
(Examples of genetic markers: restriction fragment length polymorphisms (RFLP), microsatellite markers (e.g. dinucleotide repeats), single nucleotide polymorphism (SNP))

Then, save both the disease status information and marker information of members of these families in a file. Run computer linkage analysis programs to see whether any marker(s) “co-segregate” with the disease status.

Co-segregation: follow the “movement” of each marker during meiosis to see whether the marker tends to “move” together with the disease status.

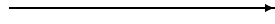
# MEIOSIS

father's chromosome 1

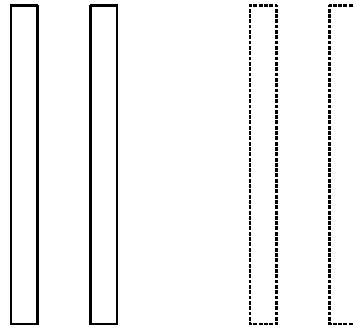


FF FM

duplication

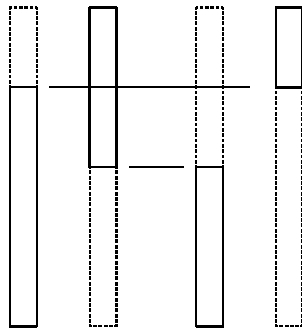


FF1 FF2 FM1FM2



2 cell divisions

crossovers



child's ch1 (half)

Assume there is a disease gene whose presence is completely responsible for the disease (1 copy: dominant disease, 2 copies: recessive disease). A marker located next to it has a larger chance to co-segregate with the disease gene, thus the disease status.

For a marker that does not co-segregate with the disease gene (e.g. a marker at a separate chromosome), it will still appear together with the disease gene, by chance alone, for half of the times.

co-segregate: non-recombinant. not co-segregate: recombinant.

The number of recombinants divided by the total number of meiosis is called the recombination fraction (rate):  $\theta$ .

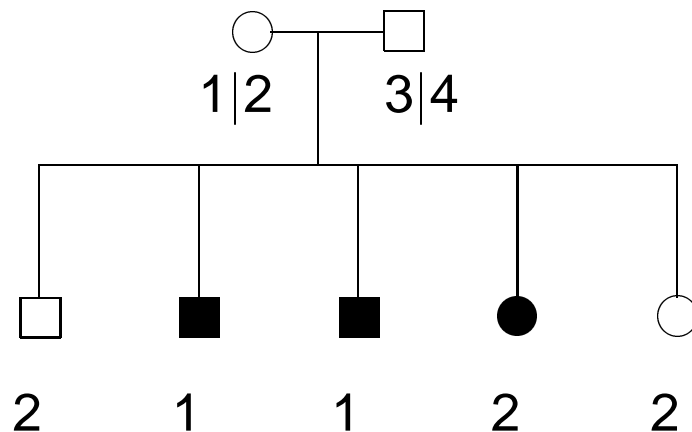
The smaller the  $\theta$ , the stronger the co-segregation, the stronger the “linkage”, the shorter the distance between the disease gene and the marker.

## Estimating $\theta$ from a nuclear pedigree data

Suppose mother is the carrier of the disease gene, and father is not.

Suppose allele 1 is linked to the disease gene. Then,  $\theta = 1/5$ .

One can see that allele 1 almost always co-segregates with the affected children, and 2 with the unaffected children. So the degree of co-segregation between this marker and the disease status is high (the “linkage signal” at this marker is high).



For real marker data from a pedigree, there are following complication issues:

(1) If both parents carry 1|2 genotype, it is not clear in a child's genotype, which allele is originated from the father, and which from the mother. It is a general problem that in the current technology, genotype is the observable, "ordered genotype" is not!

(2) If a parent's genotype is 1|1, there is no way to distinguish recombinant and non-recombinant, thus this meiosis is completely useless ("uninformative").

Computer programs are developed to handle missing information and other uncertain situations, by assigning probabilities to each situation.

The popular programs include: LINKAGE (written in Pascal), FASTLINK (in C), VITESSE (in C), GENEHUNTER (in C). For a complete list of all linkage programs, see

*<http://linkage.rockefeller.edu/soft/>*

In the currently standard approach of linkage analysis (following a paper by Newton Morton in 1955), the disease model (e.g. dominant, recessive), and the population frequencies of marker/disease gene alleles are given, then the probability (“likelihood”) of observing both the marker and the disease status in the pedigree is calculated, as a function of the recombination fraction  $\theta$ :

$$L(\theta) \propto \text{Prob}(\mathbf{g}, \mathbf{x} | \theta, \mathbf{f}, \mathbf{p})$$

The likelihood ratio (“odd”) between the likelihood with linkage and that without is:

$$\frac{L(\theta)}{L(\theta = 0.5)}$$

The logarithm of this ratio is (“log of odd” or LOD)

$$LOD = \log_{10} \frac{L(\theta)}{L(0.5)}$$

If the true  $\theta$  is indeed 0.5, the test  $t$  defined as

$$t = \max_{\theta} 2 \ln \frac{L(\theta)}{L(0.5)}$$

follows a  $\chi^2$  distribution with 1 degree of freedom (in the large sample limit).

The statistical foundation of the standard linkage analysis is the *hypothesis testing*. Or, more accurately, to reject the null hypothesis.

Null hypothesis: simple hypothesis which you know is not correct.

For linkage analysis of one marker, the null hypothesis is that the marker is not linked to the disease status, i.e. the maximum LOD occurs at  $\theta = 0.5$ .

In hypothesis testing, a test statistic  $t$  is chosen, its distribution under the null hypothesis is determined. Then the observed  $t$  from the data is obtained, and one calculates the probability that  $t$  can be larger than or equal to the observed value (tail area, or “p-value”). The smaller the p-value, the more unlikely that the null hypothesis is incorrect, and the more confident that a linkage can be claimed.

E.g.,  $\max \text{LOG} = 3$ ,  $2 \times \ln(10) \times 3 = 13.8155$ , and the p-value is 0.0002 (0.02%) if the  $\chi_1^2$  is used.

The problem with rejecting null hypothesis: there is no clue on what the alternative hypothesis is.

In the context of linkage analysis, if it is unlikely that a marker is uncorrelated with the disease status, what is its involvement with the disease status?

For a Mendelian single-gene disease, it is actually less a problem because out of all markers, there is only one region where the LOD is very high (and p-value very small). Other markers may achieve intermediate LOD values by chance, but not as high as the top peak.

For non-Mendelian “complex” diseases, no marker stands out as the clear linkage signal. Typically, there are many markers with the intermediate LOD value, which may be truly linked to one of the disease genes, or may be purely due to chance. I.e., signal and noise have the similar strength.

## Examples of non-Mendelian (complex) genetic disorders (traits):

\* psychiatric/behavioral disorders: schizophrenia, manic-depression, autism, panic-disorder, alcoholism

...

\* Neurological disorders: Alzheimer, epilepsy,...

\* Common diseases: hypertension, heart diseases, insulin-dependent diabetes, ...

Many complex diseases are not strictly “genetic”, but “genetic susceptible”: with a gene, one is perhaps more likely to be affected than the general population.

It is also universally true that many, not one, susceptibility genes are involved in a complex disease (trait). But it is not known whether the collective effect of these genes is additive.

An example: *Attention Deficit Hyperactivity Disorder in Children* (source: AJHG, Dec'98)

- \* Affect 3-5% children in US
- \* More boys than girls are affected
- \* Subtypes: inattentive, hyperactive-impulsive, both
- \* Family and adoption studies indicate genetic influence
- \* Twin studies indicate the heritability is from moderate to high (0.8, 0.89)
- \* Perhaps many genes, each has a small effect
- \* Molecular studies focus on neurotransmitter activity of dopamine (because the central role of dopamine in motor activity and reward-seeking behavior)
- \* Association with the dopamine D4 receptor gene (DRD4); with dopamine transporter gene (DAT1) ("candidate gene" approach)
- \* Knockout gene study in mice

## Two typical precautions in this type of analysis

1. Choosing different subtypes (e.g. inattentive, hyperactive) to define affected children. The issue is: genetically speaking, do different subtypes belong to the same disease?

2. Separating different families in the analysis. The issue is: is it possible that different genes (or different groups of genes) are responsible for the disease in different families? (“genetic heterogeneity”)

The large number of contributing “factors” to the disease, as well as their possible collective (interactive) effect, points to the possible use of “artificial neural networks”.

E.g. many genes.

E.g. genes and environmental factors.

In general,  $z = f(x_1, x_2 \dots)$ .

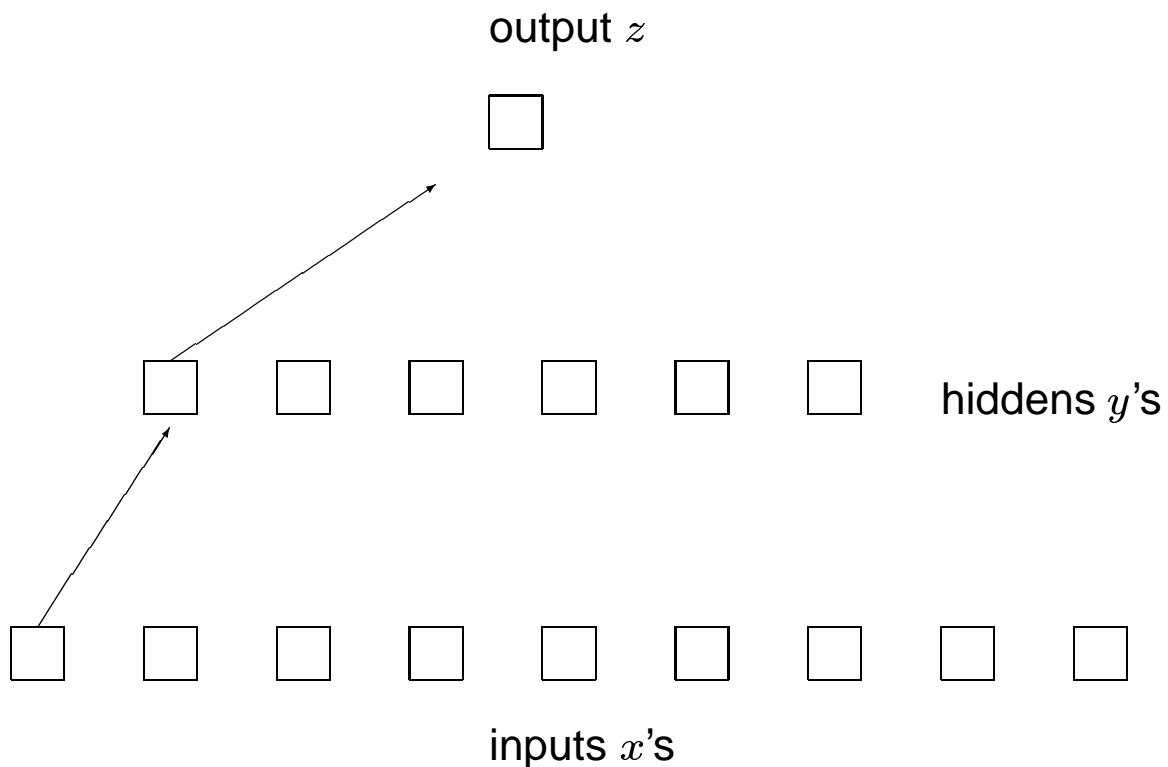
In neural network, the general function is approximated by:

$$z = L\left(\sum_j w_{j \rightarrow o} L\left(\sum_i w_{i \rightarrow j} x_i + b_j\right) + b\right)$$

or (if  $z$  is not bound by 1)

$$z = \sum_j w_{j \rightarrow o} L\left(\sum_i w_{i \rightarrow j} x_i + b_j\right) + b$$

where  $L(t) = 1/(1 + e^{-t})$  is the Sigmoid, or logistic function.



The purpose of neural network is not to reject a null hypothesis, but to find an alternative hypothesis (even though this alternative is in a form of “black box”).

Neural network is very similar to curve fitting and regression. Actually, it is a *nonlinear* regression analysis.

Similar to curve fitting using polynomial functions, where the model complexity increases with the highest order of the polynomial function used, the model complexity of the neural network increases with the number of hidden layers.

Neural network is just one example of the “data mining”, “pattern recognition”, “machine learning” ... techniques.

When the output is categorical instead of numerical, neural network is similar to “discriminant analysis”.

Neural network needs to learn by example: examples with different inputs and output values are fed to the network, and weights are adjusted during the learning.

As in other machine learning techniques, it is important to save some examples to validate the learned result.

Since neural network can approximate any nonlinear function, it is very easy to “over-fit” the data: models with higher complexity is used to fit the noise in the data. THE MOST IMPORTANT practical issue is on how to avoid the over-fitting.

Another example: *Alcoholic dependence (alcoholism)*

- \* Genetic contribution: 60% (a US study), 39% (a Swedish study).
- \* Possible subtypes: whether the person develops alcoholism before or after he/she develops other major psychiatric disorder, early/late onset, male-limited or not, etc.
- \* Possible association between the enzyme activity level of platelet “monoamine oxidase” (MAO), platelet “adenylyl cyclase” (AC), and the disorder.
- \*Candidate gene: aldehyde dehydrogenase (ALDH) which breaks down the potentially toxic acetaldehyde. Asians tend to have 1 (30-40%) or 2 copies (5-10%) of the defective allele of this gene: lower risk.
- \* People with high risk tends to have lower and slower responses in EEG after a stimulus.

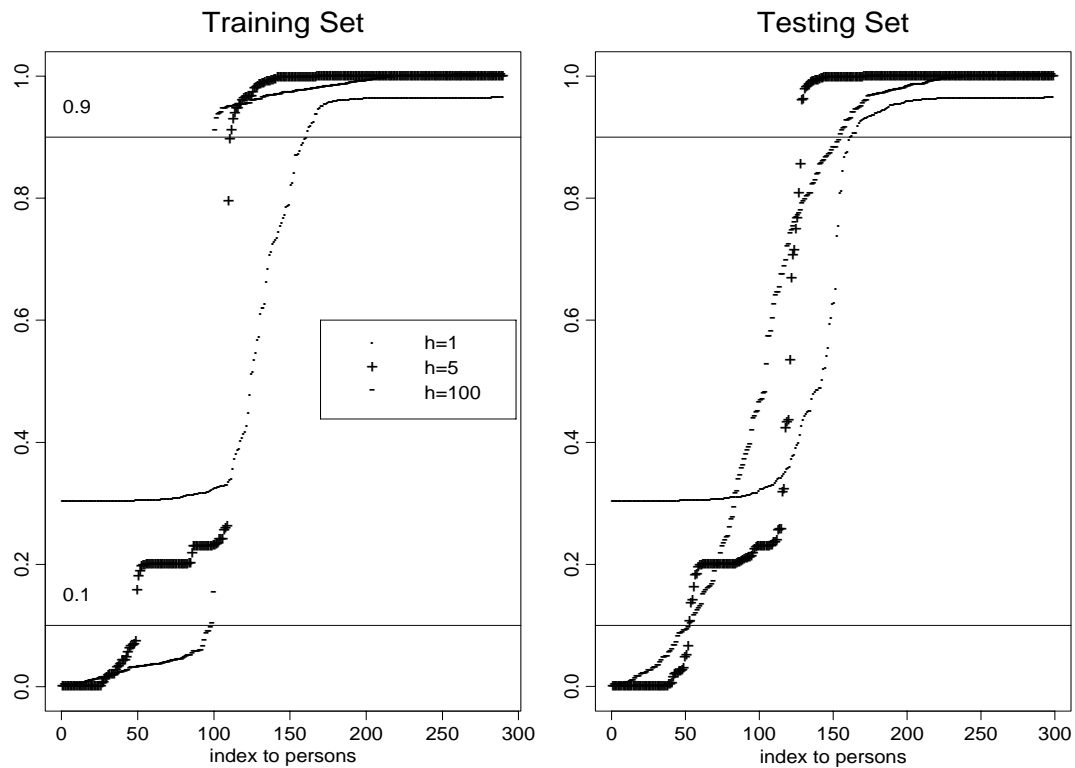
## **Prediction of affection status from intermediate phenotypes**

\* 10 inputs: measurements of EEG potential after 300 millisecc of a stimulus at 8 different locations; MAO activity, gender

\* 1 output: affection status

\* number of samples: 290 persons with known inputs and outputs, plus 299 persons with known inputs but unknown output (either no information, or never had a chance to drink, or can be classified as unaffected but with some symptoms). 290 known cases can further be split into training set and validation set.

## Learning from all 290 cases



Predicted output in ascending order for both the 290 known cases and 299 unknown cases with the number of hidden units equal to 1, 5, and 100.

## Learning from all 290 cases (**summary**)

\* For the 290 known cases, the predicted output is either close to 0 or to 1 (a gap is opened up)

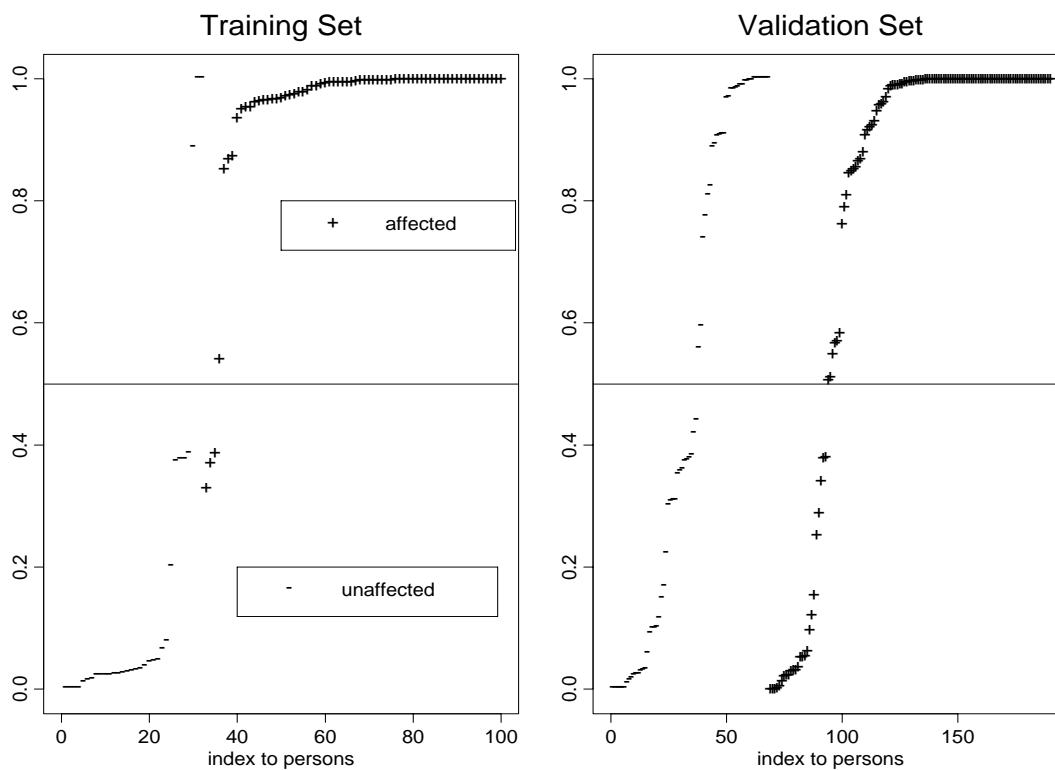
\* For 299 unknown cases, there is no way to know whether the prediction is correct or not. But we can ask how many cases whose predicted output is close to either 0 or 1?

\* No gap for the 299 unknown cases

\* The percentages of cases with the extreme values ( $< 0.1$  or  $> 0.9$ ) are:

$N_H$	1	5	10	20	100
$\alpha_{train}$	0.445	0.786	0.914	0.966	0.993
$\alpha_{test}$	0.462	0.746	0.856	0.719	0.659

## Learning from only part of the 290 known cases



Predicted output in ascending order for the learning set (100 cases) and the validation set (190 cases). Affecteds and unaffecteds are sorted and plotted by their predicted output separately. The number of hidden units is 40.

## Learning from only part of the 290 known cases (summary)

\* Again, there is a gap for the training set, but no gap for the validation set.

\* Percentage of correctly predicted cases (using 0.5 as the threshold):

	290(T)+0(V)			200(T)+90(V)			100(T)+190(V)		
$N_H$	4	10	40	4	10	40	4	10	40
t	.84	.84	.85	.845	.905	.80	.950	.850	.94
v	-	-	-	.711	.656	.767	.674	.658	.705

$N_H$  is the number of hidden units, and T(V) the number of persons in the training (validation) set.

\* More extensive runs with various parameters in neural networks lead to 1.5 - 13.5 % error rates in training set, and 21- 35% error rates in the validation set. (unpublished results)

## **An extension of linkage analysis with affected sib-pairs using neural networks:**

- \* Using affected sib-pairs is a popular method in linkage analysis
- \* The justification is that if both sibs are affected, they should both carry the same disease allele
- \* Comparing the sibs and find all “identical-by-descent” (IBD) alleles. The disease allele is supposedly one of them.
- \* Traditionally, each marker is examined at the time. No possible joint pattern was studied.
- \* If we use the IBD at each marker as an input, distinguishing affected-affected and affected-unaffected sib pairs as having two different outputs, neural network will carry out a classification (discrimination). This classification potentially considers all possible joint interactions among input variables.

## Another example: *Panic Disorder*

- \* Also known as anxiety neurosis
- \* The chance to be affected in US population is 1.5-5% (prevalence). Women are 2-3 times more likely be affected than men.
- \* Possible subtypes: “definite panic”, “probable panic”, “possible panic” (recurrent brief, spontaneous, two-symptom attack).
- \* Association with the serotonin transporter gene regulatory region

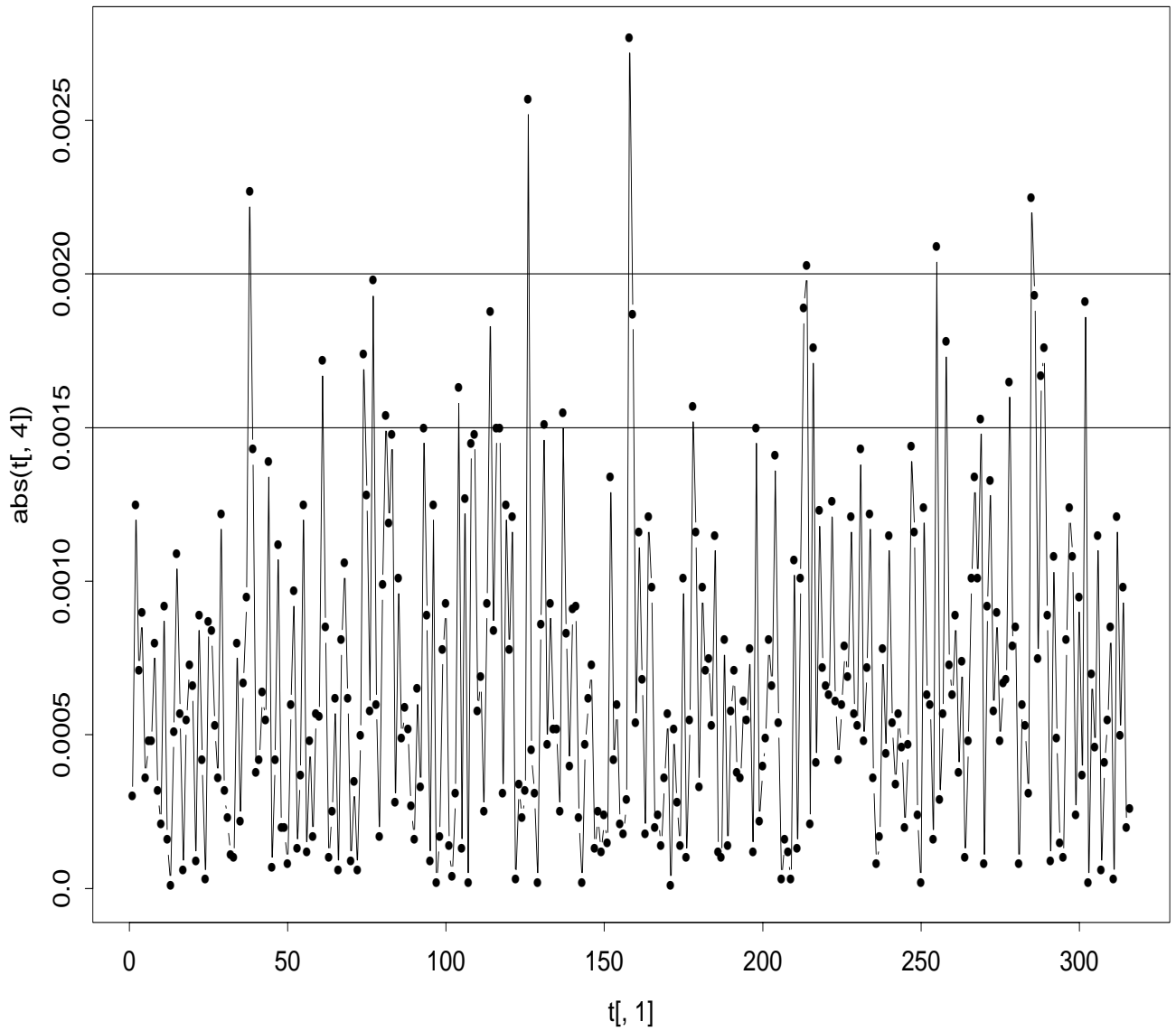
## Data

- \* The disease: Panic disorder
- \* 264 sibpairs: 182 affected-affected sibpairs, 82 affected-unaffected pairs
- \* Using 212 sample points for training, 52 for validation
- \* Number of markers (number of inputs) is 316 (on all 22 autosomal chromosomes)

## Considerations

- \* How many hidden layers should be used? Start with 1 (which describes an additive interaction among inputs), increase slowly. Or, start with a large number, gradually delete “unimportant” links and hidden layers (“pruning” technique).
- \* After the training, the contribution of each input (marker) to the output is estimated (e.g. partial derivative of the output with the input).
- \* Plot the contribution as the function of markers.

An example of the contribution for all inputs:



This study is in progress...

## General References

### **Linkage Analysis**

P Sham, *Statistics in Human Genetics* (Arnold, 1998)

J Ott, *Analysis of Human Genetics Linkage*, 3rd ed.  
(Johns Hopkins, 1999)

### **Complex Diseases**

edited by JL Haines, MA Peticak-Vance, *Approaches to Gene Mapping in Complex Human Diseases* (Wiley-Liss, 1998)

### **Neural Networks**

CM Bishop, *Neural Networks for Pattern Recognition* (Oxford, 1995)

BD Ripley, *Pattern Recognition and Neural Networks* (Cambridge, 1995)

## Notes from July 2000

**computer program for linkage:** a new program much faster than other programs is newly available: ALLEGRO (<http://www.decode.is/allegro/>) (Nature Genetics, 25:12-13 (2000) ).

**the problems with null hypothesis and likelihood framework** should be handled with a satisfactory solution in model selection framework (W. Li, paper in preparation, 2000).

**neural networks:** from the model selection point of view, neural networks may be too “complex”, and may not be selected by a model selection procedure.

**alcoholic dependence results** are now published in Li, Haghghi, Falk, Genetic Epidemiology (supplement), 17:S223-S228 (1999). although the overfitting issue is not solved in this paper.

**microarray data analysis:** the similar discriminant analysis for affected-affected and affected-unaffected sib pairs can be applied to microarray data analysis.